Classification and Feature Extraction in Man and Machine

Dissertation
zur Erlangung des Grades eines Doktors
der Naturwissenschaften
der Fakultät für Mathematik und Physik
der Eberhard-Karls-Universität zu Tübingen
vorgelegt von

Arnulf B.A. Graf

aus Lausanne (Schweiz)
2004

Tag der mündlichen Prüfung: 18.10.2004

Dekan
Prof. Dr. Peter Schmid
1. Berichterstatter
Prof. Dr. Hanns Ruder und Prof. Dr. Bernhard Schölkopf
2. Berichterstatter
Prof. Dr. Heinrich H. Bülthoff

Abstract

This dissertation attempts to shed new light on the mechanisms used by human subjects to extract features from visual stimuli and for their subsequent classification. A methodology combining human psychophysics and machine learning is introduced, where feature extractors are modeled using methods from unsupervised machine learning whereas supervised machine learning is considered for classification. We consider a gender classification task using stimuli drawn from the Max Planck Institute face database. Once a feature extractor is chosen and the corresponding data representation is computed, the resulting feature vector is classified using a separating hyperplane (SH) between the classes. The behavioral responses of humans to one stimulus, in our study the gender estimate and its corresponding reaction time and confidence rating, are compared and correlated to the distance of the feature vector of this stimulus to the SH. It is successfully demonstrated that machine learning can be used as a novel method to "look into the human head" in an algorithmic way.

In a first psychophysical classification experiment we note that a high classification error and a low confidence for humans are accompanied by a longer processing of information by the brain. Furthermore, a second classification experiment on the same stimuli but in a different presentation order confirms the consistency and the reproducibility of the subjects' responses.

Using several classification algorithms from supervised machine learning, we show that separating hyperplanes (SHs) are a plausible model to describe classification of visual stimuli by humans since stimuli represented by features distant from the SH are classified more accurately, faster and with higher confidence than the ones closer to the SH. A piecewise linear extension as in the K-means classifier seems however less adapted to model classification. Furthermore, the comparison of the classification algorithms indicates that the Support Vector Machine (SVM) and the Relevance Vector Machine (RVM), both exemplar-based classifiers, compare best to human classification performance and also exhibit the best man-machine correlations. The mean-of-class prototype learner, its popularity in neuroscience notwithstanding, is the least human-like classifier in all cases examined. These findings are corroborated by the stochastic nature of the human classification between the first and second classification experiments: elements

close to the SH are subject to more jitter in the subjects' gender estimation than elements distant from the SH.

The above classification studies also give a hint at the mechanisms responsible for the computation of the feature vector corresponding to a stimulus, in other words the feature extraction procedure which is defined by the combination of a data type with a preprocessor. Gabor wavelet filters reveal to be the most suited preprocessor when considering the image pixel data type. The biological realism of both Gabor wavelets and the image data confirms the validity of our approach. Alternatively, the information contained in the data type defined by the combination of the texture and the shape maps of each face, these maps bringing each face into spatial correspondence with a reference face, is also shown to be useful when describing the internal face representation of humans. Non-negative Matrix Factorization applied on the texture-and-shape data type is demonstrated to describe well the preprocessing of visual information in humans, and this has three implications. First, humans seem to use a basis of images to encode visual information, what may suggest that models such as kernel maps are less adapted since they do not use a basis to decompose (visual) data. Second, this basis seems to be part-based, in contrast to Principal Component Analysis which yields a holistic basis. Third, this part-based basis is spatially not too sparse, excluding Independent Component Analysis. Both for the encodings and for the basis, a medium degree of sparseness is shown to be most adapted.

Alternative approaches to model classification of visual stimuli by humans are subsequently introduced. In order to get novel insights into the metric of the human internal representation of faces, the above data is analyzed using logistic regression interpolations between the mean subjects' class estimate for a stimulus and the distance of this stimulus to the SH of each classifier. We show that a representation based upon the subjects' gender estimates is most appropriate, while the classification performance is demonstrated to be a poor measure when comparing man and machine. A novel psychophysical experiment is then designed where the hypotheses generated from machine learning are used to generate novel stimuli along a direction—the gender axis—orthogonal to the SH of each classifier. The study of the subjects' responses along these gender axes allows us then to infer the validity of the prediction given by machine learning. The results of this experiment—SVM and RVM are best while the prototype classifier is worst—validate the models given by machine learning and close the "psychophysics-machine learning" loop.

We finally show in a psychophysical experiment that it is more difficult to cast concepts from machine learning into a formalism describing the memory mechanisms of humans. However, machine learning is demonstrated to be an appropriate model for feature extraction and classification of visual stimuli in humans given the particular task we chose.

Zusammenfassung

Diese Dissertation befasst sich mit den Mechanismen, die Menschen verwenden, um Merkmale aus visuellen Reizen zu erzeugen und anschliessend zu klassifizieren. Es wird eine experimentelle Methode entwickelt, die menschliche Psychophysik mit maschinellem Lernen verbindet. Im Mittelpunkt der Arbeit steht ein Geschlechtsklassifikationsexperiment, das mit Hilfe der Kopfdatenbank des Max Planck Instituts durchgeführt wird. Hierzu werden verschiedene niedrig-dimensionale Merkmale aus den Gesichtsbildern extrahiert. Das Klassifikationsverfahren auf diesen Merkmalen ist durch eine Trennebene zwischen den beiden Klassen modelliert. Die Antworten der Versuchspersonen werden verglichen und korreliert mit der Distanz der Merkmale zur Trennebene. In dieser Arbeit wird bewiesen, dass maschinelles Lernen ein neues und wirksames algorithmisches Verfahren ist, um Einblicke in menschliche kognitive Prozesse zu erhalten.

In einem ersten psychophysischen Klassifikationsexperiment wird gezeigt, dass eine hohe Fehlerrate und ein niedriges Vertrauen der Versuchspersonen einer längeren Verarbeitung der Information im Gehirn entsprechen. Ein zweites Klassifikationsexperiment auf den selben Reizen aber in unterschiedlicher Reihenfolge, bestätigt die Konsistenz der Antworten der Versuchspersonen und die Reproduzierbarkeit der folgenden Resultate.

Es wird gezeigt, dass Trennebenen ein adäquates Modell sind, um die Klassifikation visueller Reize bei Menschen zu beschreiben. Reizmerkmale, die entfernt von der Trennebene sind, werden dabei genau, schnell und mit hohem Vertrauen klassifiziert. Es stellt sich heraus, dass Verfahren, die auf einer stückweis-linearen Trennebene basieren, weniger geeignet sind. Dahingegen beschreiben beispielbasierte Verfahren wie die Support Vector Machine oder die Relevance Vector Machine am besten das Verhalten der Versuchspersonen. Dies wird belegt durch Studien, die sowohl den Klassifikationsfehler vom Menschen und der Maschine vergleichen als auch deren Verhalten korrelieren. Der weitverbreitete Prototypenlerner schneidet am schlechtesten ab. Diese Resultate werden unterstützt durch eine Studie der stochastischen Komponente des menschlichen Klassifikationverfahrens: die Schätzung des Geschlechts ist inkonsequent zwischen dem ersten und zweiten Klassifikationsexperiment auf den Mustern nahe zur Trennebene.

Im weiteren Rahmen erlauben die in dieser Arbeit durchgeführten Stu-

dien Aussagen über die Mechanismen der menschlichen Merkmalsextraktion. Die biologisch-bewiesene Relevanz von Gaborfilterantworten erweist sich auch in dem Kontext der hier durchgeführten Studien als geeignete Kodierung von Pixeldaten. Desweiteren erweist sich die Information enthalten in der Kombination von Textur- und Form-Flussfeldern als gut geeignet zur Beschreibung der menschlichen Merkmalsextraktion. Hier werden räumliche Korrespondenzen der Bildreize miteinbezogen. Mit Hilfe dieses Datentyps kann gezeigt werden, dass Menschen für diese Aufgabe wahrscheinlich eine Bilderbasis verwenden, die aus Musterteilen besteht und nicht aus Gesamtmustern. Letztlich werden die Merkmalsextraktionsverfahren hinsichtlich ihrer Spärlichkeit untersucht, wobei sich ein mittlerer Grad an Spärlichkeit als am besten erweist.

Im weiteren werden Verfahren zur Modellierung des menschlichen Verhaltens bei Klassifikation von visuellen Reizen untersucht, die Aussagen über die Metrik der internen Gesichtsdarstellung erlauben. Dafür wird eine logistische Regression zwischen der Geschlechtseinschätzung der Versuchsperson für einen Reiz und der Distanz dieses Reizes zur Trennebene verwendet. Es wird gezeigt, dass eine Darstellung, die auf Antworten der Versuchsperson basiert, sich besser eignet, als eine Darstellung, die auf dem wahren Geschlecht basiert. Es stellt sich heraus, dass der Klassifikationsfehler ein schlechtes Mass zwischen Mensch und Maschine ist. In einem weiteren psychophysischen Klassifikationsexperiment werden die Trennebenen der Maschine verwendet, um neue Gesichtsreize zu erzeugen: diese liegen auf einer Geschlechtsachse, die senkrecht zur Trennebene steht. Die Unterscheidung durch die Versuchspersonen der Reize auf dieser Achse bestätigt die obigen Vorhersagen: die Support Vector Machine und die Relevance Vector Machine erweisen sich als besser als der Prototypenlerner, um das menschliche Klassifikationsverfahren zu modellieren. Mit diesem Experiment wird die "Psychophysik-maschinelles Lernen" Schleife geschlossen.

In einem abschliessenden psychophysischen Experiment wird gezeigt, dass es schwieriger ist, maschinelles Lernen auf das Gedächnissverhalten des Menschen anzuwenden, obwohl sich maschinelles Lernen als gut erweist, um Merkmalextraktion und Klassifikation visueller Reize bei Menschen zu modellieren.

Acknowledgments

First and foremost I take pleasure in expressing my profound gratitude to F.A. Wichmann for introducing me to methodologically sound human psychophysics. I like to acknowledge his help and guidance during the development of the ideas in this dissertation.

This research was conducted at the Max Planck Institute for Biological Cybernetics in an interdisciplinary project between the Department for Cognitive and Computational Psychophysics headed by Prof. H.H. Bülthoff and the Department for Empirical Inference for Machine Learning and Perception headed by Prof. B. Schölkopf. In a sense this dissertation continues unpublished work by Prof. B. Schölkopf while being a PhD student at the Max Planck Institute for Biological Cybernetics some years ago in the group of Prof. H.H. Bülthoff. I like to thank both directors for creating an agreeable research atmosphere and for providing me with their excellent facilities. Furthermore I would like to express my deep gratitude to Prof. H.H. Bülthoff for providing me with the opportunity to attend various conferences, summer schools and workshops which broadened my horizon in the neurosciences. In addition, I like to express my thanks to Prof. H. Ruder for accepting to serve as my thesis director at the Department of Mathematics and Physics of the Eberhard Karls University of Tübingen.

Finally I would also like to thank V. Blanz for providing me with a modified version of the MPI face database software. Many people have contributed to make my stay at the MPI scientifically beneficial. Of those, C. Wallraven, J. Hill, O. Bousquet, M. Giese, D. Cunningham, O. Chapelle, A. Gretton, M. Kleiner, C. Curio, A. Casile and M. Franz are worth special mentioning.

Contents

Chapter 1

Introduction

Das wird nächstens schon besser gehen,
Wenn Ihr lernt alles reduzieren
Und gehörig klassifizieren.

Mephistopheles in
Faust, Der Tragödie Erster Teil
Johann Wolfgang Goethe

Motivation

The aim of this dissertation is to obtain a better understanding of the mechanisms and principles underlying classification as well as feature extraction of visual stimuli by human subjects. For this, we combine machine learning and psychophysical techniques to gain insight into the algorithms used by human subjects during visual classification of faces. In this "psychophysics-machine learning" research we substitute a very hard to analyze complex system—the human brain—by a reasonably complex system—a learning machine—that is complex enough to capture some essentials of the human behavior but is amenable to close analysis, allowing us to make predictions about human behavior based on the properties of the machine. This research is focused on a novel methodology allowing to bridge the gap between human psychophysics and machine learning by extracting quantitative information from a high-level psychophysical setup. Bringing together theoretical modeling and behavioral data is arguably one of the main challenges when studying the "computational brain" [Churchland and Sejnowski, 1992]. The last decade has seen important technological advances in neuroscience from a microscopic scale (e.g. multi-unit recordings) to a macroscopic scale (e.g. functional Magnetic Resonance Imaging). However, on an algorithmic level, the methods and understanding of brain processes are still limited. Can we generate testable hypotheses about the algorithms used by humans to extract features from visual inputs and subsequently classify them? This

would be of utmost interest for human psychophysics. Can we find feature extractor and classifier pairings performing significantly better in gender classification than others? If so, this would be of interest for the next generation of human interface designs where computer systems should respond intelligently towards the agent they are interacting with. Currently high-level vision research, with its intrinsically complex stimuli, is seriously hampered by a lack of identifying methods, at the algorithmic level, as to what is happening during high-level classification. The here-presented method has the potential to overcome this obstacle.

Object classification plays a central role in visual neuroscience and it is an issue well adapted to the interdisciplinary nature of neuroscience. Indeed the study of classification involves psychophysics, physiology (single cell recordings, functional Magnetic Resonance Imaging, Electroencephalograms, Magnetoencephalograms, ...) as well as theoretical modeling as is illustrated in an overview by [Tarr and Bülthoff, 1998]. However, before a visual pattern can be recognized or classified, it must be represented in the brain. For this, a feature extractor is required in order to convert the signal arriving on the retina into a representation useful for the brain: a feature vector. In this dissertation we address the question on how to look into people's heads using a novel methodology combining psychophysical techniques and machine learning. Such a method provides insight into the *algorithms* used by man, in contrast to imaging techniques which deal essentially with functional *anatomy*. For this, the theoretical framework chosen in this dissertation is supervised and unsupervised machine learning. We hope to show that machine learning is well suited for our enterprise of explaining feature extraction and classification in human subjects given a human psychophysical classification experiment. Of course other theoretical approaches exist. Reinforcement learning [Sutton and Barto, 1998] deals with decision-making of an agent in a state space given a reward scheme. The policy chosen by the agent models the internal mechanisms of the brain implying the agent's high-level behavior. Modeling on a neuronal level and extensions of the latter to group of neurons is treated in [Dayan and Abbott, 2001]. This type of methods are closely related to the biological mechanisms encountered in the nervous system. The link to high-level classification experiments is however less obvious. Both of these approaches apply to different contexts and would thus require other studies and different experimental setups and would finally address different issues and functions of the brain.

As far as the *feature extractor* is concerned, the following questions may arise as also pointed out by [Peterson and Rhodes, 2003], and in particular by [Bülthoff and Bülthoff, 2003]. What algorithms best describe the way humans extract features from their visual inputs? Indeed the representation of objects for the purpose of recognition is a fundamental issue in biological and computer vision as has been shown by [Bülthoff and Edelman, 1992]. Furthermore, does our brain extract parts from the objects (part-based ap-

proach) or does it use them as a whole entity (holistic approach)? There have been some contradictory attempts to answer this question. A psychological study has been done by [Pelli, Farell, and Moore, 2003] where words stand for the holistic representation whereas letters for the part-based one. It was suggested that humans adopt a part-based scheme where words are inefficiently recognized. Further evidence of a part-based representation in the context of object recognition was obtained by [Biederman, 1987] using the so-called "geons" approach. However [Gauthier, Curran, Curby, and Collins, 2003] suggests a holistic processing of visual information. Similar question have been asked in the context of face recognition: do we use face patches or attributes to classify faces or do we rather consider the face as a whole? Here the holistic approach is predominant [O'Toole, Defenbach, Valentin, McKee, Huff, and Abdi, 1998, Valentin, Abdi, Edelman, and O'Toole, 1997, O'Toole, Abdi, Deffenbacher, and Valentin, 1993]. Further, is the encoding used by the brain sparse? And how about its image basis, if any?

For the *classifier*, we may ask the following questions. What algorithms describe best the way the brain classifies its visual inputs after feature extraction? Might humans use something akin to hyperplanes for classification? If so, is the learning rule as simple as in mean-of-class prototype learners (classification according to nearest prototype) or are more sophisticated algorithms better candidates? Early work trying to elucidate the principles of pattern recognition and classification of visual stimuli by humans is due to [Reed, 1972] using face-like stimuli (line drawings) and is followed by [Rosch, Mervis, Gray, Johnson, and Boyes-Braem, 1976]. Both studies rely on the theoretical side on the mean-of-class prototype learner and on the Content Model where classification is done using the similarity to all patterns in the dataset. Although the psychophysical experiments have only slightly changed since then, the evolution of learning theory has been essentially ignored by experimentalists. These baseline models for classification are still in use nowadays in psychological literature [Lamberts, 1997, Palmeri, 2001] and in physiology [Sigala and Logothetis, 2002, Sigala, Gabbiani, and Logothetis, 2002]. They show a discrepancy between the state-of-the-art modeling and the methods effectively in use in experimental research. In particular, the concept of prototype plays a central role in experimental neuroscience. It has been used for instance by [Leopold, O'Toole, Vetter, and Blanz, 2001] in a psychophysical setup to study aftereffects induced by adaptation using the MPI face database. The center of the face space, i.e. the prototype, was used to create an "anti-face", the latter being shown to play a central role in adaptation. Although there is no classification, the concept of prototype is still present. Is this justified? And is sparseness of the classification algorithm an important issue?

To answer the above questions we developed a novel methodology to bridge quantitatively the gap between psychophysics and supervised and unsupervised machine learning. An early attempt was communicated in

[Graf and Wichmann, 2004]. The psychophysical classification experiment is a behavioral study where machine learning is used to help understand human feature extraction and classification behavior. Feature extractors are modeled using methods from unsupervised machine learning (absence of class labels) whereas the classification is modeled using supervised machine learning (the class labels are used). This approach may be termed as psychophysical machine learning or in short *PSYCHO ML*. We hope to shed new light on the concepts underlying classification and feature extraction by humans and supply alternatives to the presently commonly-used methods by using elaborate theoretical models from machine learning. In this way we demonstrate the usefulness of machine learning to describe some fundamental mechanisms of the brain.

Classification and Feature Extraction: a pairing?

Classification may be argued to be used by the brain in order to discretize the continuous perceptual world [Rosch, Mervis, Gray, Johnson, and Boyes-Braem, 1976]. The visual world perceived by a human observer can be fully described by a five-dimensional, continuous function, the plenoptic function [Adelson and Bergen, 1991]: $P = P(x, y, t, \lambda, V_x)$, where P is the light-intensity distribution as a function of space (x, y) on the observer's retina, time (t) as processed by temporal filters, wavelength (λ) as sampled by the three cone types and V_x are two samples along the x (eye-eye)-axis taken by the two eyes. One task of early vision is to measure the plenoptic function such that it provides as much information about the visual world as possible. Consequently early vision is still continuous. Visual cognition, on the other hand, is discrete, where the world is perceived as a discrete set of objects, resulting from perceptual categorization. The determination of the relative quantity of visual objects can be seen as the counting of discrete elements, the latter being provided by classification. This counting ability can be argued to be the basis of the numerical and computational functions of the brain. The prefrontal cortex in monkeys was shown to participate in visual representations which could contribute to judgments of quantity and seems thus to be one of the "counting" centers of the brain [Nieder, Freedman, and Miller, 2002]. Thus classification is a fundamental issue since it may lay the foundations of a computational theory of the brain.

The ability of the brain to classify, i.e. to group objects into meaningful categories, is a fundamental cognitive process. How does the brain classify, what categories does it form and how does it represent stimuli, or what aspects of the plenoptic function are measured, extracted and thereafter grouped into classes? In the following we shall consider feature extraction and classification in its well-specified machine learning interpretation in order to avoid the richness and semantic connotations of the terms "feature

extraction" and "categorization" as used in cognitive psychology: we will refer to feature extraction and classification as a process that is purely data-driven, most akin to *perceptual* categorization in the psychological literature.

The image formed on the retina, projected onto its photoreceptors, can be represented by a high-dimensional input vector of dimension $100 - 150$ million. A crucial aspect of early vision is to reorganize this vast input space into a more manageable format. This is the feature extraction step studied in this thesis. Research suggests that early visual processing attempts to find a *sparse* code for image representation and that this enterprise is greatly helped by the redundancy of the visual input vectors [Rao, Olshausen, and Lewicki, 2002]. Moreover, in order to use machine learning classifiers, a *dimensionality reduction* of these input vectors is required. Note the terminology difference in machine learning and vision sciences: preprocessing in vision science stands for feature extraction in the machine learning literature.

From a purely theoretical point of view, there exists a clear distinction between input images, the feature extraction step and the classification method. However, mathematically certain preprocessing and classification algorithms are incompatible with each other, or contain elements of each other, or are even inseparably intermingled. The following is an attempt to describe what appear logically separable steps under separate headings: the feature extractors are drawn from unsupervised machine learning whereas the classifiers are drawn from supervised machine learning. Further, as encountered in most applications of machine learning such as computer vision [Graf, Smola, and Borer, 2003, Wallraven and Graf, 2004], it is not possible to separate preprocessing from the actual classification algorithm: a good preprocessor combined to a bad classifier can yield similar results as a bad preprocessor combined with a good classifier. This is why we here study both the (unsupervised) preprocessor combined with the (supervised) classifier.

Relation to Physics

Learning theory, in particular classification, is a fundamental problem and has attracted scientists from apparently distant disciplines such as theoretical physics, as for instance [Engel and den Broeck, 2001] where methods from statistical mechanics are applied to classification. The information conveyed and processed in the brain can be modeled using spatiotemporal electric wave patterns both on the microscopic level (neurons and their action potentials as measured by microelectrodes) and on the macroscopic level (brain states i.e. behavior as measured by functional Magnetic Resonance Imaging). The mesoscopic level aims to bridge the gap between the latter levels using the concept of wave packets [Freeman, 2003], a funda-

mental concept in quantum mechanics. This level is created by local state transitions, can be measured in electroencephalogram studies and seems to be a precursor of awareness. Furthermore, cortical imaging methods motivated an approach for modeling the dynamics of the interaction between populations of neurons based upon statistical mechanics. In this approach a probabilistic model of the dynamics of interacting neurons, instead of a classical model enumerating the neurons and their connections, has been introduced by [Sirovich, 2003]. The state of a population of neurons is then replaced by their probable states, yielding a formalism of operators and their eigentheory. This formalism bears again quite some analogies to quantum mechanics. Moreover, some eminent physicists have converted from theoretical and applied physics to the neurosciences. Amongst others Leon Cooper (Nobel Prize in Physics with J. Bardeen and J.R. Schrieffer in 1972 for their studies on the so-called BCS theory of superconductivity) has developed the Bienestock-Cooper-Munro (BCM) model for synaptic plasticity [Bienenstock, Cooper, and Munro, 1982]. He is now investigating the biological mechanisms that underlie learning and memory storage. Brian Josephson (Nobel Prize in Physics in 1973 for his theoretical predictions of the properties of a supercurrent through a tunnel barrier, in particular those phenomena which are generally known as the Josephson effects) has recently been examining from the viewpoint of theoretical physics complex phenomena in the brain, in particular language [Josephson, 2004]. He is currently attempting to understand what may loosely be characterized as intelligent processes in nature associated with some functions of the brain.

As our understanding of the mechanisms of the brain evolves, novel links to physics will emerge. Subsequently such connections will allow to apply the well-established methods of theoretical physics to the neurosciences, and ultimately link both disciplines.

Dissertation Setting

In our attempt of comparing the classification behavior of man and machine, we use high-level vision stimuli since they are meaningful and biologically important. We opt for gender classification of images from human faces, indeed a highly relevant biological task for strongly social beings like ourselves.

The stimuli are drawn from a processed (or cleaned) version of the Max Planck Institute (MPI) face database where all faces are centered in the image, have same pixel-surface area and same mean and standard deviation of the intensity. Fig.1.1 gives an overview of the methodology used to study feature extraction and classification in man and machine. The responses of humans to one stimulus, in our study the gender estimate and its corresponding reaction time (RT) and confidence rating (CR), are recorded.

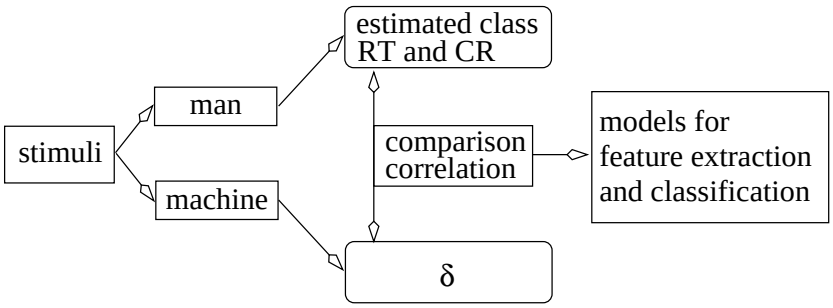


Figure 1.1: General flowchart of the experiments bridging the gap between human psychophysics and machine learning.

Machine classification is modeled using separating hyperplanes in the space of the feature vectors corresponding to the stimuli. For each subject the distance δ of each feature vector to the hyperplane is computed. The responses of man are then compared and correlated to δ. On the hand of these results, we may get a hint at the mechanisms and strategies used by humans to classify visual stimuli. Furthermore, to validate the approach introduced here, a novel psychophysical experiment is designed where the hypotheses generated from machine learning are used to generate novel stimuli along a direction orthogonal to the separating hyperplane of each classifier. These stimuli are then presented to the subjects in a gender discrimination task and their responses allow us to study the validity of the gender-axes predicted by machine learning. By doing so, we close the psychophysics-machine learning loop. Finally the mechanisms involved in the memorization of visual stimuli are studied using machine learning.

The algorithms chosen in this study both for feature extraction and classification are the most representative members of each family of algorithms sharing a similar theoretical foundation. These algorithms are enumerated in Fig.1.2 but a complete description with corresponding details will be given later in Chapters 2 and 4. One of the fundamental principles of machine learning is Occam's razor: *No more things should be presumed to exist than are absolutely necessary*, or as rephrased by Albert Einstein: *Everything should be made as simple as possible, but not simpler.* This principle has found its main application in supervised and unsupervised machine learning through regularization theory [Chen and Haykin, 2002] which enforces smoothness and simplicity on the solutions of a problem. However it may also be argued that Nature can be described in a similar way. We shall thus apply this principle in the choice and hypothesis made throughout this dissertation by keeping them as simple as possible. Although we are aware that a multitude of mechanisms may be involved in feature extraction and classification as suggested by [Ashby and Ell, 2001], we take the following

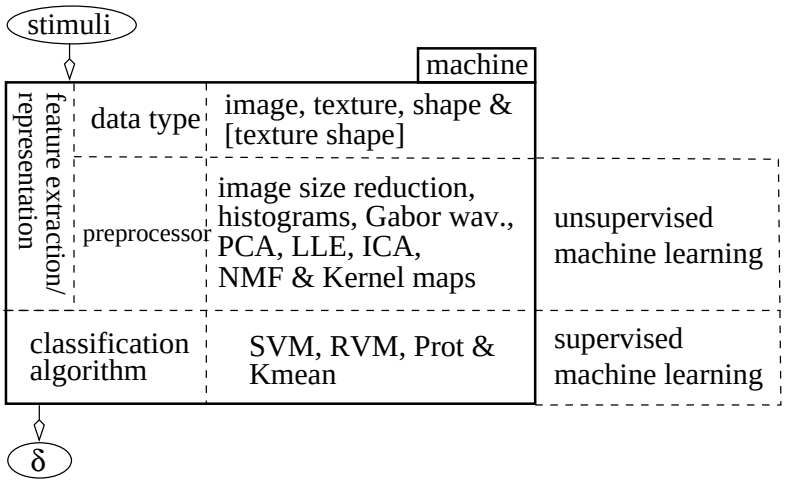


Figure 1.2: Flowchart of the machine part of the experiment with a list of the algorithms chosen for feature extraction and for classification.

approach: one model is used to explain feature extraction and one linear model is used for classification. Both models are as simple as possible, and most of them are consequently linear.

We here show and demonstrate a methodology, PSYCHO ML, and for this we use a human face database. This framework can however be adapted to other types of stimuli in a straightforward manner. The choice of the face stimuli is due to the fact that face stimuli are clearly one of the most biologically meaningful stimuli for humans. Humans see faces from the day they are born and face preference was suggested to be present before the birth of the child [Bednar and Miikkulainen, 2003], the latter remaining however able to learn new ones at all ages. This may show the coexistence of fixed and flexible internal preferences for faces in the human brain. This is chronologically the first learning i.e. classification task in our lifetime. Further, gender classification is an important biological tasks since it is a prerequisite for mating and thus one can expect humans to be good at it. Further, face recognition gains in importance in computer vision and other industrial applications for purpose of surveillance [Phillips, Grother, Micheals, Blackburn, Tabassi, and Bone, 2003]. While humans perform generally well in gender discrimination, machines have been found to be relatively bad at this task [Graf and Wichmann, 2004, Bromley and Säckinger, 1991, Gray, Lawrence, Golomb, and Sejnowski, 1995, Blackwell, Vogl, Dettmar, Brown, Barbour, and Alkon, 1997]. Ever since the beginning of machine learning, gender discrimination has thus been a central theme and a benchmark for any computer vision system. Finally, the results obtained here for faces may generalize to other object types since faces and non-faces have been argued to share common

early visual processing stages and to rely on similar mechanisms [Gauthier, Curran, Curby, and Collins, 2003]. Using functional Magnetic Resonance Imaging studies in humans and electrophysiology in monkeys, faces have been demonstrated to be not "special" and that other objects can be represented by similar neural activities [Gauthier and Logothetis, 2000].

Dissertation Structure

The dissertation is structured as follows. The MPI face database is described in Chapter 2 and the inhomogeneities in texture and shape of the faces are removed in a cleaning step. The data type and preprocessors, composing together the feature extraction step, are subsequently described. Considerations on the sparseness of the so-obtained encodings are given and their applicability in a gender discrimination task is studied. Chapter 3 deals with the analysis of the psychophysical classification experiments and represents the foundation for the studies which will follow. In Chapter 4 the methodology allowing to bridge the gap between psychophysics and machine learning is detailed, the choice of classifiers is motivated and some corresponding experimental paradigms are outlined. The classification behavior of man and machine is first compared and then correlated in Chapter 5, while stability criteria on the subjects' responses allow to validate the corresponding findings and the reproducibility of the results. These findings are discussed and related to other studies in machine learning, in psychophysics, and in neurophysiology. In Chapter 6 alternative approaches to analyze the data of man and machine are introduced, allowing us to design a novel psychophysical discrimination experiment where the hypothesis from machine are tested experimentally. The possible extension of models from machine learning to explain memorization of visual stimuli by humans is considered in Chapter 7 and Chapter 8 concludes this dissertation.

Chapter 2

The Database and its Encodings

In this chapter the Max Planck Institute (MPI) human face image database is described and inhomogeneities in texture and shape are eliminate in a cleaning step. Feature extraction is defined using various data types provided by the MPI database and various preprocessors from unsupervised machine learning. The resulting image bases are discussed, in particular their spatial sparseness. The sparseness of the encodings as much as their discriminability in gender classification studies are subsequently analyzed.

2.1 The MPI Face Database

When considering an image database of human faces, it is important to notice that faces differ in texture and shape information. The parametric face modeling technique developed by [Blanz and Vetter, 1999] allows to create novel, textured 3D faces from a database of existing 3D laserscans (geometrical and textural data) of 200 individuals (real human faces). Each face in this model is represented by a texture and shape vector thus defining a vector space of faces. In order to create this vector space, each 3D head was brought into a pixel-by-pixel correspondence with an internal average reference head in a bootstrapping procedure with the help of a dense optical flow process and fine-tuning by hand. A dimensionality reduction (Principal Component Analysis, PCA) on the optic flow data yields 200-dimensional texture and shape vectors. Both of these vectors define the face vector space. Multidimensional 3D morphing of faces, i.e. the generation of new faces can now be done simply by linearly combining elements in the face space. The morphing capacities of the MPI face database[1] have been shown to be very efficient in the context of face recognition studies [Blanz and Vetter, 2003]. Morphable

[1]To be found at `http://faces.kyb.tuebingen.mpg.de` .

models of the faces of the database are fitted to the studied face by estimating the texture and shape information, allowing thus to learn class-specific information. The advantages of the above correspondence representation over a pixel-based representation are reported in [Vetter and Troje, 1997] both on a computational and on a psychophysical ground: the texture & shape framework was shown to allow a better generalization to new faces and a better reconstruction of faces from a low-dimensional representation such as PCA.

2.2 Cleaning of the Database

We consider in this study a greyscale version of the faces in the MPI database. Indeed, there is some evidence that color cues do not affect significantly rapid scene categorization in man and monkeys [Delorme, Richard, and Fabre-Thorpe, 2000]. In the similar context of images of natural scenes, it was demonstrated that the color cue does only slightly improve the recognition memory of human subjects [Wichmann, Sharpe, and Gegenfurtner, 2002]. It was argued that this is mainly due to the fact that color increases attention and improves segmentation. Since we here want to make the gender classification task of intermediate difficulty, the removal of the color cue is justified. Frontal 256x256 pixels 8-bit grey-scale ($[0, 255]$) views of the 200 heads composing the MPI face database are generated from the laser scans. The database is gender-balanced and contains 200 Caucasian faces. It was shown by [Mamassian and Goutcher, 2001] that the human visual system uses a prior knowledge, which was modeled using a Bayesian approach by [Mamassian and Landy, 1998] in the context of line drawings, on the illumination position. In particular the light is assumed to come from above and slightly off-center. Thus the faces from the MPI face database are rendered under above and off-center illumination (elevation $\Theta = 65°$ and azimuth $\Phi = 40°$) to take into account the natural bias of the human visual system. A selection of faces as described above is shown in the first two columns of Fig.2.1. The mean and standard deviation of the intensity of the face in the image, its pixel-surface area and its location are represented in the first column of Fig.2.2. The following inhomogeneities in shape and texture can then be observed, as already mentioned in [Graf and Wichmann, 2002]:

1. the male faces are *darker* than the female ones on average
2. the male faces are *larger* than the female ones on average
3. the faces are not *centered*

We suppose that the above are the source of the problems encountered by [O'Toole, Vetter, and Blanz, 1999] when finding a discrepancy in the recognition between the male and the female data. In the cleaning of the

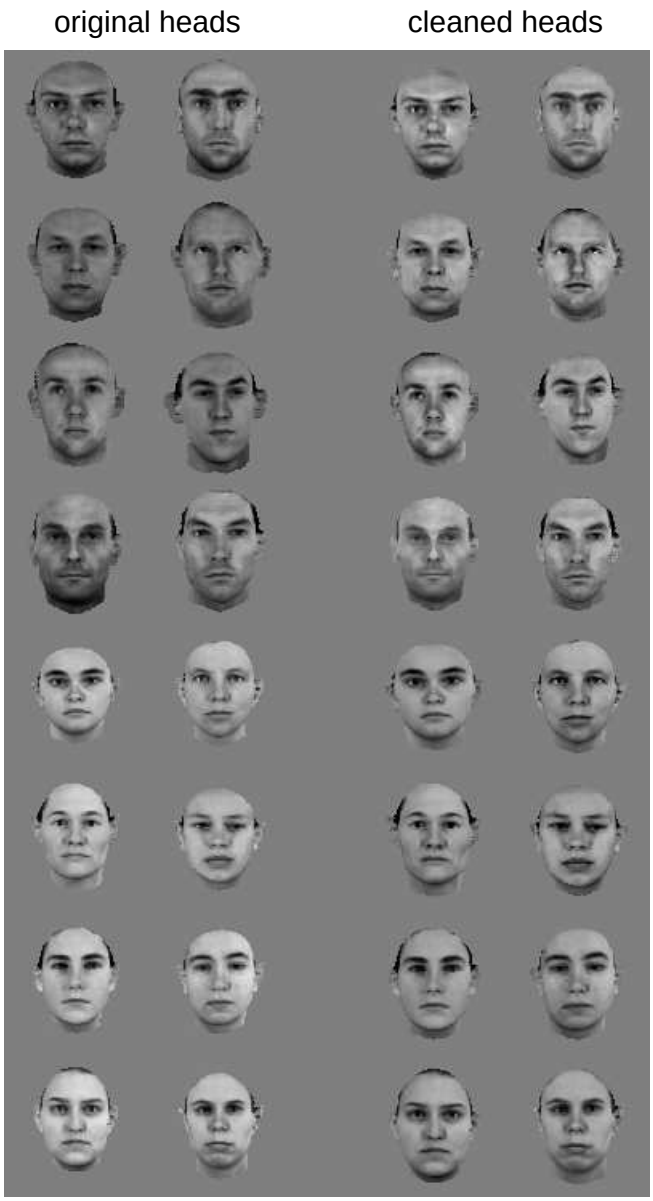


Figure 2.1: Comparison of some faces from original (two left columns) and cleaned (two right columns) databases. The first four rows represent male faces whereas the last four show female faces.

database, the above cues are eliminated since they are too obvious for man and machine and create too great homogeneities for an artificial learning machine. Moreover, these cues cannot always be considered as biologically relevant in a real environment. The *luminance* cue in determining gender is

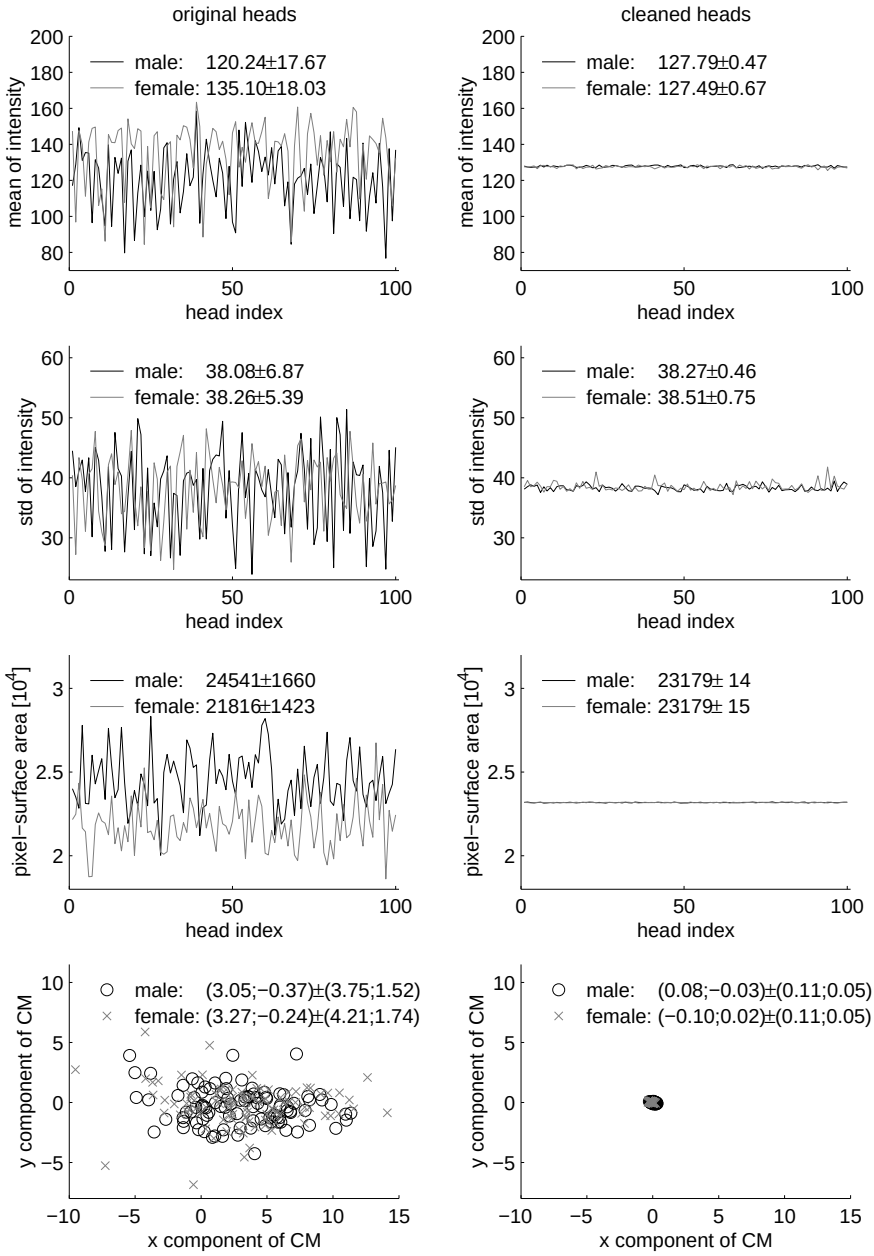


Figure 2.2: Comparison of parameters of faces from original (left column) and cleaned (right column) databases. These parameters are the mean and standard deviation of the intensity of the face in the image (first and second rows), its pixel-surface area (third row) and the offset of the center of mass (CM) of the face with respect the center of the image (fourth row).

certainly present but not crucial since the determination of gender cannot depend upon the illumination conditions of the subject i.e. recognition of the gender must be as efficient in a shady as in a bright environment. The *size* cue is obviously not important since the latter varies as function of the distance between observer and subject.

The faces are therefore processed, or cleaned, such that all faces have the same mean and standard deviation of the intensity, same pixel-surface area and are centered, the latter allowing to increase the homogeneity among the images. Setting the standard deviation of the intensity is almost a histogram equalization. The latter reveals however to be a non-unique iterative procedure which introduces visible inhomogeneities on the face and was thus not used to clean the database.

One of the central parts of this cleaning procedure is the generation of a "mask" for each face. This mask is obtained by generating a face from the MPI face database with a background set to 0 and a face in the image whose intensity is set to 255. Instead of cleaning the face images themselves, this procedure is applied to the texture and shape coefficients as explained below and shown in Fig.2.3. In other words it is assumed that the cleaning

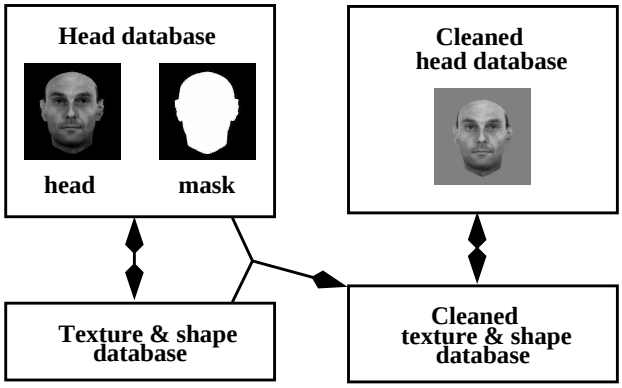


Figure 2.3: General flowchart of the cleaning procedure of the MPI face database. A head and its mask are extracted from the "raw" database. The corresponding texture and shape coefficients are computed and cleaned according to the parameters of the "raw" face image. Finally a cleaned database is generated from the cleaned coefficients.

of the heads in the images is equivalent to the cleaning of the corresponding texture and shape coefficients. The cleaning of the MPI face database is then done as follows:

1. Generate the *raw* 8 bit greyscale head images with their masks
2. Determine the pixel-surface area of each head and the position of its center with respect to the center of the image using the masks

3. Center the heads and their masks individually such that each head is centered on the center of the image

4. Resize the heads and their masks individually such that each head has the same pixel-surface area, this size corresponding to the mean pixel-surface area over all heads before resizing

5. Compute the intensity parameters for each head as follows:

   (a) compute the mean intensity of each face in the image using its mask

   (b) set the background of the image to the mean over all faces of the mean intensity of each face

   (c) adjust the mean intensity of each face to the value of the background

   (d) compute the standard deviation of the intensity of each face in the image using its mask

   (e) adjust the standard deviation of each face to the mean over all faces of the standard deviation of the intensity of each face in the image

Using the parameters computed in the above cleaning of the images, the texture and shape coefficients can then be cleaned as follows:

1. Generate the *raw* texture and shape coefficients for each head

2. Center and scale the flowfields around the center of the reference head

3. Adjust the mean and standard deviation of the texture of each head, and set the background

4. Generate the *cleaned* heads from the *cleaned* coefficients

Fig.2.1 shows the original and the cleaned faces: the effect of the cleaning step is visible. Furthermore the parameters (intensity, size and position of center) of the original and the cleaned database are compared in Fig.2.2, demonstrating quantitatively the cleaning process.

2.3 Feature Extraction

We here present various manners to represent the MPI face database prior to its use for classification by the machine. The data after feature extraction is referred to as *encoding*. A data type is extracted from the MPI face database. Subsequently the preprocessors perform feature extraction in the machine learning sense by finding different types of representations

or embeddings of the data. For the purpose of numerical tractability, all the considered preprocessing algorithms also perform dimensionality reduction in the sense that they find an encoding of the data of lower dimensionality than the original data. An overview is given in Fig.2.4. The context con-

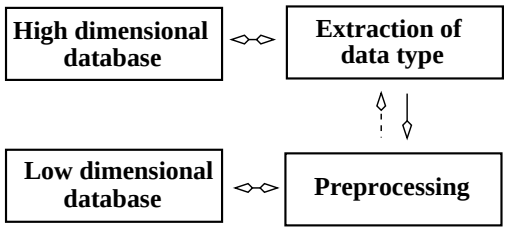


Figure 2.4: General flow chart of the representation of the MPI face database. A data type is extracted form the high dimensional database. A preprocessor is applied to this data type and yields the low dimensional database of the encodings. The dashed arrow indicates a procedure which is not possible for all types of preprocessors. The choice of the data types and preprocessors is explained in the text.

sidered here is thus different from the one encountered in early vision where preprocessing is used to make the input sparse rather than low-dimensional. From a biological point of view, the question whether we could treat the preprocessing algorithms considered below as a combined low-level (sparse encoder) *and* mid-level (dimensionality reduction) preprocessor arises. We finally discuss and study properties of the preprocessors and conclude by considerations on the sparseness of the encoding obtained from the various types of preprocessing and on the linear separability of the encodings with respect to gender classification.

2.3.1 Data Types

In the context of the present studies, we consider the following data types:

- the image vector $\mathcal{I} \in \mathbb{R}^{256^2}$ resulting from the pixel matrix of the image
- the texture vector $\mathcal{T} \in \mathbb{R}^{1 \cdot 256^2}$ of the face in the image
- the shape vector $\mathcal{S} \in \mathbb{R}^{2 \cdot 256^2}$ of the face in the image
- the vector resulting from the combinations of the texture and shape vectors of the face in the image as $[\mathcal{T}\mathcal{S}] \in \mathbb{R}^{3 \cdot 256^2}$ where $[\cdot]$ represents the vector concatenation operator. Notice that any permutation in this concatenation does not change the amount of information carried by the vector and is thus not considered.

All the information conveyed in the image, texture or shape vectors is rescaled to the range $[-1, 1]$ as below:

$$\vec{x}_i \leftarrow 2\frac{\vec{x}_i - \min_i \vec{x}_i}{\max_i \vec{x}_i - \min_i \vec{x}_i} - \vec{1} \quad \in [-1\ 1] \tag{2.1}$$

where $\vec{x}_i$ is a vector from the dataset. The above amounts to scale and translate identically all the elements of the dataset. This procedure aims at presenting the preprocessors with data of the same order of magnitude. This is especially important when combining the texture and shape information in order to give both cues the same weight since we have no *a priori* knowledge on their mutual importance.

2.3.2 Preprocessors

We here deal with the preprocessing algorithms $\mathcal{P}$ which are applied to the data types. The operators $\mathcal{P}$ compute an embedding and a dimensionality reduction of the data. Below we mention and motivate the choice of the preprocessing operators $\mathcal{P}$, some of them being described thoroughly in Appendix A:

- *Principal Component Analysis* (PCA, [Duda, Hart, and Stork, 2001]): eigenvalue decomposition of the data along the directions of largest variance in the data.
- *Locally Linear Embedding* (LLE, [Roweis and Saul, 2000]): neighborhood-preserving dimensionality reduction.
- *Independent Component Analysis* (ICA, [Cardoso, 1998, Bartlett, Movellan, and Sejnowski, 2002]): extraction of statistically independent variables from the data. For ICA I the independence of the patterns (the images) is maximized whereas for ICA II, the components of the patterns (the pixels) are made independent.
- *Non-negative Matrix Factorization* (NMF, [Lee and Seung, 1999]): decomposition of data using only non-negative values.
- *Empirical Kernel Maps* using Radial Basis kernel functions (RBF) [Schölkopf and Smola, 2002]: eigenvalue decomposition of the data using a nonlinear extension of the Gram matrix of the data.
- benchmarks in artificial computer vision: *image size reduction* and *intensity histograms* only of the face in the image.
- benchmark in biologically-inspired computer vision and a crude approximation to human/monkey early vision models: *Gabor wavelet filters* [Hubel and Wiesel, 1962] followed by a linear empirical kernel map in order to reduce the dimensionality of the feature vector.

We set the number of nearest-neighbors of LLE to be 15 according to [Graf and Wichmann, 2002]. The number of iterations of the NMF algorithm is set to 1000 for the sake of numerical tractability. The Gabor filters are considered on 3 scales and for 6 orientations.

We have 200 different faces in the database and thus the intrinsic dimensionality of the database is 200. The above preprocessors are then used to create an encoding of the data type of dimensionality 200, except for image size reduction and the histograms where this encoding is of dimension $16 \times 16 = 256$ and $2^8 = 256$ respectively as summarized in Table 2.1. The

preprocessor	$\mathcal{P}(\mathcal{I})$	$\mathcal{P}(\mathcal{T})$	$\mathcal{P}(\mathcal{S})$	$\mathcal{P}([\mathcal{TS}])$
PCA	200	200	200	200
LLE	200	200	200	200
ICA I & II	200	200	200	200
NMF	200	200	200	200
Kernel Maps	200	200	200	200
Image size reduction	256	-	-	-
Histograms	256	-	-	-
Gabor Wavelets	200	-	-	-

Table 2.1: Possible combinations of preprocessors with data types and the dimensionality of the resulting encodings.

preprocessors PCA, ICA I & II and NMF are invertible in the sense that the data can be reconstructed from the encodings (see Sec.2.3.3). In these cases, we may thus expect to have a perfect reconstruction—no reconstruction error—since we keep all 200 components of the encodings.

One of the first attempts to use PCA in the context of face recognition on the pixel information is due to [Sirovich and Kirby, 1987]. Related studies have given rise to the so-called "eigenface" representation [Turk and Pentland, 1991] and attempts have been made to biologically motivate this type to representation using for instance artificial neural networks (ANN). An overwhelming amount of literature on gender classification using PCA applied on the image data type followed these two papers. PCA and an autoencoder ANN were used in [Valentin, Abdi, Edelman, and O'Toole, 1997] to study facial analysis. The robustness of this approach (assessed using the stability of the eigenvectors) as much as its ability to perform novelty detection—in this context the recognition of faces from another race or the so-called "other race effect"—were studied. The eigenvector corresponding to large eigenvalues were shown to contain low frequency information, to be stable and allow good generalization, the contrary being true for those corresponding to small eigenvalues. The eigenvectors of small eigenvalue were shown to be unstable i.e. vulnerable to degradation. This inspired the authors to lesion i.e. perturb the autoencoder ANN and its behavior

was qualitatively compared to that of human neurological deficits. This comparison made the authors conclude on the biological plausibility of the encoder network. In the framework of this thesis, we propose a more rigorous study relating man to machine. The effect of the eigenvalues are further studied in [O'Toole, Abdi, Deffenbacher, and Valentin, 1993] in the context of face recognition. The small eigenvalues were shown to account for face recognition whereas the large ones for gender classification. Thus, the selection of a subset of the eigenvalue spectrum is task-dependent. This finding has a limited impact on this study since we consider the whole spectrum.

PCA and LLE are compared in [Graf and Wichmann, 2002] on a gender classification task using the MPI face database. While the optimal number of nearest-neighbors for LLE was determined for this database, it was shown that PCA clearly outperforms LLE. ICA I and ICA II have been shown to outperform PCA in the context of face recognition by [Bartlett, Movellan, and Sejnowski, 2002, Baek, Draper, Beveridge, and She, 2002]. The specific algorithm used to implement ICA, its architecture (ICA I or ICA II) and other algorithmic properties as much as the type of database are compared and benchmarked in [Draper, Baek, Bartlett, and Beveridge, 2002].

The kernel maps compare different stimuli using a Gaussian window (RBF function) i.e. they extract *global* similarities between different patterns. The Gabor wavelet filters on the other hand are applied on each stimulus and compare regions on each stimulus i.e. they extract *local* properties of each stimulus. Furthermore the RBF of the kernel map and the Gabor filter have very different spectral characteristics, some of the main properties of Gabor filters being reported in [Daugman, 1985]. These two preprocessors are thus fundamentally different. An elaboration of the Gabor filter model is the model by [Riesenhuber and Poggio, 2000, 2002]. This model performs feature extraction using a hierarchical tree of linear and non-linear "maximum"-like branches. The nodes of this tree are Gabor filters as considered here and we thus assume that Gabor wavelet filters represent this model in the context of the studies done in this thesis.

Other possible methods to represent the data and perform dimensionality reduction include most notably multi-dimensional Fisher Analysis (see [Duda, Hart, and Stork, 2001] and [Lu, Plataniotis, and Venetsanopoulos, 2003] for a non-linear extension applied in the context of face recognition) which is intrinsically different from the above methods since it relies on the labels of the data points i.e. it is a supervised preprocessing. We do not consider this method for preprocessing in the context of this study since we need to clearly separate the (unsupervised) preprocessing step from the actual (supervised) classification procedure.

As an extension to the above choice of preprocessors, we could consider their combinations, the latter being popular in artificial computer vision and data analysis. Gabor wavelets, PCA and ICA have been combined by [Liu and Wechsler, 2003] in order to form an independent Gabor feature

detector for face recognition using a feature-based probabilistic classifier. The idea is to use ICA to reduce the redundancy of the information given by the Gabor filter. The framework for ICA I and ICA II has also been extended by [Bartlett, Movellan, and Sejnowski, 2002] in the context of face recognition by considering the combined ICA recognition system consisting of the two ICA representations. This combination was shown to yield better recognition than its components individually. The ICA problem has been solved using non-negativity constraints on the sources similarly to NMF in [Plumbley, 2003] and yields the nonnegative ICA algorithm. This algorithm is based on nonlinear PCA and has been tested in the context of image and audio analysis. The more stages the preprocessor has, the more difficult it is to infer biological plausibility. Moreover combining many preprocessors may result in overfitting and may thus violate the general principle of "keeping it simple", also known as Occam's razor [Duda, Hart, and Stork, 2001]. We shall thus not consider combinations of preprocessors, unless unavoidable such as for the linear Kernel map following Gabor filters or the PCA step prior to ICA (see Appendix A).

Finally, following Occam's principle, we do not consider non-linear extensions of the above preprocessors i.e. we do not kernelize them. Indeed, such an extension would introduce a difficulty in the interpretation of the forthcoming results of this study since it relies on the mapping to a possibly unknown high-dimensional feature space. This approach has however been successfully used in computer vision (see [Schölkopf, Smola, and Müller, 1998] for PCA).

2.3.3 Discussion

The preprocessing/decomposition algorithms PCA, ICA I & II and NMF are linear and can thus be written in the form:

$$\mathcal{X} = BE \tag{2.2}$$

where $\mathcal{X}$ is the original data matrix, B is the basis matrix and E the matrix of encodings. Once B and E are determined by the algorithm, the original data may be reconstructed. In other words, these algorithms are invertible and the quality of the decomposition can be assessed by computing the reconstruction error. In Appendix E the reconstruction error, a sample of original and reconstructed heads and a subset of the basis vectors are displayed for the various data types. Moreover, for PCA, the eigenvalue spectrum and the cumulative variance are also displayed. For convenience, Fig.2.5 presents the first four basis elements of the above preprocessors for the image data as also shown in Appendix E. This figure allows to assess and visualize the sparseness in space of the preprocessing algorithms i.e. the sparseness of the basis vectors. We further discuss below the main results for each type of preprocessing and define the reconstruction error as the

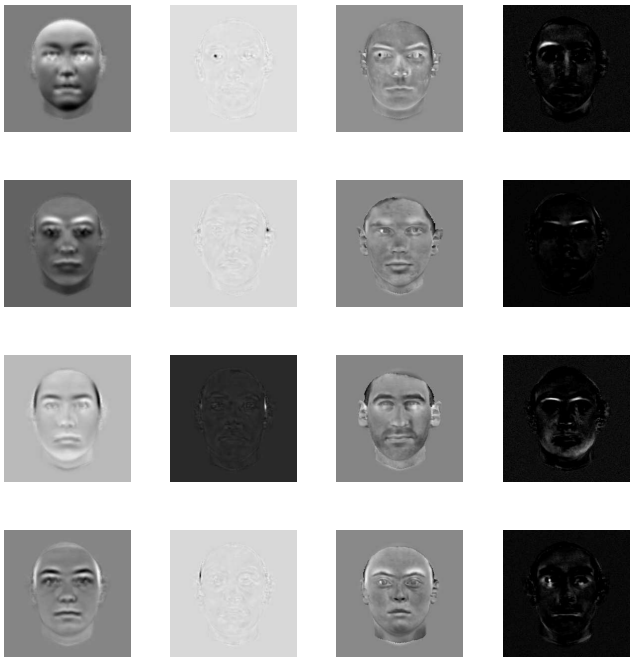


Figure 2.5: Four faces of the basis for the image data type for PCA, ICA I & II and NMF (first to fourth column).

Euclidean distance between two faces in their pixel representation. This distance is however different from the perceptual distance between two faces[2] which can only be assessed using psychophysical experiments as done for instance in [Vetter and Troje, 1997]. Thus we also plot some original and reconstructed images in Appendix E in order to cover both the Euclidean and perceptual distances.

For *PCA*, the eigenvalue spectrum is decreasing and has no flat regions for all data types. The PCs find (orthogonal) directions which explain the data well. The cumulative variance plot for the shape data indicates that only a few number of PCs explain almost all the data. This is not the case for the image and texture data. The combination of texture and shape data then yields, as could be expected, an intermediate behavior. PCA is thus well suited to describe the shape data. As could be expected, for all data types the reconstruction error is nil when considering all the 200 PCs. The decreasing behavior of the reconstruction error is similar for all data types. This perfect reconstruction is confirmed when comparing the original and reconstructed images. When considering the face basis obtained from PCA, we see that for all data types, this basis is *holistic* and non-sparse

[2]For example, the perceptual distance between a face in light and dark conditions is close to nil whereas the Euclidean one is high.

(see also Fig.2.5). In the case of the image or texture data, the intensity of the image is globally changed. In the case of the shape data, the face is globally distorted. For the texture and shape data, a combination of both behaviors appears. It is difficult to draw further conclusions since the unconstrained nature of the encodings allow for complicated addition and subtractions from these basis images to recreate one face from the database.

For *ICA I*, the reconstruction error is nil for all data types, suggesting thus a perfect reconstruction of the data. This is confirmed by considering the original and reconstructed heads. The basis for all data types is clearly sparse and *part-based* in a highly-localized manner (see also Fig.2.5). In other words, the parts are reduced to very small surface subsets of the image. For the image and texture data, local intensity spots illustrate this part-based behavior. For the shape data, the part-based behavior translates to small deformation of the face and to strong changes of the position of the face in the image. For the combination of texture and shape information, an intermediate behavior is observed.

As for ICA I, the reconstruction is perfect for *ICA II* as testified by the value of the reconstruction error and the images from the original and reconstructed dataset. The basis of faces is here however clearly *holistic* (see also Fig.2.5) and similar considerations as for PCA apply here. However, the holistic nature of the basis vectors is here much more pronounced than for PCA and yields quite similar results as Vector Quantization [Duda, Hart, and Stork, 2001, Haykin, 1999]. Such a basis consists almost of "prototypes" of whole faces, and an extreme sparseness in the encoding can be expected in a winner-takes-all manner. Such a basis bears some similarities with the so-called "grandmother" cells discussed in [Barlow, 1972].

For *NMF*, we have a close to perfect reconstruction of the data as testified by the reconstruction error[3] and the images of the original and reconstructed heads. The quality of the reconstruction depends upon the choice of the maximum number of iterations, here 1000. This number could be increased in order to improve even further the reconstruction performance. However this would result in unacceptable high computational costs. The basis for all data types is clearly *part-based* and sparse (see also Fig.2.5). The pixel surface of the parts is here however larger than for ICA I, and carries thus more biologically-meaningful content. These parts correspond to what psychophysicist would call "features", for instance the eyebrows, the eyes or the nose. Similar considerations as for ICA I apply here, although the loss of centering is here less apparent.

In summary, the linear preprocessors can be described as follows. The basis images of PCA and ICA II are holistic whereas those corresponding to ICA I and NMF are part-based. The spatial sparseness—the regions of the

[3]The norm of the difference between the original and reconstructed image vectors is of the order of ~ 10, which is negligible for a vector of size $\sim 10^5$.

face used in the basis images—is very high for ICA I and lower for NMF. Finally, while the holistic basis images of PCA are slightly blurred, those of ICA II are almost patterns of the dataset.

The remaining non-linear preprocessors shown here cannot be formalized as PCA, ICA or NMF, i.e. they do not allow a decomposition of the data in an encoding and a basis matrix. Thus no basis of images exits, nor a reconstruction error. While *LLE* and *kernel maps* are applied to all data types, *image size reduction*, *histograms* and *Gabor wavelet filters* can only be applied on the image data. The real part of Gabor wavelet filters on 3 scales and for 6 orientations are displayed in Fig.2.6 and the magnitude of their application on one image of the database, the corresponding images being downsampled. The elliptic shape of the filters can be seen and repre-

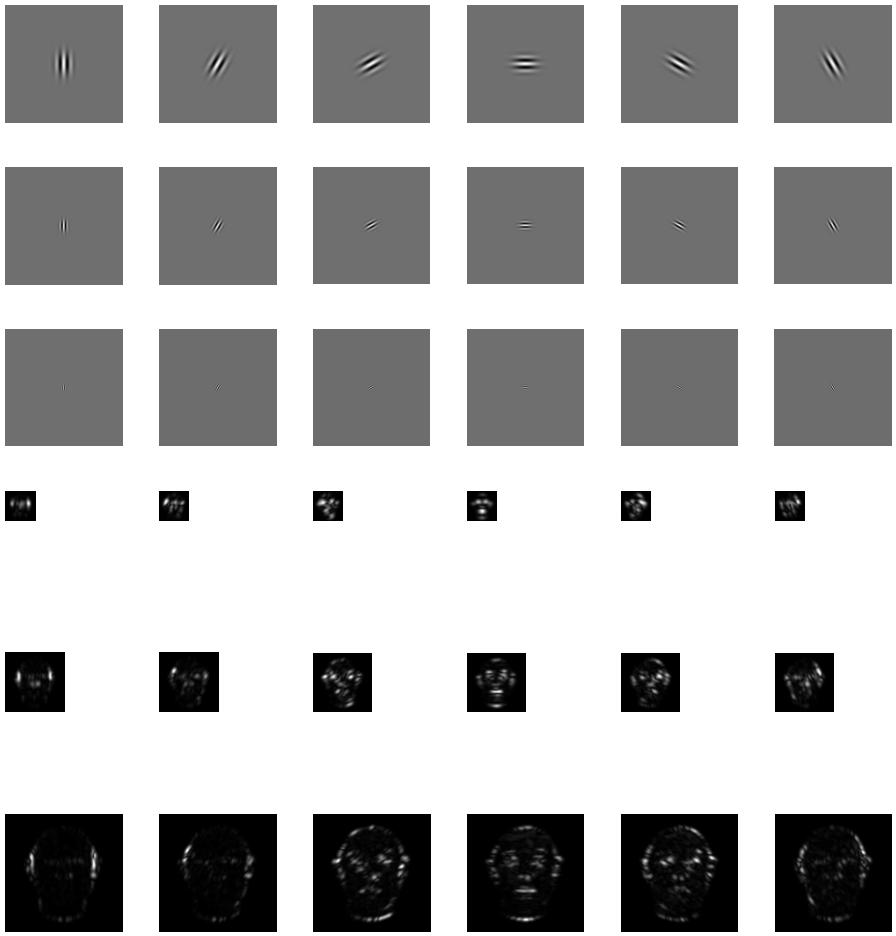


Figure 2.6: Real part of Gabor wavelet filters on 3 scales and for 6 orientations (first to third rows) and magnitude of their downsampled application to an image (fourth to sixth rows).

sents orientation selectivity. The higher the scale i.e. the smaller the spatial localization of the filter, the better a face can be recognized. At such scales, a Gabor filter acts as an edge detection algorithm: it finds the contours in the image. At lower scales the filtered segmented image gets more blurred, this face being used to proceed to downsampling.

2.4 Sparseness of Encodings

Sparseness is argued to be a biologically plausible description of efficient neural codes [Olshausen and Field, 1996, Willmore and Tolhurst, 2001]. In some of the above cases, the sparseness of the basis was apparent by considering its images: NMF and ICA I led to part-based (spatially sparse) bases whereas PCA and ICA II yielded a holistic (spatially non-sparse) basis. What about the sparseness of the corresponding encoding, i.e. of the low-dimensional representation? We study below this sparseness using the developments of [Willmore and Tolhurst, 2001]. The sparseness of a vector $\vec{x} \in \mathbb{R}^M$ is computed using its kurtosis as:

$$s = \left[\frac{1}{M} \sum_{i=1}^{M} \left(\frac{x_i - \mu}{\sigma} \right)^4 \right] - 3 \tag{2.3}$$

where μ, σ are respectively the mean and the standard deviation of $\vec{x}$. A sparseness of $s = 0$ indicates a Gaussian distribution of x_i whereas $s > 0$ represents a supergaussian (highly peaked distribution) and $s < 0$ a subgaussian distribution (distribution with flat regions). A distribution with many small values and only a few large ones yields a high value of s, whereas a uniform distribution has a small value of s. In oder to accommodate large variations in the order of magnitude of s, it is useful to define sparseness by applying a transfer function to the above definition as:

$$S = \log(|s| + 1) \tag{2.4}$$

Assume $\mathcal{X} \in \mathbb{R}^{p \times n}$ is the data matrix of the encodings given by the application of $\mathcal{P}$ on the data type, p being the number of patterns and n their dimension. In other words each row of $\mathcal{X}$ is a stimulus and each column is a vector representing a given dimension of all the encodings. In order to get an intuitive description of the concept of sparseness for the data matrix $\mathcal{X}$, we follow the lines of [Willmore and Tolhurst, 2001] and define each dimension of the data, i.e. each column of $\mathcal{X}$, as a *neuron*. We then have n neurons, each one producing p responses, and have the following two possible different definitions of sparseness:

1. the vector $\vec{x}$ is a column of $\mathcal{X}$. The resulting sparseness S_l is called *lifetime sparseness* and describes codes in which each neuron's lifetime

response distribution has high kurtosis i.e. the firing rate of each neuron is sparse. In other words, each neuron gets rarely activated, but when it fires, it produces a response of large magnitude. A lifetime-sparse coding is represented by neurons that respond to stimuli using spikes which occur rarely but convey a large amount of information. Lifetime sparseness can be seen as the sparseness in the n components of the encodings.

2. the vector $\vec{x}$ is a row of $\mathcal{X}$. The resulting sparseness S_p is called *population sparseness* and indicates that a small subset of neurons from a large population is active at any time in response to a stimulus. Also different subsets of neurons are activated by different stimuli. A population-sparse coding is energetically efficient since few neurons fire at any time; few neurons are necessary to convey the information relative to a stimulus. Population sparseness can be interpreted as the sparseness in the number p of the encodings.

The lifetime and population sparseness as defined above are represented in Fig.2.7 for all preprocessors and data types. It can be seen that both sparseness definitions of S are not correlated. The highest sparseness is achieved by ICA II both for S_l and S_p in all cases, indicating that this preprocessor yields the sparsest representation of all the considered ones. On the image data type PCA, the kernel map and Gabor filters have $S_p \gg S_l$ which indicating energetically-efficient encodings. For the other data types, PCA, the kernel map and ICA I also have $S_p \gg S_l$. This suggests that the components of the patterns are sparsely represented rather than the patterns themselves. This could be expected for ICA I by construction and for PCA this corroborates the findings of [Willmore and Tolhurst, 2001]. For LLE, we have $S_p \simeq S_l$ and neither of the codes is really sparse. NMF shows an intermediate degree of sparseness both for S_l and S_p. Furthermore, $S_l \simeq S_p$, indicating an encoding which saves energy while having an intermediate sparseness of the firing rate. This fact may hint at the fact that, overall, NMF may be a good candidate to encode visual information.

2.5 Discriminability of Encodings

The encodings of the faces i.e. the datasets obtained by combining the data types with the preprocessors shall be used in the context of this study for gender classification tasks. It is thus cautious to check whether the resulting two classes are not too much overlapping or intermingled, in which case the preprocessor shall not be considered further. For this, we proceed to a *Fisher Linear Discriminant Analysis* (FLDA, [Duda, Hart, and Stork, 2001]). FLDA seeks a direction in the dataset most efficient for discrimination—the Fisher direction—by maximizing a discrimination score

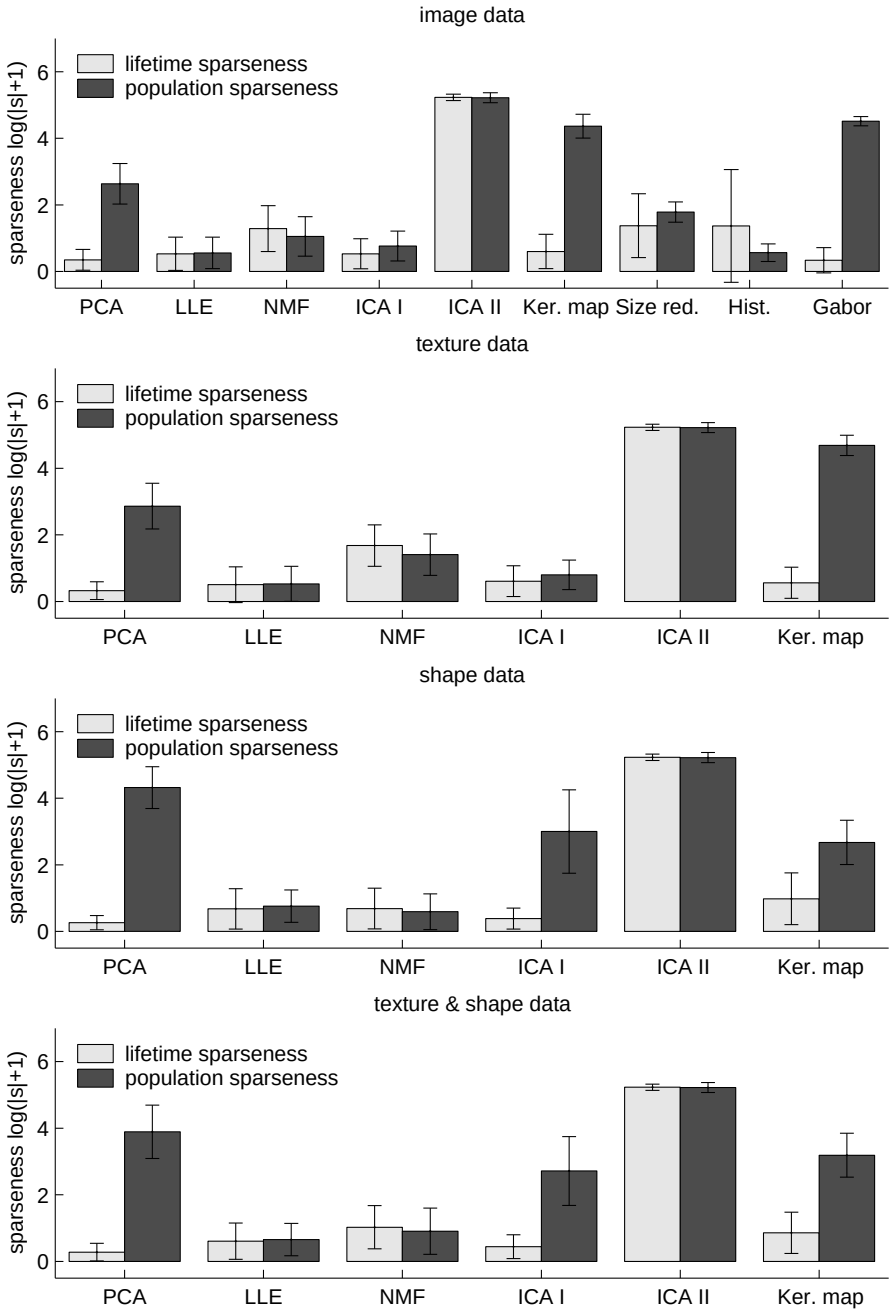


Figure 2.7: Lifetime and population sparseness for each data type and preprocessor.

based upon variances in each class and between each class. FLDA finally projects the data on this line, yielding a one-dimensional representation al-

lowing to visualize discriminability. Clearly the direction of the Fisher line for which the projected data is well separated is the result of an optimization process, which reveals to be solvable analytically.

For this, we first compute the means (or center-of-mass) of each class:

$$\vec{\mu}_{\pm 1} = \frac{\sum_{i|y_i=\pm 1} \vec{x}_i}{\sum_{i|y_i=\pm 1} 1} \tag{2.5}$$

where $\vec{x}_i$ for $i = 1, \ldots, p$ are the encodings of each pattern and y_i their class. The latter allow to determine the within-class scatter matrix:

$$S_w = \sum_{k=\pm 1} \sum_{i|y_i=k} |\vec{x}_i - \vec{\mu}_k\rangle\langle\vec{x}_i - \vec{\mu}_k| \tag{2.6}$$

and the in-between class scatter matrix:

$$S_b = |\vec{\mu}_{+1} - \vec{\mu}_{-1}\rangle\langle\vec{\mu}_{+1} - \vec{\mu}_{-1}| \tag{2.7}$$

Using the lines of [Duda, Hart, and Stork, 2001], the direction of best separability is then given by:

$$\vec{w} = S_w^{-1}(\vec{\mu}_{+1} - \vec{\mu}_{-1}) \tag{2.8}$$

The projection of a data point $\vec{x}$ on this line is then given by $\langle\vec{w}|\vec{x}\rangle$. The above allow the computation of the Rayleigh quotient or Fisher linear discriminant:

$$J(\vec{w}) = \frac{\langle\vec{w}|S_b\vec{w}\rangle}{\langle\vec{w}|S_w\vec{w}\rangle} \tag{2.9}$$

The above is a measure of the discriminability of the two classes. There exists a non-linear extensions of the above using the kernelization of the dual space expression of FLDA [Mika, Rätsch, and Müller, 2001].

In practice, computing J over the whole dataset would result in overfitting and thus bad generalization ability. We thus evaluate J using a 5-fold cross-validation scheme as below:

1. randomize data
2. perform a cross-validation scheme as:
   (a) use the training set to compute $\vec{w}$
   (b) use the testing set to compute S_w and S_b
   (c) use $\vec{w}$, S_w and S_b to assess J
   (d) project the testing data onto $\vec{w}$
3. compute mean and standard deviation of J

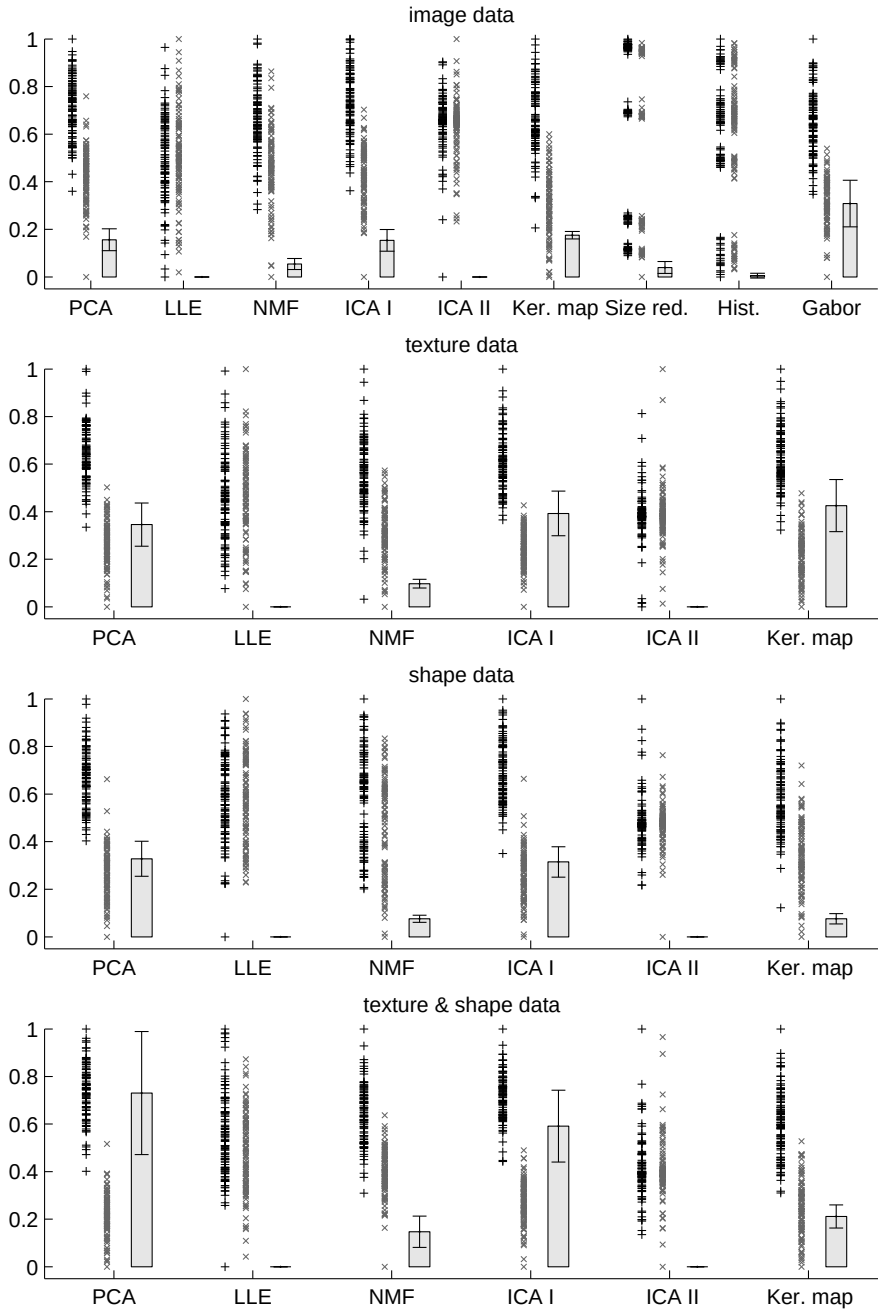


Figure 2.8: Projection of the dataset along the Fisher direction (data rescaled to $[0, 1]$) and Rayleigh quotient J for each data type and preprocessor.

The projected classes and the corresponding Rayleigh quotients are represented in Fig.2.8 for each data type and preprocessor. The score J is a measure of how well the data is linearly separable in the space spanned by the preprocessor. When considering the image data, Gabor wavelets allow the best discriminability of the two classes. We can notice that overall PCA and ICA I show for all data types the best discriminability: the value of J is high and the projections of the data are distinct. These preprocessors are thus globally most well suited for classification studies, especially on the data types different from the image one. The discriminability of kernel maps, although similar to the one of PCA and ICA I on the image and texture data, is less good on the shape and texture & shape data. The discrimination for NMF, although possible, is not as good as for the above preprocessors for all data types. For the image data, size reduction and histograms are poorly adapted for classification studies. Most importantly, LLE and ICA II show no discriminability at all for all cases: $J = 0$ and the projections of the classes are totally overlapping. It was also verified that applying classifiers on these datasets resulted in a classification error of $\sim 50\%$, which is chance (results not reported here). These types of preprocessors are thus not adapted for classification studies when considering the face database. In particular, classification using a linear decision function is not possible i.e. the data seems to be linearly not separable in the space span by these preprocessors. They are thus not considered any further in this dissertation.

Chapter 3

Human Classification Behavior

This chapter deals with the description and analysis of two psychophysical classification experiments. Human subjects are asked to classify images of faces according to their gender and their responses are subsequently analyzed. Both experiments present the same stimuli but in a different order, allowing us to assess the consistency of the subjects' responses and the reproducibility of the results which will follow in the next chapters of this dissertation.

3.1 Classification Experiment I

In this first psychophysical experiment, 55 human subjects were asked to classify sequentially 152 from a possible 200 faces from the MPI face database (see Chapter 2) according to their gender. We recorded three responses: the estimated gender (i.e. female/male) with the corresponding reaction time (RT) and subsequently a confidence rating (CR= $1, 2, 3$) on a scale from 1 (unsure) to 3 (sure). Stimuli were presented against the mean luminance ($50\,\mathrm{cd/m^2}$) of a carefully linearized Clinton Monoray CRT driven by a Cambridge Research Systems VSG 2/5 display controller. Neither male nor female faces changed the mean luminance. Subjects viewed the screen binocularly with their head stabilized by a headrest. The temporal envelope of stimulus presentation was a modified Hanning window (a raised cosine function with a raising time of $t_{transient} = 500$ms and a plateau time of $t_{steady} = 1000$ms, for a total presentation time $t = 2000$ms per face) to avoid aftereffects of the stimuli on the subjects' retinae. After the presentation, an empty screen with mean luminance was presented for 1000ms before the presentation of the following stimulus. Subjects were asked to classify as fast as possible to obtain perceptual, rather than cognitive, judgments. Most of the time they responded well before the presentation of the stimulus had ended (mean RT over all stimuli and subjects was approximately

900ms). The subjects were however not constrained on the time needed to indicate their confidence. All subjects had normal or corrected-to-normal vision and were paid for their participation. No feedback upon the correctness of the subjects' answers was provided. Most of them were students from the University of Tübingen and all of them were naive to the purpose of the experiment. A training phase of 8 faces (4 male and 4 female faces) precedes the actual classification experiment in order to acquaint the subject with the stimuli and with the setup: the corresponding results are discarded. Details about the experimental setup may be found in appendix D. This classification paradigm bears some similarities with the one considered in [O'Toole, Defenbach, Valentin, McKee, Huff, and Abdi, 1998], the latter using however less well-controlled stimuli and less adequate timings, both for the stimulus presentation and for the recording of the subjects' responses.

We present below the evaluation of the data from this psychophysical classification experiment. Analysis of the classification performance of humans is based on signal detection theory as presented in Appendix C. When averaged across subjects, we get the discriminability d' and the male bias $\log(\beta)$ as reported in the first row of Fig.3.1 for all subjects, for male and for female subjects. The values of d' indicate that the classification task is comparatively easy, despite the fact that the images of the faces have no hair, make-up ..., although without being trivial (no ceiling effect). We also observe a strong male bias (a large number of females classified as males but very few males classified as females). There are thus more misclassifications for female stimuli than for male ones, similarly to what was observed by [O'Toole, Defenbach, Valentin, McKee, Huff, and Abdi, 1998]. There is no significant difference in discriminability or in male bias across the subjects' gender and thus, in the rest of this study, we consider all subjects.

The plots of the second row of Fig.3.1 show the correlations of RT and classification error, classification error and CR, and RT and CR for all stimuli, for only the male and only the female stimuli. The error bars represent the standard error around the mean. First, RTs are longer for incorrect answers than for correct ones, reflecting a longer processing of difficult information. Second, a high CR is correlated with a low classification error and thus subjects have veridical knowledge about the difficulty of individual responses—this is certainly not the case in many low-level psychophysical settings. Third, the RT decreases as the CR increases, i.e. stimuli easy to classify are also classified rapidly. It may thus be concluded that a high error (or equivalently a low CR) implies higher RTs. This may suggest that patterns difficult to classify need more computation, i.e. longer processing, by the subjects' brain than patterns easy to classify. Moreover subjects are less confident when classifying difficult stimuli. We now consider the above analysis for all stimuli and for each gender separately. When considering all data, we get, as expected, the mean trend in the subjects' responses. We see that when making no error, the subjects respond faster for male than

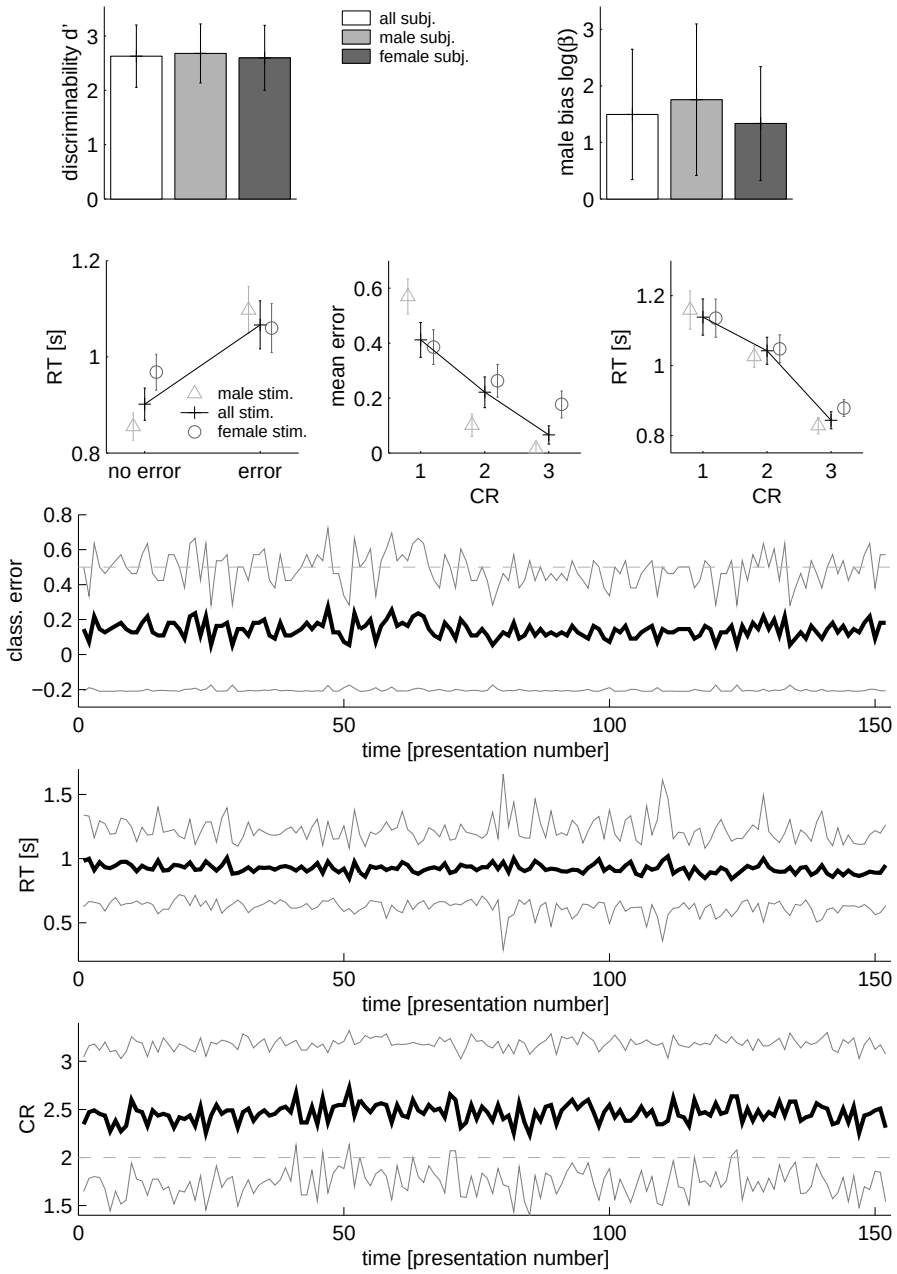


Figure 3.1: Human classification behavior. First row: discriminability and male bias. Second row: mutual dependencies of the subjects' responses. Third, forth and fifth rows: mean temporal evolution of these responses, the thick curve indicating the mean, the grey one the standard deviation and the horizontal lines indicate chance level.

for female stimuli. Moreover, the subjects take longer to make a mistake for male than for female stimuli. In other words, they only make a mistake for a male face after long cerebral processing. For male stimuli, when subjects are confident, their decision is almost always correct. This effect is not as clear for female stimuli. However, unsure subjects mainly misclassify male stimuli, whereas female stimuli are better classified. Finally, for high CR, male stimuli are classified faster than female ones. It may be concluded that male faces are easier to classify (less error, faster and higher confidence) than female ones. This corroborates the above finding of a male bias in the subjects' responses.

One of the crucial assumptions for our enterprise of comparing man to machine is the *stability* of the internal representation of faces in the subjects' brain. In other words, the exposition of the subjects to face stimuli should not modify their internal representation acquired during their lifetime. Consequently there should be no transient learning dynamics on the timescale of this experiment. Indeed, would the subjects' responses vary as the experiment proceeds, this would indicate that they create or refine their internal face representation. The modeling of the human classification behavior would then require different types of classification algorithms than those presented in Chapter 4, namely *on-line* learning algorithms. The stability of the internal face representation of the subject is thus a key hypothesis of the present study. The third, fourth and fifth row of Fig.3.1 address this point by showing the mean of the subjects' responses (thick line) and the corresponding standard deviation (thin line) as function of the time of appearance of the stimuli i.e. their presentation index, the horizontal lines indicating the chance level. It may be concluded for each response that there is indeed no learning during the experiment. The subjects do not get "experts" and do not learn some features allowing to enhance their classification performance, RT or CR. The subjects also do not get tired or loose their concentration. The internal representation of faces in the subjects' brain seems thus to be stable.

We analyze here also the subjects' responses on a stimulus-by-stimulus basis by averaging the subjects' responses for each face across all subjects. The subjects' responses are then for each stimulus the probability P_{male} to classify this stimulus as male, the corresponding reaction time RT and confidence rating CR. We then obtain Fig.3.2 where each cross stands for one head stimulus. We note that when P_{male} is extremal (i.e. $P_{male} \rightarrow 0$ or 1), subjects answer fast and are also confident. In other words the subjects seem to know when stimuli are very typically male ($P_{male} \rightarrow 1$) or female ($P_{male} \rightarrow 0$). On the other hand, when $P_{male} \sim 0.5$, the subjects are unsure, which is reflected by a high RT and a low CR. Finally, we here corroborate the relations between the subjects' responses deduced from Fig.3.1.

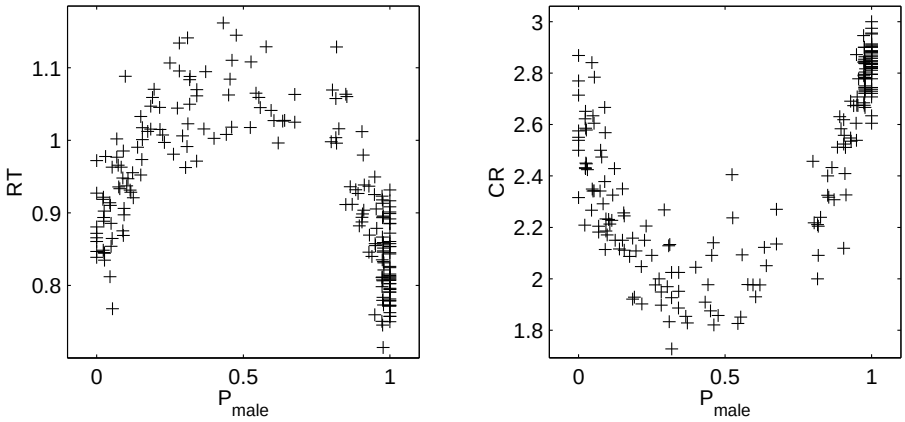


Figure 3.2: Human classification behavior on a stimulus-by-stimulus basis. Each of the 200 crosses stands for a stimulus and represents the RT, respectively the CR, associated with a stimulus as function of the probability P_{male} that this stimulus is classified as male (left and right respectively).

3.2 Classification Experiment II

In order to assess further the stability of the subjects' internal representation of the faces, we consider their responses to stimuli and proceed to a second classification experiment subsequently to the first one. This psychophysical experiment is identical to the first one and the stimuli are the same except that they are shown in a different order. The same types of responses are gathered. Fig.3.3 compares the tied rank of the responses of the subjects between the first and the second classification experiment on a stimulus basis i.e. for each stimulus, the subject's responses are averaged such that each cross in the plots stands for a stimulus. We compute Spearman's rank correlation coefficients r (linear correlation between the tied rank of one variable and the tied rank of the other) between the subject's responses (classification error, RT and CR) for the first and for the second classification experiment. The mean value of r and its standard deviation are obtained by averaging over 1000 random pooling of 90% of the stimuli in a bootstrapping manner.

The analysis of these scatter plots indicates that the crosses corresponding to the stimuli are close to diagonal, although not perfectly aligned with it, and the high value of r confirms this tendency for all responses. These results suggest the overall consistency and stability over time of the average responses over all subjects for each stimulus. Consequently, we here have a second hint at the stability over time of the subjects' internal representation of faces. Furthermore, we can conclude on the reproducibility of the human classification behavior for this database of faces. This validates the results

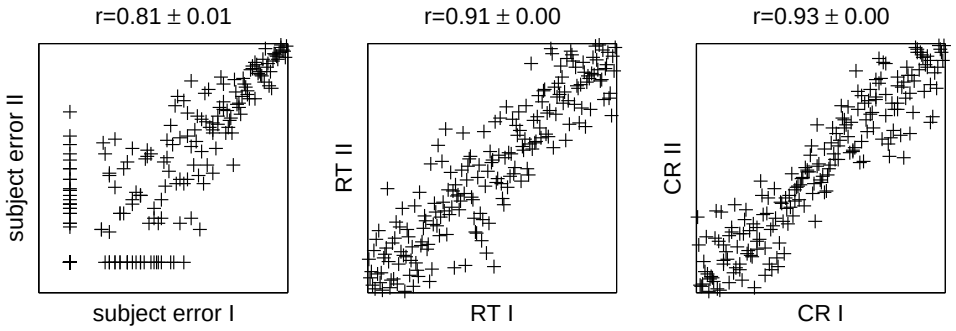


Figure 3.3: Stability of the subject's responses between the first and the second classification experiment. Each of the 200 crosses stands for a stimulus and represents the tied rank of a subject's response (classification error, reaction time or confidence rating) in the first and second classification experiment, the scales ranging from 1 to 200.

obtained in this dissertation.

One of the central issues addressed in this experiment is the *stochastic* nature of the human classification behavior. In particular, the inconsistency, or jitter, in the subjects' classification to some specific stimuli will be of central importance in the studies done in Chapter 5 comparing classification in man and machine. Indeed, under the hypothesis that hyperplanes between the classes account for classification, inconsistency in the labelling of elements near these hyperplanes will indicate a classifier which might model the classification behavior of humans: elements near to the hyperplane are difficult to classify and thus subject to inconsistency in labelling. The horizontal and vertical aggregations for the subject error (first plot of Fig.3.3) indicate such stimuli which have been correctly classified in one experiment and incorrectly in the other: these are the main outliers from the diagonal trend in the scatter plots. These elements are the patterns difficult to classify and are responsible for the relatively low value of r for the subject error. Most importantly, these stimuli are "marginal" patterns that illustrate the jitter of the subjects' classification on some specific stimuli.

3.3 Experimental Details

Every subject had to classify 152 faces in the first psychophysical experiment, and thereafter again 152 faces in the second one, making a total of 304 trials per subject. The duration a subject spent for both experiments, including instruction time, was on average 40 minutes. Considering all 55 subjects, this amounts to 16720 trials or more than 37 hours of experimental

time.

Before running the final experiment as described above, a number of *pilot studies* had to be conducted. In a first set of 21 pilot subjects, the parameters relative to presentation of the stimuli were tuned in order to make the subject's categorical judgment accurate enough while avoiding to a certain extent cognitive processing in the decision process, i.e. too long presentation and response times. The monitor was also calibrated and its parameters were set. Furthermore, the geometrical setup of the room (position of the observer, of the monitor ...) was measured and the corresponding values kept in the subsequent experiments. These settings were tested on a second set of 46 pilot subjects. At this stage the stimuli were still shown without the adjustment of the standard deviation of the intensity of the face as described in chapter 2. This fact allowed an easier gender discrimination since some elements of the faces such as beards or pimples were still strongly apparent. The stimuli were then modified accordingly at the end of these studies and the second classification experiment, with the same parameters as the first one, was added.

Chapter 4

Machine Classification Behavior

This chapter presents the methodology chosen to bridge the gap between the classification algorithms from supervised machine learning and human psychophysical experiments. Some classification algorithms are presented and their choice is motivated. Finally some experimental methods allowing us to use these classifiers are outlined.

4.1 From Machine Learning to Psychophysics

We place ourselves in the context of supervised machine learning. We thus assume being given an empirical labeled dataset $\mathcal{D} = \{(\vec{x}_i, y_i)\}_{i=1}^{p}$ where $\vec{x}_i \in \mathbb{R}^n$ are the patterns and $y_i = \pm 1$ the target values i.e. the class of the patterns. In this study we look for a family of classification algorithms having a common formulation, albeit different principles of classification. We also seek algorithms that give an intuitive measure characterizing their classification behavior and allowing to compare them to the classification behavior of humans. The biological plausibility of algorithm in such a formalism can then be assessed using comparisons and correlations with human classification behavior.

We propose the dual formulation of a learning algorithm as a common ground for the present studies. For this we consider linear separating hyperplane algorithms given by an offset (bias) b and a normal (weight) vector defined as a linear combination of the patterns of the dataset:

$$\vec{w} = \sum_{i \in \mathcal{S}} \alpha_i \vec{x}_i \tag{4.1}$$

where $\mathcal{S}$ is a subset of $\{1, \ldots, p\}$ and $\vec{\alpha}$ is a vector resulting from the specific classification algorithm. Duality indicates that the normal vector is

expressed as a linear combination of patterns of the dataset. We also impose sparseness on the learning algorithm in the sense that for classification, only a small set of patterns is necessary to compute the separating hyperplane. In other words, we consider sparse dual space classification. The elements used to compute this hyperplane are termed the expansion vectors or also the *representations*. They capture the essence of the dataset (i.e. they represent it), and only the latter are needed for classification. They are a minimal representation of the dataset and are of two types:

- they belong to the dataset: $\vec{r}_j \in \mathcal{D}, j = 1, \ldots, r$. In this case the classifier is referred to as *exemplar-based*.
- they are generated from elements of the dataset $\vec{r}_j \in \mathcal{G}(\mathcal{D}), j = 1, \ldots, r$ where $\mathcal{G} \neq id$ is the generation operator of the representations given the dataset. These generated patterns do in general not belong to the dataset and the corresponding classifier is not exemplar-based.

These types of algorithms allow an easy extension to nonlinear decision function through the use of a kernel function by replacing the scalar products arising in the algorithm by a nonlinear kernel function as: $\langle \vec{x} | \vec{y} \rangle \leftarrow K(\vec{x}, \vec{y})$. These kernels yield in most cases better classification performance of the algorithms and are thus widely-spread in the machine learning community. However, their use implies first a loss of interpretability of the results since the data lies then in a high dimensional feature space which is in most cases not explicitly known [Schölkopf and Smola, 2002]. Second, the kernel function itself as much as its parameters would then have to be determined, which is still an open problem. Using kernel functions would finally yield a huge amount of degrees of freedom to the studies of this dissertation. We will thus not consider further the concept of kernel function and will follow the principle of Occam's razor stating to "keep it as simple as possible". Linear classifiers, despite their apparent simplicity, will be shown in this dissertation to model well the classification of visual stimuli by humans.

Linear separating hyperplane (SH) algorithms have an easy geometrical interpretation: they are planes in a high dimensional space. Their classification behavior can then be characterized using the signed distance of a pattern $\vec{x}$ to the SH as

$$\delta(\vec{x}) = \frac{\langle \vec{w} | \vec{x} \rangle + b}{\|\vec{w}\|} \tag{4.2}$$

Notice that $|\delta|$ reflects the construction rule of the classification hyperplane rather than the generalization ability of the algorithm. Most importantly, the measure $\delta(\vec{x})$ is the machine counterpart of the human classification error, the reaction time (RT) and the confidence rating (CR) as defined in the previous chapter. Indeed, the *classification error* of machine is given by:

$$\epsilon(\vec{x}) = \frac{|sign(\delta(\vec{x})) - y(\vec{x})|}{2} \in \{0, 1\} \tag{4.3}$$

where y is the class of $\vec{x}$. A "probabilistic" output of a classifier is then obtained using a sigmoidal logistic function $\sigma(x)$:

$$\begin{cases} P(y=+1|\vec{x}) = \sigma(\delta(\vec{x})) = \frac{1}{1+\exp(-\delta(\vec{x}))} \in \mathbb{R} \\ \\ P(y=-1|\vec{x}) = 1 - P(y=+1|\vec{x}) \in \mathbb{R} \end{cases} \tag{4.4}$$

Since the function σ is monotonically increasing, we can state:

> The confidence rating CR of a subject for a stimulus $\vec{x}$ should positively correlate with $|\delta(\vec{x})|$.

Furthermore, we may also deduce:

> The probability to make a classification error for a stimulus $\vec{x}$, or the subject's classification error, should negatively correlate with $|\delta(\vec{x})|$.

Finally, one may also expect the following:

> The reaction RT of a subject for a stimulus $\vec{x}$ should negatively correlate with $|\delta(\vec{x})|$.

The identical definitions of classification error, RT and CR for machine reflect the fact that all these parameters are tightly related given the above assumptions. The relation between these parameters for man was already demonstrated in Chapter 3. Moreover it would not be very useful to compare for instance the execution times of the various algorithms to get a measure for the RT since these are hardware and implementation dependent. Further the values of the subjects' RT also indicates that responses are far over the lower temporal limit of the visual system as suggested by [Fabre-Thorpe, Delorme, Marlot, and Thorpe, 2001]. We are thus in a range above threshold, which allows us to proceed to further studies. In order to make the measure δ meaningful, we have to assume that the subject's internal representation of the stimuli is learnt i.e. the classes are built and in the context of SH algorithms δ is independent of time. In other words, the subject is in a testing phase, and learning has already been done previously. We hypothesize that using the SH formalism, it is possible to bridge the gap between a human psychophysical classification experiment and supervised machine learning.

4.2 Hyperplane Classifiers

Here we motivate the specific choice of the linear SH classification algorithms made in this study. A thorough analysis of these algorithms is presented in Appendix B. The main idea of this thesis is not to present an exhaustive

study of all SH classifiers and to compare them to humans. Instead, we try to isolate families of hyperplane classifiers that represent a distinct classification behavior and then consider one of the elements of this family. The biological plausibility of a classification mechanism is then inferred rather than a specific algorithm implementing it. In this study we consider the four SH classification algorithms presented below and represented in Fig.4.1 on a two-dimensional toy example.

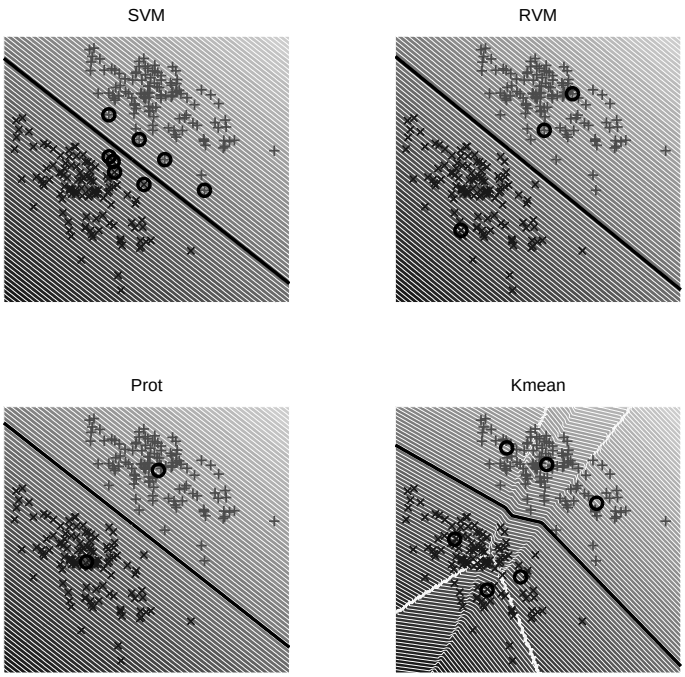


Figure 4.1: Comparison of the linear classifiers in dual form. The thick line represents the SH, the thick circles the representations and the thin lines the contours of the function $f(\vec{x}) = \langle \vec{w} | \vec{x} \rangle + b$.

The *Support Vector Machine* (SVM, [Vapnik, 2000, Schölkopf and Smola, 2002]) is a state-of-the-art maximum margin classification algorithm rooted in statistical learning theory. SVMs have a rather intuitive geometrical interpretation: they classify by maximizing the margin separating both classes while minimizing the classification errors. This trade-off between maximum margin and misclassifications is controlled by a parameter C which is mainly set by cross-validation. The dual space parameter $\vec{\alpha}$ is obtained by maximizing $\sum_i \alpha_i - \frac{1}{2} \sum_{ij} y_i y_j \alpha_i \alpha_j \langle \vec{x}_i | \vec{x}_j \rangle$ subject to $\sum_i \alpha_i y_i = 0$ and $0 \leq \alpha_i \leq C$. The offset is computed as: $b = \langle y_i - \langle \vec{w} | \vec{x}_i \rangle \rangle_{i|0<\alpha_i<C}$.

Probabilistic Bayesian classification is represented by the *Relevance Vector Machine* (RVM, [Tipping, 2001]). It optimizes the expansion coefficients of a SV-style decision function using a hyperprior which favors

sparse solutions. The RVM classifies patterns by maximizing a conditional probability of class membership $P(\vec{y}|X,\vec{\beta})$ given the data $X = \{\vec{x}_i\}_{i=1}^{p}$ and some hyperparameter $\vec{\beta}$. The class membership $P(\vec{y}|X,\vec{\alpha})$ is modeled using a Bernoulli distribution. Sparseness for $\vec{\alpha}$ is introduced using a Gaussian distribution for $P(\vec{\alpha}|\vec{\beta})$. Learning then amounts to maximizing $P(\vec{y}|X,\vec{\beta}) = \int P(\vec{y}|X,\vec{\alpha})P(\vec{\alpha}|\vec{\beta})d\vec{\alpha}$ with respect to $\vec{\beta}$, allowing the computation of $\vec{\alpha}$, and thus also $\vec{w}$ and b. Since this integral cannot be solved analytically, the Laplace approximation (local approximation of the integrand by a Gaussian) is used for solution, yielding an iterative update scheme for $\vec{\beta}$. Contrary to SVMs, RVMs do not allow an easy geometrical interpretation, although an attempt to visualize their classification behavior is presented in [Graf, Bousquet, and Rätsch, 2004a].

Common classifiers in neuroscience, cognitive science, psychology and philosophy—Platon already talked of the most typical example of each object to live in the world of "ideas"—are variants of the *Prototype classifier* (Prot, [Reed, 1972]). Their popularity is partly due to their utmost simplicity: they classify according to the nearest mean-of-class prototype. In the simplest form all dimensions are weighted equally but variants exist that weight the dimensions inversely proportional the class variance along the dimensions. As we cannot estimate class variance along all 200 dimensions from only 200 stimuli, we chose to implement the simplest Prot with equal weight along all dimensions. Defining the prototypes $\vec{p}_{\pm} = \frac{\sum_i \vec{x}_i(y_i \pm 1)}{\sum_i (y_i \pm 1)}$, the weight vector is then expressed as: $\vec{w} = \vec{p}_+ - \vec{p}_-$ and the offset as: $b = \frac{\|\vec{p}_-\|^2 - \|\vec{p}_+\|^2}{2}$. Note that due to the homogeneity of the faces in the MPI face database this is very likely close to the "best" possible prototype classifier [Graf and Wichmann, 2002].

An extension of the prototype classifier is to consider multiple prototypes in each class computed using the K-means clustering algorithm. By combining these prototypes with a nearest-neighbor classifier, we obtain the *Kmeans classifier* (Kmean, [Duda, Hart, and Stork, 2001]). We use the benchmark algorithm in the unsupervised learning, namely Kmeans, and adapt it to a supervised learning context. The number of means K is assumed to be the same for both classes and its value is determined using cross-validation. The SH obtained here is piecewise linear. Kmean represents thus the family of piecewise linear SH algorithms, the latter being a natural extension of the single SH ones. The extension of the prototype algorithm to a multi-prototype has also been considered by [Edelman, 1995] in the context of the "chorus of prototype" approach. The latter is however less transparent than the Kmean algorithm introduced above and cannot be applied to the formalism proposed in this dissertation.

We define as *representations*—the thick circles in Fig.4.1—the patterns corresponding to $\alpha_i \neq 0$ for SVM and RVM i.e. the Support Vectors (SVs) and the Relevance Vectors (RVs). The prototypes and means are the rep-

resentations of Prot and Kmean respectively since only the prototypes and the means are needed for classification. However, here we have $\alpha_i \neq 0 \quad \forall i$ i.e. all points of the dataset are used to compute these representations, and their number is always 2 for Prot and $2K$ for Kmean. The sparseness of the classifiers is then defined using the number of the representations. Clearly SVM and RVM are exemplar-based since the SVs and the RVs are elements from the dataset whereas Prot and Kmean are not because the prototypes and means are computed using the elements of the dataset but do not belong to it.

SVM, Prot and Kmean are *distance* models whereas RVM are *probabilistic* models [Reed, 1972]. Distance models are dependent upon the choice of the metric. We restrict ourselves to the Euclidean metric for the sake of the interpretability of the results and in order to get a linear SH. Other metrics, such as a Minkowski metric, could be considered, similarly to what is done when kernelizing an algorithm. Similarity measures and a weighted Minkowski norm yield the in psychology widespread General Context Model (GCM) and its derivatives [Lamberts, 1997, Palmeri, 2001, Nosofsky, 1991, Knapp and Anderson, 1984]. Whereas in prototype learning classification of a pattern $\vec{x}$ is done using $\|\vec{x} - \vec{p}_{\pm}\|$ and $\vec{p}_{\pm}$ the prototype of each class, the GCM proceeds to classification according to $\sum_{i|y_i=\pm 1} \|\vec{x} - \vec{x}_i\|$. The GCM has thus no sparseness and does not allow a hyperplane formulation, what makes it not suited for the studies in this dissertation. Prot and GCM belong to the same family of distance-based classifiers where classification is done using the distance of a new pattern to all existing ones, either directly as for GCM or indirectly through the prototype. The prototype learner is finally chosen to represent the "psychology-inspired" models. A whitened version of the prototype classifier, namely Fisher linear discriminant classifier (FLD, [Fisher, 1936, Mika, Rätsch, Weston, Schölkopf, and Müller, 2003]), will be considered in Chapter 6.

There are other manners to extend the Kmeans clustering algorithm to a classifier. Indeed, any classifier can be applied to the means obtained by Kmeans clustering. The application of a nearest-neighbor rule as above yields a piecewise linear decision function. Applying SVM, RVM or Prot on these means would yield a linear SH. Such an approach can yield a first "sparsification" stage to any classifier, albeit the concept of sparseness is not well-defined for such an ensemble of (sparse) classifiers. However in the context of this study, this would make any conclusions on the biological plausibility of a classification mechanism hazardous since two mechanisms are used in the classification stage. Furthermore, such an approach would violate Occam's razor and is thus not consider any further. Finally, in order to take into account that the male and female classes are not balanced (recall the male bias in the subjects' responses as shown in Chapter 3), one could consider a different number of means in each class, say K_+ and K_- for the male and female class respectively, and in particular one could set

$K_+ > K_-$. However such an approach would increase the number of free parameters of the algorithm (K_+ and K_- instead of K) and would thus be prone to overfitting while violating Occam's razor.

The K-nearest neighbors classification algorithm [Duda, Hart, and Stork, 2001] has a piecewise linear SH with a large number of segments, and may result in severe overfitting and poor generalization ability. Further the number of such segments cannot be determined beforehand and the sparseness of such an algorithm is usually very low. Since the K-nearest neighbors algorithm belongs to the same family as Kmean, it is not considered further.

One of the most popular classifiers are Artificial Neural Networks (ANN, see [LeCun, Bottou, Orr, and Müller, 1998]). The latter are not adapted for the present study since they do not yield a (sparse) dual space formulation and the corresponding SH. However, under some conditions, their output may be interpreted as a posterior probability of class membership, and thus be quite similar to δ. The absence of a dual space representation also implies that the concept of representation (for instance the SVs, the RVs, the prototypes or the means) does not exist for ANN, this concept being of importance as will be shown in Chapter 5. Moreover, ANN should mainly be used in cases where there are more patterns than dimensions ($p \gg n$, [Bottou, 2003]). Although this is increasingly the case when considering nowadays databases since their size is growing faster than the computational power, this is not the cases for this study since we have $p = n$. Moreover, ANN have some intrinsic limitations: they are polynomial approximators and they have many free parameters such as the learning rate or the number of hidden units. For wrongly chosen parameters, the ANN will not converge while training, will get stuck in local minima or will overfit on the training dataset. The lack of a thorough theory to select these parameters makes them not usable in the context of the present study. Finally SVMs can be thought of as a more principled version of two-layered feedforward ANNs [Haykin, 1999, Schölkopf, Burges, and Vapnik, 1995, Vapnik, 2000].

Perceptrons are arguably one of the first attempts to model neurons [Rosenblatt, 1958]. Although they can be modeled as proceeding to classification using a SH [Cristianini and Shawe-Taylor, 2000], they are not sparse. The same is true for Adaboost which is an increasingly popular classifier based upon boosting i.e. an efficient combination of weak learners [Freund and Schapire, 1995]. Since both Perceptrons and Adaboost can be interpreted as maximum margin classifiers (a maximum margin is a property of these algorithms but no notion of margin appears in their definition, see [Graepel, Herbrich, and Williamson, 2001] and [Schapire, Freund, Bartlett, and Lee, 1998] respectively), SVMs are assumed to represent them in the following studies.

Finally RVMs represent the family of (sparse) Bayesian inference classification using Gaussian Processes (GPs, [Williams and Barber, 1998]). A mean field approach as derived from Statistical Physics has been applied

to binary classification with GPs in [Opper and Winther, 2000] and SVMs have been shown to be closely related to GPs.

4.3 Classification with Spiking Neurons

We mention below two learning, and thus also classification methods, derived from spiking neural networks: *SpikeNET* and the *Liquid State Machine*. These algorithms are supposed to imitate cortical microcircuits. It is certainly biologically-plausible or at least biologically-inspired to model single neurons using spike trains on the lowest level and then to extend the model to populations of neurons [Gerstner and Kistler, 2002]. Although not as straightforward as in the context of machine learning, learning and thus also classification can be considered using spiking neuron models [Gerstner and Kistler, 2002, Feng, Sun, Buxton, and Wei, 2003]. One huge difference to machine learning is the introduction of *time* though the learning dynamics. Although such approaches are promising for future research, the lack of a clear general theory, the absence of a usable implementations and the difficulty to fit these methods into the here-introduced general framework stimulus-man-machine excludes them from the studies considered in this dissertation. Further, as they are, these algorithms do not fit into the context of the studies proposed here since they do not allow a (sparse) dual formulation and cannot be interpreted using hyperplanes for classification.

The first spike following a stimulus has been shown to have important properties in the visual cortex such as orientation selectivity [Delorme, 2003]. SpikeNET [Delorme, Gautrais, Rullen, and Thorpe, 1999, Delorme and Thorpe, 2003] allows to model large networks of spiking integrate-and-fire neurons and is an example of a computational model where a biologically-inspired concept yields high computational performance. In this type of spiking neuron model, only the neurons emitting a spike are further processed while the remaining ones are ignored. The underlying assumption is that only a small set of strongly excited neurons fire at least one spike in a small time delay. This low computational cost is argued to be biologically-plausible in [Thorpe, 2002]. This classification based upon the first spikes corresponding to a stimulus allows to introduce naturally the concept of RT, and this is, from the perspective of this thesis, the greatest advantage of this formulation. In this framework, face processing (detection and localization) was investigated by [Rullen, Gautrais, Delorme, and Thorpe, 1998] where computational studies showed that visual processing using only one spike per neuron seems to be possible, reducing thus computational costs.

A Recurrent Neural Network (RNN) is a neural network with some (delayed) feedback loops [Haykin, 1999]. These loops yield to rich range of possible dynamical evolutions and rise issues about for instance attractors and their stability. In the Liquid State Machines (LSM) [Maass, Natschläger,

and Markram, 2002b], Integrate-and-Fire neurons are used in a RNN creating thus a dynamical system with high-dimensional states argued to be a model of a generic cortical microcircuit. Only the neurons that read out the information of the RNN are trained by a classifier. LSM combine high-dimensional dynamical systems with statistical learning theory [Maass, Natschläger, and Markram, 2003], with Bayesian inference or a linear classifier [Natschläger and Maass, 2004]. The output neuron can give a stable response despite the fact that its input signals are transient and of high dimension. One of the main domains of application of LSM is the real-time processing of time-varying inputs such as visual stimuli [Maass, Legenstein, and Markram, 2002a]. At this point, we can compare ANNs and LSMs. ANNs are a polynomial approximator: they create high dimensional highly nonlinear decision functions. This complexity is most probably not biologically-plausible. LSMs on the contrary put the stimuli in a random system of states and then retrieve the states using a simple decision function. This simple manner to retrieve information is robust and thus more biologically-plausible.

4.4 Tricks of the Trade

As most practitioners in machine learning may know, applying the classification algorithms directly on the raw data, the encodings as defined in Chapter 2, may lead to convergence problems and numerical instabilities. To avoid these, in the context of this study the data is *centered* and *normalized* prior to classification. The centering step $\vec{x} \leftarrow \vec{x} - \frac{1}{p}\sum_i \vec{x}_i$ makes especially sense when followed by a normalization step $\vec{x} \leftarrow \frac{\vec{x}}{\|\vec{x}\|}$. Centering can also be argued as being biologically meaningful when considering the saccades of the occulomotor system [Yarbus, 1967]. Notice that for PCA, LLE and ICA, the data is already centered. Since we only consider linear SH algorithms, normalizing the inputs is equivalent to making the algorithms classify in a normalized (feature) space. Classification in normalized spaces has been shown to be advantageous for linear classifiers [Herbrich and Graepel, 2001], and in particular for SVMs [Graf and Borer, 2001, Graf, Smola, and Borer, 2003]. When normalizing, the data points lie on a unit hypersphere, resulting thus in the loss of one dimension of the data out of 200 or 256 depending on the data type. To assess this loss of information, a Fisher analysis (see Chapter 2) is done on the centered and normalized encodings in Fig.4.2. By comparing Fig.4.2 to Fig.2.8, it can be seen that normalization affects only slightly the separability of the classes and the same conclusions as for the non-normalized encodings apply. Thus the loss of one degree of freedom though normalization is negligible for the sake of classification. Finally, normalization is a sensible approach, particularly since it allows to get the parameter of machine, namely δ, in the same range for all types of data,

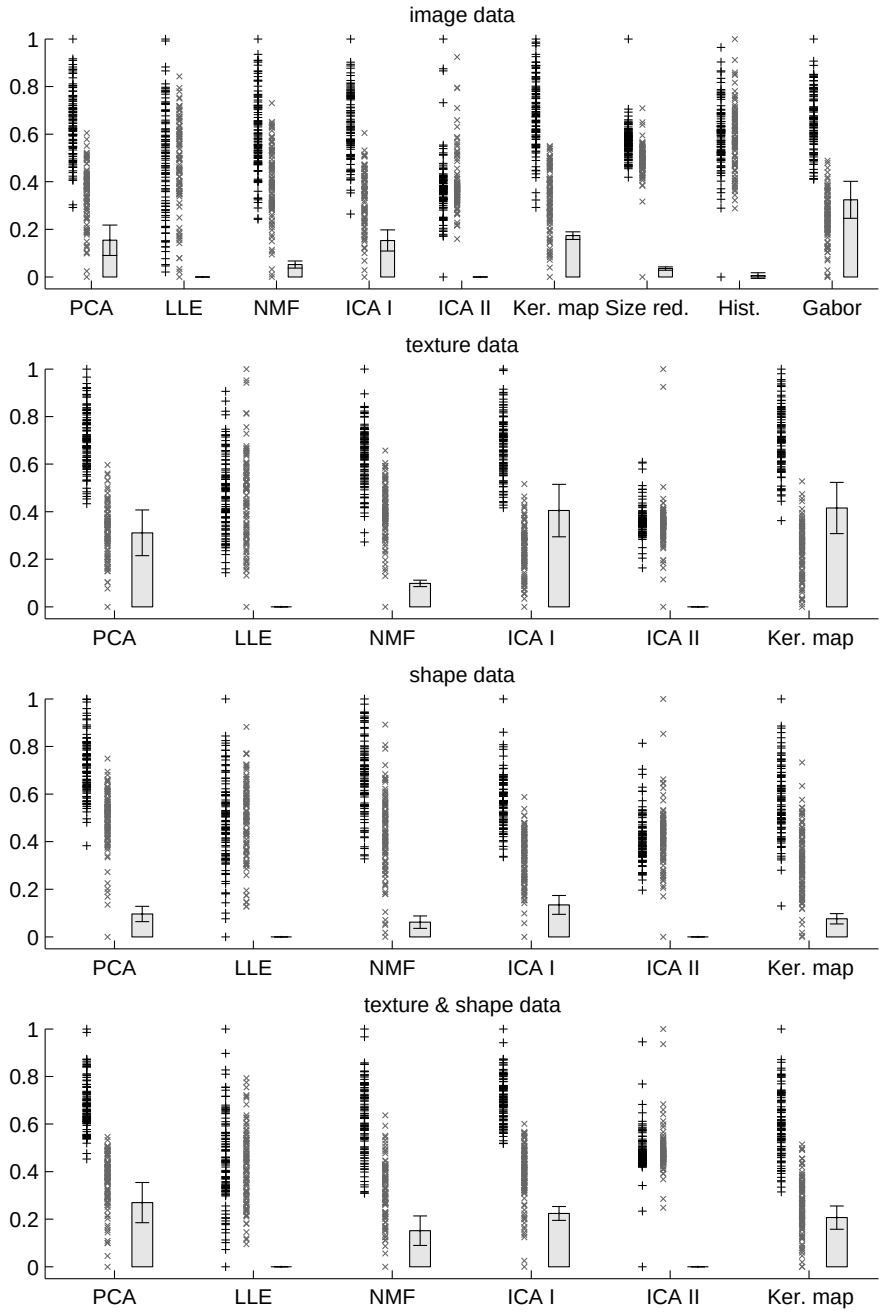


Figure 4.2: Projection of the centered and normalized dataset along the Fisher direction (data rescaled to $[0, 1]$) and Rayleigh quotient J for each data type and preprocessor.

preprocessors and classifiers. As an alternative to normalization, *whitening* can be considered after centering the data. In this case, every component of a vector is divided by the standard deviation over the whole dataset of this component. The effect on the discriminability of the classes is reported using a Fisher analysis as in Fig.4.3. It can be seen that whitening destroys or strongly reduces the discriminability of the data in all cases. Whitening thus changes the nature of the vectors of the database and is therefore not considered further in the present study.

SVMs and Kmeans have one free parameter: the regularization parameter C and the number of means K. This parameter is set automatically by the machine using an M-fold cross-validation procedure [Duda, Hart, and Stork, 2001]. Indeed, setting this number by hand to a user-defined value may result in overfitting and thus in a loss of generalization of the algorithm. Cross-validation consists of splitting the dataset into M equal-sized subsets of patterns. Then M classifiers are created by training on all possible $M-1$ subsets and the corresponding classification errors are computed on the remaining subset. The M-fold cross-validation classification error on the whole dataset is then the average of the M classification errors obtained for each subset. The parameter of the algorithm which minimizes this error is finally chosen. Cross-validation is also used to estimate the classification error of an algorithm over a dataset. When the classification algorithm has also a free parameter, a double M-fold cross-validation scheme is used in a two-stage manner. In the cross-validation studies considered in this study, the order of the cross-validation M and the range of K or C can be varied. The set of values for cross-validation with SVM is set to $C \in \{1, 2, \ldots, 10, 13, 15, 18, 20, 30, 40, 50, 100, 500, 1000\}$. If different values of the parameter give the same error, the minimal value of C is taken such as to enforce a large margin rather than reduce the classification errors. The range of possible values of K is limited to $2, \ldots, 10$ in order to avoid "overfitting" and avoid to have more means in each class than data points, which is not trivial since in the subject dataset (see Chapter 5) the classes are not at all balanced (strong male bias). If different values of K give the same error, the minimal value is chosen in order to minimize the number of representations, and consequently the complexity of the piecewise linear decision function. Note that there is an alternative way to determine the optimal K using the eigenvalues of the covariance matrix of the data. The eigenvalues above a user-defined threshold are used to determine the number of means. Since here again we need a heuristic to determine this threshold, we do not gain much from this approach compared to cross-validation.

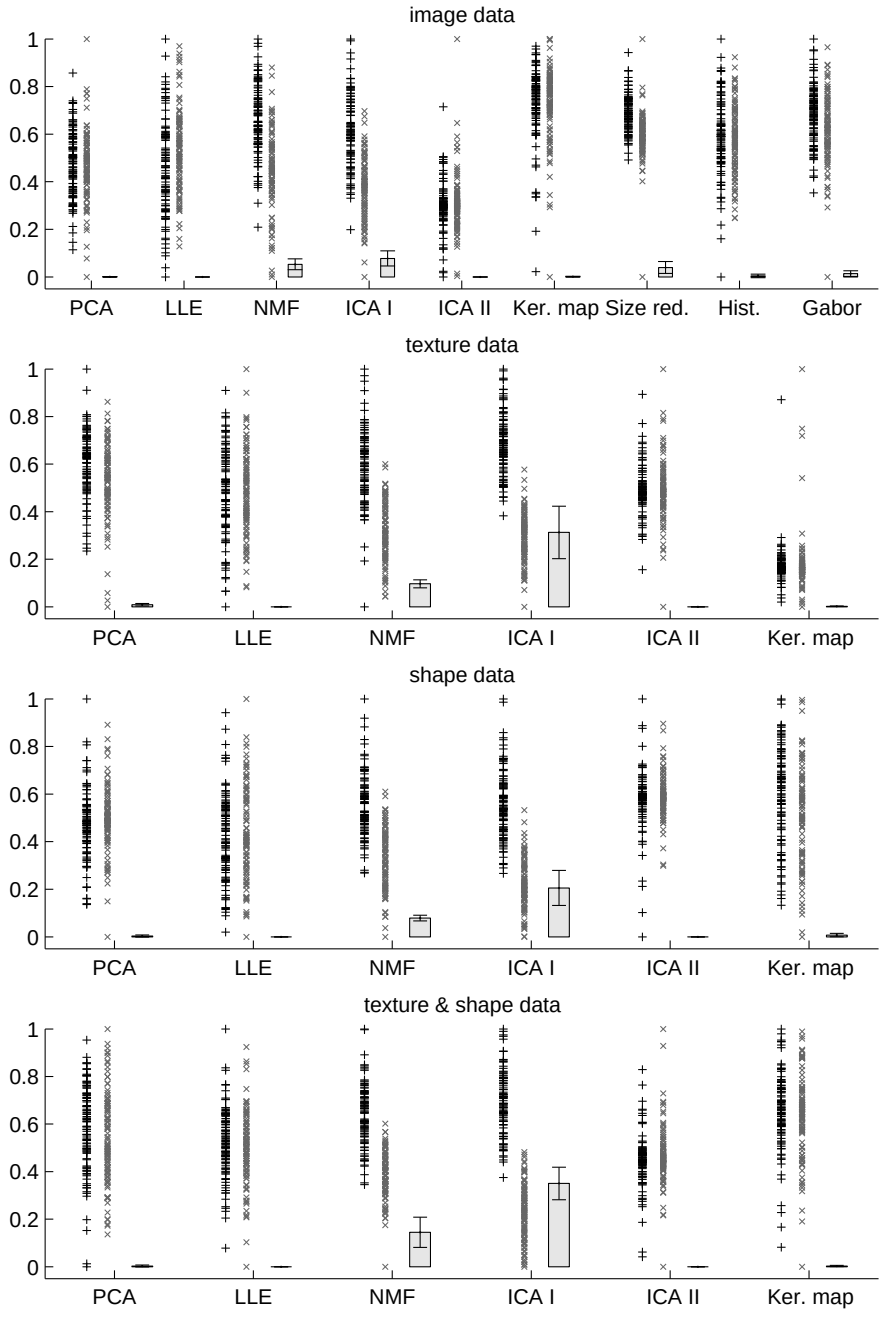


Figure 4.3: Projection of the centered and whitened dataset along the Fisher direction (data rescaled to $[0, 1]$) and Rayleigh quotient J for each data type and preprocessor.

Chapter 5

Classification Behavior of Man and Machine

We attempt to describe in this chapter the mechanisms used by human subjects for feature extraction of visual stimuli and their subsequent classification. For this, the classification performance of man and machine are compared and their classification behaviors are correlated for various feature extractors and classifiers from machine learning. The corresponding results are then corroborated using studies on the stochastic nature of human classification. General conclusions and a discussion are finally given while a literature review closes this chapter.

5.1 Overview

In this section we present, describe and motivate the methods and techniques used to study the classification behavior of man and machine. The results of these studies allow to compare man and machine for a given feature extractor (data type and preprocessor) and classifier, and may allow us to infer on the mechanisms and strategies used by human subjects to classify visual stimuli. The comparison of the classification *performance* of man and machine is only a weak indicator of which combination may or may not be adapted to describe human classification. The actual comparison between the classification *behavior* is done in the correlation studies between the man and machine. Studies on the *stochastic* nature of human classification follow. For all these studies we introduce a methodology where machine learning is used to extract quantitative measures from a psychophysical setup, allowing to bridge the gap between machine learning and psychophysics. On the basis of these studies, we hope to unravel the mechanisms used by the human subjects to extract features from visual stimuli and to perform their subsequent classification.

The face stimuli presented to man are generated using both the shape *and* the texture information. When comparing the classification performance of

man and machine, both information must also be shown to machine. In this case, we thus only consider as data types the image data $\mathcal{I}$ and the texture & shape data $[\mathcal{TS}]$. Indeed, it would be unfair to compare man who was shown all the information content of the stimuli and machine which would only have been shown the texture or shape data. However, the correlation studies can be done for all data types since in this case we are not comparing the classification performance of man and machine *per se*, but rather try to unravel the mechanisms of classification in man and the underlying feature extraction process. Thus allowing the machine to work only on the texture or shape data allows to assess whether this type of data may allow high correlations between the classification behavior of man and machine.

In the studies below, we consider two types of stimulus datasets: the *true* and the *subject* datasets. The patterns in both datasets are the encodings corresponding to each feature extractor—the pairing of a data type with a preprocessor. The true dataset is constituted by the encodings combined with the true labels of the stimuli—their true gender as given by the MPI face database. The subject dataset is composed of the encodings combined with the labels of the stimuli as estimated by the subjects in the first psychophysical classification experiment, in other words the estimated gender. This dataset represents what we assume to be the subjects' internal representation of the face space. Both the true and the subject datasets contain an ordered list of 152 stimuli out of a possible 200, this list being different for every subject. Thus both datasets can be seen as the subject's personal datasets.

In the next three sections of this chapter (the classification, correlation and stability studies), the methodology is first introduced and the results are subsequently presented in two steps. In a first step the methodology is illustrated on PCA applied to the texture & shape data type, as already done in a previous study by [Graf and Wichmann, 2004], since this combination will be shown *a posteriori* to illustrate best the difference between the various classification algorithms presented in Chapter 4. Furthermore this feature extractor can be considered *a priori* as a benchmark since it uses PCA, a widely-spread preprocessor whose biological plausibility has been hypothesized [Turk and Pentland, 1991, O'Toole, Abdi, Deffenbacher, and Valentin, 1993, Vetter and Troje, 1997, O'Toole, Defenbach, Valentin, McKee, Huff, and Abdi, 1998]. Finally, using the texture & shape data solves the problem of correspondence between faces (a nose is mapped to a nose, ...) and is thus intrinsically adapted to the stimuli considered in this study. For the other data types and preprocessors, the reader is referred to Appendix F for the complete set of plots. In a second step, summary plots allow the comparison of the various feature extractors and classifiers. The result are finally discussed and a literature review concludes this chapter.

5.2 Classification Performance of Man and Machine

5.2.1 Methodology

We compare the classification errors of man and machine for each subject individually both on the true and subject datasets. These errors allow a crude comparison between the combinations of the various data types, preprocessors and classifiers and give a first hint on the mechanisms humans might have used for classification. It also allows to compare the true and the subject dataset i.e. the real representation of the stimuli and the subject's internal one.

For *humans*, the classification error on the true dataset is simply obtained by considering the mean and standard deviation over all stimuli of the difference between the estimated class to the true one. The classification error on the subject dataset cannot be computed directly since the subject's labels are not known beforehand. This error is thus estimated using a method derived from cross-validation. For each stimulus shown to the subject, we compute the mean error the other subjects made on this stimulus by defining as an error when the other subjects responded differently than the considered subject. In other words we compare the subjects gender responses on common stimuli and compute the mean consistency between subjects. We then compute the mean and standard deviation of this error over all presented stimuli.

For *machines* the mean and standard deviation of the classification error is obtained, both for the true and the subject datasets, using a single 5-fold cross-validation for RVMs and Prots and a double 5-fold cross-validation to determine C for SVMs and K for Kmeans[1]. We also determine the mean and standard deviation of the number of representations $\sharp(SV)$, $\sharp(RV)$ and K_{opt} which are a measure of sparseness of the classifier (see Chapter 4). Furthermore $\sharp(SV)$ is also an indicator of the generalization ability of the SVM classifier.

Since every subject gets a different set of 152 randomly chosen faces from the 200 available, the mean and standard error over all subjects of the mean and standard deviation of the classification errors of man and machine and of the number of representations are finally computed.

5.2.2 Results

When classifying the *true* dataset for PCA applied on the texture & shape data type, Fig.5.1 shows that none of the classifiers yields a classification performance comparable to that of humans. The prototype classifier, pop-

[1]We also studied 7 or 10-fold cross-validation schemes. These gave similar results to the 5-fold one, however these schemes were much more computationally expensive.

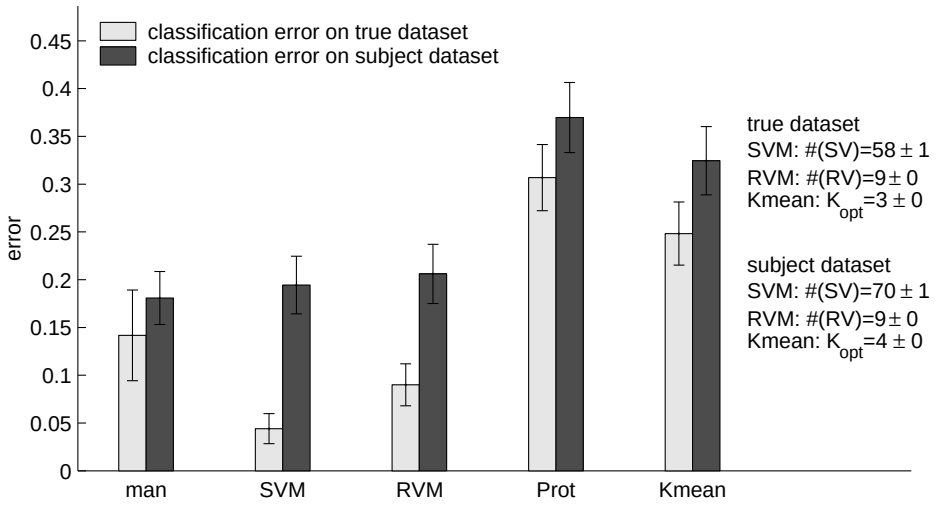


Figure 5.1: Comparison of the classification performance of man and machine on the true and subject dataset using a cross-validation scheme.

ular in neuroscience, psychology and philosophy, performs on average much worse than humans and thus cannot possibly be used by humans for classification given the linear PCA representation applied to the texture & shape data type. The better classification performance of Kmeans compared to the simple prototype classifier may be explained by the piecewise linear decision function. The low classification errors of SVMs and RVMs indicate that the two classes—the genders—can be separated well given the PCA representation on the texture & shape data. An intriguing fact is that SVMs and RVMs perform better than man. The subjects were presented with human faces with some high-level features such as hair, beards, or glasses removed. However, such features were likely used by the subjects to create their representation of gender-space during their lifetime. The subjects are thus trained on one type of data and tested on another. The machines on the other hand are trained and tested on the same type of stimuli: stimuli from the MPI face database. This may explain the quite disappointing performance of man in such a biologically-relevant task compared to machine. This result is corroborated by [Moghaddam and Yang, 2000] comparing the classification performance of man and SVMs on a different face database and by [Golomb, Lawrence, and Sejnowski, 1991, Gray, Lawrence, Golomb, and Sejnowski, 1995, Blackwell, Vogl, Dettmar, Brown, Barbour, and Alkon, 1997] where humans are compared to Artificial Neural Networks (see Section 5.6).

The classification error on the *subject* dataset as shown in Fig.5.1 represents the ability of the classifier to learn what we, based on the responses

of the subjects, presume to be their internal representation of face-space. The machines have more difficulty in learning the datasets with the subject's labels than the one with the true labels, which may suggest that the subject's internal representation of the stimuli is not optimal for the machine and makes classification a harder task for machine. Given our aim of re-creating the subjects' decision boundaries using artificial classifiers, this makes Kmeans a mediocre, and the prototype learner a rather poor candidate using the PCA representation. Furthermore we can see that humans have a similar classification performance to SVMs and RVMs, making the latter good candidates for modeling the classification performance of the human subjects on for visual stimuli.

For both datasets, we notice that $\sharp(SV) \gg \sharp(RV)$. This shows that SVMs need much more representations for classification than RVMs. This high number of SVs indicates that the classification task is difficult. However the generalization ability of SVMs is still good as shown by the low classification errors in Fig.5.1. The low number of K_{opt} for both datasets indicates that the classes have few intrinsic clusters and can be assumed to be smooth manifolds.

As a first step allowing to isolate groups of models for feature extraction and classification, we study the classification performance on the subject dataset since the latter is the most meaningful for the studies to come. As mentioned before, such a study is only relevant for the image and texture & shape data types. Fig.5.2 shows these plots for the data types, the preprocessors and the classifiers, the horizontal lines indicating the performance of man. The image data—or the pixel information—can be seen as the input arriving on the retina, making this data type *a priori* biologically meaningful. Moreover Gabor wavelets have been found to describe the receptive fields in V1 in physiological studies [Hubel and Wiesel, 1962] and have been successfully used in computational models of spatial vision such as in [Itti, Koch, and Braun, 2000]. We here experimentally verify that such Gabor wavelets are also a sensible choice for face processing since they are the only preprocessor on the image data with similar classification error as humans. With this feature extractor, SVMs performs most similarly to man and, surprisingly, Prots and RVMs perform similarly bad.

The texture & shape data yields overall classification performances closer to those of humans than on the image data. This may suggest that the information contained in the texture & shape data is useful for human subjects to build their internal representation of faces as already suggested by [Troje and Bülthoff, 1996, Vetter and Troje, 1997]. A similar finding was obtained by [O'Toole, Vetter, and Blanz, 1999] where it was shown that humans rely on both the texture and the shape information to classify faces according to their gender. PCA, ICA I and NMF are most adapted in this respect. SVMs and RVMs are the classifiers which best reflect human classification performance whereas Prots and Kmeans are much worse.

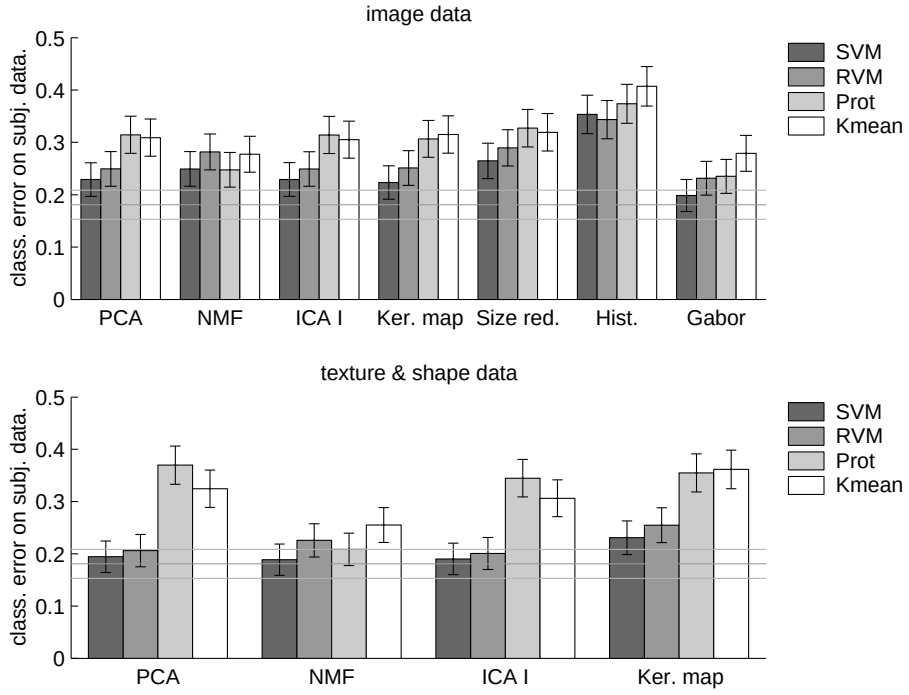


Figure 5.2: Summary plots showing the mean classification error on the subject dataset for each classifier, data type and preprocessor. The horizontal lines represent the mean and standard deviation of the classification performance of man on the subject dataset.

5.3 Classification Behavior of Man and Machine

5.3.1 Methodology

Although useful to assess the classification performance of a classifier, a cross-validation scheme is not appropriate when correlating the classification behavior of man and machine mainly because of the lack of interpretability. Indeed, in order to validate the separating hyperplane (SH) model for classification, we postulate that we can model the classification behavior of man using a single SH and not M SHs as required by a cross-validation scheme. Consequently we train the hyperplane algorithms on the whole dataset in order to obtain a single SH. In the case of SVMs and Kmeans, in order to determine the optimal value of C and K respectively, we need to proceed to a single 10-fold cross-validation on the classification error. However, the classifier is still trained on the whole dataset using this optimized value of C or K, yielding thus a single SH. Notice that this cross-validation step is done on the classification error and not on the man-machine correlations as defined below.

We *first* study the mean and standard deviation of the following errors of the various algorithms for each subject:

1. the training error on the subject dataset

2. the classification error on the subject dataset determined using the unseen stimuli with, as labels, the sign of the mean of the other subjects' responses for each of these unseen stimuli

3. the classification error on the true dataset computed using the unseen stimuli with their true labels

where the unseen stimuli are the remaining 48 stimuli out of the 200 which have not been seen by man, and thus neither by machine. These values are averaged and the standard error is computed over all subjects. As before, for each subject, $\sharp(SV)$, $\sharp(RV)$ and K_{opt} are a measure of the sparseness of the classifier and $\sharp(SV)$ allows to assess the generalization ability of the SVM classifier. Their mean and standard deviation over subjects is then computed. In this first study, and in particular through the training error, we assess the "quality" of the parameters of the SH and thus the domain of validity of the conclusions drawn from the correlation studies to be presented below. Furthermore, this training error indicates the ability of the classifier to recreate the subject's classification behavior and in particular the subject's internal representation of the stimuli. Finally, the comparison of the two classification errors with the ones obtained previously using a cross-validation scheme allows to assess the necessity of a cross-validation scheme in the present studies.

From this point on, we shall only consider the SHs obtained on the subject dataset since only these SHs reflect what we hypothesize to be the internal representation of the subjects. In other words, for each subject a personal SH is computed using the labels estimated by this subject. The distance δ to this SH of each stimulus presented to this subject is then computed for each classification algorithm. In the case of Kmeans this distance is computed using the piece of hyperplane constructed using the mean of each class nearest to the considered stimulus. The distance δ represents the classification behavior of machine (see Chapter 4).

Second, the histograms (frequency of occurrence) of δ for each classifier over all datasets shown to the subjects, i.e. the ensemble of all δ obtained for all subjects, are computed for two groups of stimuli: the representations (the Support Vectors, the Relevance Vectors, the Prototypes and the Means) and the non-representations (the remaining patterns presented to the subject). These plots allow to visualize the actual geometrical configuration of these stimuli with respect to the SH. In particular, they indicate the geometrical regions most useful for classification for each algorithm i.e. the position relative to the SH of the elements the classifier actually uses for

classification. Further, these plots also allow to visualize the effect of the combined effect of the classifier and the feature extractor on the distribution of the stimuli in the dataset, and in particular on the discriminability of the two classes (see also Chapter 2). These histograms thus also give a one-dimensional visualization of the data manifold resulting from the pairing feature extractor—classifier.

Third, the classification behavior of man and machine is correlated with all parameters of man and machine averaged over subjects and over sets of stimuli. In other words, the mean of the subject's classification error, RT, and CR of the first psychophysical classification experiment for a given set of stimuli are correlated to the mean distance of these stimuli to the SH for the four types of hyperplane classifiers. The average and standard deviation over stimuli are computed over the following sets for each subject:

1. the set of correct and incorrect gender responses: $E = 0, 1$

2. the RTs which have been discretized over three bins as follows:

$$\alpha + \frac{i-1}{3}(\beta - \alpha) \leq RT(i) \leq \alpha + \frac{i}{3}(\beta - \alpha), \quad i = 1, 2, 3$$

   where $\alpha = \min(RT)$ and $\beta = \max(RT)$.

3. the three sets of CRs: $CR = 1, 2, 3$

The mean and standard error over the subjects is then computed from the mean and standard deviations of the above parameters. This study allows to assess a global averaged classification behavior i.e the correlations between the classification behavior of man and machine on a wide scale.

Fourth, we finally assess the correlation of the classification behavior of man and machine on a stimulus-by-stimulus basis: the parameters of man and machine are averaged only over subjects. In other words, we compute, for each stimulus and classifier, the relation between the mean value of the distance $|\delta|$ to the SH over all subject datasets and the mean response of the subjects (classification error, RT or CR from the first psychophysical classification experiment) for this stimulus. The resulting scatter plots relating man on the ordinate axis to machine on the abscissa are shown for each classifier and for each type of response, each of the 200 scatter points representing one stimulus. To quantitatively assess this correlation, we perform a non-parametric rank correlation analysis using the tied rank of the subject's response and of $|\delta|$ across the set of stimuli by computing Spearman's rank correlation coefficient r. The mean value of r and its standard deviation are obtained using a bootstrap method by averaging over 1000 random poolings of 90% of the 200 stimuli.

5.3.2 Results

The first row of Fig.5.3 assesses the classification errors on the true and on the subject datasets using PCA combined with the texture & shape data type without using a cross-validation scheme. Comparing this figure to

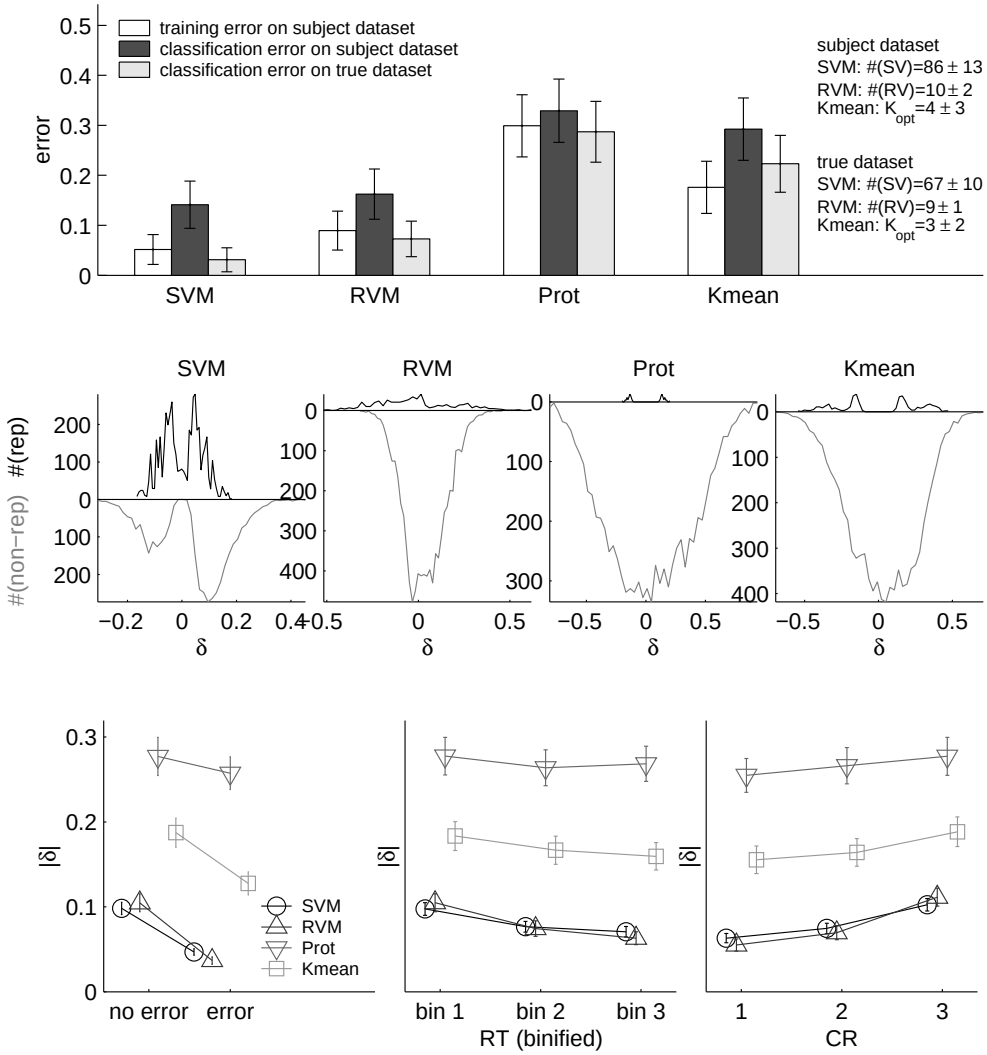


Figure 5.3: Comparisons of training and classification errors of machine without a cross-validation scheme (first row). Histograms of distances of (non-)representations to the SH (second row). Correlation of classification behavior of man and machine with parameters averaged over subjects and sets of stimuli (third row).

Fig.5.1, we conclude that a complete cross-validation scheme is superfluous

for both datasets, although the actual classification errors may slightly differ. This is however not of concern in the correlation studies since we are here more interested in the comparison between quantities rather that in their actual value. Further, the same conclusions as before apply for the number of representations of the classifiers. The training error on the subject dataset is, as it should be, smaller than the corresponding classification error. The training error indicates that SVMs and RVMs are well suited to learn the subject's internal representation given the PCA representation since both have a low training error. However Prots and Kmeans are badly adapted for this task.

The second row of Fig.5.3 compares the histograms of δ for each classifier between the representations and the non-representations for PCA applied to the texture & shape data type. For SVMs the distribution of the representations is peaked near the margins i.e. the representations are localized near or on the margins. For RVMs, the representations are almost uniformly distributed across the dataset with a slight peak around $\delta = 0$ i.e. RVMs use patterns sampled from the whole dataset for classification. Furthermore, for RVMs and SVMs there are more representations on the female side ($\delta < 0$) i.e. on the side where most errors are made (the subjects have a male bias i.e. more females are classified as males then the contrary as shown in Chapter 3). In other words, these algorithms put their representations where classification is difficult—around the SH, in the margin stripe or in the difficult class. Such a behavior may be argued to be meaningful. The contrary is true for Prots and Kmeans where the representations are only put in the middle of the classes where classification is easy. The non-representations are spread almost uniformly on a large subset throughout the dataset for Prots and Kmeans. For RVMs this subset is much smaller and is located around the SH. SVMs, according to their classification principle, create a margin stripe between classes with as few patterns inside as possible.

The studies on the geometrical position of the representation with respect to the SH for all feature extractors using the above histograms are given in Appendix F. Essentially the same conclusions as above apply, although these histograms dependent on the geometry of the space given by the feature extractor. This suggests that, despite the fact that the dimensionality of the encodings is equivalent to the number of patterns, we get a rich plethora of spaces spanned by the encodings corresponding to the various feature extractors. Generally, RVMs "sample" best the dataset since their representations are spread throughout the dataset, although with an increase towards the SH. This peak is absent for preprocessors such as the size reduction or the histograms and for data types such as the shape data, these feature extractors being anyhow shown below to be poor candidates to model feature extraction in humans. In other words, a good feature extractor allows RVMs to put its representations around $\delta = 0$ and RVMs can

use best all information available in the dataset by sampling the whole of it. On the contrary, SVMs use patterns near or on the margin to proceed to classification for all feature extractors. Both RVMs and SVMs concentrate their representations on the elements difficult to classify i.e. elements with low $|\delta|$. The worst case is represented by Prots and Kmeans which only use easy elements, which is reflected by the histograms of the representations that are essentially two peaks near the middle of the classes for all feature extractors.

The classification behaviors of man and machine are correlated for PCA and the texture & shape data type as feature extractors and results are summarized in the third row of Fig.5.3 with parameters averaged over subjects and stimuli. We observe, first, that the error of the subjects is high for $|\delta|$ low, suggesting that elements near the SH are more difficult to classify. Second $|\delta|$ is low for high RTs: the elements near the SH seem to require more processing in the subjects' brain resulting in a higher RT. Third, the high CR for high $|\delta|$ indicates that the subject is sure when stimuli are far from SH. Thus elements far from the SH are classified more accurately, faster and with higher confidence than those near to the SH. This intuitive behavior is a first hint at the validity of the assumption that hyperplanes may be used to account for classification by humans. SVMs and RVMs choose among existing elements to build their SH and they maximize a margin, respectively a conditional probability, and have a low δ as already seen in the above histograms. This may imply that they classify in an efficient manner. Prots and Kmeans generate new elements to compute their hyperplane and have a high δ (see also the histogram plots), what may hint at a robust classification. On the basis of these plots, it is difficult to compare the classifiers among each other. The following correlation analysis solves this problem.

Fig.5.4 presents the scatter plots corresponding to an analysis on a face-by-face basis for each classifier and for each type of response when averaging only over subjects given our choice of the feature extractor (PCA and the texture & shape data type). This plot is the central plot for the correlation analysis between man and machine. SVMs and RVMs show most correlation between the subject's response and $|\delta|$. The prototype algorithm again behaves in the least human-like manner of the four classifiers. The correlation between the classification behavior of man and machine indicates for RVMs, SVMs and to some extent Kmeans, that heads far from the SH are more easily processed by humans: the subjects' brain needs to do more processing (higher RT) to classify stimuli close to the decision hyperplane, while stimuli far from it (high $|\delta|$) are classified more accurately (low error) and with higher confidence (high CR). Thus, as already mentioned above, modeling the classification behavior of humans using hyperplanes seems to be a plausible approach. Further, the poor correlation for Kmeans indicates that it is unlikely that the subjects are using this type of piecewise linear decision function.

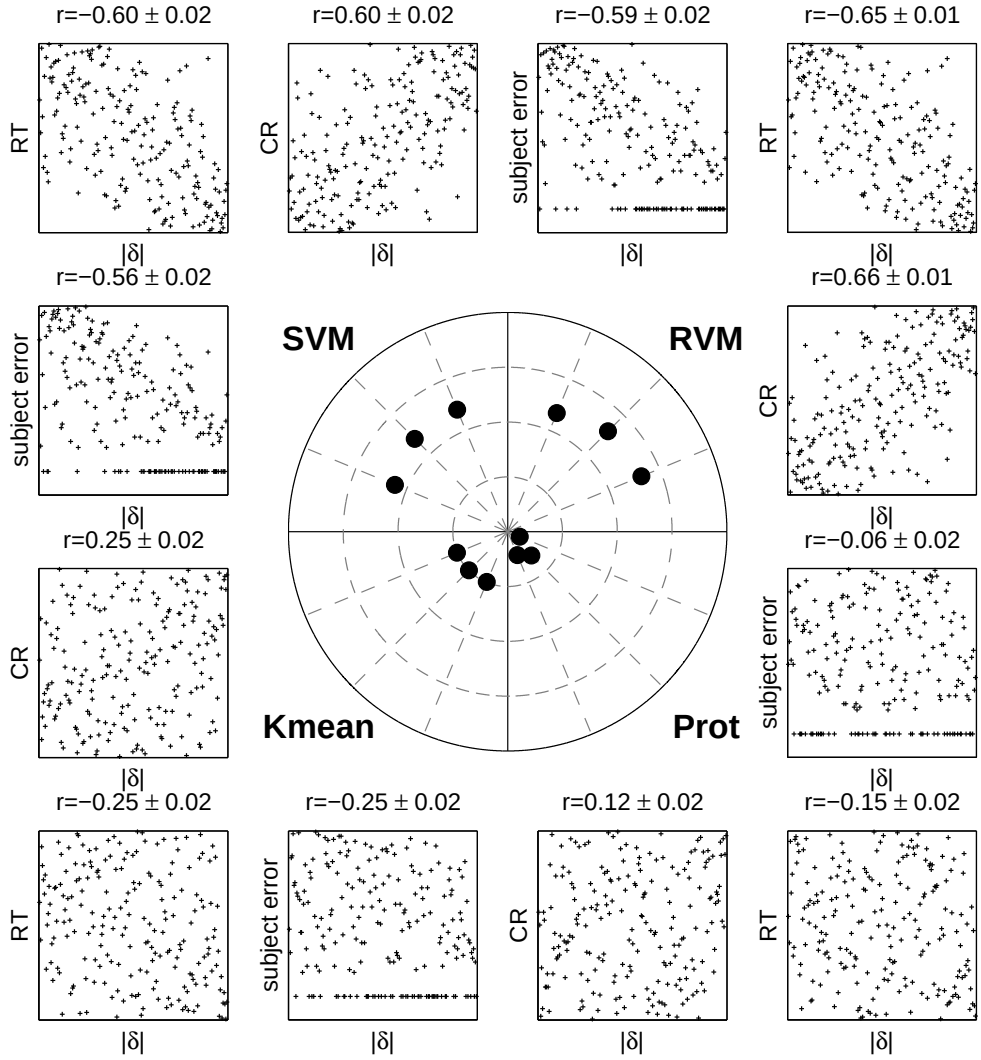


Figure 5.4: Correlation of classification behavior of man and machine with parameters averaged over subjects. On the borders: scatter plots and correlation coefficients relating on a stimulus basis the classification behavior of man (classification error, RT or CR) to the classification behavior of machine (distance $|\delta|$ of a stimulus to the SH). The tied rank of the variables are plotted and the scales range thus from 1 to 200. In the center: polar representation of the absolute value of the correlation coefficient for each classifier and human response. The origin corresponds to $|r| = 0$ and the outer circle to $|r| = 1$.

Except for Kmean, both the RT and the CR show slightly more man-machine correlation than the subjects' classification error, although all these

performance measures are related as was shown in Chapter 3. In other words, if there is a good man-machine correlation for one of the subjects' responses, we may expect to get roughly a similarly good correlation for the other responses. The above finding is consistent with the fact that RTs have been reported to play a central role in neuronal modeling where, for instance, the RT of neurons—the time of the first spikes—can be used to model face processing [Rullen, Gautrais, Delorme, and Thorpe, 1998] and orientation selectivity in the visual cortex [Delorme, 2003].

In a second attempt to isolate a smaller group of feature extractors and classifiers, we proceed to a summary analysis of the mean correlation $\langle |r| \rangle$ between the human responses and machine. Fig.5.5 shows theses summary plots for each data type, preprocessing algorithm and classifier. We also plot a selectivity "threshold" allowing to eliminate some potential candidates.

As far as the *data type* is concerned, the shape representation is less relevant to model human classification and the texture more; the image representation is an intermediate case. The shape thus helps less than the texture to discriminate genders for the considered face stimuli. The texture alone tells more about the gender of the stimuli than the image i.e. using the correspondences of the MPI face database increases the man-machine correlations. The combination of shape and texture information decreases the relevance of the latter. Further, adding the shape information also increases the classification error compared to the one obtained for the texture data alone (see Appendix F). Thus considering the shape information in addition to the texture decreases the classification performance and the man-machine correlation. This is rather unintuitive since one may expect that adding information would improve at least the classification performance of a classifier. This may suggest that an exhaustive representation is less adapted than an efficient one. Furthermore, this effect may be due to the fact that the texture and shape vectors, both rescaled to $[-1, 1]$, each have a unit weight in the texture & shape vector (uniform weighting as mentioned in Chapter 2). Further studies could assess the effect of this linear cue combination on the man-machine correlations presented here. This would give an insight on the actual importance of the texture and of the shape cues in gender classification tasks.

For the *preprocessors*, if we consider the image data, histograms and image size reduction exhibit, as could be expected, the lowest correlation coefficients, making them poor candidates to model feature extraction. Gabor wavelet filters on the other hand provide best correlation among all preprocessors. This makes again the combination of Gabor wavelet filters and the pixel data a biologically plausible feature extractor as already mentioned in the classification performance studies. Furthermore we show that the combination of PCA on image data with a prototype learner as considered in [Valentin, Abdi, Edelman, and O'Toole, 1997] seems least likely to be used by humans. PCA, ICA I, NMF and the Kernel maps behave similarly well

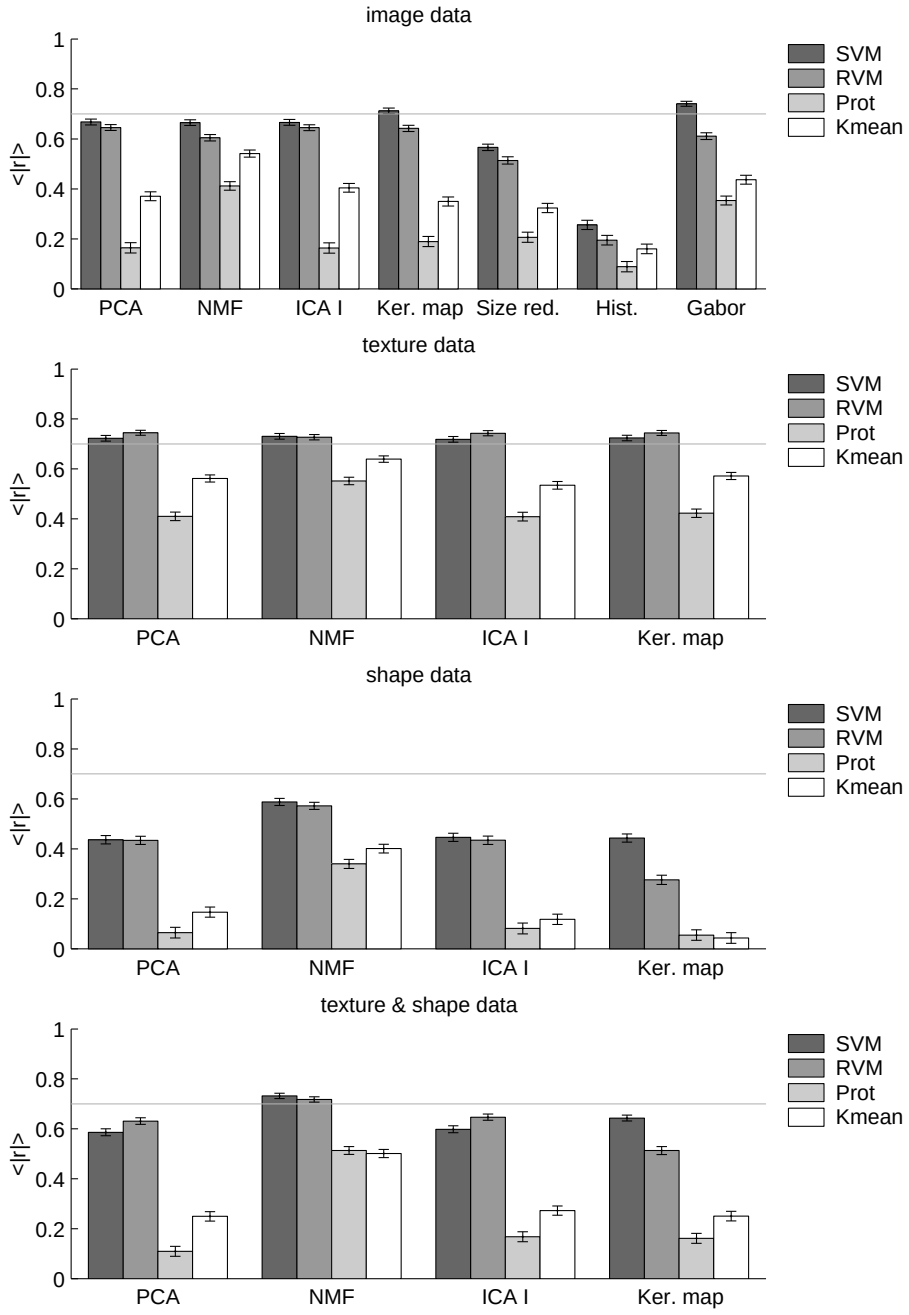


Figure 5.5: Summary plots showing the mean over the subjects' responses of the absolute value of the correlation coefficient $\langle |r| \rangle$ relating the classification behavior of man to machine for each classifier, data type and preprocessor. The horizontal line represents the threshold set at 0.7.

on the texture data whereas NMF outperforms the others on the texture & shape data type. Given the relevance of the texture & shape data type, NMF thus seems overall to be a good candidate to model preprocessing. Kernel maps are also good candidates on the image and texture data, suggesting that mechanisms similar to Gaussian windows may be used by humans to compare stimuli. Both NMF and ICA I rely on a part-based representation. However the extreme sparseness of the ICA I basis (see Chapter 2) may make it a worse candidate than NMF to model feature extraction. In conclusion, a moderately sparse part-based representation as for NMF may be a good model for feature extraction of visual stimuli. As for the detailed study for PCA on the texture & shape data type, SVMs and RVMs best correlate human to machine classification behavior and clearly outperform Prots and Kmeans. Prototype learning is thus ruled out despite the findings of [Reed, 1972] on human classification of faces and those of [Rosch, Mervis, Gray, Johnson, and Boyes-Braem, 1976] in naming tasks. The bad performance of Prots is however confirmed by [O'Toole, Defenbach, Valentin, McKee, Huff, and Abdi, 1998] where it has been argued that prototype learning could not be a model for classification. Further, linear SHs seem to be a plausible model for classification by humans as indicated by the high values of $\langle |r| \rangle$ for all the feature extractors. The bad correlation of Kmeans again indicates that a piecewise linear SH seems not to be appropriate to model human classification.

At this point we may assess the sparseness of the classifiers determined using the number of elements needed to compute the SH. The sparseness S of a dual space classifier is determined using the number of representations r and the total number of elements of the dataset $t = 152$ as follows:

$$S = \log(\frac{t}{r}) \tag{5.1}$$

We clearly have: $\lim_{r \to t} S = 0$ and $\lim_{r \to 0} S = \infty$. Fig.5.6 presents this measure of sparseness for the four classifiers investigated on the subject dataset and for all the considered feature extractors. The sparseness for the Prots is constant since this algorithm uses by construction one prototype in each class. Also it is difficult to conceive an algorithm with a higher degree of sparseness, although for RVMs one RV may be enough to proceed to classification for some easy toy datasets. Kmean is very sparse whereas SVMs are least sparse, RVMs having an intermediate behavior. These behaviors apply throughout data types and preprocessors. The sparseness of the considered classifiers is thus robust and stable across feature extractors. Since the highly-sparse Prots and Kmeans have been shown to perform worst in every respect in the previous studies, the sparseness of a classifier does not seem to be a good criterion to indicate whether an algorithm can be used to describe the mechanisms used by subjects to classify visual stimuli. Indeed the highly non-sparse SVMs have been shown to perform well in our

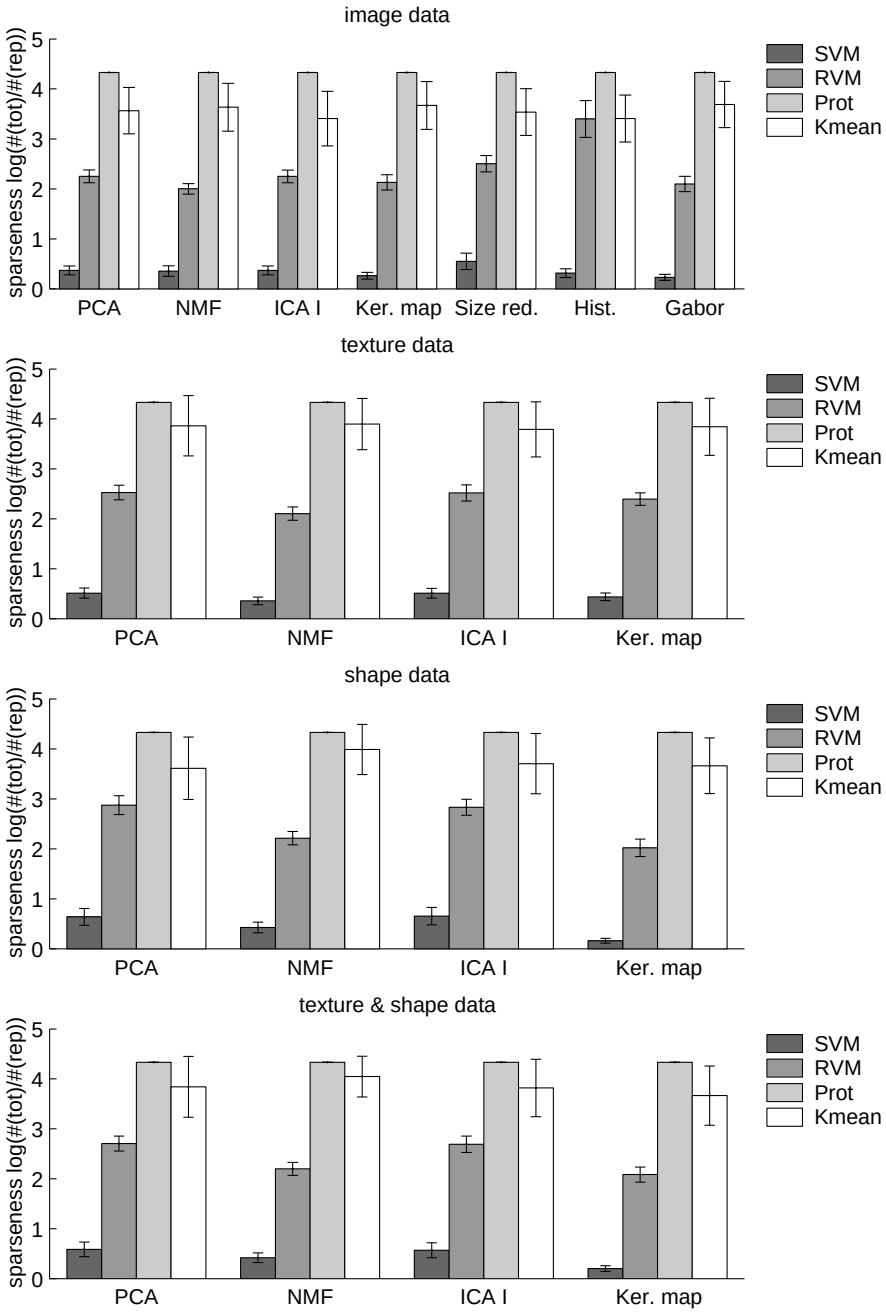


Figure 5.6: Sparseness of each classifier on the subject dataset for the various data types and preprocessors.

approach. It is however important to see the above results on sparseness under the following perspective: whereas SVMs and RVMs use elements from

the dataset as representations, the representations for Prot and Kmean are computed using *all* patterns of the dataset. If one pattern is added to the dataset or one pattern of the dataset is modified, both the prototypes and the means have to be recomputed. On the other hand, depending upon the "location" of these patterns in the space of the feature extractor, the SVs and the RVs may not be modified. SVMs and RVMs may then be argued to be more robust both to noise in the patterns and to novel patterns, which is not the case for Prot or Kmean. This is certainly an important point when modeling the classification behavior of humans, since it is unlikely that humans recompute their internal decision function for every new stimulus.

5.4 Stochastic Classification Behavior of Man

5.4.1 Methodology

We here study the stochastic nature of the classification behavior of man, in particular the jitter in the subject's classification error on the patterns near the separating hyperplane. We also conduct the correlation analysis relating man to machine on a stimulus basis as done above, however for both the first and second psychophysical classification experiment. The implications of these studies are two fold.

First, under the assumption of the SH model for classification, patterns near the SH should be subject to more jitter (instability in the subjects' responses) than those far from the SH. We use this stochastic effect to investigate yet another time the mechanisms used by humans for feature extraction and classification. For this, we compute the error difference $\Delta = |e1 - e2|$ between the subject's classification error in the first ($e1$) and second experiment ($e2$) for each stimulus. This error difference is then correlated with the distance $|\delta|$ to the SH obtained from the first classification experiment for each stimulus using the subject dataset. In the case of a meaningful model of human classification, we expect Δ to be high near the SH (i.e. for low $|\delta|$). A classifier showing no dependency between Δ and $|\delta|$ can then be argued as being not meaningful. In other words, we here study the inconsistencies of the subject's gender estimate as function of the distance of the corresponding stimuli to the SH. This study is only done for the subject classification error since the RT or CR associated with a stimulus are not expected to be submitted to jitter: a pattern close to the SH, i.e. difficult to classify, may be associated to either of the classes (jitter of the classification error), but it will always take long for subject to decide (constant high RT between both experiments) and the subject will never be confident (constant low CR).

Second, from these studies we also get a measure of the consistency of all the subjects' responses and can thus assess the stability of the subject's internal representation of the face stimuli (see also Chapter 3). Indeed the comparison of the man-machine correlation coefficients computed using the

subjects' responses from the first and second psychophysical classification experiments is then an indicator of the stability of human classification behavior and of the reproducibility of the results of these studies.

At this point it should be noted that the noise in the subjects' labeling is stochastic. In other words, both in the first and in the second classification experiments, the labels are endowed with noise. We do not have access to unnoised responses, and consequently to an unnoised SH. In the studies of this section we thus compare two noised responses, and hope to use this fact to corroborate the previous findings on the mechanisms used by humans to classify visual stimuli.

5.4.2 Results

The man-machine correlation coefficients corresponding to the first and second psychophysical classification experiments are compared in Fig.5.7 for PCA applied on the texture & shape data. Considering the polar plot, we

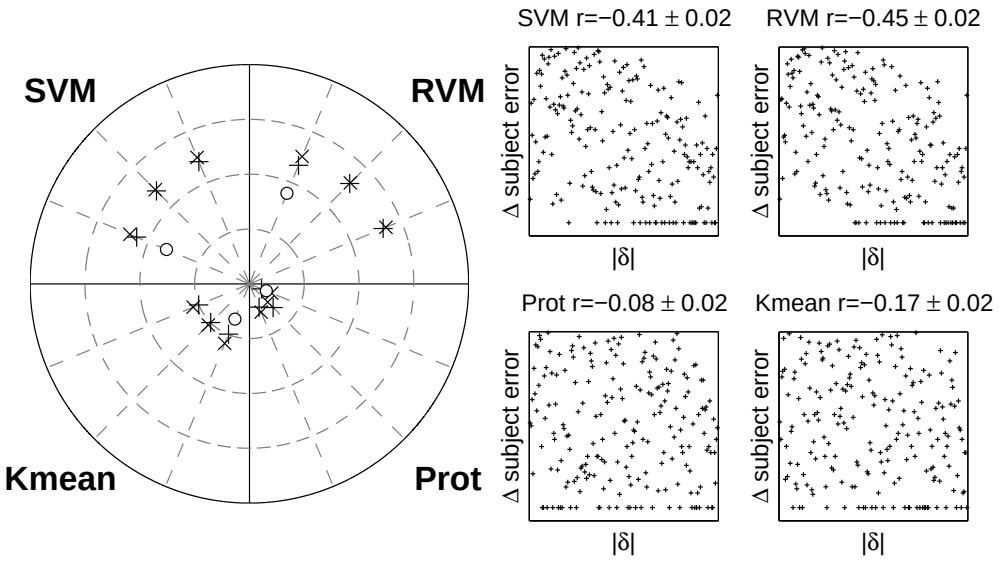


Figure 5.7: On the right: scatter plots and correlation coefficients on a stimulus basis of the difference Δ in the subject's classification error between the first and the second classification experiment and the distance $|\delta|$ of the stimuli to the SH. On the left: polar representation of the absolute value of the correlation coefficients for each classifier and human response for the first "+" and second "x" classification experiment, the "$\circ$" on the error axis representing this coefficient for the plots on the right. The axes are scaled as in the previous correlation plots.

can conclude that the correlation coefficients relating the responses of man

and machine on a stimulus basis (i.e. responses averaged across all subjects) are similar, if not identical, between the first and the second classification experiment. In other words these correlations are consistent over time, which is a consequence of Fig.3.3 in Chapter 3 where it was shown that the subjects' responses were stable on a stimulus basis (responses averaged across subjects). This suggests the stability of the subjects' internal representation of the faces and the reproducibility of the subjects' answers. The latter is a fundamental point in every scientific investigation. Furthermore the scatter plots indicate that SVMs and RVMs exhibit the highest correlation between the difference Δ in the subject's classification error between the first and second classification experiment and the distance to the SH. In both cases, most jitter in the subjects' gender estimate is observed near the SH i.e. sensitive patterns lie near the SH. This is a meaningful behavior given the SH model of classification and the mechanisms underlying RVMs and SVMs seem to be most suited to model human classification given our choice of feature extractors. The Prot and Kmean classifiers, as before, seem least adapted for this enterprise.

As a final step in assessing the feature extractors and classifiers best suited to model the classification of humans, we consider the stability of the subject's classification error—the correlation between Δ and $|\delta|$—for all data types, preprocessors and classifiers in Fig.5.8 using a selectivity threshold. As for the previous correlation analysis, these plots indicate that the shape data is less adapted to model human classification. Confirming the previous findings, Gabor filters on the image data and NMF on the texture & shape data type are the preprocessors which show best man-machine correlations. Under the hypothesis of hyperplane classification, SVMs and RVMs are the most adapted classifiers (most jitter near SH) and Prots and Kmeans the least adapted ones (jitter uniformly distributed). The good man-machine correlations of Kernel maps on the image and the texture & shape data type, what can also be observed in Fig.5.5, may also suggest that humans may use something akin to Gaussian windows to compare stimuli.

5.5 Summary & Discussion

From the studies of this chapter we can deduce that humans may use a mechanism akin to SVMs and RVMs for classification of visual stimuli. As for the feature extractors, the table below relates the combinations of data type and preprocessor which relate best man to machine in the three studies considered in this chapter. Gabor wavelet filters on the image data and NMF applied on the texture & shape data type are shown to compare and correlate overall best with human classification behavior. They may be considered as good candidates to model feature extraction of visual information by humans.

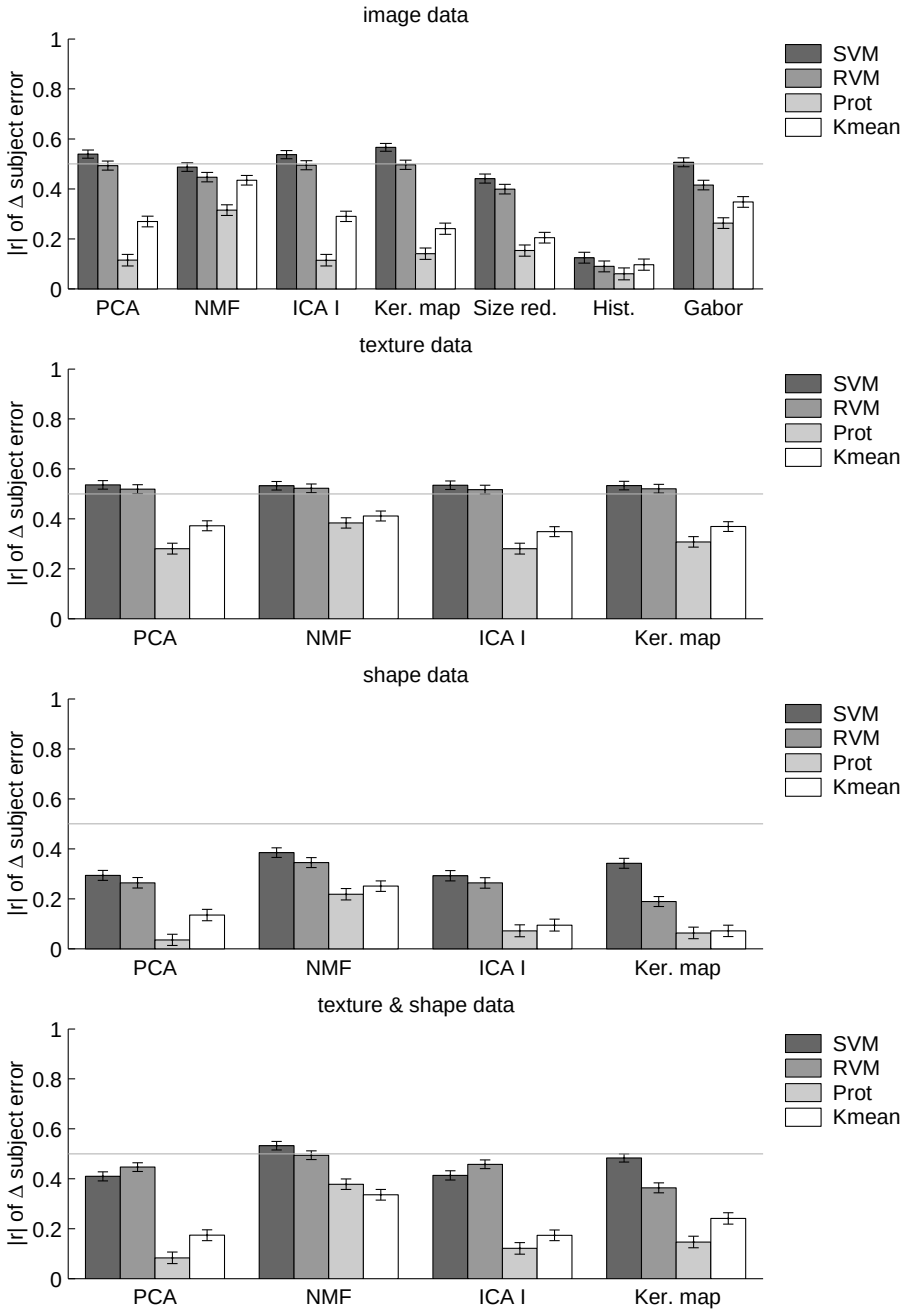


Figure 5.8: Summary plots showing the absolute value of the correlation coefficient r relating the difference Δ in classification errors of the subjects to $|\delta|$ for each classifier, data type and preprocessor. The horizontal line represents the threshold set at 0.5.

	class. error	correlation	stoch. class.
$\mathcal{I}$	Gabor wav.	Gabor wav.	Gabor wav.
	-	Kernel map	Kernel map
	-	-	PCA
	-	-	ICA I
$\mathcal{T}$	-	NMF	NMF
	-	PCA	PCA
	-	ICA I	ICA I
	-	Kernel map	Kernel map
$\mathcal{S}$	-	-	-
$[\mathcal{TS}]$	NMF	NMF	NMF
	PCA	-	-
	ICA I	-	-

Table 5.1: Best combinations of data type and preprocessor for the three types of studies done in this chapter.

As far as the *classifiers* are concerned, hyperplanes reveal to be suited to describe classification of visual stimuli in the human brain. Stimuli near this hyperplane are classified by the subjects less accurately, with higher reaction time and lower confidence than those far from it. In other words, the brain needs to do more processing for elements difficult to classify. Piecewise linear decision functions seem however not suited for modeling the classification behavior of humans. At this point it may be useful to stress the point that the optimal parameters of SVMs and Kmeans (C and K respectively) were set by a cross-validation scheme on the classification error and *not* on the man-machine correlation. Thus the computation of the SH for all four classifiers did not take the human responses into account since this would strongly bias the man-machine correlations.

Simple and very sparse algorithms such as the prototype learner and the Kmean classifier reveal to be poor candidates to model classification in the human brain. However, more elaborate and less sparse classifiers such as SVMs and RVMs seem to be more appropriate. In other words too much simplicity is bad, as already stated by Einstein (see Chapter 1). Furthermore one may speculate that the classification of visual stimuli by human subjects is done in a probabilistic manner, for instance by maximizing the conditional probability of a class membership as done by RVMs, or in a statistically optimal manner, for instance by maximizing of the margin separating both classes as done by SVMs. In other words, our results seem to indicate that exemplar-based classification algorithms may be well-suited to describe the classification behavior of humans. Although the entire dataset is needed for training, only the representations are used to define the SH, and thus also

to classify novel patterns. While the representations for SVMs, the SVs, are gathered near the margin, the RVMs spread their representations, the RVs, throughout the dataset. Therefore RVMs make a more exhaustive use of the dataset for classification and do not "waste" the data outside of the margin stripe when classifying novel patterns as is the case for SVMs. A clustering combined with a nearest-neighbor classifier and a mean-of-class prototype classification using a uniform distribution of the dual space coefficients were shown to be poor candidates for modeling classification in humans. One of the main problems arising when using prototype learners is the absence of update rule for the computation of the prototypes, which may yield a lack of "exploration" of the database since classification is solely done using 2 non-optimized patterns. Boosting the prototype classifier brings the algorithm to "divide & conquer" as shown in [Graf, Bousquet, and Rätsch, 2004a], yielding an interesting extension of the mean-of-class prototype learner, and solving the problem 7.12 stated in [Schölkopf and Smola, 2002]. The prototype formalism was also extended in [Graf, Bousquet, and Rätsch, 2004a] to allow more sophisticated update rules, among others SVMs and RVMs. In a nutshell, the SH is computed using its normal vector, the latter being defined by two patterns. These patterns are then constrained to belong to the convex hull of the dataset and are then defined as the "generalized" prototypes. In this way, most of the SH algorithms can be cast into the prototype learning formalism, the prototype being however not a simple mean-of-class but is computed according to the classification algorithm represented by the SH. Finally, a whole family of algorithms was cast in this prototype framework, allowing thus to derive a novel and powerful visualization of their classification behaviors and allowed us to gain novel insights into the principles of classification.

In order to model *feature extraction* in the human brain, Gabor wavelets on the image data and the combination of the texture & shape type with NMF seem most appropriate to account for the processing of visual information in classification tasks. The image formed on the retina is clearly most similar to the image data type and Gabor wavelets seem to be implemented in the brain [Hubel and Wiesel, 1962]. The fact that this feature extractor is revealed by the studies of this chapter as a good candidate to model the encoding of visual information is an *a posteriori* proof of the validity of our approach. While the image data is clearly biologically relevant since the image on the retina is comparable to a pixel representation, the texture & shape data type suggests that a spatial correspondence principle may also be useful when describing the internal face representation of humans. The optimal preprocessor NMF suggests that humans may use a part-based basis to represent their visual information. Furthermore, sparseness in the encoding seems less important, and sparseness of the images of the basis may be moderate. Indeed, too much sparseness in the basis as for ICA I reveals to give a poor description of human feature extraction. In other words, while

the encoding has low sparseness, a part-based basis seems most plausible to be used by humans, thus ruling out PCA with its holistic basis. A similar result was obtained by [Pelli, Farell, and Moore, 2003] in the context of word recognition: words (holistic representation) are recognized worse then letters (part-based representation). Furthermore the plausibility of local features was also verified in a different context, namely viewpoint invariant recognition by [Tarr, Bülthoff, Zabinski, and Blanz, 1997] and part-based elements are the basis for the "geons" theory of [Biederman, 1987]. However, our findings are opposite of those obtained by [O'Toole, Defenbach, Valentin, McKee, Huff, and Abdi, 1998, Valentin, Abdi, Edelman, and O'Toole, 1997, O'Toole, Abdi, Deffenbacher, and Valentin, 1993] where PCA is used as a (holistic) representation for faces and of the findings by [Gauthier, Curran, Curby, and Collins, 2003] where object processing in humans has been studied using a psychophysical setup combined with electrophysiology. Also notice that one of the main drawbacks of PCA on a theoretical ground is the indeterminacy of the directions (eigenvectors) having similar eigenvalues: this degeneracy may hamper the use of PCA to efficiently encode visual information. It can also be noticed that PCA and ICA I have similar classification behaviors: their classification performance and their man-machine correlations are roughly identical, although their basis images are totally different (holistic versus part-based). This similarity could be expected since when working with linear classifiers, PCA and ICA I are identical up to a rotation and a translation—ICA I has a PCA preprocessing step when whitening of the data. Kernel maps, although good candidates for some data types, are outperformed by NMF, hinting at the fact that humans may use a basis of images to encode visual information. Furthermore NMF exhibits a lower Fisher J separability score between classes than those for PCA, ICA I and Kernel maps. There thus seems to be no relation between the geometrical discriminability between classes and the relevance of a preprocessor for the corresponding classification tasks. Nature may thus not implement a feature extraction separating classes in an optimal way with respect to the Fisher score. Finally, histograms and neighborhood-preserving maps such as LLE have been shown to be less suited to model feature extraction in the human brain.

Obviously our results rely on some crucial assumptions. Indeed we assume the representativeness of the face space spanned by the MPI face database. In other words, the question is how well the MPI face database samples the real internal representation of faces in the human brain. In particular, when rejecting the prototype learner as a plausible candidate for human classification, we assume that the mean face of our human subjects' is close to the sample mean of our database. Clearly, a larger face database would be welcome, but is not trivial as we need the texture and shape flowfield maps. Furthermore, there is the different learning regime. Machines were trained on the dataset proper, whereas humans were assumed

to have extracted the relevant information during their lifetime, and they were tested on faces with some cues removed. In order to avoid the above assumptions, future research will deal with the use of novel low-level stimuli. In such a framework the subject will have to learn and classify: there will be a training and a testing phase for the subject. The modeling aspect will then deal with on-line learning and reinforcement learning [Sutton and Barto, 1998].

5.6 Some Related Studies

There have been some previous attempts to model classification and feature extraction in the brain using psychophysical setups. In these studies man and machine are compared, and their classification behaviors are not correlated. As shown in this dissertation, comparing man and machine is not sufficient to account for a true description of the classification mechanisms used by humans. The correlation framework introduced in this dissertation allows to pinpoint these mechanisms much better.

The first attempts to compare the classification performance of man and machine in the context of gender classification were done using Artificial Neural Network (ANN) classifiers applied on a PCA representation of the image information using an antoencoder network. ANN were shown to classify better than man, although not much, using the so-called SEXNET architecture [Golomb, Lawrence, and Sejnowski, 1991]. Further studies as done in [Gray, Lawrence, Golomb, and Sejnowski, 1995] using various resolutions of the stimuli images but without the PCA stage indicate that the gender classification problem using images of faces seems to be linearly separable since a simple Perceptron yielded results similar to a multi-layer ANN. Face stimuli with different expressions, orientations, noise and masks were considered by [Blackwell, Vogl, Dettmar, Brown, Barbour, and Alkon, 1997] in the above context: man versus ANN on the images directly (no feature extraction). At low noise levels, man outperformed machine whereas at high noise levels, humans were much worse than ANN. While the results for stimuli corrupted by noise are corroborated by the findings in this dissertation, adding noise to the face stimuli considered in this dissertation would certainly be a future direction to take to investigate feature extraction in the human brain. In conclusion, from the studies comparing man to ANN on a face database, humans were shown to perform better than machine. This finding is contrary to what was reported in [Bromley and Säckinger, 1991] where human experts and ANNs are tested on digits from the postal service database USPS. This discrepancy may be due to the nature of the stimuli and to the fact that for digits, a 10-class classification problem is considered instead of a dichotomic one as done for gender classification studies.

Gender classification on face images was done by [Moghaddam and Yang,

2002] on 21x12 pixels images using the state-of-the-art SVMs. In this approach no feature extractor was necessary since the image vectors were small enough. SVMs where shown to outperform classical methods such as Fisher, Nearest-neighbor or RBF classifiers. This study is purely computational and SVMs were mainly considered in the nonlinear case, i.e. with nonlinear kernel functions, to achieve best classification performance. SVMs have been used in the same context on the PCA and LLE of real-sized 256x256 pixels images from the MPI face database [Graf and Wichmann, 2002]. PCA was shown to be superior to LLE for this classification task, similar to the findings of this dissertation: LLE does not allow class separability and makes it a bad preprocessor for classification studies. Further, the classification performance of SVMs and humans has been compared by [Moghaddam and Yang, 2000]. The stimuli were considered at two resolutions: a low one (21x12 pixels) and a high one (80x40 pixels). In both cases SVMs performed better than humans, as also observed in this dissertation. Whereas the performance of SVMs was shown to be constant over the two resolutions, human subjects classified much better at high resolution, as could be expected.

The face recognition vendor test [Phillips, Grother, Micheals, Blackburn, Tabassi, and Bone, 2003] is an evaluation of available face recognition systems for industrial large-scale real-world applications. Although not comparing directly man to machine, it exhibits some behaviors of machine which are quite human-like. Machine was evaluated among others on its reaction time i.e. the time for novelty detection. The RT of subjects was also shown in this dissertation to be of high importance since it had a good man-machine correlation. It was also shown that male faces are more easily recognized as female ones—there are more mistakes for female faces—implying thus a male bias and corroborating the findings of this dissertation.

In the context of the "other-race effect", it was shown that subjects are better at recognizing people from their own race than people from other races [Goldstone, 2003, Furl, Phillips, and O'Toole, 2002], an explanation for this fact being that one is more accustomed to people of the same race. Some basic classifiers using PCA, Fisher Discriminant Analysis or Gabor jets were trained on images of people from one race and tested on images of people from another one. The biological plausibility of the classifier was inferred using the other-race assumption. It was concluded that the other-race effect was only present for algorithms which created "distorted" or warped face representations to emphasize features allowing classification. Although not tested, SVMs have this feature: they find a decision function which maximally discriminates the classes, i.e. maximizes the margin between the classes. The classification algorithms do not share a common ground as the ones used here (hyperplane classification) and there is no clear distinction between the feature extraction step and the classifier itself. These facts make any conclusions with respect to the mechanisms involved in the brain quite tenuous. This study can be seen as an attempt to bridge the gap

between psychophysics and theoretical modeling. This approach is however less fruitful than the one introduced in this dissertation since it does not correlate man and machine.

Learning categories from examples was studied in man and machine by [Fass and Feldman, 2003] using concepts based upon the minimum description length (MDL) principle, a concept which can be linked with Bayesian inference. The study differs from this dissertation on one main point: there was a learning and testing phase for the subjects. MDL was shown to provide a good description of the subject's performance. Furthermore, human subjects were supposed to use "complexity-minimization principles" and "cognitive codes". In this dissertation, we have provided a more thorough account on the former and latter using supervised and unsupervised machine learning respectively.

One of the first attempts to model classification is due to [Reed, 1972] where human faces are classified. Some baseline classifiers such as the the General Context Model (GCM) and the prototype learner (see Chapter 4) are used to describe the classification behavior of man. It was concluded that humans use a prototype classification rule, contrary to the finding of this dissertation.

5.7 And what about Neurophysiology?

We proceed below to a brief review of some findings of neurophysiology and/or functional Magnetic Resonance Imaging (fMRI) which can be related to the studies done here. We try to draw parallels in methodology and/or findings between studies done in man and monkey.

The mechanisms underlying classification of visual stimuli and the corresponding feature extraction have been investigated in man and monkey by [Sigala, Gabbiani, and Logothetis, 2002]. Similarity-rating tasks were performed on line drawings of faces and fish with four parameters, two of them being diagnostic ones i.e. useful for classification. Humans and monkeys were shown to behave similarly. A comparison to some baseline classification models used in the psychology community revealed that the prototype learner was least adapted to model classification and feature extraction and that a linear boundary between the classes could account for classification. These two findings are corroborated by some of the results of this dissertation. Further considerations on the physiological level as done in [Sigala and Logothetis, 2002], revealed that neurons in the inferior temporal (IT) cortex respond more strongly (i.e. have a larger activity) to the diagnostic features than to the non-diagnostic ones. Thus categorization of visual stimuli seems to be done by neurons in IT on the basis of the diagnostic features. Both studies rely on the comparison between man, monkey and machine. The approach introduced in this dissertation would allow a promising extension:

correlate man, monkey and machine using the distance of the stimuli to the separating hyperplane i.e. the decision function. One of the main criticism of this study is that the theoretical methods used to model classification in machine are derived from [Reed, 1972, Nosofsky, 1991]. The developments of machine learning, both supervised and unsupervised, are valuable replacements of such methods with more thorough theoretical foundations.

Classification in monkeys has also been investigated by [Freedman, Riesenhuber, Poggio, and Miller, 2001, 2002] in the context of a category-matching task of images of cats and dogs using a morphing software allowing to create new stimuli from these two classes, and in particular stimuli on the boundary between these classes. The importance of the elements between the classes has been demonstrated in the studies of this dissertation (see Section 5.5). Multiple visual features of the stimuli were used by monkey. The class of the visual stimuli was related to the neural activity in the lateral prefrontal cortex (PF) whose circuitry is assumed malleable, allowing thus learning and memorization. It was concluded that neurons in this area code for classification. Further studies [Freedman, Riesenhuber, Poggio, and Miller, 2003] comparing IT to PF revealed that IT seems to be responsible for the analysis of shapes as already obtained by [Sigala, Gabbiani, and Logothetis, 2002, Sigala and Logothetis, 2002] while PF seems to encode classes, memory effects and the relation to behavior. This study makes also use of the computational model by [Riesenhuber and Poggio, 1999, 2000] which models object recognition in IT cortex using the combination of a hierarchical model, receptive fields, linear combinations and nonlinear "maximum" operation gates. This model describes feature extraction, however not the classification stage. A general discussion of these findings can be found in [Riesenhuber and Poggio, 2002]. Although these studies are related to the ones done here, there is no direct comparison or correlation of the classification behavior of monkey and machine. Further, while the physiological experiments deal with classification, the model deals with feature extraction, making a comparison quite hazardous.

Feature extraction has been considered in monkeys by [de Beeck, Wagemans, and Vogels, 2001] using low-dimensional parametrized shapes by studying the neural basis of the low-dimensional representations corresponding to these stimuli. Similarities between shapes are then parametrized in a low-dimensional space obtained using Multi Dimensional Scaling (MDS). Psychophysical studies and single-cell recordings in IT suggest that the order between stimuli is kept while their distance is distorted. As seen in this dissertation, these results rely on the choice of the feature extractor. MDS belongs to the same family as LLE, which was shown to be not adapted for classification studies, at least given our face stimuli. A promising extension of the studies of [de Beeck, Wagemans, and Vogels, 2001] would thus be to consider other feature extractors, in particular Gabor wavelet filters and NMF which were shown in this dissertation to work best in the context of

visual classification in the human brain.

The above physiological studies in monkeys (single-unit recordings) make the following question arise: if we understand what happens in monkeys, how does it translate to humans? It was shown by [Tsao, Freiwald, Knutsen, Mandeville, and Tootell, 2003] using fMRI that man and monkey have similar brain architectures for the processing of visual objects. In particular, it was shown that monkeys have face-selective patches which are similar in size and number with those found in humans. Furthermore the role of learning, a concept tightly linked with classification, of such face-patches is shown to have some similarities in man and monkey.

Chapter 6

Other Approaches to Model Classification in Humans

In this chapter we present alternative approaches to the ones considered so far in this dissertation to model classification of visual stimuli by humans. The classification system composed of a linear feature extractor and a linear classifier can be visualized and allows feature ranking without use of *prior* information. A novel manner to analyze the data from the previous psychophysical experiment using machine learning is introduced. Finally, the hypotheses generated from machine learning are used to generate novel stimuli which are shown to the subjects in a novel psychophysical discrimination experiment.

6.1 Overview

We attempt to understand visual classification by human subjects. For this, we consider as in the previous chapters, the combination of human psychophysics with machine learning in the context of gender classification of images of human faces. We first choose one (linear) feature extractor and four (linear) classifiers. The studies relating man and machine introduced in Chapter 5 are then done for these choices. Because we combine a linear feature extractor with linear classifiers, we have a linear classification system allowing us to visualize the *decision image* corresponding to the normal vectors of the separating hyperplanes (SHs) of the classifiers trained on the true gender labels of the stimuli as well as on the gender labels assigned by the human subjects. On the basis of these decision images, feature ranking is discussed. These results are compared to those obtained by Recursive Feature Elimination (RFE).

Along the lines of [Graf, Wichmann, Schölkopf, and Bülthoff, 2004c], we also apply RFE to the PCA components until both man and machine show equal classification performance since the classification performance of some machines outperforms that of human subjects. This results in three

sets: (i) the distances of the stimuli to the separating hyperplane (SH) of the machines trained on the true labels, (ii) the distances of the stimuli to the SH trained on the subject labels, and (iii) the distances in the RFE-reduced-dimensionality space of the stimuli to the SH trained on the true labels. We perform a logistic regression on the average proportion of correct classification against the three sets of stimulus-to-SH distances. It will be shown that humans subjects and machines often classify faces quite differently. However some machines can re-create the internal representation of faces for human subjects very well.

Using the novel concept of decision image, we predict that the female-to-maleness transition along certain directions in face-space—normal to the SH of the linear classification system closest in behavior to human subjects—should be faster than those along other directions. A psychophysical discrimination experiment using novel stimuli computed using the normal vectors to the SH for each classifier corroborates this prediction as shown in [Wichmann, Graf, Simoncelli, Bülthoff, and Schölkopf, 2004]. The results of this experiment validate the models given by machine learning to help understand the classification of visual stimuli by humans. This experiment finally allows us to close the psychophysics-machine learning loop.

6.2 Some Algorithms from Machine Learning

For the studies of this chapter, we need to generate novel stimuli from the low-dimensional decision space where classification is performed. We choose to use the image data type to be able to visualize the faces corresponding to the novel patterns of the decision space without using the morphing algorithm. Furthermore, we require the preprocessor to be invertible, excluding thus Gabor wavelets although they were shown to model best human classification behavior in Chapter 5. Since we have shown that PCA and ICA induce by construction similar classification behaviors, this reduces the candidates for preprocessing to PCA and NMF. Because of its benchmark role in unsupervised machine learning, we choose PCA, despite the fact that the studies of Chapter 5 indicate that NMF may be slightly more adapted to model visual classification in humans on the basis of the man-machine correlations. Extending the studies of this chapter using NMF is clearly a future line of research. The PCA data is the already centered and since we will need to reconstruct the face stimuli given their low-dimensional representation, we do not normalize the data.

The results of Chapter 5 hint at the fact that it is unlikely that humans use something akin to a piecewise linear decision function as studied for the K-means classifier. Instead we here study the *Fisher linear discriminant classifier* (FLD, [Fisher, 1936, Mika, Rätsch, Weston, Schölkopf, and Müller, 2003]) which finds a direction in the dataset allowing the best sepa-

ration of the two classes. This direction is then used as the normal vector of the separating hyperplane, with the offset being optimal in the least mean square error sense as: $b = -\langle \vec{w} | \frac{\vec{p}_+ + \vec{p}_-}{2} \rangle$ where $\vec{p}_\pm$ are the prototypes of each class. The vector $\vec{\alpha}$ is the dominant eigenvector of $M\vec{\alpha} = \lambda N \vec{\alpha}$. The between-class Gram matrix is defined as: $M = |\vec{m}_- - \vec{m}_+\rangle\langle\vec{m}_- - \vec{m}_+|$ and the within-class Gram matrix as: $N = CC^T - \sum_{i=\pm} \sharp(y_i = \pm 1)|\vec{m}_i\rangle\langle\vec{m}_i|$. The parameters in these expressions are given as: $C_{ij} = \langle \vec{x}_i | \vec{x}_j \rangle$ the Gram matrix of the dataset and $\vec{m}_\pm = \frac{1}{\sharp(y_i = \pm 1)} C\vec{1}_\pm$ where $\vec{1}_+$ a vector of size p with value 0 for $\vec{x}_i | y_i = -1$ and value 1 for $\vec{x}_i | y_i = +1$ and $\vec{1}_- = \vec{1} - \vec{1}_+$. The normal vector obtained here is identical to the one computed by applying the prototype classifier on the whitened data. In fact FLD is arguably a more principled whitened variant of the prototype classifier. Indeed, the FLD weight vector can be written as $\vec{w} = S_w^{-1}(\vec{p}_+ - \vec{p}_-)$ where S_w is the within-class covariance matrix of the data. Consequently, if we disregard the constant offset b, we can write the decision function as $\langle \vec{w} | \vec{x} \rangle = \langle S_w^{-1}(\vec{p}_+ - \vec{p}_-) | \vec{x} \rangle = \langle S_w^{-1/2}(\vec{p}_+ - \vec{p}_-) | S_w^{-1/2}\vec{x} \rangle$, which is a prototype classifier using the prototypes $\vec{p}_\pm$ after whitening the space with $S_w^{-1/2}$ as pointed in [Wichmann, Graf, Simoncelli, Bülthoff, and Schölkopf, 2004]. FLD was not considered in Chapters 4 and 5 since it belongs to the category of the prototype classifiers. In the studies of this chapter, we also consider the Support Vector machine (SVM), the Relevance Vector Machine (RVM) and the (classical) mean-of-class prototype classifier (Prot).

6.3 Classification in Man and Machine

We here redo the studies presented in Chapter 5 to assess the effect of a non-normalized input space and to study the classification behavior of FLD. The plots comparing the classification performance of the classifiers to humans are shown in Fig.6.1. The classification error on the true dataset is similar between humans, SVM and FLD. Prot on the other hand has a much higher classification error than humans on both datasets. The subject dataset makes classification a harder task as illustrated by its higher classification errors when compared to the true one. Given these results, it is highly unlikely that humans use something akin to prototype classifiers, at least given the PCA representation.

The first row of Fig.6.2 shows that SVM and FLD have no training error. These classifiers are thus able to recreate the subjects' internal representation of the decision space associated to face stimuli. This results corroborates the previous finding that the classification error of humans, SVM and FLD are similar. The second row of Fig.6.2 shows the distribution of the (non-)representations—in the case of FLD, they are defined as $\vec{f}_\pm = \frac{\sum_{i|sign(\alpha_i)=\pm 1} \alpha_i \vec{x}_i}{\sum_{i|sign(\alpha_i)=\pm 1} \alpha_i}$. Both SVM and FLD separate the non-

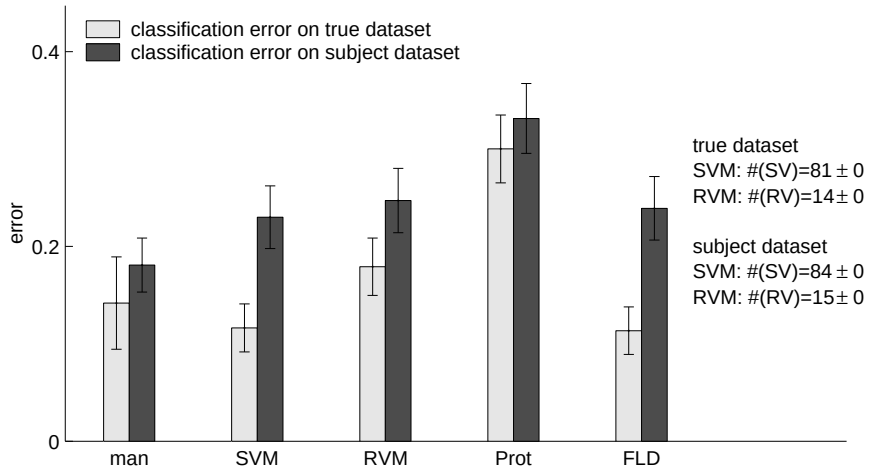


Figure 6.1: Comparison of the classification performance of man and machine on the true and subject dataset using a cross-validation scheme.

representations such as to form two peaks around the SH while RVM and Prot have this data more spread out through the dataset. Compared to the non-representations, the representations of SVM, RVM and FLD are near the SH. The last row of Fig.6.2 shows that FLD has least correlation between man and machine for the three responses of the subjects. This fact is corroborated by the correlation plots of Fig.6.3. The first row is another visualization of a previous results: SVM and FLD can recreate perfectly the subjects internal representation while RVM is less good at it. The prototype classifier is by far the worst at this task. Considering the following rows, we notice that the correlation man-machine for FLD is even worse that for the prototype classifier. The same applies when studying the stability of the subjects' responses in Fig.6.4. By comparing the above plots obtained without normalizing the PCA space to the corresponding ones in a normalized PCA space (see Appendix F), we notice that the normalization only affects slightly the results and does not affect the comparison between the classifiers.

The prototype classifier has a high training error and a low man-machine correlation. The prototype classifier can not learn the subjects' internal representation and thus it may be argued that no conclusions can be drawn from the low man-machine correlation. On the other hand, FLD has no training error, indicating that it can perfectly recreate the internal representation of humans. Thus, from its low man-machine correlation, it maybe concluded that humans do not use a mechanism akin to FLD for classification. Of course, these results hold given the PCA representation.

The bad man-machine correlation of FLD could be expected. Indeed,

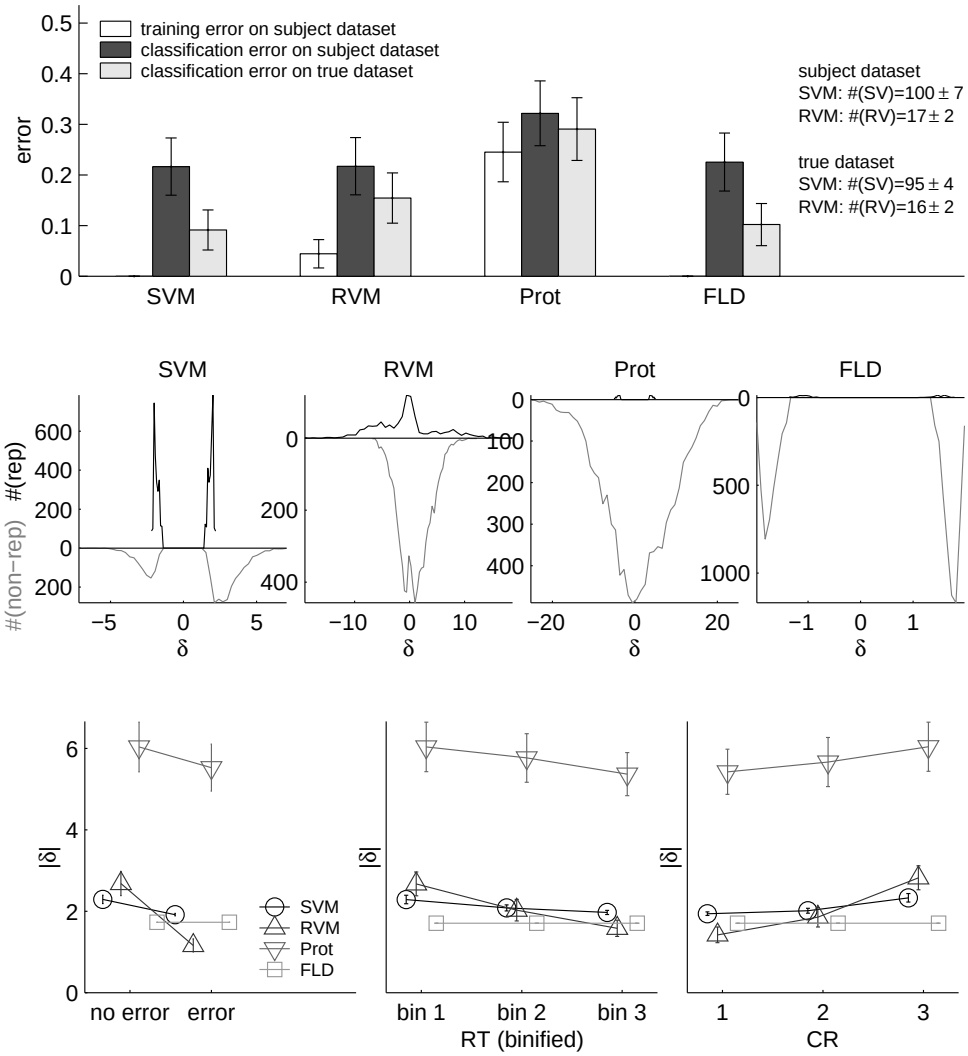


Figure 6.2: Comparisons of training and classification errors of machine without a cross-validation scheme (first row). Histograms of distances of (non-)representations to the SH (second row). Correlation of classification behavior of man and machine with parameters averaged over subjects and sets of stimuli (third row).

for 152 or 200 pattern in 200 dimensions, an algorithm such as FLD overfits and separates by construction the data in a binary way: $\delta(\vec{x}) = \pm d$ for some $d \in \mathbb{R}$. When considering the subject dataset, each of the 152 stimuli has a value of $|\delta| = d_i$ for $i = 1, \ldots, 55$. Averaging on a stimulus basis, we may expect to get roughly the same mean $\langle|\delta|\rangle$ for each stimulus. This then results in low values of the man-machine correlation.

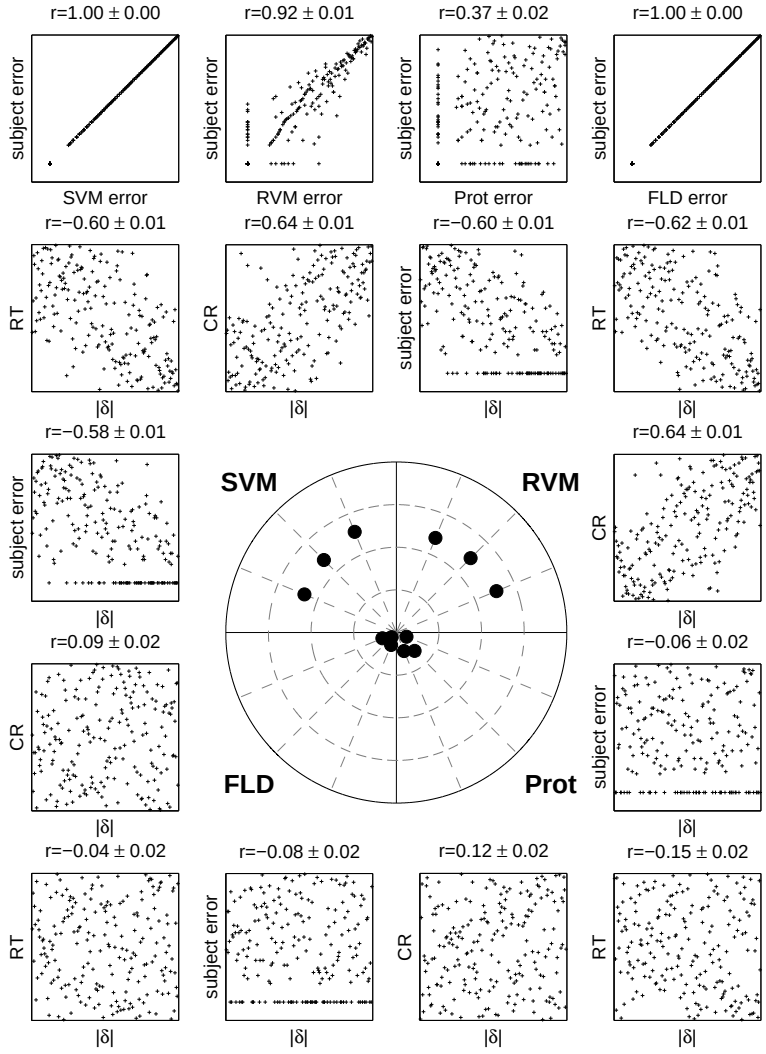


Figure 6.3: Correlation of classification behavior of man and machine with parameters averaged over subjects. First row: the subjects' error as function of the training error of machine on a stimulus-by-stimulus basis. On the borders: scatter plots and correlation coefficients relating on a stimulus basis the classification behavior of man (classification error, RT or CR) to the classification behavior of machine (distance $|\delta|$ of a stimulus to the SH). The tied rank of the variables are plotted and the scales range thus from 1 to 200. In the center: polar representation of the absolute value of the correlation coefficient for each classifier and human response. The origin corresponds to $|r| = 0$ and the outer circle to $|r| = 1$.

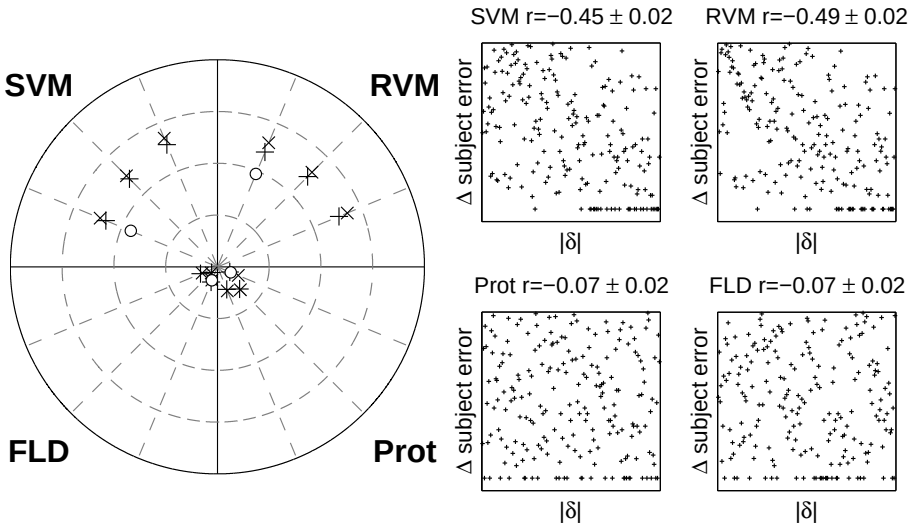


Figure 6.4: On the right: scatter plots and correlation coefficients on a stimulus basis of the difference Δ in the subject's classification error between the first and the second classification experiment and the distance $|\delta|$ of the stimuli to the SH. On the left: polar representation of the absolute value of the correlation coefficients for each classifier and human response for the first "+" and second "x" classification experiment, the "$\circ$" on the error axis representing this coefficient for the plots on the right. The axes are scaled as in the previous correlation plots.

One manner to avoid this problem is to consider the mean SH over all subjects defined by $\vec{w} = \langle \vec{w}_i \rangle$ and $b = \langle b_i \rangle$ where the SH for each subject is defined by $\vec{w}_i$ and b_i. The distances δ of the stimuli to this mean SH are then computed. For FLD, this will result in non-binary values for δ. Using this approach, the classifiers are trained either on all 200 stimuli of the true dataset or on the 152 stimuli presented to each subject. There is no testing phase and a single separating hyperplane (SH) is obtained for each classifier, as opposed to multiple SHs as obtained when doing cross-validation to estimate the classification error as done in Chapter 5 or in [Graf and Wichmann, 2004]. The histograms of the distance $\delta(\vec{x})$ of the stimuli to the SH are used to describe classification for each classifier and dataset as shown in Fig.6.5. For all classifiers, the training error on the true dataset is smaller than on the subject dataset. The histograms are also wider for the subject dataset than on the true one, suggesting that the subjects' labels make learning a harder task for the machines as already noticed previously. All classifiers except the prototype classifier have a low training error on the subject dataset showing their ability to reproduce the subjects' class labeling. The larger training error of the prototype classifier is also reflected

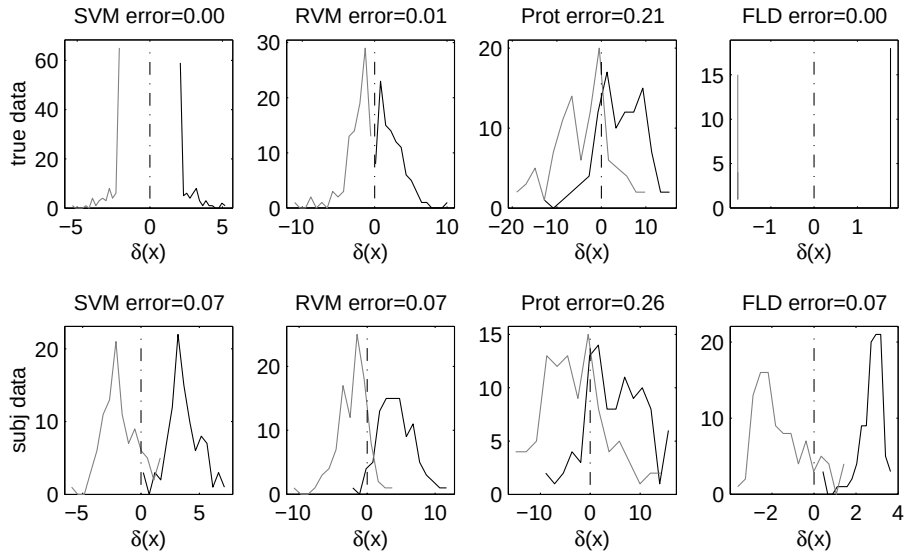


Figure 6.5: Histograms of the distance $\delta(\vec{x})$ of the faces to the SH for each classifier for the true and the subject dataset. The corresponding training errors are indicated in the titles. The light line indicates the female stimuli whereas the dark one the male stimuli.

by the overlap of the histograms for male and female stimuli. For FLD, we see that, as expected, the histogram of δ on the true dataset is binary because of the overfitting done by FLD on 200 patterns in 200 dimensions. On the other hand, this histogram on the subject dataset, computed using the mean SH, spreads these two peaks into distribution of non-zeros supports.

6.4 The Decision Images

We introduce below the concept of decision image (see also [Wichmann, Graf, Simoncelli, Bülthoff, and Schölkopf, 2004, Graf, Wichmann, Schölkopf, and Bülthoff, 2004c]) and show its relation to another important issue related to classification: *feature ranking*. We subsequently compare our findings with those obtained using a principled method from machine learning, namely Recursive Feature Elimination (RFE).

The data matrix $X \in \mathbb{R}(200 \times 256^2)$ is defined by the 200 faces of size 256x256 from the MPI face database. Using the same experimental procedures as previously described in Chapter 3, 55 human subjects are asked to classify a random subset of 152 of the 200 faces and the subjects' gender estimate $\hat{y}$ is recorded for each presented face with the convention that $\hat{y} = -1$ for a female face and $\hat{y} = +1$ for a male face. We consider no dimensionality reduction and thus keep all 200 components of the PCA when applied

to the data matrix X. This implies that the reconstruction of the data is perfect and we can write: $E = \bar{X}B^T \Leftrightarrow \bar{X} = EB$ where $E \in \mathbb{R}(200 \times 200)$ is the matrix of the encodings (each row is a PCA vector in the space of reduced dimensionality), $B \in \mathbb{R}(200 \times 256^2)$ is the orthogonal basis matrix and $\bar{X}$ the centered data matrix. The combination of the encoding matrix E with the true class labels y of the MPI database yields the *true* dataset, whereas its combination with the class labels $\hat{y}$ of the subjects yields the *subject* dataset.

To model classification in human subjects, we use methods from supervised machine learning. In particular, we examine linear classifiers where classification is done using a SH defined by its normal vector $\vec{w}$ and offset b. Furthermore we require the classifier to be in dual form: the normal vector can be written as a linear combination of the patterns $\vec{x}_i \in \mathbb{R}^{200}$ (the i^{th} row of E) as $\vec{w} = \sum_i \alpha_i \vec{x}_i$.

The combination of a linear feature extractor $\bar{X} = EB$ and a linear classifier $y(\vec{x}) = \langle \vec{w} | \vec{x} \rangle + b$ yields a linear classification system: $\vec{y} = \vec{w}^T E^T + \vec{b}$ where $\vec{b} = b\vec{1}$. We define the *decision image* as the vector $\vec{W} \in \mathbb{R}^{256^2}$ effectively used for classification as: $\vec{y} = \vec{W}^T \bar{X}^T + \vec{b}$. The decision image is an image vector incorporating the feature extraction and the classification algorithms and can be used for a compact and efficient visualization. The classification of an image is then done by comparing it to this decision image. We then have $\vec{w}^T E^T = \vec{W}^T \bar{X}^T \Leftrightarrow \vec{w}^T B^{-T} \bar{X}^T = \vec{W}^T \bar{X}^T$ where B^{-1} is the pseudo-inverse of B. From the last condition, we obtain a definition of the decision image $\vec{W} = B^{-1}\vec{w} \in \mathbb{R}^{256^2}$. In the case of PCA where $B^{-1} = B^T$, we simply have $\vec{W} = B^T \vec{w}$.

Fig.6.6 shows the decision images $\vec{W}$ for the four classifiers, SVM, RVM, Prot and FLD. The decision images in the first row are those obtained if the classifiers are trained on the true dataset; those in the second row if trained on the subject dataset, marked on the right hand side of the figure by "true data" and "subj data", respectively. Decision images are represented by a vector pointing to the positive class and can thus be expected to have male attributes (the negative, $-\vec{W}$, of it looks female). The inspection of the decision images is instructive since it allows to proceed to *feature ranking*: both dark and light regions are more important for classification than the grey regions. For the prototype learner, the eye and beard regions are most important. SVM, RVM and FLD have somewhat more "holistic" decision images. Equally instructive is the comparison of the optimal decision images of the classifiers in the first row and those in the second row where the machines attempt to re-create the decision boundaries of the subjects. The decision images for the subject dataset are slightly more "face-like" and less holistic than those obtained using the true labels; the eye and mouth regions are more strongly emphasized. This trend is true across all classifiers. This suggest that human subjects base their gender classification strongly on the eye and mouth regions of the face—clearly a sub-optimal strategy as revealed

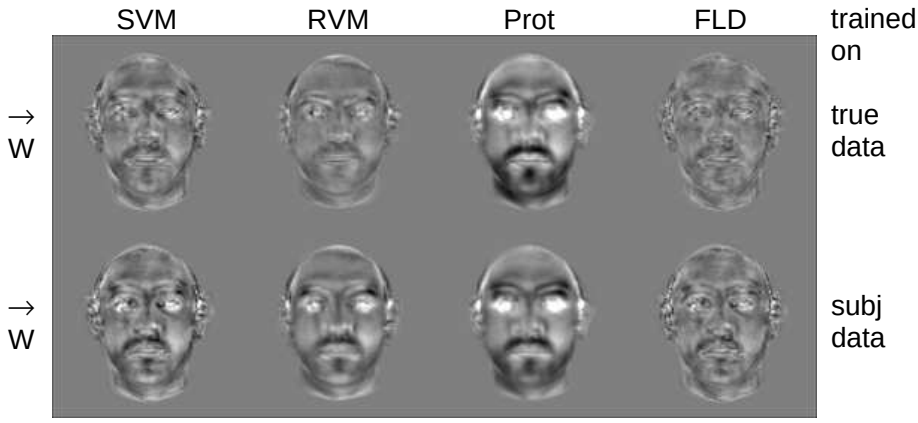


Figure 6.6: Decision images $\vec{W}$ for each classifier for both the true and the subject dataset; all images are rescaled to $[0, 1]$ and their means set to 128 for illustration purposes (different scalers for different images).

by the more holistic true dataset SVM, RVM and FLD decision images. A decision image thus represents a way to extract the visual cues and features used by human subjects during visual classification without using *a priori* assumptions or knowledge about the task at hand. It allows one to proceed to feature ranking by highlighting the regions of the image most useful for classification. In addition this method does not require multiple training and testing iterations as is typical for feature ranking methods in machine learning such as recursive feature elimination (see below).

We can also define the *generalized portraits* $\vec{W}_{\pm}$. This term was introduced by [Vapnik and Lerner, 1963] with the idea in mind that when trained on a set of portraits of members of a family, one would obtain a "generalized" portrait which captures the essential features of the family as a superposition of all family members. The generalized portraits $\vec{W}_{\pm}$ can be seen as "summary" faces in each class reflecting the decision rule of the classifier. They can be viewed as an extension of the concept of a prototype: they are the prototype of the faces the classifier bases its decision on. We note that $\vec{w}$ can be written as: $\vec{w} = \sum_i \alpha_i \vec{x}_i = \sum_{i|sign(\alpha_i)=+1} \alpha_i \vec{x}_i - \sum_{i|sign(\alpha_i)=-1} |\alpha_i| \vec{x}_i$. This allows to define the generalized portraits as $\vec{W}_{\pm}$ which are computed by inverting the PCA transformation on the patterns $\vec{w}_{\pm} = \frac{\sum_{i|sign(\alpha_i)=\pm 1} \alpha_i \vec{x}_i}{\sum_{i|sign(\alpha_i)=\pm 1} \alpha_i}$. The vector $\vec{w}_{\pm}$ is constrained to be in the convex hull of the data in the respective class in order to yield a "viewable" portrait. The generalized portraits for the SVM, RVM and FLD together with the Prot, where the prototype is the same as the generalized portrait, are shown in Fig.6.7. Again, on the subject data, the mean SH across all subjects is considered. The generalized portraits can be associated with the correct class: $\vec{W}_{+}$ are

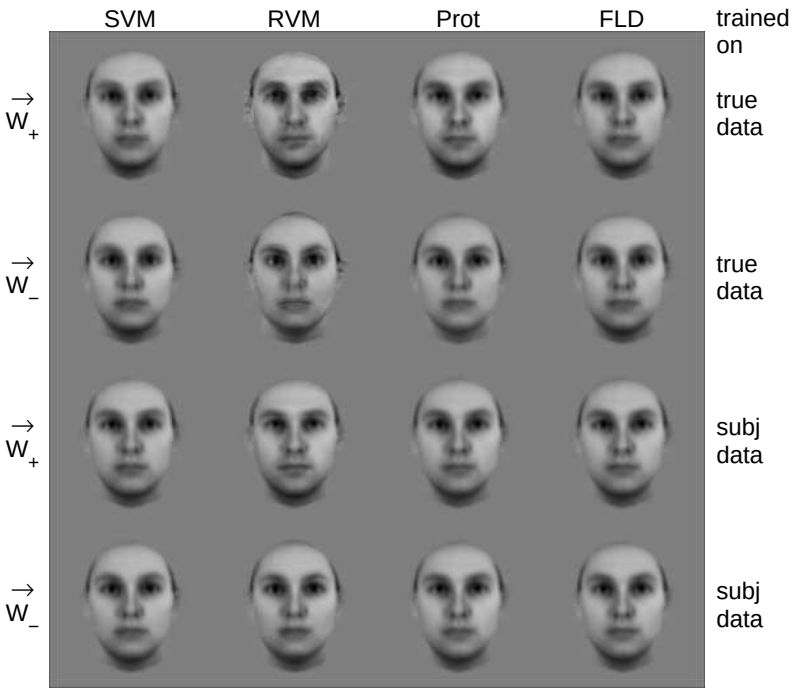


Figure 6.7: Generalized portraits $\vec{W}_\pm$ for each classifier for both the true and the subject dataset; all images are rescaled to $[0, 1]$ and their means set to 128 for illustration purposes (different scalers for different images). [Unfortunately the downsampling (low-pass filtering) of the faces necessary to fit them in the figure makes all the faces somewhat more androgynous than they are viewed at full resolution.].

males whereas $\vec{W}_-$ are females. The SVM and the FLD use patterns close to the SH (see Fig.6.2) for classification and hence their decision images appear androgynous, whereas Prot and RVM tend to use patterns distant from the SH resulting in more female and male generalized portraits. The comparison of the optimal, true, generalized portraits to those based on the subject labels shows that classification has become more difficult: the generalized portraits have moved closer to each other in gender space, narrowing the distance between the classes and thereby diminishing the gender typicality of the generalized portraits for all classifiers.

We have previously studied the decision images and mentioned their function in feature ranking. Recursive Feature Elimination (RFE, [Guyon, Weston, Barnhill, and Vapnik, 2002]) is a benchmark method in machine learning used to rank features. We here use RFE to rank features and in the next section to study the classification error of the machine classifiers when successively removing components from the PCA space. RFE is used to find

the components of the stimuli in the PCA space—which correspond to the features in the image space—most useful for each classifier to perform gender classification. The classification performance of humans is then used to define the size of the set of most important features, i.e. the dimensionality of the decision space. To apply RFE, we first compute the normal vector $\vec{w}$ for each classifier using the entire true dataset. Feature are removed successively one-by-one by eliminating the least important feature $k = \arg\min_i w_i^2$ from all the patterns of the dataset at each iteration and by training the classifier on the remaining ones. These remaining features are those most important as a group and not necessarily individually. At each step of the RFE algorithm, the training error of each algorithm is computed as shown in Fig.6.8. If classification performance was a meaningful measure, the reduced

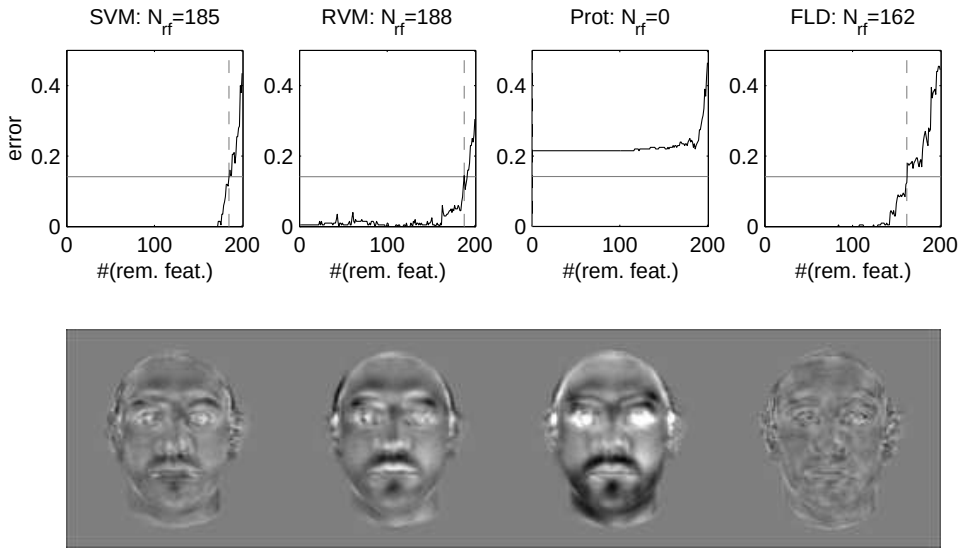


Figure 6.8: First row: training error of machine as function of the number of removed features, the horizontal line indicating the mean subject error and the vertical one the corresponding number of removed feature N_{rf} indicated in the title. Second row: decision images obtained after RFE rescaled to the range $[0, 1]$ and with means set to 128.

set of features yielding the same classification error for machine and humans should result in machines similar in behavior to our human subjects. For each classifier these features of $\vec{w}$ are kept whereas the remaining ones are set to zero. This issue is studied further in the next section.

The decision image corresponding to this "reduced" normal vector are displayed in Fig.6.8 for each classifier. Similarly to the previous decision images on the true and subject datasets, a low intensity (dark region) indicates a feature used to determine female faces whereas a high intensity

(a light region) shows regions of the face used for a male decision. Both regions are of importance for classification. For the prototype learner, the eye and beard regions are most important whereas the other classifiers appear more "holistic". These decision images, although computed using the true dataset, result in machines with the subjects' classification error. These RFE-decision images are roughly similar to the ones obtained on the true and on the subject dataset and yield no additional clues on the classification mechanisms used by man and machine. However, as will be shown in the next section, the decision space induced by RFE is very distinct from the one obtained for the true and the subject datasets.

6.5 Man-Machine Analysis Using Logistic Regression

We here give a novel approach aiming at comparing linear classifiers to human visual classification as presented in [Graf, Wichmann, Schölkopf, and Bülthoff, 2004c]. We attempt to gain more understanding of the classification of visual stimuli by human subjects using techniques from machine learning. Ultimately we would like to understand (the metric of) the human internal representation of faces.

After preprocessing the $p = 200$ face stimuli using Principal Component Analysis (PCA) resulting in the input patterns $\vec{x}_i, i = 1, \ldots, p$, we train a Support Vector Machine (SVM), a Relevance Vector Machine (RVM), a prototype classifier (Prot) and a Fisher linear discriminant classifier (FLD) on the true labels of the stimuli (true dataset) as well as on the labels assigned by the human subjects (subject dataset). On the true dataset some machines outperform the human subjects on our gender discrimination task as far as the classification performance is concerned (see Fig.6.1). To equate the performance of human subjects and machines in terms of classification error, we apply Recursive Feature Elimination (RFE) to the PCA components of the stimuli and remove their components one-by-one until both man and machine show equal classification performance. This results in three sets of distances $\delta(\vec{x}_i)$ of the stimuli $\vec{x}_i$ to the separating hyperplanes (SHs) of each classifier:

1. the set of distances $\Delta_{true} = \{\delta_{true}(\vec{x}_i)\}_{i=1}^{p}$ of the stimuli to the SH computed using the machines trained on the true dataset

2. the set of distances $\Delta_{subj} = \{\delta_{subj}(\vec{x}_i)\}_{i=1}^{p}$ to the stimuli to the mean SH across all subjects computed using the machines trained on the subjects dataset

3. the set of distances $\Delta_{RFE} = \{\delta_{RFE}(\vec{x}_i)\}_{i=1}^{p}$ of the RFE-reduced-dimensionality stimuli to the SH computed using the machines trained on the RFE-reduced-dimensionality stimuli with the true labels

Averaging across all the 55 subjects, we can assign a proportion correct to every one of our stimuli: this is the probability that a given stimulus will be correctly classified by human subjects. Thus we can perform a logistic regression on the average proportion correct of a stimulus across subjects against the three sets of stimulus-to-SH distances. If any of the machines trained on one of the three sets of distances Δ_{true}, Δ_{subj} and Δ_{RFE} may have captured more than just the input-output (classification error) mapping of the human subjects but instead captured some aspects of the human internal representation of faces, then the distance of a face to the SH should reflect the classification difficulty. Thus a regression of a monotonic function against the Δ-sets on the x-axis and the classification probabilities of the human subjects on the y-axis should yield a good fit, an "averaged" psychometric function.

The "subjects' outputs" are the mean probability $P(\hat{y} = +1|x)$ that a stimulus x is classified as male across all our subjects. Looking at the probabilities we find that they are almost uniformly distributed over the interval $[0, 1]$, that is, there are some faces that are correctly classified as females by all subjects ($P(\hat{y} = +1|x) = 0.0$), some faces that are correctly classified as males by all subjects ($P(\hat{y} = 1|x) = 1.0$) and almost all values in between appear such as faces that half the human subjects classified as females and half as males ($P(\hat{y} = 1|x) = 0.5$). This situation is typical for virtually all psychophysical tasks where human performance is a smooth, monotonic function of task difficulty. If the distance of a face to the SH reflected the classification difficulty, then a regression of a monotonic function against the different sets of distances (Δ_{true}, Δ_{subj} and Δ_{RFE}) on the x-axis and the classification probabilities $P(\hat{y} = +1|x)$ on the y-axis should yield an "averaged" psychometric function. We fit "psychometric functions", henceforth simply referred to as *logistic regressions*, using the constrained maximum-likelihood methods described in [Wichmann and Hill, 2001a]. Goodness-of-fit is assessed using a deviance-variant D/p, the log-likelihood ratio statistic normalized by the number of data points (the $p = 200$ faces).

We plot in Fig.6.9 the logistic regression for the set of distances Δ_{true} computed using the true labels. The relative high values of the deviance D between the logistic regression function and the data indicates that the combination of the distance set Δ_{true} for all studied classifiers does not reflect the essential structure of the human internal representation of faces.

In Fig.6.10 we show the logistic regression for the set of distances Δ_{RFE} computed in the RFE-reduced-dimensionality face space and using the true labels. The even higher values of D indicate that this type of set of distances Δ_{RFE} is even worse for modeling the human internal representation of faces than the previous distance type Δ_{true} using the true labels. First we recall that all the machines, despite classifying in the space of reduced dimensionality, have same classification error as humans who are shown stimuli from the full space. Second, the logistic regression of the subjects' gender

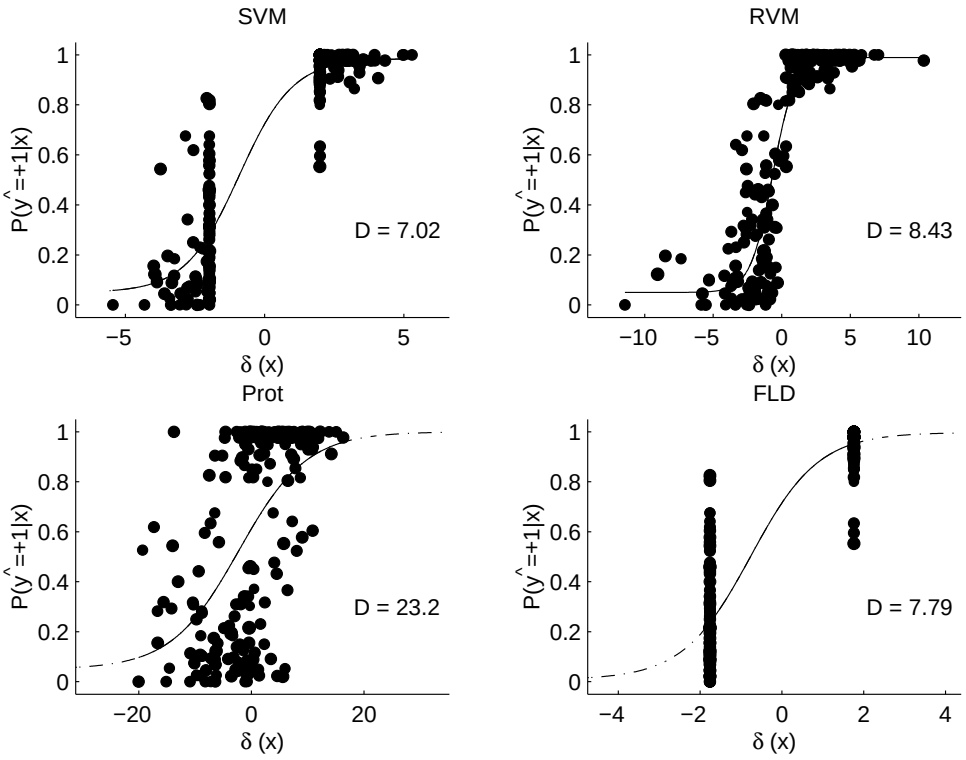


Figure 6.9: Logistic regression of the subjects' probability to answer male as function of the distance of the stimuli to the SH for Δ_{true}.

estimate to Δ_{RFE} of machine is very poor, and even worse than human responses compared to machine classifying using the full input of the true dataset Δ_{true}. This may suggest that the classification error *per se* is not a good measure. Moreover, PCA may not be the right representation for stimuli given a gender classification task of human faces, unlike the results of [Turk and Pentland, 1991, O'Toole, Abdi, Deffenbacher, and Valentin, 1993].

The best fit between the subjects' data and one of our sets of distances is obtained in Fig.6.11 for the distance set Δ_{subj} computed using the subjects' gender labels. The low values of D for SVM and FLD are striking: both machines are not only able to re-create the decision boundaries of human subjects in terms of classification error (i.e. 0% training error, implying 14% classification error on true labels as the subjects as shown in Fig.6.1) but they appear to capture the human internal representation of faces to a remarkable degree. For this task, RVM is also a rather good candidate. However, the prototype classifier again fails at this, as already noticed in [Graf and Wichmann, 2004]. Furthermore, the prototype learner is the classifier where

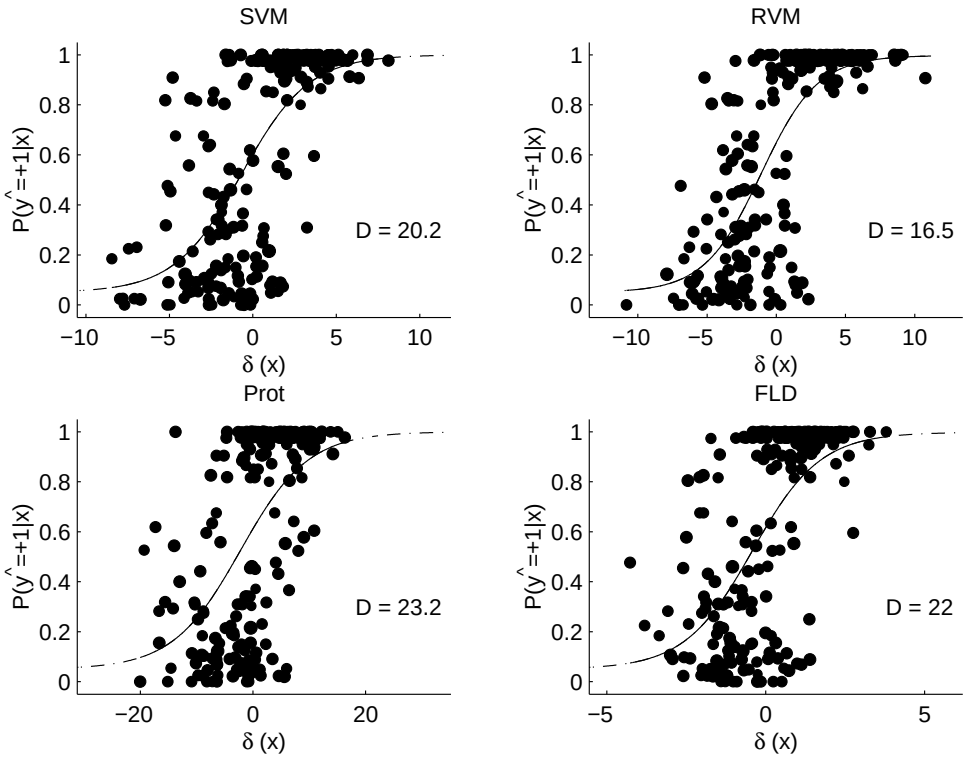


Figure 6.10: Logistic regression of the subjects' probability to answer male as function of the distance of the stimuli to the SH for Δ_{RFE}.

for each of the three sets of distance, the goodness-of-fit D of the logistic regression to the data is consistently the worst.

The good performance of FLD here seems to be in contradiction with its bad man-machine correlation as shown at the beginning of this chapter. This may be explained by the fact that here we take the mean SH over all subjects to compute δ, whereas in the correlation studies δ was computed for each subject and then averaged over stimuli. The good logistic regression for the data computed using FLD demonstrates the efficiency of taking the mean SH to compute δ. On the other hand, the other classifiers behave as in the correlation studies: SVM and RVM behave most human-like whereas the prototype classifier is least appropriate to model human classification. This illustrates that FLD, more than any of the other three classifiers, is sensitive to the (data) analysis—classification, correlation and regression studies. FLD can thus be considered as a poor candidate to model the classification of visual stimuli by humans.

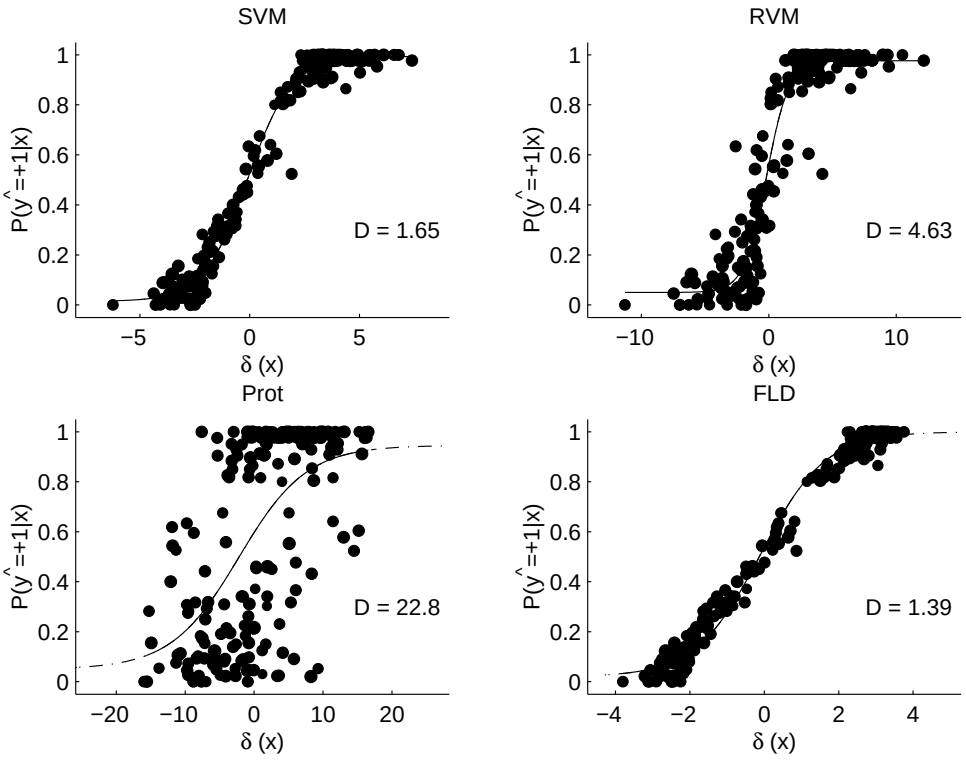


Figure 6.11: Logistic regression of the subjects' probability to answer male as function of the distance of the stimuli to the SH for Δ_{subj}.

6.6 Going Orthogonal, and Closing the Loop

We here introduce a novel psychophysical setup where machine learning is applied to perception in order to account for human gender discrimination as presented in [Wichmann, Graf, Simoncelli, Bülthoff, and Schölkopf, 2004]. We use the vector $\vec{w}$ of each classifier to generate testable hypothesis about the classification mechanisms used by humans: we generate novel stimuli by adding (or subtracting) various "amounts" ($\lambda \frac{\vec{w}}{\|\vec{w}\|}$) to the origin of the PCA space. The novel image stimuli, $I(\lambda)$, are generated by inverting the PCA transformation as $I(\lambda) = PCA^{-1}\left(\lambda \frac{\vec{w}}{\|\vec{w}\|}\right)$. The correlation studies at the beginning of this chapter reported that the subjects' responses to the faces—proportion correct, reaction times and confidence ratings—correlated very well with the distance of the stimuli to their separating hyperplane (SH) for support and relevance vector machines (SVMs, RVMs) but not for simple prototype (Prot) classifier or for FLD. If these correlations really implied that SVM and RVM capture some crucial aspects of human internal face representation the following prediction must hold: already for small

$|\lambda|$, $I_{SVM}(\lambda)$ and $I_{RVM}(\lambda)$ should look male/female whereas $I_{Prot}(\lambda)$ and $I_{FLD}(\lambda)$ should only be perceptually male/female for larger $|\lambda|$. In other words: the female-to-maleness axis of SVM and RVM should be closely aligned to those of our subjects whereas that is not expected to be the case for FLD and Prot. A psychophysical gender discrimination experiment confirms our predictions: the female-to-maleness axis of the SVM and, to a smaller extent, RVM, are more closely aligned with the human female-to-maleness axis than those of the prototype (Prot) and a Fisher linear discriminant (FLD) classifier. In other words, from the analysis of the machines we make predictions for human subjects which we subsequently test psychophysically. By doing so, we close the man-machine loop, and demonstrate that machine learning is a suitable method to model the classification of visual stimuli, at least on the considered face database.

In the studies of this section only the subject dataset is considered since the latter reflects what we hypothesize to be the subjects' internal representation of faces and we consider the mean SH across all subjects for each classifier. We first visualize some faces along a direction orthogonal to the SH of each classifier. The patterns in the PCA space are defined as:

$$\vec{x}(\lambda, \vec{z}) = \vec{z} + \lambda \frac{\vec{w}}{\|\vec{w}\|}$$

where $\vec{z}$ is a point of the PCA space. The face images are then obtained by inverting the PCA transformation. Fig.6.12 represents such faces computed from the origin of the PCA space, $\vec{z} = 0$, i.e. from a neutral genderless face. When starting from the prototypes of each class ($\vec{z} = \vec{p}_{\pm}$), we obtain the faces of Fig.6.13. These faces lie along an axis assumed to allow the sharpest categorical decision in the decision space of each classifier. The four considered classifiers yield different "optimal" decision directions as shown by the novel stimuli generated on them. This effect is strongest for the extremal values $\lambda = \pm 18$ where we obtain female and male "caricatures" along an axis defined by each classifier.

In order to yield a more quantitative measure of the class transition between female and male stimuli, we compute an estimate of the probability of male answer $P(\hat{y} = +1|x)$ for each of the novel face stimulus given λ. For each novel stimulus, the 25 patterns in the PCA space with most similar δ are considered—considering the closest ones would be less meaningful since δ was shown to be most appropriate to model classification in humans [Graf and Wichmann, 2004]. An average over the subjects' responses over these "nearest" patterns is then computed to get an estimate of the class probability for the novel patterns. The function relating $P(\hat{y} = +1|x)$ for the novel patterns to λ can then be displayed as done in Fig.6.14. The prototype learner exhibits the slowest class transition, the latter being also non-monotonic. This hints again at the fact that this classifier is poorly adapted to model visual classification by humans. All the other classifiers

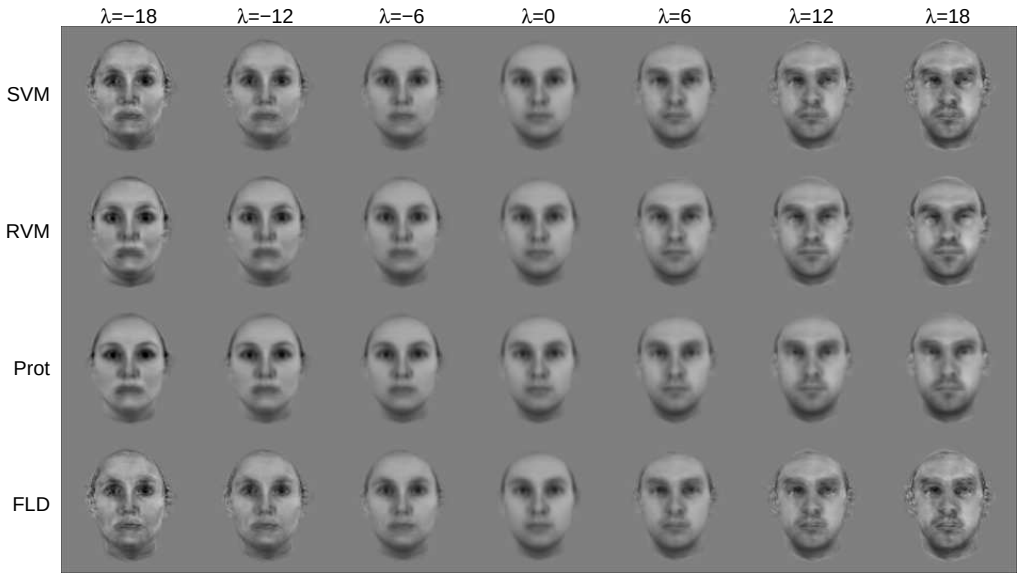


Figure 6.12: Novel face stimuli generated on a direction orthogonal to the SH staring from the origin of the PCA space.

have a steep monotonic transition. SVM and FLD have the strongest slope, indicating a sharp perceptual boundary between both classes as was already observed for Δ_{subj} in the logistic regression studies of the previous section. The male bias of the human subjects observed in Chapter 3 explains that $P(y_{est} = +1|\vec{x})$ does not drop to 0 in the female class ($\lambda < 0$). For SVM, RVM and FLD we notice that the transition happens for values of λ in the range $[-3, 3]$. The above functions are obtained by estimating the subjects' responses on the novel stimuli generated along a direction orthogonal to the SH using the subjects' gender estimates on the stimuli of the dataset. Below, we proceed to a psychophysical experiment where the subjects actually classify these novel stimuli for values of λ in the range $[-3, 3]$.

Four observers—one (FAW) with extensive psychophysical training and three naïve subjects paid for their participation—took part in a standard, spatial (left versus right) two-alternative forced-choice (2AFC) discrimination experiment. Subjects were presented with two faces $I(-\lambda)$ and $I(\lambda)$ and had to indicate which face looked more female. Stimuli were presented against the mean luminance ($50\,\mathrm{cd/m^2}$) of a carefully linearized Clinton Monoray CRT driven by a Cambridge Research Systems VSG 2/5 display controller. Neither male nor female faces changed the mean luminance. Subjects viewed the screen binocularly with their head stabilized by a headrest. The temporal envelope of stimulus presentation was a modified Hanning window (a raised cosine function with rise and fall times of 500 ms and a

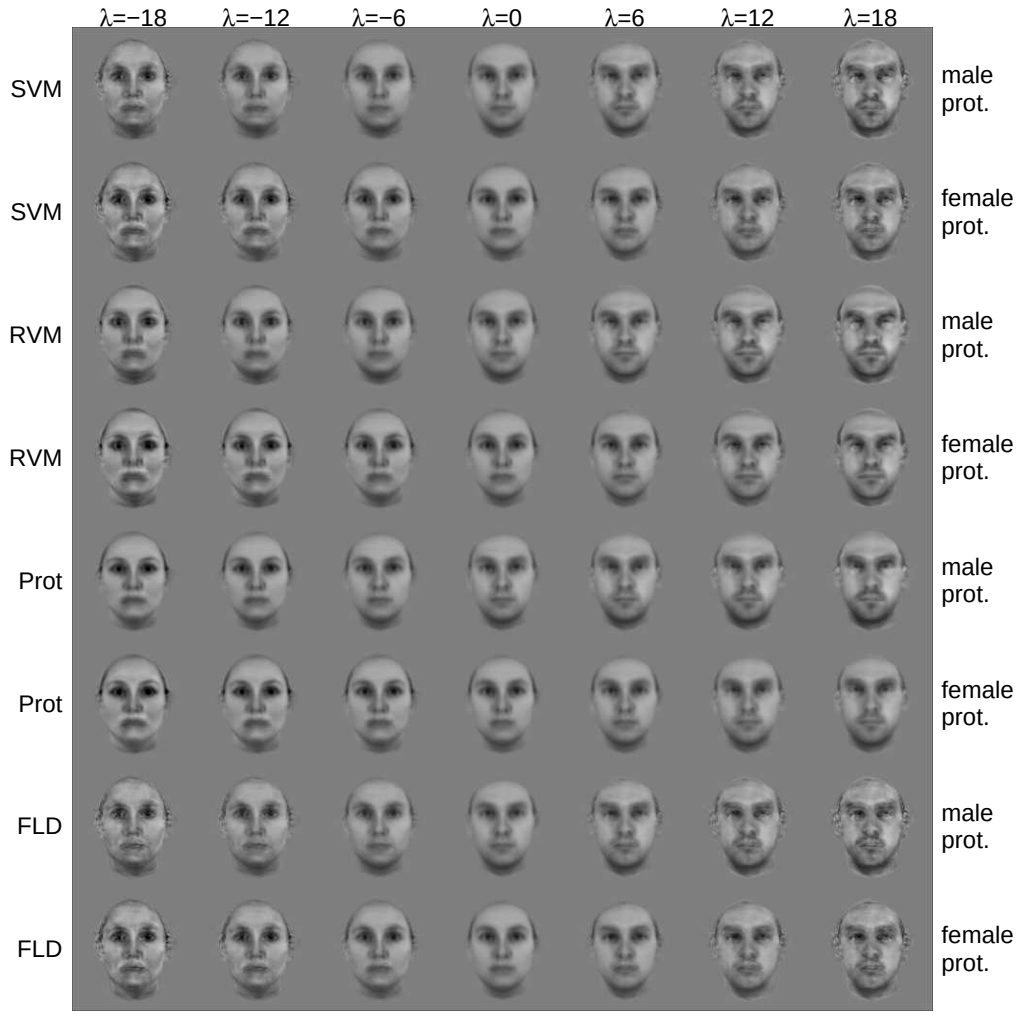


Figure 6.13: Novel face stimuli generated on a direction orthogonal to the SH staring from the prototypes of each class of the PCA space.

plateau time of 1000 ms). The probability of the female face being presented on the left was 0.5 on each trial and observers indicated whether they thought the left or right face was female by touching the corresponding location on a Elo TouchSystems touch-screen immediately in front of the display; no feedback was provided.

Trials were run in blocks of 256 in which eight repetitions of eight stimulus levels, $\pm\lambda_1 \ldots \pm \lambda_8$, for each of the four classifiers were randomly intermixed. The naïve subjects required approximately 2000 trials (roughly 8 blocks) before their performance stabilized; thereafter they did another five to six blocks of 256 trials. All results presented below are based on the trials

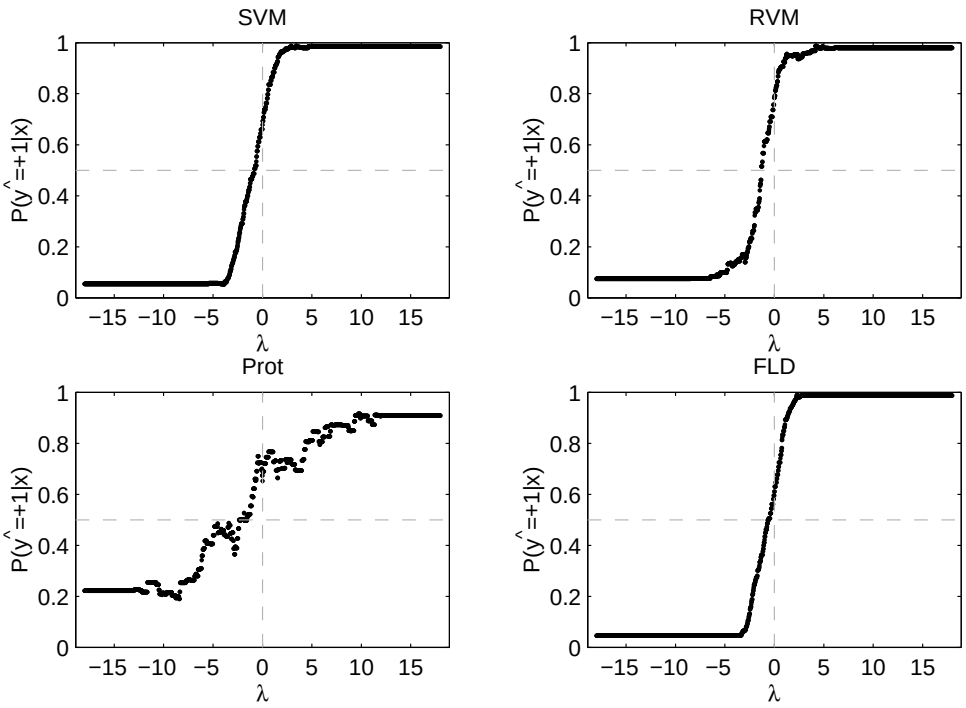


Figure 6.14: Estimate of the probability $P(\hat{y} = +1|x)$ that a face, lying on a direction orthogonal to the mean SH over all subjects, is classified as male by the subjects as function of λ.

after training; all training trials were discarded.

Fig.6.15a shows the raw data and fitted psychometric functions for one of the observers. Proportion correct gender identification on the y-axis is plotted against the stimulus level λ on the x-axis on semi-logarithmic coordinates. Psychometric functions were fitted using the `psignifit` toolbox for Matlab which implements the constrained maximum-likelihood method described in [Wichmann and Hill, 2001a]. 68%-confidence intervals (CIs), indicated by horizontal lines at 75 and 90% correct in Fig.6.15a, were estimated by a bootstrap method also implemented in `psignifit` [Wichmann and Hill, 2001b]. The raw data appear noisy because each data point is based on only eight trials. However, none of the fitted psychometric functions failed various Monte Carlo based goodness-of-fit tests [Wichmann and Hill, 2001a].

To summarize the data we extracted the λ required for two performance levels ("thresholds"), 75 and 90% correct, together with their corresponding 68%-CIs. Fig.6.15b–e shows the thresholds for all four observers normalized by λ_{SVM} (the "threshold elevation" with respect to the SVM). Thus

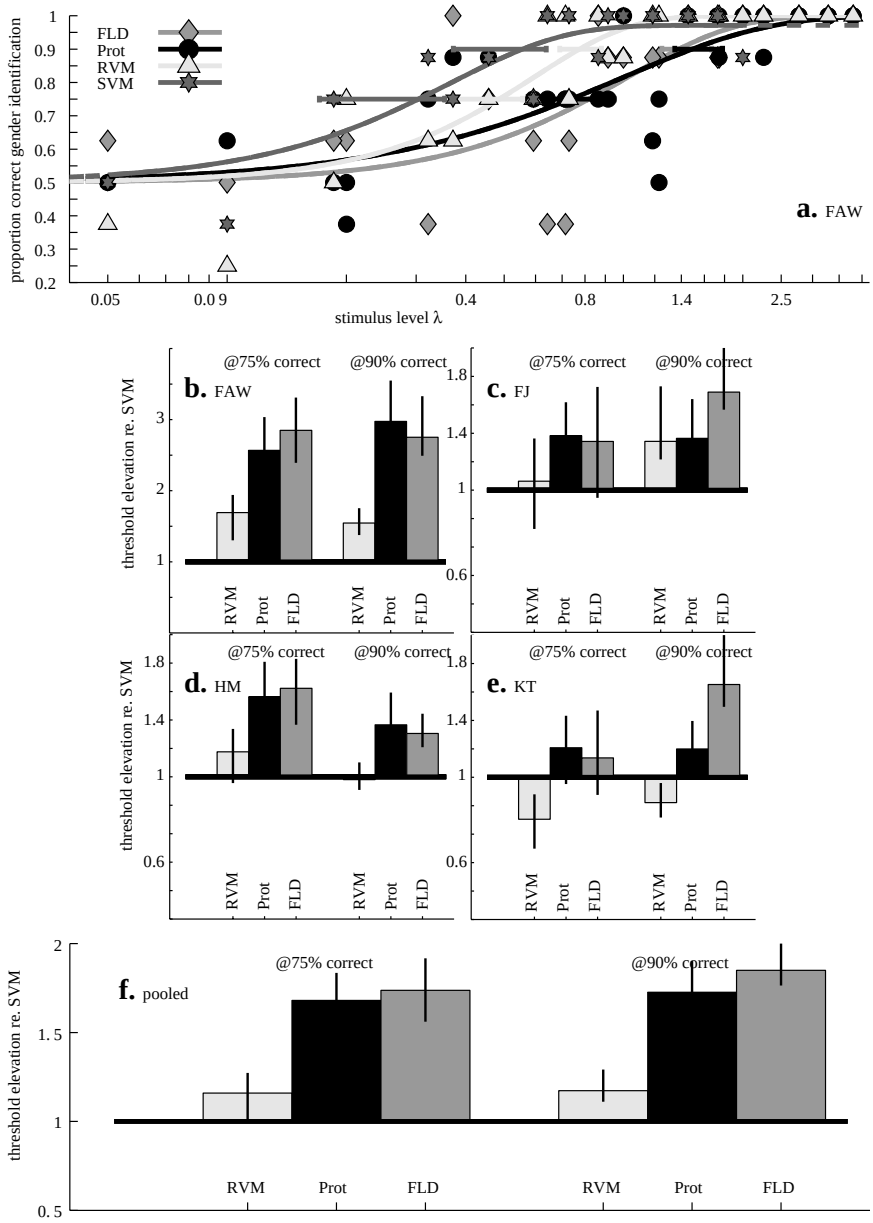


Figure 6.15: a. Shows raw data and fitted psychometric functions for one observer (FAW). b–e. For each of four observers the threshold elevation for the RVM, Prot and FLD decision image relative to that of the SVM; results are shown for both 75 and 90% correct together with 68%-CIs. f. Same as in b–e but pooled across observers.

values larger than 1.0 for RVM, Prot and FLD indicate that more of the

corresponding decision images had to be added for the human observers to be able to discriminate females from males. In Fig.6.15f we pool the data across all observers. As the main trend, poorer performance for Prot and FLD compared to SVM and RVM, is apparent for all four observers. The difference between SVM and RVM is small; going along the direction of both Prot and FLD, however, results in a much "slower" transition from female-to-maleness.

The psychophysical data are very clear: all observers require a larger λ for Prot and FLD; the length ratio ranges from 1.2 to nearly 3.0, and averages to around 1.7 across observers. In the pooled data all the differences are statistically significant but even at the individual subject level all differences are significant at the 90% performance level, and five of eight are significant at the 75% performance level. It thus appears that SVM and RVM capture more of the psychological face-space of our human observers than Prot and FLD. From our results we cannot exclude the possibility that some other direction might have yielded even steeper psychometric functions, i.e. faster female-to-maleness transitions, but we can conclude that the decision images of SVM and RVM are closer to the decision images used by human subjects than those of Prot and FLD. This is exactly as predicted by the correlations between proportion correct, RTs and confidence ratings versus distance to the hyperplane reported at the beginning of this chapter—high correlations for SVM and RVM, low correlations for Prot and FLD.

Comparing the numerical estimate of the probability $P(\hat{y} = +1|x)$ as function of λ (see Fig.6.14) to the one obtained above in the psychophysical experiments (see Fig.6.15a), we notice that human subjects seem to be slightly more sensitive to gender discrimination—a smaller range for λ is needed for discrimination—as would be suggested by the corresponding numerical estimate. Moreover, the predictions of both studies are identical for SVM, RVM and the prototype learner. The results for FLD are, however, different and this issue is discussed in the next section.

6.7 Summary & Discussion

We studied the classification of face stimuli using linear methods from machine learning for feature extraction (Principal Component Analysis, PCA) and for classification (the Support Vector Machine SVM, the Relevance Vector Machine RVM, the prototype classifier Prot and the Fisher linear discriminant classifier FLD) together with data from human psychophysical classification experiments.

The combination of a linear feature extractor and a linear classifier allowed us to visualize the *decision images* of a classifier corresponding to the vector normal to the SH of each classifier. Decision images can be used for feature ranking by determining the regions of the stimuli most useful for

classification simply by analyzing the distribution of light and dark regions in the decision image. In addition we defined the *generalized portraits* to be the prototypes of all faces used by the classifier to obtain its classification. For the SVM and the RVM, this is the weighted average of all the support vectors and relevance vectors respectively (i.e. the representations of these algorithms), and for the prototype classifier it is the prototype itself. The generalized portraits are, like the decision images, another useful visualization of the combination feature extractor-classifier. Feature ranking is also performed using a benchmark method from machine learning: Recursive Feature Elimination (RFE). The decision image on the true and subject datasets and obtained using RFE look quite similar, although the decision space they define is very different as shown by the studies on the sets Δ.

If trained on the true labels, some machines perform the classification task quite similarly to humans in terms of classification performance. However, they classify faces very differently from human subjects as was shown by the poor logistic regression fits for the distance set Δ_{true} with the mean human probability to classify a stimulus as male.

SVM and FLD can, however, re-create the decision boundary and the internal representation of faces for human subjects very well indeed if trained on the subjects' labels as shown by the excellent logistic regression for Δ_{subj}. RVM is already quite a bit worse, and the prototype learner is as bad as for the true labels.

Finally, equating the classification performance of man and machine through RFE makes machines even less human-like than if trained on the true labels. This shows that even if man and machine perform a task equally well—i.e. same classification error—this does not imply anything about their internal workings. This was shown by the very poor logistic regression fits to Δ_{RFE} in the RFE-reduced-dimensionality face space. In addition this may indicate that PCA may be less biologically meaningful than it is sometimes assumed [Turk and Pentland, 1991, O'Toole, Abdi, Deffenbacher, and Valentin, 1993].

Clearly our results need to be interpreted with caution: the learning regimes for man and machine were very different. Human subjects were trained during their lifetime and tested in this study using a smallish sample of a specific set of face stimuli. On the other hand, the machines were only trained on these stimuli. Finally the above results rely on the linearity of the classification system formed by the feature extractor and the classifier; incorporating non-linearities is clearly a future direction of research.

Using the decision images, a serie of psychophysical investigation yielded one of the central result of this chapter: the corroboration of the machine-learning-psychophysics research methodology. In the machine-learning-psychophysics research we substitute a very hard to analyze complex system (the human brain) by a reasonably complex system (learning machine) that is complex enough to capture essentials of the human subjects' behavior but

is nonetheless amenable to close analysis. Machines were first used to model and explain the classification behavior of man in Chapter 5 and at the beginning of this chapter. From the analysis of the machines we then derive predictions for human subjects which we subsequently test psychophysically. By doing so, we close the man-machine loop, and demonstrate that machine learning is a suitable method to model the classification of visual stimuli, at least on the considered face database.

Given the success in predicting the female-to-maleness steepness of the $\vec{w}_{SVM}$ axis we believe that the decision image $\vec{W}_{SVM}$ captures some of the essential characteristics of the human decision algorithm. This result is perhaps counter-intuitive given the SVM's androgynous generalized portraits; still, this "classify-close-to-the-margin" algorithm appears closest to the algorithm used by the human subjects in classifying faces. Because we can show the relevance of the decision image of SVMs for human observers psychophysically, this provides us with a much quicker alternative to psychophysical feature extraction techniques such as the "bubbles" [Gosselin and Schyns, 2001] or the noise classification image [Ahumada, 2002] techniques.

While the results from SVM, RVM and the prototype classifier are coherent in all the studies of this chapter, the results for FLD are quite different. First, FLD shows low man-machine correlations when δ is computed on a subject basis. However, the logistic regressions between man and machine are good, as much as the interpolated ones when estimated the subjects' responses on stimuli orthogonal to the SH of each classifier. In the actual psychophysical experiment on such stimuli, FLD does however not exhibit a human-like behavior. This quite inhomogeneous behavior of FLD is due to its overfitting of the PCA data. It may be interpreted that on the basis of the last psychophysical experiment, the man-machine correlation on a subject basis is valid and that FLD is simply not suited to describe human classification of visual stimuli. Furthermore, the inconsistency of the results provided by FLD is another hint that this classification algorithm may not be adapted in the context of the present studies.

Chapter 7

Applying Machine Learning to Model Human Memory

Following the studies of the classification behavior of humans, we may ask the following question: how are visual stimuli memorized by the human subjects? Can machine learning also account for the memory behavior of humans using the MPI face database? To answer these questions, we introduce here a novel methodology allowing to embed machine learning in a human psychophysical set of memory experiments using a feedback loop architecture and an online generation of novel stimuli.

7.1 Overview & Methodology

The memory experiments are structured as follows. After a first gender classification experiment identical to the one presented in Chapter 3, the representations—the Support Vectors (SVs), the relevance vectors (RV), the prototypes (Prots) and the means of Kmeans (Means)—along with some novel stimuli computed using the SVs are computed for each subject using the classifiers introduced in Chapter 4. These patterns are randomly intermingled with others drawn directly from the MPI face database. The memory behavior of the subjects—that is the ability of subjects to retrieve these representations again from memory— is then studied using a seen/unseen (or old/new) experiment on this set of patterns. Finally a second gender classification experiment investigates the classification of the subject on the whole MPI face database and on the representations. Presenting the subject with its personal dataset of representations, computed using the first classification experiment, constitutes the actual feedback loop architecture. The novelty of this approach is the introduction of this feed-back loop embedding an artificial classifier from machine learning in a psychophysical framework in order to obtain insight into the classification and memory process. We may also note that the concept of memory is tightly related to the concept

of generalization as already pointed out in [Posner and Keele, 1968]. A first account on this methodology was communicated in [Graf, Wichmann, Bülthoff, and Schölkopf, 2004b].

The memory-after-classification experiment aims to assess the various learning algorithms on the basis of the elements allowing to compute their separating hyperplane (SH): the representations. That is, the memory experiment is performed in order to assess whether for instance the SVs or the prototypes are more easily classified as "seen before". This would give evidence of an "internal image" as a decision variable during classification and thus provide insights into the structure of internal memory. Presenting subjects with unseen stimuli such as the prototypes is an alternative, albeit more direct, manner to study the aftereffects associated with face images as done by [Leopold, O'Toole, Vetter, and Blanz, 2001]. In our approach the aftereffects are directly induced and not merely extrapolated.

7.1.1 Database and Feature Extraction

The emphasis is here put on the classification algorithm and less on the feature extractors defined in Chapter 2. The MPI face database of 200 stimuli is split into a dataset Σ of 160 patterns and into a dataset Υ of 40 patterns. Since 8 stimuli from Σ are shown in the first presentation aimed at acquainting the subjects with the setup, Σ is composed of $160 - 8 = 152$ stimuli. The PCA preprocessor on the texture and shape data type is chosen as a feature extractor. Indeed, when designing the experiment, there was no *a priori* knowledge about the preprocessor best suited to describe human classification behavior (see Chapter 5). PCA was then chosen since it is a benchmark preprocessor in unsupervised machine learning and accounts in literature such as [Turk and Pentland, 1991] hint at its biological relevance. The texture and shape data type is here the most suited choice since novel stimuli will be generated online—this data type was specifically intended for these type of studies since it allows to take advantage of the morphing capacities of the MPI face database through the face modeler by [Blanz and Vetter, 1999].

7.1.2 Classification Experiment I

This first classification experiment is the experiment presented in Chapter 3 and analyzed in Chapter 5. We recall that a gender-balanced subset of 152 from the possible 200 faces of the MPI face database are presented to the subjects using the timing parameters of Table 7.1. The subjects then perform a gender classification experiment on sequentially-presented faces of males and females. Their gender estimate along with its reaction time (RT) and confidence rating (CR) are recorded. No feedback is provided.

	$t_{transient}$ [s]	t_{steady} [s]	t_{is} [s]
classification experiments I & II	0.5	1	1

Table 7.1: Parameters of the Hanning window for the presentation of the stimuli to the subjects for the first and second classification experiments. The rising and falling time are represented by $t_{transient}$ whereas t_{steady} is the plateau time and t_{is} is the time between stimuli.

7.1.3 Online Computation of Representations

For the classifiers we consider, as before, the SVM, the RVM, the prototype learner Prot and the Kmeans classifier Kmean (see Chapter 4). All the classifiers are trained on the (non-normalized) input PCA data, which is by definition centered. In order to limit the complexity of the Kmean classifier, the number of means for each class is set to $K = 3$, which was to shown *a posteriori* to be a reasonable value by the studies of Chapter 5, the trade-off parameter C of the SVM being set by cross-validation on the training set. In order to allow (fast) convergence of the SVM and RVM algorithms, the data is mapped into a normalized feature space as proposed by [Graf, Smola, and Borer, 2003] by using a normalized linear kernel function defined as: $K(\vec{x}, \vec{y}) = \frac{\langle \vec{x} | \vec{y} \rangle}{\|\vec{x}\| \|\vec{y}\|}$. The representations are then the elements of the input space corresponding to the SVs and RVs computed in this normalized feature space. Thus all representations are in the (non-normalized) input space and can consequently be viewed by inverting the PCA transformation and applying the face modeler (morphing process). This would not have been the case when working in a normalized input space since every vector would then be scaled, which results in the loss of one degree of freedom of the data.

For each subject, the classifiers are trained on the PCA of the texture and shape data type of the face stimuli presented to them. These vectors are combined with the subject's estimate of the gender of this stimulus, forming the subject dataset as defined in Chapter 5. The corresponding representations (SVs, RVs, Prots and Means) are then computed and some novel stimuli (nmSVs and pSVs, see below) are generated as listed in 7.2. Seen and unseen stimuli as well as male and female stimuli are balanced and randomly mixed.

All the *seen* stimuli are patterns from the original MPI face database. The SVs are separated into two classes: the margin SVs for which $\alpha_i < C$, mSVs, and the remaining ones for which $\alpha_i = C$, rSVs, yielding SV = mSV $\cup$ rSV. If not enough SVs or RVs are present[1], their class is filled with elements from mSVs or Σ.

For the *unseen* stimuli, the elements from Υ are faces from the original

[1]This may be true for some subjects, especially for the very sparse RVM algorithm.

29 unseen stimuli	29 seen stimuli
2 nmSV	2 rSV
2 pSV	5 mSV
2 Prot	5 RV
6 Mean	17 $[\Sigma - \{SV, RV\}_{\text{shown before}}]$
17 Υ	

Table 7.2: Seen and unseen stimuli computed for each subject using the responses of the first classification experiment.

MPI face database. All the other unseen stimuli are novel types of stimuli which are computer-generated online during the experiment such as the representations for the prototype and for the Kmeans classifiers. Furthermore, two novel types of SVs are generated. First, the new margin SVs, nmSV, are the mean of the margin SVs (mSV) for each class. Since the mSVs lie on two hyperplanes—the margins of the SVM algorithm—these averages also lie on these margins. They are another instantiation of the prototype algorithm, this time applied only to a small subset of geometrically well-defined patterns. Second, the actual prototypes, pSVs, of the SVM algorithm are computed as:

$$\vec{p}_{\pm}^{SV} = \frac{\sum_{i|sign(\alpha_i)=\pm 1} \alpha_i \vec{x}_i}{\sum_{i|sign(\alpha_i)=\pm 1} \alpha_i}$$

assuring that the prototypes lie in the convex hull of the patterns of each class individually as suggested by [Graf, Bousquet, and Rätsch, 2004a].

Some of the above representations are displayed in fig.7.1 for a two-dimensional toy example. By construction the SVs are spread on or around the margins. On this toy dataset it is easy to guess the location of the nmSVs and the pSVs: they are thus not drawn. The Prots are in the middle of the classes, whereas the Means are distributed in each class. This toy example illustrates an important fact: the prototype of each class is not confounded with one of the Means i.e. $Prots \notin Means$. The RVs are spread throughout the dataset. The above descriptions were already done on the MPI face database in Chapter 5 where the histograms of δ gave similar descriptions.

The above representations are determined online while the subject takes a rest of approximately 2 minutes after the first classification experiment and is given instructions about the memory experiment. During this time the novel stimuli are generated on the subject dataset using the morphing process and the representations from Σ are defined. Fig.7.2 shows a selection of these representations computed used the true dataset. At this point it is important to stress that classification is done in the PCA space of the texture and shape data type. In this space, and using the true dataset, there is just one rSVs and we only plot 6 from the 75 SVs. On the subject dataset, as

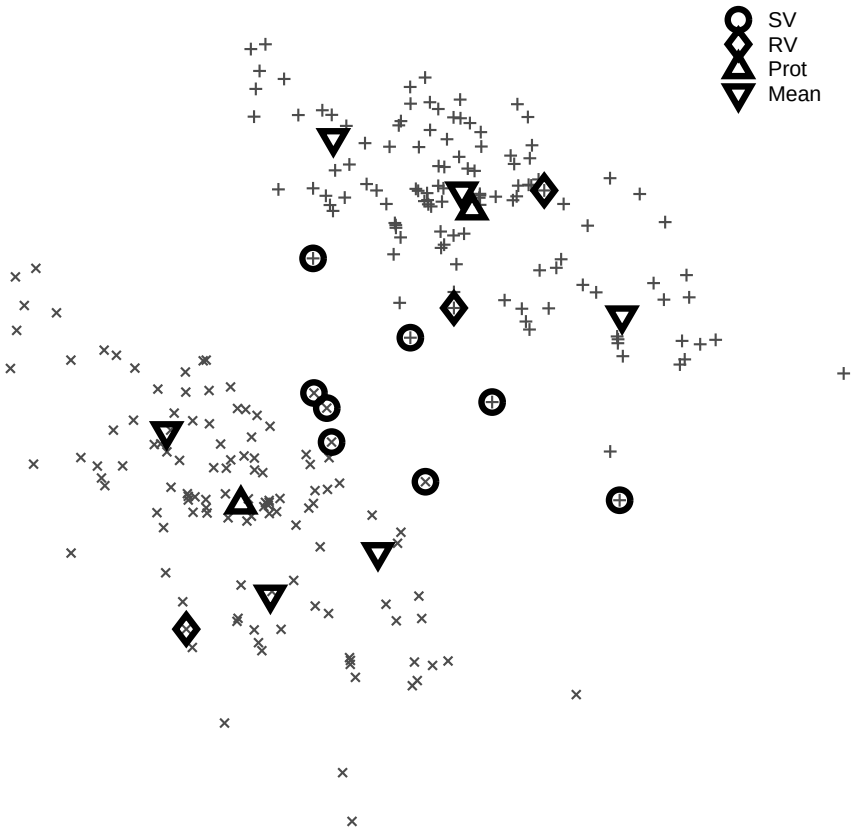


Figure 7.1: Two-dimensional toy dataset with following representations: SV, RV, Prot and Mean.

used in the experiments, the situation is different since the subjects' labels makes classification a harder task (see Chapter 5) and one expects more elements to lie in the margin stripe (the rSVs) and not only on the margin (the mSVs). The pSVs are different from the nmSVs because of the convex hull condition imposed in their computation. The nmSVs are quite similar to the Prots since the nmSVs are an average done over 75 from the 200 available patterns. We also represent 6 of the 10 RVs. Finally the means are different from the Prots as already mentioned above, and seem to sample better the face space in each class.

7.1.4 Memory Experiment

In the actual memory experiment, human faces are presented sequentially. The subjects are asked to decide whether the faces are seen or unseen, i.e. whether they are old or new respectively. The subjects are instructed

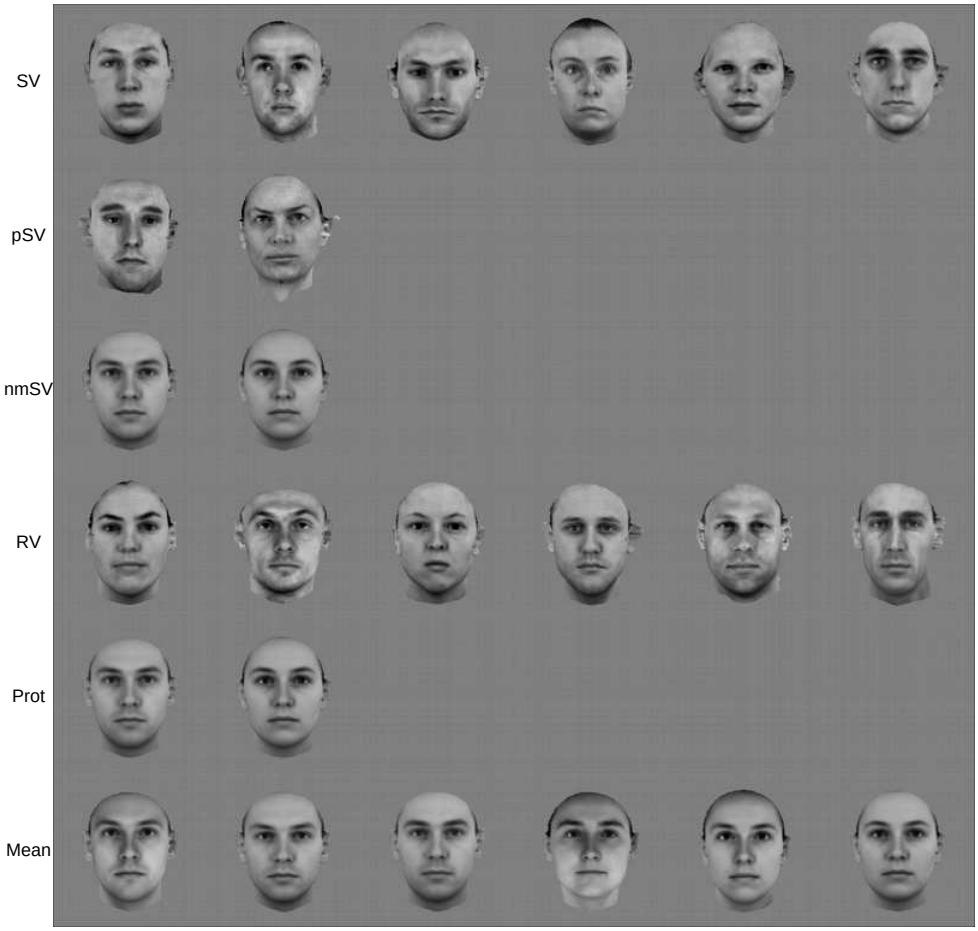


Figure 7.2: Plotting of representations computed using the true dataset.

that there are 29 seen and unseen faces and that they can take their time to make their decision. After the seen/unseen response, subjects are asked to rate the distinctiveness of the presented faces as $1, 2, 3$ from low to high. The seen-unseen responses and the distinctiveness given by the subjects are the only parameters, and we do not deal with reaction times. The timing issues of the presentation of the stimuli are reported in Table 7.3. Pilot studies revealed that the above task was close to impossible for the average subject. Thus, in order to make the above task easier, the 29 seen stimuli are first presented using the timing parameters of Table 7.3 in order to "refresh" the subjects' memory on the stimuli they have already been seen during the first classification experiment. Furthermore, the subjects have been instructed before the experiment that if they perform well in the memory experiment, they would get a money reward which was granted to subjects obtaining a discriminability $d' \geq 1.5$ for the seen/unseen task. However, no feedback

	$t_{transient}$ [s]	t_{steady} [s]	t_{is} [s]
presentation	0.5	3	1
memory experiment	0.5	2	1

Table 7.3: Parameters of the Hanning window for the presentation of the stimuli to the subjects for the presentation of the seen stimuli and for the actual memory experiment. The rising and falling time are represented by $t_{transient}$ whereas t_{steady} is the plateau time and t_{is} is the time between stimuli.

was given during the memory experiment.

7.1.5 Classification Experiment II

After the memory experiment, a second classification experiment is considered. This experiment is similar to the first one (see Table 7.1 for timing issues related to the presentation of the stimuli) except that the stimuli presented here are the complete set of 200 faces from the MPI face database combined with the representations computed using the subject's responses in the first classification experiment. In general there are more than $200 + 6$ Means $+ 2$ Prots $+ 2$ nmSVs $+ 2$ pSVs $= 212$ stimuli presented since some stimuli are SVs and RVs at the same time. The experiment keeps track of the type of stimulus (Σ, SVs, ...) classified by the subject.

7.2 Results

The first classification experiment has been extensively studied in Chapter 5. We analyze below the memory and the second classification experiments.

7.2.1 Memory experiment

The human responses—the seen/unseen estimate and the distinctiveness—are shown in Fig.7.3 for all subjects. The first plot in the *first* row of Fig.7.3 of the seen/unseen discriminability d' indicates that subjects can memorize the stimuli although, as indicated above, this task is not easy. However the value of d' is quite disappointing considering the fact that the subjects were exposed to exactly these 29 seen faces in the presentation just before the memory experiment. Furthermore there seems to be no seen/unseen bias. In the second plot, we notice that stimuli having a high distinctiveness are accompanied by a low seen/unseen error. There is no significant difference in this effect if the male, female or all the data is considered. The third plot shows that a seen stimulus is usually attributed a high distinctiveness whereas an unseen one is mostly accompanied by a low distinctiveness. The

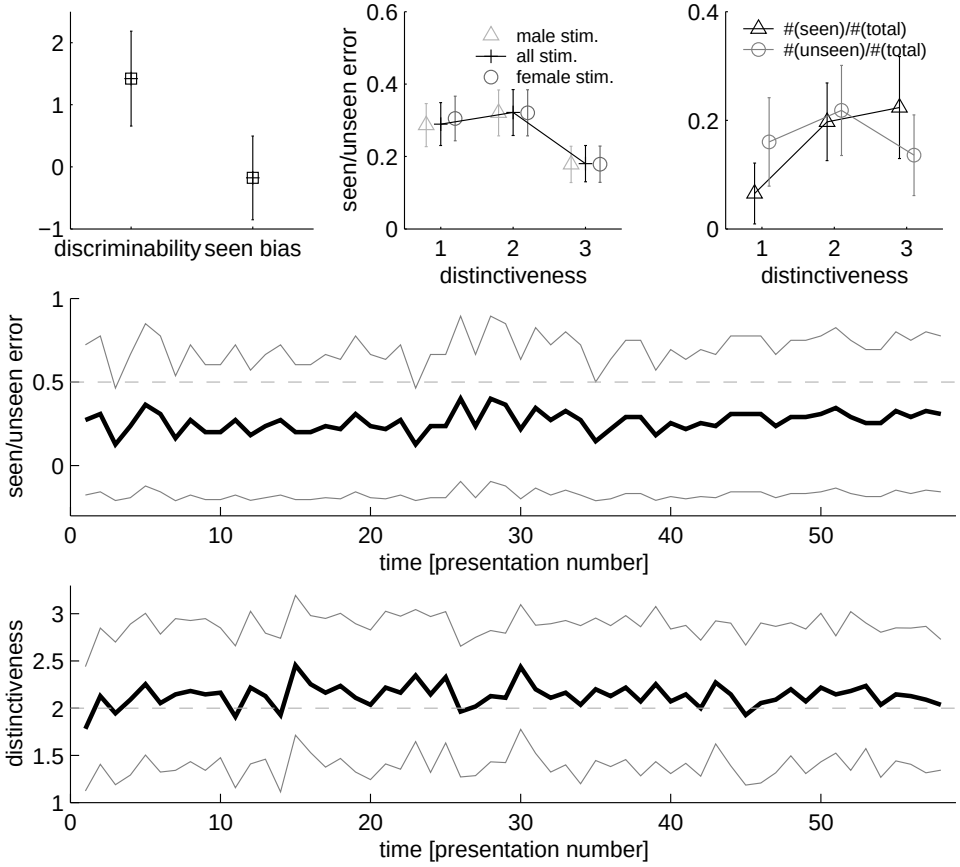


Figure 7.3: Analysis of the all the subjects' responses for the memory experiment.

second and *third* rows show the mean temporal evolution of the subjects' responses. From these curves it can be concluded that there are no learning and fatigue effects in the subjects: their answers stay stable over time. The seen/unseen error, although high as already noticed above, is below chance and there is a slight bias to give a high distinctiveness.

The above results allow us to proceed to the actual analysis of the memory experiment. In particular, the subjects' responses (seen/unseen and distinctiveness) are studied on various sets of stimuli:

- the representations: SV (mSV, rSV), RV, Prot, Mean and the novel SV (nmSV and pSV)
- the non-representations: Σ (without the SVs and the RVs) and Υ

and the results are represented in Fig.7.4, where a good subject is defined as a subject who satisfies $d' \geq 1.5$ for the seen/unseen task. The *first* row

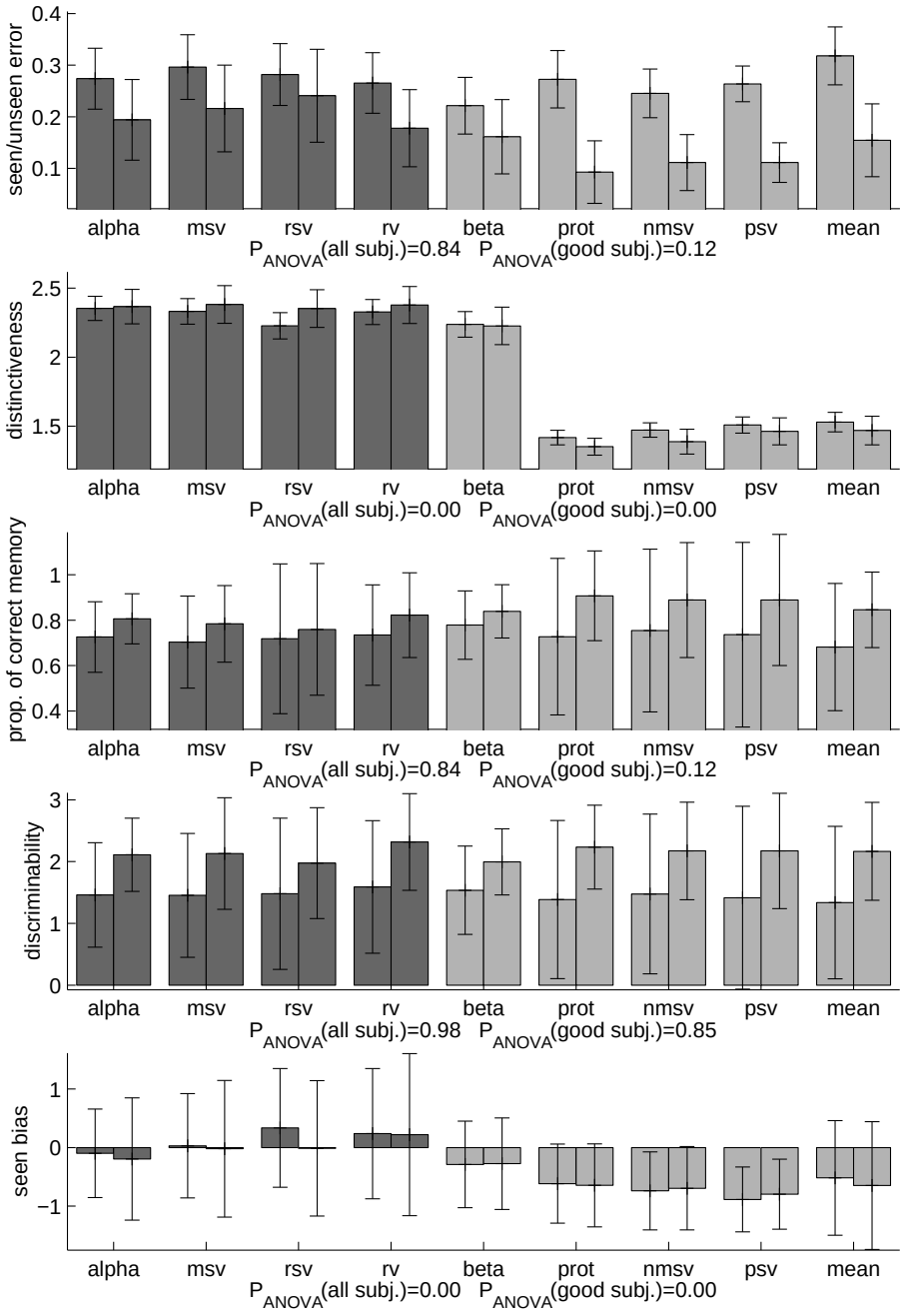


Figure 7.4: The subjects' responses for each set of stimuli in the memory experiment. The dark bars indicate seen stimuli whereas the light ones stand for unseen stimuli. For each set of stimuli, the left bar shows the response for all subjects and the right one only for the good ones.

exhibits the seen/unseen error for each set of stimuli. When considering all the subjects, there is no significant trend as confirmed by an ANOVA analysis for each response across subjects. Machine learning may help to explain classification, however it fails to account for memory effects given this database and experimental setup. For the good subjects, a slight effect may however be observed. The novel computer-generated patterns—Prot, Mean, nmSV and pSV—are usually more easily recognized as unseen than the stimuli drawn from the original MPI face database. This effect is also observed for the distinctiveness plotted in the *second* row where for both types of subjects the stimuli drawn for the original dataset, namely Σ, Υ, the SV and the RV, are given a high distinctiveness by the subjects. The novel computer-generated stimuli—Prot, Mean, nmSV and pSV— on the other hand are given a low distinctiveness by the subjects. The subjects thus have a veridical knowledge about the original of the presented stimuli, and this is reflected in the distinctiveness response and the seen/unseen error to a certain extent. One reason for this might be that the morphing algorithm removes cues within the faces—such that the face blemish for example—and that these cues are used and recognized by the subjects. The quality of this algorithm thus still seems insufficient for controlled scientific studies on computer-generated face stimuli.

As an alternative quantity to the seen/unseen error, the *third* and *fourth* rows show the proportion of correct memory answers and the discriminability d' corresponding to the seen/unseen estimates. Although as could be expected, the good subjects are better for both quantities than all of them (the combination of good and bad subjects), there is again no significant trend as shown by the ANOVA analysis—it is almost certain that the responses for all stimuli sets are equal. The *fifth* row shows that seen stimuli have a seen bias whereas unseen ones have an unseen bias, as could be expected. This trend is less strong for the good subjects, which makes them less biased observers.

The above analyses seem to indicate that the subjects' responses are constant across all types of stimuli (representations or non-representations). We thus not proceed to any further machine learning analysis on this memory data as we may not expect to gain any meaningful insights from the data at hand.

7.2.2 Classification Experiment II

The second classification experiment is similar to the first one, except that here subjects also classify the representations computed using their responses of the first experiment. This is the actual man-machine feedback loop: the responses of the first gender classification experiment are used to define stimuli (SV and RV) and generate novel ones (Prots, Mean, nmSV and pSV) which are then classified in a second experiment, again using a gender clas-

sification task. The subjects' responses, here the classification error, the RT and the CR, are computed for each set of stimuli and are presented in Fig.7.5. The results of the first row—the classification error—are reflected by those of the fourth representing the corresponding discriminability d'. The first three rows (gender estimate, RT and CR) show that the subjects have the worst behavior (high error, high RT and low CR) on the pSV and the best one on the Σ stimuli, which is corroborated by Fig.7.6 (see below). We may expect to get roughly similar responses for the first classification experiment (exp. I) and on the Σ stimuli of the second one. However, we see here that they are different, although not significantly: subjects seem to behave better (lower error, lower RT and higher CR) in the second than in the first experiment on the same stimuli. This may have two explanations. First, the set Σ is the set of stimuli shown in the first experiment with the SVs and the RVs removed. The subjects' classification on the SVs and the RVs seems thus worse than on the other elements. These representations are thus difficult to classify, what seems intuitive for the SVs since they lie near the margin. Second, subjects may exhibit a slight accustomization to the stimuli. This also corroborates the assumption of Chapter 5 about the existence of an internal jitter in the subject's representation of the stimuli. This jitter is here shown to increase the classification performance. Generally subjects behave better on the seen stimuli than on the unseen ones. In the fifth row, we see that the computer-generated stimuli (Prot, Mean, nmSV and pSV) have no significant male bias, as opposed to the stimuli drawn from the original MPI face database. The face modeling algorithm seems thus to remove gender-specific cues, as corroborated by the above finding that these faces also have low distinctiveness.

We now analyze the results of this second classification experiment using machine learning. As explained in Chapter 4, the data is normalized in order to yield comparable ranges for distance $|\delta|$ of the stimuli to the SH when comparing classifiers, the representations being also computed in this normalized input space. The classifiers are trained on the data from the first classification experiment and $|\delta|$ is computed for the (non-)representations. Fig.7.6 is a summary plot showing the subjects' responses on the (non-)representations as function of $|\delta|$ for each classifier, the data being averaged over sets of stimuli defined by the (non-)representations. The same trends as for the studies done for the first class experiments can be observed: stimuli far from the SH (high $|\delta|$) are in general classified with a small error, a low RT and a high CR. The subjects behave best (low error, low RT and high CR) on the seen stimuli from Σ for all classifiers. This may be explained by the fact that these stimuli are far from the SH in all cases. The subjects behave worst on pSVs, although they are not the closest stimuli to the SH. This suggests that yet another instantiation of the concept of prototype proves to be not biologically meaningful, thus corroborating the findings of Chapter 5. Finally, the positions of the (non-)representations are

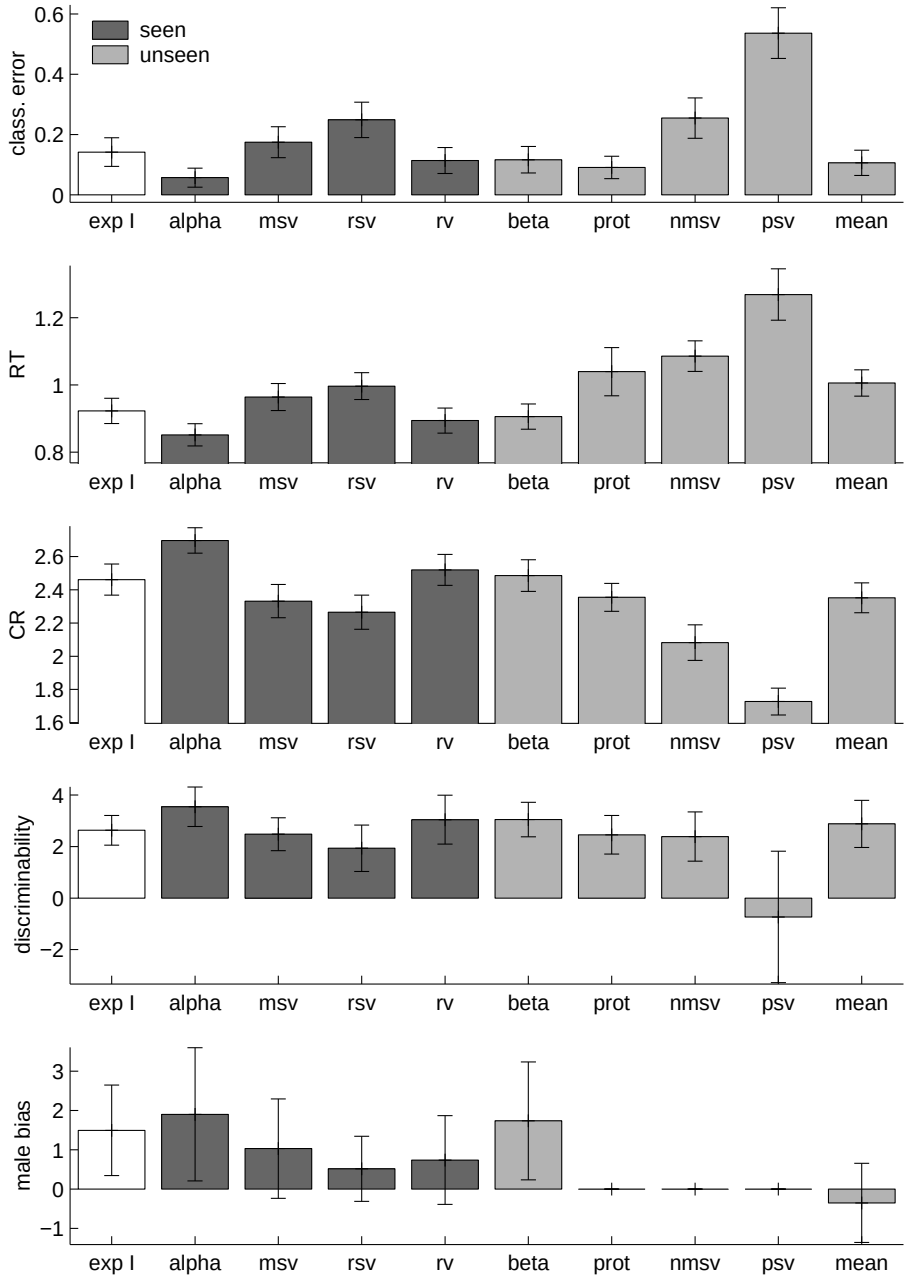


Figure 7.5: The subjects' responses for each set of stimuli in the second classification experiment. The dark bars indicate seen stimuli whereas the light ones stand for unseen stimuli, the white ones standing for the first classification experiment.

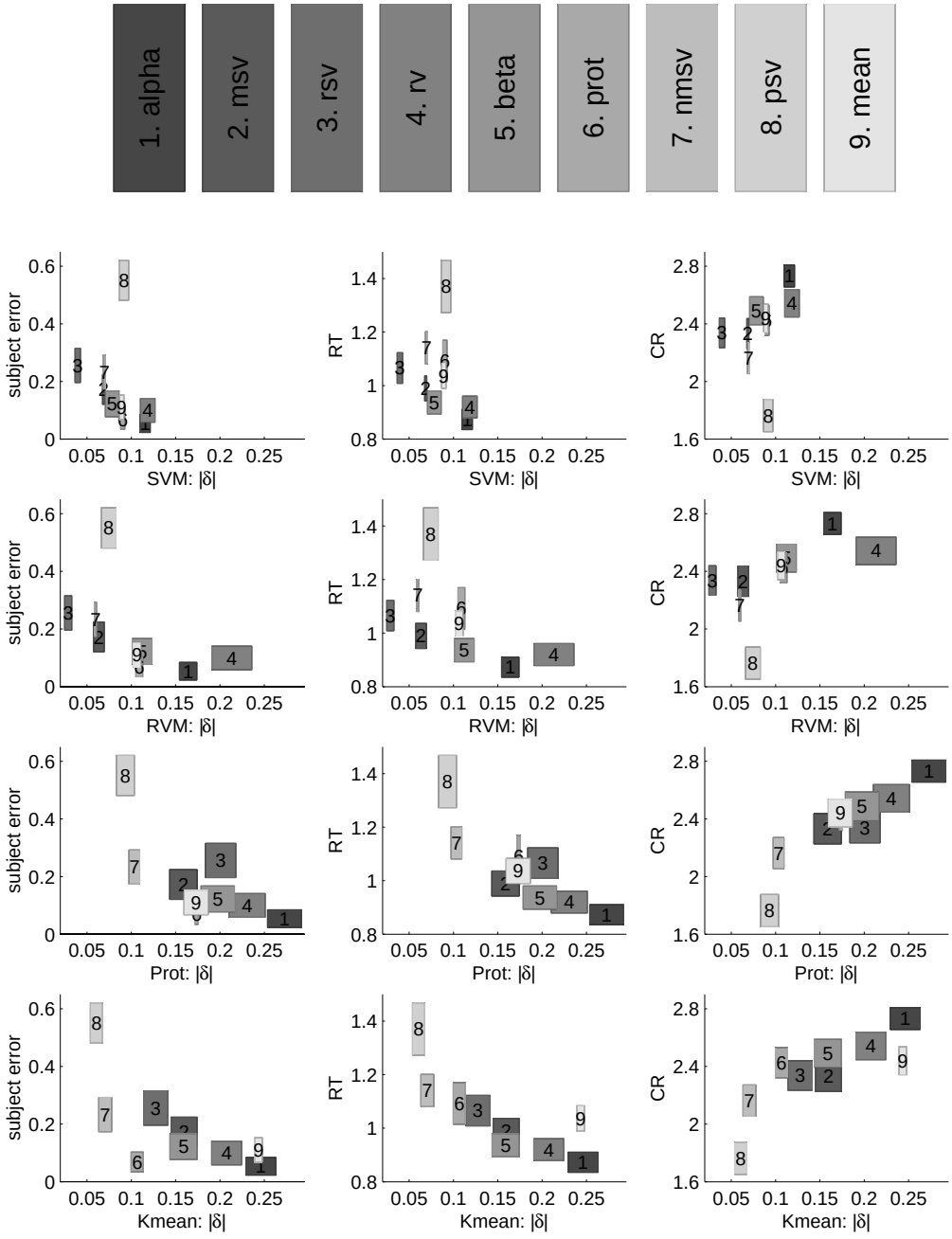


Figure 7.6: The subjects' responses (gender estimate, RT and CR) for each set of stimuli as function of their distance $|\delta|$ to the SH in the second classification experiment.

strongly dependent upon the type of classifier, and they reflect the different classification mechanisms.

7.3 Summary & Discussion

The classification experiment of Chapter 5 is a behavioral study where machine learning is used to help understand human classification. The studies done in this chapter aim at assessing the strategies used by humans in memorization tasks. As an extension of the methods used to model classification, we show that machine learning is not suited to account for the memory behavior of humans, at least within the PCA feature space spanned by the human face database. In particular, the representations from machine learning computed using a classification experiment could not account for the subject's seen/unseen memory behavior. However, this results may also be due to the fact that the number of stimuli memorized by the subjects (i.e., their short-/mid-term memory capacity) is quite low, making thus these kind of studies difficult[2]. These poor results may be due to the fact that the subjects' have quite a poor memorization capacity. Furthermore, these results may also hint at the fact that humans may use something akin to a "mixed" approach where for instance SVs could be used for training whereas Prots are used for classification.

Using methods similar to [Sigala and Logothetis, 2002], [Peters, Gabbiani, and Koch, 2003] investigate memory capacity in classification tasks. In psychophysical investigations, it is shown that models do not seem to need a high memory capacity, and that a small set of representations may be enough. However, the results of [Peters, Gabbiani, and Koch, 2003] are based upon the matching of the classification performance of man and machine, which was shown in Chapter 5 and will be shown in Chapter 6 to be not adapted. In the studies of this chapter, we note that the issue of memory capacity is a problem both in man and machine.

As far as the MPI face database is concerned, morphing of stimuli using the face modeler is shown to be cognitively perceivable by the subjects and to interfere with their classification behavior by producing a noticeable effect on recognition performance. This problem could be avoided by considering only novel face stimuli computed as linear combinations of the existing ones. The subject would then only be confronted with computer-generated stimuli.

Our results rely on one crucial assumption: there is no additional learning in both the classification and in the memory experiments. The subjects have already acquired their internal face representation and are just tested when asked to classify a stimuli as male/female or as seen/unseen. Fur-

[2]Note that the reason for the large number of stimuli is due to the constraints of sufficiently sampling stimulus space in order to be able to model it using machine learning methods.

thermore, in the memory experiment, we investigate simultaneously two hypothesis that do not have to be necessarily correlated:

- humans have representations—how do humans learn?
- humans memorize representations—what do humans learn?

The results of the studies of this chapter then imply that either humans do not classify using representations, or they do not store them in their memory or, of course, they do neither. From the low seen/unseen discriminability despite the presentation prior to the memory experiment, we may conclude that humans are certainly bad at memorizing a relatively large number of human faces, at least in their short-term memory. This may be due to the high homogeneity of the considered face stimuli in terms of visual similarity, which was even increased by the cleaning of the MPI face database as presented in Chapter 2. Low-level stimuli may then be used in this type of experimental setup to overcome this problem. Low-level random dot stimuli have been shown to be useful to investigate the combination of classification and memory mechanisms [Knapp and Anderson, 1984] using models based upon the storage and addition of "memory traces". Distributed memory storage was introduced as an alternative to probabilistic and exemplar-based classifiers such as General Context Models and prototype classifiers which store exemplars. It was shown that for a a low number of patterns, novel stimuli are classified by the subjects according to their similarity to the learnt patterns, similarly to what is done in the General Context Model. However, for a high number of stimuli, prototype classification was demonstrated to be most relevant. This yields a a "two-regime" classification scheme which is certainly meaningful for low-level stimuli, its application to high-level stimuli revealing to be more difficult.

One of the first attempts to study the concept of prototype in recognition (memory) tasks is due to [Posner and Keele, 1968]. The prototype and its "distortions" are shown to be more easily classified as other stimuli from the corresponding class. In other words, after memorizing a set of distortions, the subjects recognize the prototype better than other patterns of the corresponding class, despite the fact that the subjects have never seen the prototypes before. Furthermore, the authors also argue that patterns close to the prototype allow best generalization. In the studies of this chapter we do not get this results: the prototypes are as (badly) classified as the other (non-)representations. The authors of [Posner and Keele, 1968] also rise another fundamental issue: is the prototype computed on;line during learning or is it only retrieved when performing recognition? This question is answered in [Posner and Keele, 1970]: the prototypes are retrieved during learning and are most stable over time. Furthermore [Posner and Keele, 1970] investigate the difference in forgetting between the prototypes and its distortion and the patterns of the dataset. The prototypes are shown to be

less prone to forgetting over time than the other stimuli. This corroborates for [Posner and Keele, 1970] the crucial role the prototype classifier plays for learning.

Exemplar-based model have been studied in [Dailey, Cottrell, and Busey, 1999] to model human memory using an old/new and a distinctiveness paradigm. In such models, it is assumed that subjects store explicitly some representations, which is similar to the context of the models of this chapter. The subjects' recognition errors have been investigated on face stimuli and morphs inbetween them. However, the subjects were trained both on the morphed stimuli and the original one, making a comparison to the studies of this chapter not meaningful. The main results of [Dailey, Cottrell, and Busey, 1999] is that outliers in the dataset should be emphasized by increasing their prior and the width of their kernel function.

Perceptual classification and recognition memory have been investigated by [Nosofsky, 1991] using schematic human faces in another context as the one proposed here: the role of features such as nose, mouth or eyes, is investigated for classification and recognition tasks. The modeling part is less elaborate than the one considered here: Multi-Dimensional Scaling—a method belonging to the same family as LLE—is used as a feature extractor to span a space where General Context Models and prototype classifiers combined with a weighted Euclidean norm are used for classification. However the theoretical Ansatz is similar to the one proposed here: exemplars are stored in memory and are retrieved from it for classification or recognition. The main conclusion of this study is that classification and memory mechanisms rely on different set of mechanisms. We get a similar finding: while machine learning is well adapted to model classification, it seems less suited to explain the mechanisms of memory in our experimental setting. This finding is also corroborated by [O'Toole, Abdi, Deffenbacher, and Valentin, 1993] where the PCA representation is shown to be appropriate for the classification of human faces, but not in recognition tasks such as new versus old face. Moreover the exemplar-based approach was shown by [Nosofsky, 1991] to predict the subject's classification, a result which is corroborated by the results of this dissertation stating that SVMs or RVMs may be used to model the classification behavior of human subjects (see Chapter 5).

Most of the above previous attempts to model perceptual classification and recognition memory in man are rather different to the one considered in this study since subject learn novel stimuli and are then tested on them. Modeling, in the sense of machine learning, would then involve on-line learning algorithms. It seems that from the point of view of machine learning, further methodological advances are needed to be able to fully understand the processes behind human memory performance—this chapter has provided an experimental paradigm which might help to validate these further developments.

Chapter 8

Conclusions

We developed a novel methodology—PSYCHO ML—allowing to quantitatively bridge the gap between human psychophysics and machine learning to gain insight into the algorithms used by human subjects during visual classification of faces. In this "machine-learning-psychophysics" research we substitute a very hard to analyze complex system—the human brain—by a reasonably complex system—a learning machine—that is complex enough to capture some essentials of the human behavior but is amenable to close analysis, allowing us to make predictions about human behavior based on machine properties. The psychophysical classification experiment is a behavioral study and machine learning is used to help understand human classification behavior. Machine learning allows us to "look into the human brain" on an algorithmic level and extract quantitative information from a psychophysical experiment by using unsupervised machine learning to model feature extraction and supervised machine learning for classification. Once a feature extractor is chosen and the corresponding data representation computed, the corresponding feature vector is classified using a separating hyperplane (SH) between the classes. The responses of humans to one stimulus—the class estimate, its corresponding reaction time and confidence rating—are correlated to the distance of the feature vector of this stimulus to the hyperplane. These comparisons and correlations between man and machine then give a hint at the mechanisms and strategies used by human subjects to classify visual stimuli.

In our study, a gender classification task is considered and the stimuli are drawn from a processed version of the Max Planck Institute (MPI) human face database where the faces are centered in the image, have same pixel-surface area and same mean and standard deviation of the intensity. Such a task is clearly of high biological relevance, and this is certainly one of the main arguments in favor of using this database, its choice being further motivated in Chapter 1.

The classification behavior of man is studied in two *psychophysical ex-*

periments. In the first classification experiment, each subject is shown sequentially a random subset of the face database. The subjects are then asked to classify the faces according to their gender and the subjects' gender estimates are recorded as much as the corresponding reaction times and confidence ratings. It can then be observed that a high classification error and a low confidence rating for humans are accompanied by a longer processing of the relevant information by the subjects' brain i.e. a longer reaction time. In a second classification experiment, each subject is shown a second time the same subset of the face database as already seen in the first experiment, however in a different presentation order. The classification task is identical. This experiment is a verification of the consistency of the subject's responses. These studies validate the concepts, the reproducibility of the results and the setting of the experiments, and allow to proceed to further analysis. Moreover, a jitter in the subject's gender estimate for "difficult" stimuli is observed, this fact being used below when using machine to interpret the human data.

Separating hyperplanes (SHs) are shown to be a plausible model to describe *classification* of visual stimuli by humans since elements far from the SH are classified more accurately, faster and with higher confidence than those near to the SH. A piecewise linear extension as for the K-means clustering algorithm combined with a nearest-neighbor classifier (Kmean) seems however less adapted to model classification. Support Vector Machines (SVMs) and Relevance Vector Machines (RVMs) compare best to human classification performance and also exhibit the best man-machine correlations. The mean-of-class prototype (Prot), its popularity in neuroscience notwithstanding, is the least human-like classifier in all cases examined. A probabilistic model (such as RVM) or a statistically optimal one (such as SVM) seem to better capture the human classification behavior than the simple Prot and Kmean, suggesting also that exemplar-based models may describe best the classification behavior of human subjects. Elements near to the SH as used by SVM or RVM for classification seem to be better suited for the purpose of classification than elements in the middle to the classes as used by Prot or Kmean. In other words, neither the patterns "easy" to classify—stimuli far from the SH (the male or female caricatures)—nor the most typical ones—stimuli in the center of the classes (the prototypes or means)—seem to be useful for classification. However, the patterns difficult to classify—the androgynous faces close to the SH—can be assumed to be critical for classification. Considering the stochastic nature of the subjects' class estimation between the first and second classification experiments (the jitter in the gender estimate as mentioned above), one can expect that stimuli close to the SH are subject to more jitter than those distant of the the SH. The analysis of the corresponding man-machine correlations finally confirms the above findings: humans may use mechanisms akin to SVMs or RVMs, but are unlikely to use prototype classifiers.

Applying the above studies on the encodings corresponding to the stimuli obtained for various *feature extractors* (i.e. the combination of a data type with a preprocessor), Gabor wavelet filters seem to be a well adapted model for preprocessing on the image pixel data type for encoding visual information. This may be expected, and thus also confirms the validity of our approach, since the data on the retina is clearly an image and Gabor wavelet filters have been shown in numerous studies such as [Hubel and Wiesel, 1962] to be biologically plausible. A less intuitive result is the good performance of the texture-and-shape data type which has an in-built knowledge of the spatial correspondence between regions of the images. This seems to indicate that, as an alternative approach to using the raw pixel data, the human subjects may use the information contained in the combination of the texture and the shape maps of each face to build their internal representation of visual stimuli. On the texture-and-shape data type, Non-negative Matrix Factorization is demonstrated to describe well the preprocessing of visual information in humans, and this has three implications. First, humans seem to use a basis of images to encode visual information, what may suggest that models such as kernel maps are less adapted since they do not use a basis to decompose (visual) data. Second this basis seems to be part-based, in contrast to Principal Component Analysis which yields a holistic basis. Third, this part-based basis is spatially not too sparse, ruling out Independent Component Analysis which has a maximally sparse basis. Finally, histograms and neighborhood-preserving methods such as Locally Linear Embedding do not seem adapted to model feature extraction in humans, at least given the MPI face database.

The above findings have some implications on *sparseness* issues in the representation and processing of visual stimuli in the human brain. A high sparseness of the classification algorithm does not seem to be relevant for classification. In other words the SH of SVMs and RVMs, both exemplar-based classifiers, is computed using a relatively large subset of patterns from the dataset. This fact may account for good classification performance and a high robustness of the decision function. However, for the encodings and for the image basis, a medium degree of sparseness is shown to be most adapted. For the basis this translates into a part-based basis where "regions" from the face are highlighted. Thus, a bit of sparseness in the basis and encoding is good, but too much is bad.

Another important issue related to classification is feature ranking. The latter is studied using the linearity of the classification system composed of a linear feature extractor combined with a linear classifier. The resulting decision images are a direct and principled way to visualize the regions of the face stimuli most useful for classification both for man and machine without use of *prior* information. These images are compared to those obtained using Recursive Feature Elimination (RFE), a benchmark method from machine learning. While all these decision images look quite similar,

they induce very different decision spaces as shown by the following studies. We then study the metric of the human internal representation of faces using a logistic regression to interpolate between the subjects' class estimate for a stimulus and the distance from this stimuli to the SH. It is shown that if trained on the true labels some machines perform the classification task quite similarly to humans in terms of classification performance but they classify faces very differently from human subjects. On the other hand, machines can re-create the decision boundary and the internal representation of faces for human subjects very well if trained on the subjects' labels. When machine are trained in a space of reduced dimensionality as obtained by RFE, we notice that equating the classification performance of man and machine through RFE makes machines even less human-like than if trained on the true labels. This shows that even if man and machine perform a task equally well—i.e. same classification error—this does not imply anything about their internal workings, corroborating one of the results of this thesis: the classification performance is not enough to infer on the mechanisms used by humans. In these studies, SVM and RVM behave best, while the prototype learner is again the worst candidate to model classification in humans.

Using the decision images, a novel psychophysical experiment is designed where the hypotheses generated from machine learning are used to generate novel stimuli along a direction—the gender axis—orthogonal to the SH of each classifier. The correlation studies of this dissertation reported that the subjects' responses to the faces correlated very well with the distance of the stimuli to their SH for SVMs and RVMs but not for the simple prototype classifier. If these correlations really implied that SVMs and RVMs capture some crucial aspects of human internal face representation, their gender axis should be closely aligned to those of our subjects whereas that is not expected to be the case for Prot. A psychophysical gender discrimination experiment confirms these predictions. In other words, from the analysis of the machines we make predictions for human subjects which we subsequently test psychophysically. By doing so, we close the man-machine loop, and demonstrate that machine learning is a suitable method to model the classification of visual stimuli, at least for the MPI face database.

Finally, we investigate the mechanisms responsible for the memorization of visual stimuli. After a first classification experiment some sets of stimuli—the representations and the non-representations—are determined for each subject. The subjects' labelling of these stimuli as seen/unseen may be a clue for their relevance during classification. It is shown that given the MPI face database and the particular task we chose, no effect can be observed: all stimuli are memorized similarly. We can then conclude that it is difficult to cast concepts from machine learning into a formalism describing the memory mechanisms of humans. However, machine learning is successfully used to model feature extraction and classification of visual stimuli in humans.

Bibliography

E.H. Adelson and J.R. Bergen. The plenoptic function and the elements of early vision. In M.S. Landy and J.A. Movshon, editors, *Computational Models of Visual Processing*, chapter 1, pages 3–20. MIT Press, 1991.

A.J. Ahumada. Classification image weights and internal noise level estimation. *Journal of Vision*, 2:121–131, 2002.

F.G. Ashby and S.W. Ell. The neurobiology of human category learning. *Trends in Cognitive Sciences*, 5(5):204–210, 2001.

K. Baek, B.A. Draper, J.R. Beveridge, and K. She. PCA vs ICA: A comparison on the feret data set. In *Joint Conference on Information Sciences*, 2002.

H.B. Barlow. Single units and sensation: A neuron doctrine for perceptual psychology? *Perception*, 1:371–394, 1972.

M.S. Bartlett, J.R. Movellan, and T.J. Sejnowski. Face recognition by independent component analysis. *IEEE Transactions on Neural Networks*, 13(6):1450–1464, 2002.

J.A. Bednar and R. Miikkulainen. Learning innate face preferences. *Neural Computation*, 15:1525–1557, 2003.

I. Biederman. Recognition-by-components: A theory of human image understanding. *Psychological Review*, 94:115–147, 1987.

E. Bienenstock, L.N. Cooper, and P. Munro. A theory for the development of neuron selectivity: Orientation specificity and binocular interaction in visual cortex. *The Journal of Neuroscience*, 2:32–48, 1982.

K.T. Blackwell, T.P. Vogl, H.P. Dettmar, M.A. Brown, G.S. Barbour, and D. L. Alkon. Identification of faces obscured by noise: Comparison of an artificial neural network with human observers. *Journal of Experimental and Theoretical Artificial Inteligence*, 9:491–508, 1997.

V. Blanz and T. Vetter. A morphable model for the synthesis of 3d faces. In *Siggraph99*, pages 187–194. ACM Press, 1999.

V. Blanz and T. Vetter. Face recognition based on fitting a 3d morphable model. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 25(9):1–12, 2003.

L. Bottou, 2003. Private discussion.

J. Bromley and E. Säckinger. Neural-network and k-nearest-neighbor classifiers. Technical report, AT&T, 1991.

H.H. Bülthoff and S. Edelman. Psychophysical support for a two-dimensional view interpolation theory of object recognition. *Proceedings of the National Academy of Scienes USA*, 89:60–64, 1992.

I. Bülthoff and H.H. Bülthoff. Image-based recognition of biological motion, scenes and objects. In M.A. Peterson and G. Rhodes, editors, *Perception of Faces, Objects, and Scenes: Analytic and Holistic Processes*, pages 146–176. Oxford University Press, 2003.

J.-F. Cardoso. Blind signal separation: statistical principles. *Proceedings of the IEEE*, 9(10):2009–2025, 1998.

C.-C. Chang and C.-J. Lin. *LIBSVM: a library for support vector machines*, 2001. Software available at `http://www.csie.ntu.edu.tw/~cjlin/libsvm`.

O. Chapelle and B. Schölkopf. Incorporating invariances in nonlinear support vector machines. In *Advances in Neural Information Processing Systems 14*, pages 609–616. MIT Press, 2002.

Z. Chen and S. Haykin. On different facets of regularization theory. *Neural Computation*, 14:2791–2846, 2002.

P.S. Churchland and T.J. Sejnowski. *The Computational Brain*. MIT Press, 1992.

C. Cortes and V.N. Vapnik. Support-vector networks. In *Machine Learning*, 20, pages 273–297. Kluwer Academic Publishers, Boston, 1995.

N. Cristianini and J. Shawe-Taylor. *An Introduction to Support Vector Machines and other kernel-based learning methods*. Cambridge University Press, 2000.

M.N. Dailey, G.W. Cottrell, and T.A. Busey. Facial memory is kernel estimation (almost). In *Advances in Neural Information Processing Systems 11*, pages 24–30. MPI Press, 1999.

J.G. Daugman. Uncertainty relation for resolution in space, spatial frequency, and orientation optimized by two-dimensional visual cortical filters. *Journal of Optical Society of America A*, 2(7):1160–1169, 1985.

P. Dayan and L.F. Abbott. *Theoretical Neuroscience*. MIT Press, 2001.

H. Op de Beeck, J. Wagemans, and R. Vogels. Inferotemporal neurons represent low-dimensional configurations of parametrized shapes. *Nature Neuroscience*, 4(12):1244–1252, 2001.

A. Delorme. Early cortical orientation selectivity: How fast inhibition decodes the order of spike latencies. *Journal of Computational Neuroscience*, 15:357–365, 2003.

A. Delorme, J. Gautrais, R. Van Rullen, and S. Thorpe. Spikenet: A simulator for modeling large networks of integrate and fire neurons. *Neurocomputing*, 26-27:989–996, 1999.

A. Delorme, G. Richard, and M. Fabre-Thorpe. Ultra-rapid categorization of natural scenes does not rely on colour cues: a study in monkeys and humans. *Vision Research*, 40:2187–2200, 2000.

A. Delorme and S. Thorpe. Spikenet: An event-driven simulation package for modeling large networks of spiking neurons. *Network: Comput. Neural Syst.*, 14:613–627, 2003.

B.A. Draper, K. Baek, M.S. Bartlett, and J.R. Beveridge. Recognizing faces with PCA and ICA. In *Computer Vision Image Understanding—Face Recognition*, 2002.

R.O. Duda, P.E. Hart, and D.G. Stork. *Pattern Classification*. John Wiley & Sons, second edition, 2001.

S. Edelman. Representation, similarity, and the chorus of prototypes. *Minds and Machines*, 5:45–68, 1995.

A. Engel and C. Van den Broeck. *Statistical Mechanics of Learning*. Cambridge University Press, 2001.

M. Fabre-Thorpe, A. Delorme, C. Marlot, and S. Thorpe. A limit to the speed of processing in ultra-rapid visual categorization of novel natural scenes. *Journal of Cognitive Neuroscience*, 13(2):171–180, 2001.

D. Fass and J. Feldman. Categorization under complexity: A unified MDL account of human learning of regular and irregular categories. In *Advances in Neural Information Processing Systems 16*. MIT Press, 2003.

J. Feng, Y. Sun, H. Buxton, and G. Wei. Training integrate-and-fire neurons with the informax principle II. *IEEE Transactions on Neural Networks*, 14(2):326–336, 2003.

R.A. Fisher. The use of multiple measurements in taxonomic problems. *Annals of Eugenics*, 7(2):179–188, 1936.

D.J. Freedman, M. Riesenhuber, T. Poggio, and E.K. Miller. Categorical representation of visual stimuli in the primate prefrontal cortex. *Science*, 291:312–316, 2001.

D.J. Freedman, M. Riesenhuber, T. Poggio, and E.K. Miller. Visual categorization and the primate prefrontal cortex: Neurophysiology and behavior. *Journal of Neurophysiology*, 88:930–942, 2002.

D.J. Freedman, M. Riesenhuber, T. Poggio, and E.K. Miller. A comparison of primate prefrontal and inferior temporal cortices during visual categorization. *The Journal of Neuroscience*, 23(12):5235–5246, 2003.

W.J. Freeman. The wave packet: An action potential for the 21st century. *Journal of Integrative Neuroscience*, 2(1):3–30, 2003.

Y. Freund and R.E. Schapire. A decision-theoretic generalization of on-line learning and an application to boosting. In *European Conference on Computational Learning Theory*, pages 23–37, 1995.

N. Furl, P.J. Phillips, and A.J. O'Toole. Face recognition algorithms and the other-race effect: Computational mechanisms for a developmental contact hypothesis. *Cognitive Science*, 26:797–815, 2002.

I. Gauthier, T. Curran, K.M. Curby, and D. Collins. Perceptual interference supports a non-modular account of face processing. *Nature Neuroscience*, 6(4):428–432, 2003.

I. Gauthier and N.K. Logothetis. Is face recognition not so unique after all? *Cognitive Neuropsychology*, 17:125–142, 2000.

W. Gerstner and W.M. Kistler. *Spiking Neuron Models: Single Neurons, Populations, Plasticity*. Cambridge University Press, 2002.

R.L. Goldstone. Do we all look alike to computers? *Trends in Cognitive Sciences*, 7(2):55–57, 2003.

B.A. Golomb, D.T. Lawrence, and T.J. Sejnowski. SEXNET: A neural network identifies sex from human faces. In *Advances in Neural Information Processing Systems 3*, pages 572–577, 1991.

F. Gosselin and P.G. Schyns. Bubbles: a technique to reveal the use of information in recognition tasks. *Vision Research*, 41:2261–2271, 2001.

T. Graepel, R. Herbrich, and R.C. Williamson. From margin to sparsity. In *Advances in Neural Information Processing Systems 13*, pages 210–216, 2001.

A.B.A. Graf and S. Borer. Normalization in support vector machines. In *Pattern Recognition (DAGM)*, LNCS 2191, pages 277–282. Springer, 2001.

A.B.A. Graf, O. Bousquet, and G. Rätsch. Prototype learning revisited. *Journal of Machine Learning Research*, 2004a. To be submitted.

A.B.A. Graf, A.J. Smola, and S. Borer. Classification in a normalized feature space using support vector machines. *IEEE Transactions on Neural Networks*, 14(3):597–605, 2003.

A.B.A. Graf and F.A. Wichmann. Gender classification of human faces. In *Biologically Motivated Computer Vision (BMCV)*, LNCS 2525, pages 491–501. Springer, 2002.

A.B.A. Graf and F.A. Wichmann. Insights from machine learning applied to human visual classification. In *Advances in Neural Information Processing Systems 16*, pages 905–912. MIT Press, 2004.

A.B.A. Graf, F.A. Wichmann, H.H. Bülthoff, and B. Schölkopf. Classification and memory behaviour of man revisited by machine. In *CSHL Meeting on Computational & Systems Neuroscience (COSYNE)*, page 72, 2004b.

A.B.A. Graf, F.A. Wichmann, B. Schölkopf, and H.H. Bülthoff. Comparing linear classifiers to human visual classification. In *Advances in Neural Information Processing Systems 17*, 2004c. Submitted.

M.S. Gray, D.T. Lawrence, B.A. Golomb, and T.S. Sejnowski. A perceptron reveals the face of sex. *Neural Computation*, 7(6):1160–1164, 1995.

I. Guyon, J. Weston, S. Barnhill, and V.N. Vapnik. Gene selection for cancer classification using support vector machines. *Machine Learning*, 46:389–422, 2002.

S. Haykin. *Neural Networks: a Comprehensive Approach*. Prentice Hall, second edition, 1999.

R. Herbrich and T. Graepel. A PAC-Bayesian margin bound for linear classifiers: Why svms work. In *Advances in Neural Information System Processing 13*, 2001.

D. Hubel and T. Wiesel. Receptive fields, binocular interaction and functional architecture in the cat's visual cortex. *Journal of Physiology*, 160: 106–154, 1962.

A. Hyvärinen. Fast and robust fixed-point algorithms for independent component analysis. *IEEE Transaction on Neural Networks*, 1999.

A. Hyvärinen and E. Oja. A fast fixed-point algorithm for independent component analysis. *Neural Computation*, 9:1483–1492, 1997.

A. Hyvärinen and E. Oja. Independent component analysis: Algorithms and applications. *Neural Networks*, 1999.

L. Itti, C. Koch, and J. Braun. Revisiting spatial vision: Towards a unifying model. *Journal of the Optical Society of America A*, 17(11):1899–1917, 2000.

B.D. Josephson. How we might be able to understand the brain. In *Proceedings International Conference on Complex Systems*, 2004.

A.G. Knapp and J.A. Anderson. Theory of cateogization based on distributed memory storage. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 10(4):616–637, 1984.

U.H.-G. Kressel. Pairwise classification and support vector machines. In B. Schölkopf, C.J.C. Burges, and A.J. Smola, editors, *Advances in Kernel Methods - Support Vector Learning*. MIT Press, 1999.

K. Lamberts. Process models of categorization. In K. Lamberts and D. Shanks, editors, *Knowledge, Concepts, and Categories*, chapter 10. MIT Press, 1997.

Y. LeCun, L. Bottou, G.B. Orr, and K.-R. Müller. Efficient backprop. In *Neural Networks, Tricks of the Trade*, Lecture Notes in Computer Science 1524. Springer Verlag, 1998.

D.D. Lee and H.S. Seung. Learning the parts of objects by non-negative matrix factorization. *Nature*, 401:788–791, 1999.

D.D. Lee and H.S. Seung. Algorithms for non-negative matrix factorization. In *Advances in Neural Information Processing Systems*, pages 556–562, 2000.

D.A. Leopold, A.J. O'Toole, T. Vetter, and V. Blanz. Prototype-referenced shape encoding revealed by high-level aftereffects. *Nature Neuroscience*, 4(1):89–94, 2001.

C. Liu and H. Wechsler. Independent component analysis of gabor features for face recognition. *IEEE Transactions on Neural Networks*, 14(4):919–928, 2003.

J. Lu, K.N. Plataniotis, and A.N. Venetsanopoulos. Face recognition using kernel direct discriminant analysis algorithms. *IEEE Transactions on Neural Networks*, 14(1), 2003.

W. Maass, R.A. Legenstein, and H. Markram. A new approach towards vision suggested by biologically realistic neural microcircuit models. In *Biologically Motivated Computer Vision (BMCV)*, LNCS 2525, pages 282–293. Springer, 2002a.

W. Maass, T. Natschläger, and H. Markram. Real-time computing without stable states: A new framework for neural computation based on perturbations. *Neural Computation*, 14:2531–2560, 2002b.

W. Maass, T. Natschläger, and H. Markram. A model for real-time computation in generic neural microciruits. In *Advances in Neural Information Processing Systems 15*, pages 229–236. MIT Press, 2003.

P. Mamassian and R. Goutcher. Prior knowledge on the illumination position. *Cognition*, 81:B1–B9, 2001.

P. Mamassian and M.S. Landy. Observer biases in the 3d interpretation of line drawings. *Vision Research*, 38:2817–2832, 1998.

B.W. Mel. SEEMORE: Combining color, shape and texture histogramming in a neurally-inspired approach to visual object recognition. *Neural Computation*, 9(4):777–804, 1997.

S. Mika, G. Rätsch, and K.-R. Müller. A mathematical programming approach to the kernel fischer algorithm. In *Advances in Neural Information Processing Systems 13*. MIT Press, 2001.

S. Mika, G. Rätsch, J. Weston, B. Schölkopf, and K.-R. Müller. Constructing descriptive and discriminative non-linear features: Rayleigh coefficients in kernel feature spaces. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 25(5):623–628, 2003.

B. Moghaddam and M.-H. Yang. Gender classification with support vector machines. Technical report, Mitsubishi Electric Research Laboratory, 2000.

B. Moghaddam and M.-H. Yang. Learning gender with support faces. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(5):707–711, 2002.

T. Natschläger and W. Maass. Information dynamics and emergent computation in recurrent circuits of spiking neurons. In *Advances in Neural Information Processing Sysems 16*, 2004.

A. Nieder, D.J. Freedman, and E.K. Miller. Representation of the quantity of visual items in the primate prefrontal cortex. *Science*, 2002.

R.M. Nosofsky. Tests of an exemplar model for relating perceptual classification and recognition memory. *Journal of Experimental Psychology: Human Perception and Performance*, 17(1):3–27, 1991.

B.A. Olshausen and D. Field. Emergence of simple-cell receptive field properties by learning a sparse code for natural images. *Nature*, 381:607–609, 1996.

M. Opper and O. Winther. Gaussian processes for classification: Mean-field algorithms. *Neural Computation*, 12(11):2655–2684, 2000.

A.J. O'Toole, H. Abdi, K.A. Deffenbacher, and D. Valentin. Low-dimensional representation of faces in higher dimensions of the face space. *Journal of the Optical Society of America A*, 10(3):405–411, 1993.

A.J. O'Toole, K.A. Defenbach, D. Valentin, K. McKee, D. Huff, and H. Abdi. The perception of face gender: The role of stimulus structure in recognition and classification. *Memory and Cognition*, 26:146–160, 1998.

A.J. O'Toole, T. Vetter, and V. Blanz. Three-dimensional shape and two-dimensional surface reflectance contributions to face recognition: an application of three-dimensional morphing. *Vision Research*, 39:3145–3155, 1999.

T.J. Palmeri. The time course of perceptual categorization. In U. Hahn and M. Ramscar, editors, *Similarity and Categorization*, chapter 11. Oxford University Press, 2001.

D.G. Pelli, B. Farell, and D.C. Moore. The remarkable inefficiency of word recognition. *Nature*, 423:752–756, 2003.

R.J. Peters, F. Gabbiani, and C. Koch. Human visual object categorization can be described by models with low memory capacity. *Vision Research*, 43(21):2265–2280, 2003.

M.A. Peterson and G. Rhodes, editors. *Perception of Faces, Objects, and Scenes: Analytic and Holistic Processes*. Oxford University Press, 2003.

P.J. Phillips, P. Grother, R.J. Micheals, D.M. Blackburn, E. Tabassi, and M. Bone. Face recognition vendor test 2002: Evaluation report. Technical report, DARPA, 2003.

J.C. Platt, N. Cristianini, and J. Shawe-Taylor. Large margin DAGs for multiclass classification. In *Advances in Neural Information Processing Systems 12*. MIT Press, 2000.

M. D. Plumbley. Algorithms for nonnegative independent component analysis. *IEEE Transactions on Neural Networks*, 14(3):534–543, 2003.

M.I. Posner and S.W. Keele. On the genesis of abstract ideas. *Journal of Experimental Psychology*, 77(3):353–363, 1968.

M.I. Posner and S.W. Keele. Retention of abstract ideas. *Journal of Experimental Psychology*, 83(2):304–308, 1970.

W.H. Press, S.A. Teukolsky, W.T. Vetterling, and B.P. Flannery. *Numerical Recipes in C*. Cambridge University Press, second edition, 1992.

R.P.N. Rao, B.A. Olshausen, and M.S. Lewicki, editors. *Probabilistic Models of the Brain*. MIT Press, 2002.

S.K. Reed. Pattern recognition and categorization. *Cognitive Psychology*, 3: 382–407, 1972.

M. Riesenhuber and T. Poggio. Hierarchical models of object recognition in cortex. *Nature Neuroscience*, 2:1019–1025, 1999.

M. Riesenhuber and T. Poggio. Models of object recognition. *Nature Neuroscience*, 3 (suppl.):1199–1204, 2000.

M. Riesenhuber and T. Poggio. Neural mechanisms of object recognition. *Current Opinion in Neurobiology*, 12:162–168, 2002.

E. Rosch, C.B. Mervis, W.D. Gray, D.M. Johnson, and P. Boyes-Braem. Basic objects in natural categories. *Cognitive Psychology*, 8:382–439, 1976.

F. Rosenblatt. The perceptron: a probabilistic model for information storage and organization in the brain. *Psychological Review*, 65(6):386–408, 1958.

S.T. Roweis. Finding the first few eigenvectors in a large space, 1996.

S.T. Roweis and L.K. Saul. Nonlinear dimensionality reduction by locally linear embedding. *Science*, 290:2323–2326, 2000.

R. Van Rullen, J. Gautrais, A. Delorme, and S. Thorpe. Face processing using one spike per neurone. *BioSystems*, 48:229–239, 1998.

L.K. Saul and S.T. Roweis. Think globally, fit locally: Unsupervised learning of low dimensional manifolds. *Journal of Machine Learning Research*, 4: 119–155, 2003.

R.E. Schapire, Y. Freund, P. Bartlett, and W.S. Lee. Boosting the margin: a new explanation for the effectiveness of voting methods. *The Annals of Statistics*, 26(5):1651–1686, 1998.

B. Schölkopf, C. Burges, and V.N. Vapnik. Extracting support data for a given task. In *Proceedings, First International Conference on Knowledge Discovery & Data Mining*. AAAI Press, 1995.

B. Schölkopf, C.J.C. Burges, and A.J. Smola, editors. *Advances in Kernel Methods: Support Vector Learning*. MIT Press, 1999.

B. Schölkopf and A.J. Smola. *Learning with Kernels*. MIT Press, 2002.

B. Schölkopf, A.J. Smola, and K.-R. Müller. Nonlinear component analysis as a kernel eigenvalue problem. *Neural Computation*, 10:1299–1319, 1998.

N. Sigala, F. Gabbiani, and N.K. Logothetis. Visual categorization and object recognition in monkeys and humans. *Journal of Cognitive Neuroscience*, 14(2):187–198, 2002.

N. Sigala and N.K. Logothetis. Visual categorization shapes feature selectivity in the primate temporal cortex. *Nature*, 415:318–320, 2002.

L. Sirovich. Turbulence and the dynamics of coherent structures. *Quarterly of Applied Mathematics*, XLV:561–590, 1987.

L. Sirovich. Dynamics of neuronal populations: Eigenfunction theory; some solvable cases. *Network: Computation in Neural Systems*, 14:249–272, 2003.

L. Sirovich and M. Kirby. Low-dimensional procedure for the characterization of human faces. *Journal of the Optical Society of America A*, 4(3): 519–524, 1987.

R.S. Sutton and A.G. Barto. *Reinforcement Learning: An Introduction*. MIT Press, 1998.

M.J. Tarr and H.H. Bülthoff, editors. *Object Recognition in Man, Monkey, and Machine*. MIT Press, 1998.

M.J. Tarr, H.H. Bülthoff, M. Zabinski, and V. Blanz. To what extent do unique parts influence recognition across changes in viewpoint? *Psychological Science*, 8(4):282–289, 1997.

J.B. Tenenbaum, V. de Silva, and J.C. Langford. A global geometric framework for nonlinear dimensionality reduction. *Science*, 290:2319–2323, 2000.

S. Thorpe. Ultra-rapid scene categorization with a wave of spikes. In *Biologically Motivated Computer Vision (BMCV)*, LNCS 2525, pages 1–15. Springer, 2002.

M.E. Tipping. The relevance vector machine. In *Advances in Neural Information Processing Systems 12*, pages 652–658. MIT Press, 2000.

M.E. Tipping. Sparse bayesian learning and the relevance vector machine. *Journal of Machine Learning Research*, 1:211–214, 2001.

M.E. Tipping. *"SparseBayes": A Matlab Implementation of Sparse Bayesian Learning*, 2002.

N.F. Troje and H.H. Bülthoff. Face recognition under varying poses: The role of texture and shape. *Vision Research*, 36(12):1761–1771, 1996.

D.Y. Tsao, W.A. Freiwald, T.A. Knutsen, J.B. Mandeville, and R.B.H. Tootell. Faces and objects in macaque cerebral cortex. *Nature Neuroscience*, 6(9):989–995, 2003.

M. Turk and A. Pentland. Eigenfaces for recognition. *Journal of Cognitive Neuroscience*, 3(1), 1991.

D. Valentin, H. Abdi, B. Edelman, and A.J. O'Toole. Principal component and neural network analyses of face images: What can be generalized in gender classification? *Journal of Mathematical Psychology*, 41:398–413, 1997.

V.N. Vapnik. *The Nature of Statistical Learning Theory*. Springer, second edition, 2000.

V.N. Vapnik and A. Lerner. Pattern recognition using generalized portrait method. *Automation and Remote Control*, 24:774–780, 1963.

T. Vetter and N.F. Troje. Separation of texture and shape in images of faces for image coding and synthesis. *Journal of the Optical Society of America A*, 14(9), 1997.

C. Wallraven and A.B.A. Graf. Image classification with svms using spatio-temporal feature representations. *Journal of Machine Learning Research*, 2004. Submitted.

J. Weston and C. Watkins. Support vector machines for multi-class pattern recognition. In *ESANN'99*, 1999.

F.A. Wichmann, A.B.A. Graf, E.P. Simoncelli, H.H. Bülthoff, and B. Schölkopf. Machine learning applied to perception: Decision images for classification. In *Advances in Neural Information Processing Systems 17*, 2004. Submitted.

F.A. Wichmann and N.J. Hill. The psychometric function: I. fitting, sampling and goodness-of-fit. *Perception and Psychophysics*, 63(8):1293–1313, 2001a.

F.A. Wichmann and N.J. Hill. The psychometric function: II. bootstrap-based confidence intervals and sampling. *Perception and Psychophysics*, 63(8):1314–1329, 2001b.

F.A. Wichmann, L.T. Sharpe, and K.R. Gegenfurtner. The contribution of color to recognition memory for natural scenes. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 28(3):509–520, 2002.

T.D. Wickens. *Elementary Signal Detection Theory*. Oxford University Press, 2002.

C.K.I. Williams and D. Barber. Bayesian classification with gaussian processes. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(12):1342–1351, 1998.

B. Willmore and D. Smyth. Methods for first-order kernel estimation: Simple-cell receptive fields from responses to natural scenes. *Network: Computation in Neural Systems*, 14:553–577, 2003.

B. Willmore and D.J. Tolhurst. Characterizing the sparseness of neural codes. *Network: Computation in Neural Systems*, 12:255–270, 2001.

A.L. Yarbus. *Eye Movements and Vision*. Plenum Press, New York, 1967.

Appendix A

Data Representation

A.1 Overview

In this appendix we consider algorithms relying on unsupervised learning methods which are used to find different manners to represent and embed data. These algorithms as thus referred to a feature extractors since they are assumed to extract the most relevant information from the data. At the same time, these algorithms perform dimensionality reduction, a useful property when used as a stage anterior to classification algorithms since current techniques experience difficulties when dealing with high-dimensional datasets. In particular we study below Principal Component Analysis, Locally Linear Embedding, Independent Component Analysis, Non-negative Matrix Factorization, Empirical Kernel Maps and Gabor wavelet filters.

We assume $\vec{x}_1, \ldots, \vec{x}_p \in \mathbb{R}^n$ is the original data where p is the number of patterns and n their dimensionality. We then define the original data matrix as: $\mathcal{X} = (\vec{x}_1 | \ldots | \vec{x}_p) \in \mathbb{R}^{n \times p}$. In the following, we assume $n > p$. Let the data in the reduced space, also referred to as encoding, be written as $\vec{y}_1, \ldots, \vec{y}_p \in \mathbb{R}^k$ where $k < n$ stands for the dimension of the low-dimensional projection space of the reduced data. We define the matrix of the encoding as: $\mathcal{Y} = (\vec{y}_1 | \ldots | \vec{y}_p) \in \mathbb{R}^{k \times p}$. Note that, except for PCA, the presented feature extractors are not a map $\Phi : \mathbb{R}^n \to \mathbb{R}^k$. In other words, most of these preprocessors are local, and considering a new pattern amounts to recompute the whole algorithm to obtain features.

A.2 Principal Component Analysis

Principal Component Analysis (PCA, see [Duda, Hart, and Stork, 2001, Haykin, 1999]) is considered as a benchmark feature extractor. PCA determines axis in the data space along which the data has largest variance. These directions are called Principal Components (PCs). The encoding is then the projection of the data on these PCs. PCA seeks to find a new basis

to represent the original data such that the coefficients in this new basis are uncorrelated i.e. they cannot be linearly predicted from each other; in other words the PCs are orthonormal. We compute the mean $\vec{\mu} \in \mathbb{R}^n$ over the patterns of the dataset as:

$$\mu_i = \frac{1}{p} \sum_{j=1}^{p} \mathcal{X}_{ij} \qquad i = 1, \ldots, n \tag{A.1}$$

which allows to center the data as:

$$D = \mathcal{X} - \vec{\mu}\vec{1}^T \tag{A.2}$$

where $\vec{1}^T$ is a line vector of ones of size p. The sample covariance matrix of the data is then expressed as:

$$C = \frac{1}{p-1} DD^T \in \mathbb{R}^{n \times n} \tag{A.3}$$

Since the above matrix is real and symmetric, it can be orthogonally diagonalized (spectral theorem). Assume its real eigenvalues are sorted according to decreasing value: $\lambda_1 \geq \ldots \geq \lambda_n$, the corresponding orthonormal eigenvectors $\vec{v}_i$ forming the columns of the orthogonal matrix $V = (\vec{v}_1 | \ldots | \vec{v}_n) \in \mathbb{R}^{n \times n}$. The following can then be written:

$$C = V \Lambda V^T \tag{A.4}$$

where $\Lambda = diag(\vec{\lambda}) \in \mathbb{R}^{n \times n}$. Since $rank(C) = rank(D) \leq p$, at most p eigenvalues of C are different than 0: $\lambda_{p+1} = \ldots = \lambda_n = 0$. We define $\bar{V} = (\vec{v}_1 | \ldots | \vec{v}_k) \in \mathbb{R}^{n \times k}$ as the matrix of the first k eigenvectors. The dataset of reduced dimensionality can then be computed from the original one by projection as:

$$\mathcal{Y} = \bar{V}^T D \tag{A.5}$$

The columns of $\mathcal{Y}$, i.e. the encoding, are the new data vectors which are linear combinations of the original ones. The original data can then be approximated by:

$$\tilde{\mathcal{X}} = \bar{V}\mathcal{Y} + \vec{\mu}\vec{1}^T \tag{A.6}$$

If $\vec{x}$ is a new pattern in the space of the original data and $\vec{y}$ its corresponding projection, we have:

$$\vec{y} = \bar{V}^T(\vec{x} - \vec{\mu}) \quad \Leftrightarrow \quad \vec{x} = \bar{V}\vec{y} + \vec{\mu} \tag{A.7}$$

The PCA algorithm is the only feature extractor presented in this appendix which allows to represent/project a new pattern *without* having to reconsider all the other patterns. The PCA algorithm as presented above is a straightforward eigenvalue problem. A three-layered linear artificial neural

network of the latter, known as an *autoencoder* network, can alternatively be considered. The latter requires again the computation of the correlation matrix and is based upon a weight update based on gradient descent. Notice that the matrix V can be seen as rotation matrix around the origin i.e. a matrix of change of bases. In an extended way, $\bar{V}$ can be interpreted similarly. It is thus important that the data is centered around the origin by subtracting its mean in a first place before performing these rotations.

The computations above mainly hinge upon the determination of the correlation matrix C. For large values of n, as it is mainly the case, the determination of the eigenvalues is computationally expensive or intractable. A solution to this problem is provided in the case where $k \leq p$ by the Singular Value Decomposition (SVD) algorithm [Press, Teukolsky, Vetterling, and Flannery, 1992]. Another method, sometimes referred to as *Snap Shot Method* [Roweis, 1996, Sirovich, 1987] or linear *Kernel PCA* [Schölkopf, Smola, and Müller, 1998], searches to express the eigenvectors as linear combinations of the data vectors. It allows to find the first few leading eigenvectors in a high dimensional space. It has the computational advantage over classic PCA that it does not require the computation of a correlation matrix between the dimensions of the input but between the patterns, avoiding thus the computation of the pseudo-inverse as required by the SVD decomposition.

Using the same notations as above, we define the centered patterns as:

$$\vec{z}_i = \vec{x}_i - \vec{\mu} \quad \forall i = 1, \ldots, p \tag{A.8}$$

and write the covariance matrix of the data as:

$$C = \frac{1}{p-1} \sum_{i=1}^{p} \vec{z}_i \vec{z}_i^T \tag{A.9}$$

We then express the eigenvectors of C as a linear combination of the patterns:

$$\vec{v}^j = \sum_{i=1}^{p} \alpha_i^j \vec{z}_i \quad \forall j = 1, \ldots, n \tag{A.10}$$

The eigenvalue problem of C is then written as:

$$C\vec{v}^j = \lambda^j \vec{v}^j \quad \forall j \tag{A.11}$$

Inserting equ.A.9 and equ.A.10 into the above, we get:

$$\sum_{kl} \vec{z}_k G_{kl} \alpha_l^j = \lambda^j \sum_m \alpha_m^j \vec{z}_m \quad \forall j \quad \text{where} \quad G_{ij} = \frac{1}{p-1} \vec{z}_i^T \vec{z}_j \tag{A.12}$$

where G is the Gram matrix of the data. The above implies:

$$G\vec{\alpha}^j = \lambda^j \vec{\alpha}^j \quad \forall j \tag{A.13}$$

Finally, we can rewrite the above in matrix notation. The Gram matrix of the patterns is defined as:

$$G = \frac{1}{p-1} D^T D \in \mathbb{R}^{p \times p} \tag{A.14}$$

Since the above matrix is real and symmetric, it may be orthogonally diagonalized (spectral theorem), yielding the real ordered eigenvalues $\lambda_1 \geq \ldots \geq \lambda_p$ and the following:

$$G = V \Lambda V^T \tag{A.15}$$

where $\Lambda = diag(\vec{\lambda}) \in \mathbb{R}^{p \times p}$ and $V = (\vec{v}_1 | \ldots | \vec{v}_p) \in \mathbb{R}^{p \times p}$ is the orthogonal matrix of eigenvectors. The latter allows to write the matrix of change of base as:

$$U = \mathbb{GS}[DV] \in \mathbb{R}^{n \times p} \tag{A.16}$$

where $\mathbb{GS}$ stands for the Gram-Schmidt orthonormalization process. Note that the eigenvalues are the same for the classic PCA and the Snap Shot PCA, the eigenvectors however differ by a matrix multiplication with the original data. When defining the dimensionality of the reduced projection space as $k \leq p$, the first k columns of U represent the matrix of change of base $\bar{U} \in \mathbb{R}^{n \times k}$. Finally, the encoding is expressed as:

$$\mathcal{Y} = \bar{U}^T D \tag{A.17}$$

The approximation of the original data is then written as:

$$\tilde{\mathcal{X}} = \bar{U} \mathcal{Y} + \vec{\mu} \vec{1}^T \tag{A.18}$$

Again, the following relation between patterns from the original and reduced space are valid:

$$\vec{y} = \bar{U}^T (\vec{x} - \vec{\mu}) \quad \Leftrightarrow \quad \vec{x} = \bar{U} \vec{y} + \vec{\mu} \tag{A.19}$$

The nonlinear extension of PCA using the kernel trick, KPCA, is based upon the above formulation where in the computation of the Gram matrix the scalar product is replaced by a kernel function [Schölkopf, Smola, and Müller, 1998, Schölkopf, Burges, and Smola, 1999].

A.3 Locally Linear Embedding

Locally Linear Embedding (LLE, see [Roweis and Saul, 2000, Saul and Roweis, 2003]) can be considered as a nonlinear neighborhood-preserving extension of PCA to perform feature extraction i.e. find an embedding of the data. Each pattern of the high-dimensional space of the data is expressed as a linear combination of its K nearest neighbors, yielding thus an local linear embedding. The patterns are then projected into a low-dimensional space while preserving this embedding. Fig.A.1 gives a schematic view of the

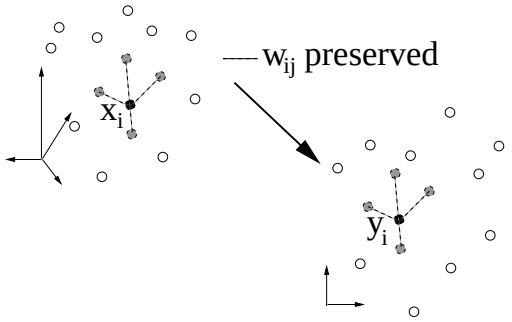


Figure A.1: Schematic view of the LLE algorithm.

LLE algorithm. In order to compute the weights w_{ij} of the local embedding, the following error function is to be minimized:

$$\epsilon = \frac{1}{2}\sum_{i=1}^{p} \|\vec{x}_i - \sum_{j=1, j\neq i}^{p} w_{ij}\vec{x}_j\|^2 \tag{A.20}$$

subject to the constraints:

$$\begin{cases} \sum_{j=1}^{p} w_{ij} = 1 \qquad \text{(convexity constraint)} \\ w_{ij} = 0 \text{ if } \vec{x}_j \notin \mathcal{N}(\vec{x}_i) \end{cases} \tag{A.21}$$

where $\mathcal{N}(\vec{x}_i)$ is the neighborhood of $\vec{x}_i$. Assume $\vec{\eta}_j^i \in \mathcal{N}(\vec{x}_i), j = 1, \ldots, K$. The problem above then reduced to the minimization of the error function:

$$\epsilon = \frac{1}{2}\sum_{i=1}^{p} \|\vec{x}_i - \sum_{j=1}^{K} w_{ij}\vec{\eta}_j^i\|^2 \quad \text{subject to} \quad \sum_{j=1}^{K} w_{ij} = 1 \tag{A.22}$$

The above problem may be solved analytically. Indeed, the Lagrangian associated to this optimization problem can be expressed as:

$$\mathcal{L} = \frac{1}{2}\sum_{i=1}^{p}\sum_{j,k=1}^{K} w_{ij}w_{ik}C_{jk}^i - \sum_{i=1}^{p}\lambda_i(\sum_{j=1}^{K} w_{ij} - 1) \tag{A.23}$$

where λ_i are the Lagrangian multipliers and $C_{jk}^i = \langle \vec{x}_i - \vec{\eta}_j^i | \vec{x}_i - \vec{\eta}_k^i \rangle$ is the Gram matrix of the data. The saddle point of the above Lagrangian ($\frac{\partial \mathcal{L}}{\partial w_{ij}} = 0$) yields:

$$\lambda_i = \sum_{k=1}^{K} w_{ik}C_{jk}^i \quad i = 1, \ldots, p \quad \forall j = 1, \ldots, K \tag{A.24}$$

From the expressions above, we deduce the weights of the local embedding:

$$w_{ij} = \frac{\sum_{k=1}^{K}(C_{kj}^i)^{-1}}{\sum_{k,j=1}^{K}(C_{kj}^i)^{-1}} \quad i = 1, \ldots, p \quad j = 1, \ldots, K \tag{A.25}$$

where $(C^i_{kj})^{-1}$ is the kj^{th} element of the inverse of the matrix $C^i \in \mathbb{R}^{K\times K}$. For each of the elements $\vec{x}_i$ of the original dataset, we search a corresponding embedding element $\vec{y}_i \in \mathbb{R}^k, k \leq p-1$ in a low-dimensional space minimizing the embedding cost function:

$$\Phi = \frac{1}{2}\sum_{i=1}^{p} \|\vec{y}_i - \sum_{j=1}^{p} w_{ij}\vec{y}_j\|^2 \tag{A.26}$$

using the previously-determined weights and subject to the following constraints:

$$\begin{cases} \sum_{i=1}^{p} \vec{y}_i = 0 \\ \frac{1}{p}\sum_{i=1}^{p} \vec{y}_i \otimes \vec{y}_i = \mathbb{I}(k) \end{cases} \tag{A.27}$$

where $\mathbb{I}(k)$ as the identity matrix of size k. The first constraint expresses the fact that the data in the low dimensional space is centered around the origin and the second constraint requires the reduced data to be orthonormal. These constraints avoid thus having an infinite number of solutions. We can rewrite the cost function as:

$$\Phi = \frac{1}{2}\sum_{i,j=1}^{p} M_{ij}\langle\vec{y}_i|\vec{y}_j\rangle \tag{A.28}$$

where $M_{ij} = \delta_{ij} - w_{ij} - w_{ji} + \sum_{k=1}^{p} w_{ki}w_{kj}$, δ_{ij} being the Kronecker symbol. The matrix $M \in \mathbb{R}^{p\times p}$ is symmetric and with the decomposition:

$$M = V\Lambda V^T \tag{A.29}$$

where $\Lambda = diag(\vec{\lambda}) \in \mathbb{R}^{p\times p}, \lambda_1 \leq \ldots \leq \lambda_p$ is the diagonal matrix of eigenvalues and $V = (\vec{v}_1|\ldots|\vec{v}_p) \in \mathbb{R}^{p\times p}$ the orthogonal matrix whose columns are the corresponding eigenvectors. The first $k+1$ eigenvectors, i.e. columns of V, are considered. One of these eigenvectors is a vector with same values throughout and corresponds to an eigenvalue of 0. This vector is discarded and the remaining k vectors correspond to the desired embedding defining the matrix $\bar{V} \in \mathbb{R}^{p\times k}$. Finally, this yields the encoding as:

$$\mathcal{Y} = \bar{V}^T \tag{A.30}$$

Since LLE is invariant to rotation, scaling and translation of the patterns, it may be seen as more biologically relevant than PCA which is is view-dependent. However, these invariances also imply that LLE is not invertible.

The parameter K is the only free parameter of the LLE algorithm. The scalar product in the expression of C^i_{jk} can be replaced by a kernel function (kernel trick), yielding the kernel matrix K^i_{jk}. The computations above are then still valid when replacing C^i_{jk} by K^i_{jk}, yielding the KLLE algorithm. Using a kernel function to compute the Gram matrix i s a nonlinear extension of the LLE algorithm and accommodates for strongly varying manifolds

underlying the data. The above eigenproblem of the matrix M should however not be kernelized since we want the data $\vec{y}_i$ to be in a well-defined space of reduced dimension i.e. the mapping should be explicitly known which is not the case for most feature spaces associated to kernel functions. The neighborhood $\mathcal{N}$ is mostly defined using the Euclidean norm. As an extension to the LLE algorithm as presented above, we could consider the use of a weighted Minkowsky norm with a similarity measure to determine the neighborhoods.

LLE recovers global nonlinear structure from locally linear interpolations. In this respect LLE is similar to Self-Organizing Feature Maps (Kohonen maps) as described in [Duda, Hart, and Stork, 2001] and to multidimensional scaling (MDS). The latter computes embeddings which preserve pairwise distances between patterns over the whole dataset (and not locally) along straight lines. Moreover, the extension of MDS which computes these distances along the shortest paths on the manifold underlying the data is called Isomap [Tenenbaum, de Silva, and Langford, 2000]. All these methods belong to a same family—we choose LLE as its representative member.

A.4 Independent Component Analysis

Independent Component Analysis (ICA, see [Cardoso, 1998, Hyvärinen and Oja, 1999]) can be seen as an unsupervised method to reduce the redundancy in the data by extracting statistically independent signals from it. In other words, it aims at expressing a set of random variables as linear combinations of statistically independent variables, the distribution of the latter being assumed to be non-Gaussian. This feature can also be used to perform dimensionality reduction and feature extraction. Early processing by the brain of sensory data is argued to be explainable by principles similar to ICA. For the sake of simplicity of the presentation, we assume the original data to be sphered using for instance PCA. We consider the linear mixing model which yields a decomposition of the data as follows:

$$\begin{aligned} \hat{S} &= WX \\ X &= BS \end{aligned} \tag{A.31}$$

where $X = \mathcal{X}^T$ is the matrix of the original data (observed signals), $S \in \mathbb{R}^{k \times n}$ the matrix of mutually independent sources (unobserved signals) and $\hat{S}$ its estimate, $W \in \mathbb{R}^{k \times p}$ the unmixing matrix and $B = \mathcal{Y}^T$ the mixing matrix and in this context also the matrix of the data in the reduced space of dimension $k < n$. The principle of ICA is to recover the estimate of the unobserved signals $\hat{S}$ from the observed mixture of the sources X using the assumption of mutual independence between the sources. Both the sources and their mixing are unknown. Below we show how to exhibit W using only X and the corresponding computation of S and B.

The differential entropy of a random vector $\vec{s}$ is defined as:

$$H(\vec{s}) = -\int f(\vec{s}) \log f(\vec{s}) d\vec{s} \tag{A.32}$$

where $f(\vec{s})$ is the probability density function of the vector $\vec{s}$, the latter being seen as an approximation of one of the sources i.e. a row of $\hat{S}$. The variable $\vec{s}$ is assumed of zero mean (i.e. centered) and of unit variance (i.e. whitened). The mutual information between the signals s_i measures the dependency between these random variables and is defined as:

$$I(\vec{s}) = H(\vec{s}) - \sum_{i=1}^{k} H(s_i) \tag{A.33}$$

ICA seeks to minimize the mutual information in order to uncorrelate the signals, or also to maximize the independence between the signals. When $I = 0$ the components of $\vec{s}$ are independent. In the assumption of a linear mixing process, minimizing the mutual information $I(\vec{s})$ is equivalent to maximizing the sum of the negentropy as: $\sum_{i=1}^{k} J(\vec{s}_i)$ as shown by [Hyvärinen and Oja, 1999]. The negentropy is defined as:

$$J(\vec{s}) = H(\vec{s}_{gauss}) - H(\vec{s}) \tag{A.34}$$

where $\vec{s}_{gauss}$ is a Gaussian random variable of the same covariance matrix as $\vec{s}$. The negentropy, like the kurtosis, is a measure of the nongaussianity of a distribution. One can equivalently say that ICA finds directions of maximum non-Gaussianity i.e. maximum negentropy. The crucial point is the estimation of the negentropy, and the following objective function has been proposed:

$$J(\vec{s}) \approx \left(E\{G(\vec{s})\} - E\{G(\vec{\nu})\}\right)^2 \tag{A.35}$$

where $E\{\cdot\}$ is the expectation, G is a non-quadratic function called the contrast function and $\vec{\nu}$ is a Gaussian variable of zero mean and unit variance. We define $\vec{s}_i = \langle \vec{w}_i | \vec{x}_j \rangle \,|_{j=1}^{n} = [WX]_i$ as the i^{th} row of $\hat{S}$, $\vec{w}_i$ as the i^{th} row of the unmixing matrix W and $\vec{x}_j$ as the j^{th} column vector of the original data X. Writing the above in terms of the unmixing matrix W, we get that minimizing the mutual information is equivalent to the following problem:

$$\begin{aligned} &\max \textstyle\sum_{i=1}^{k} J([WX]_i) \\ &\text{subject to } E\{[WX]_i\} = 1 \quad \forall i = 1, \ldots, k \end{aligned} \tag{A.36}$$

The resolution of the ICA problem using the above approximations reduces thus to an optimization problem and yields the unmixing matrix W. This optimization is done using a fast fixed-point iterative scheme as presented in [Hyvärinen, 1999, Hyvärinen and Oja, 1997]. The correctness of the choice

of the contrast function is verified by computing the reconstruction error of the original data (heuristic procedure).

Upon first consideration, it may be concluded that B is the matrix of the encoding and S the matrix whose columns represent the corresponding basis vectors. Recall that applying PCA on X or X^T yields similar results (classic versus Snap Shot PCA). However for ICA the results on X or X^T differ strongly as shown below. Thus we consider two different manners to apply ICA in this setting, one yielding a holistic and the other a part-based basis as discussed in [Bartlett, Movellan, and Sejnowski, 2002] in the context of face recognition. The PCA algorithm on the data matrix $X \in \mathbb{R}^{p\times n}$ yields the data of reduced dimensionality $R \in \mathbb{R}^{p\times k}$ where $k < n$, the mean $\vec{\mu}$ and the matrix of change of basis $U \in \mathbb{R}^{n\times k}$. The above are related as: $R = (X - \vec{1}^T\vec{\mu})U$. We consider below the case where $k = p$ i.e. we use ICA to represent the data and not to reduce its intrinsic dimensionality. We have the two following possibilities to compute ICA and obtain the unmixing matrix $W \in \mathbb{R}^{p\times p}$:

1. **ICA I**—non-sparse encoding and sparse basis: ICA is performed on the rows of U^T and yields:

$$\begin{aligned} S &= WU^T \\ B &= RW^{-1} \end{aligned} \tag{A.37}$$

   The patterns are treated as random variables and their components as outcomes or trials. We have here independence of the patterns: given a component, it is not possible to predict one pattern given another one. The basis S is thus statistically independent, what is not the case for the encoding B. The basis is thus sparse and non-global and the encoding non-sparse.

2. **ICA II**—sparse encoding and non-sparse basis: the ICA algorithm is run on the rows of R^T and yields:

$$\begin{aligned} S &= (W^{-1})^T U^T \\ B &= RW^T \end{aligned} \tag{A.38}$$

   Here the components are treated as random variables and the patterns as outcomes. We have independence of the components in the sense that it is not possible to predict a component given another one on the same pattern. Here the encoding B is statistically independent and sparse and the basis S is non-sparse.

In both cases the matrix $\mathcal{Y} = B^T$ is the matrix representing the encoding i.e. the data in the reduced space and S is the basis matrix. The reconstructed data is then given as $\hat{X} = BS + \vec{1}^T\vec{\mu}$.

A.5 Non-negative Matrix Factorization

The Non-negative Matrix Factorization (NMF, see [Lee and Seung, 1999]) algorithm performs a decomposition of the data into non-negative terms as follows:

$$\mathcal{X} = WH \tag{A.39}$$

where $W \in \mathbb{R}^{n \times k}$ is the matrix of the basis vectors of the representation and $H = \mathcal{Y}$ the matrix of the data in the space of reduced dimensionality whose columns are the encodings corresponding to each pattern of the original data. This decomposition is similar to PCA and ICA (see above). The encoding H consists of the coefficients of the linear combination of the basis vectors of W allowing to reconstruct a pattern from the original data. In the case of PCA, the constraints on the decomposition are that the columns of W be orthonormal and that the rows of H be orthogonal. Non-negative matrix factorization [Lee and Seung, 1999] is an alternative manner to factorize data where the matrices W and H are constrained to be non-negative. In this manner, cancellations among the data provided by summing positive and negative coefficients as for PCA are avoided, allowing thus only additive contributions of the encodings. The corresponding basis function can then be expected to represent "parts" of the original data $\mathcal{X}$, these parts being added to reconstruct $\mathcal{X}$. Sparseness in the basis is thus achieved. The update rule for NMF is as follows:

1. scale the data $\mathcal{X}$ to $[0, 1]$
2. initialize W and H to random values in $[0, 1]$
3. update H and W until convergence as:

$$\begin{aligned} H_{ij} &\leftarrow H_{ij} \textstyle\sum_k W_{ki} \frac{\mathcal{X}_{kj}}{(WH)_{kj}} \\ W_{ij} &\leftarrow W_{ij} \textstyle\sum_k \frac{\mathcal{X}_{ik}}{(WH)_{ik}} H_{jk} \\ W_{ij} &\leftarrow \frac{W_{ij}}{\sum_k W_{kj}} \end{aligned} \tag{A.40}$$

Alternative algorithms to perform NMF as much as the corresponding cost functions are presented in [Lee and Seung, 2000]. NMF only works for non-negative data (it is indeed rather difficult to express a negative value using a linear combination of positive ones with positive coefficients), so the original data $\mathcal{X}$ is first scaled to $[0, 1]$. It is thus not necessary to subtract the mean to the original data (centering) as done for PCA. Further the biological plausibility of NMF is argued to be the sparse basis W representing parts of the patterns of the original data, the non-negativity of the firing rate of neurons, the constancy of the sign of the strength of synaptic connections and

finally the sparseness of the neural code. The quality of the reconstruction is dependent upon the maximum number of iterations of the algorithm.

A.6 Empirical Kernel Maps

The empirical kernel map is an eigenvalue decomposition of a nonlinear extension of the Gram matrix of the original data. It is thus very similar to KPCA. We define as the empirical kernel map with respect to the data $\{\vec{x}_i\}_{i=1}^p$ the mapping:

$$\begin{array}{rccl} \vec{\varphi}_p: & \mathbb{R}^n & \rightarrow & \mathbb{R}^p \\ & \vec{x} & \rightarrow & \vec{\varphi}_p(\vec{x}) = K^{-\frac{1}{2}} \left(K(\vec{x}_1, \vec{x}), \ldots, K(\vec{x}_p, \vec{x})\right)^T \end{array} \tag{A.41}$$

where the matrix $K_{ij} = K(\vec{x}_i, \vec{x}_j)$ for $i, j = 1, \ldots, p$ is the kernel matrix relative to the positive definite kernel function $K(\vec{x}, \vec{y})$ (for more details see [Schölkopf and Smola, 2002]). This type of mapping may be used to represent and embed data. Moreover, it is also an easy way to kernelize any algorithm, in particular the linear classifiers presented in appendix B. The use of other metrics in algorithms (for classification and/or feature extraction) can also be implemented using the above kernel maps.

We consider here feature extraction and restrict ourselves to the context of finding a possibly non-linear embedding using empirical kernel maps and of reducing the dimensionality of the original dataset. For this, we write the kernel matrix of the data as (see also [Chapelle and Schölkopf, 2002]):

$$K = V \Lambda V^T \tag{A.42}$$

where the columns of the orthogonal matrix V are the eigenvectors and Λ is a diagonal matrix of the eigenvalues of K. We then compute the feature map $F \in \mathbb{R}^{p \times p}$ as:

$$F = K^{-\frac{1}{2}} K = \Lambda^{-\frac{1}{2}} V^T K = \Lambda^{\frac{1}{2}} V^T \tag{A.43}$$

such that the kernel matrix can be expressed as:

$$K = F^T F \tag{A.44}$$

The columns of $F = \mathcal{Y}$ represent thus the data in a space of dimensionality p. The above algorithm is not strictly-speaking a dimensionality reduction method since it can only produce data in p dimensions; it is not possible to choose a dimension $k \neq p$ as for PCA, NMF, LLE or ICA.

In the application considered in this thesis, we shall only consider Radial Basis Function (RBF) kernels, also called Gaussian windows, of the type $K(\vec{x}, \vec{y}) = \exp\left(-(\frac{\|\vec{x}-\vec{y}\|}{c})^2\right)$. These kernels are normalized i.e. $K(\vec{x}, \vec{x}) = 1$ and they are biologically-plausible since they resemble to some extent

receptive fields and can be embedding in a neural network scheme, the so-called RBF networks (see [Haykin, 1999]). These kernels allow a multi-scale analysis through the choice of the width parameter c. For large values of c, we have the linear case i.e. the matrix $K_{ij} = K(\vec{x}_i, \vec{x}_j) \sim \delta_{ij}$ is the identity matrix. For small values of c all the patterns are orthogonal i.e $K_{ij} \sim 1$. Both cases are limit cases and we propose below a automated way to determine an "interesting" value of c. We define the range of the non-diagonal values of the matrix K by K_{min} and K_{max}. The parameter c is then chosen such that $0 \leq K_{min} \leq T(K) \leq K_{max} \leq 1$ for the whole dataset, where $T(K)$ is the upper triangular part of K, as below:

1. compute the distance matrix d for the whole dataset given by $d_{ij} = \|\vec{x}_i - \vec{x}_j\|$
2. compute the kernel matrix $K_{ij} = \exp\left(-(\frac{d_{ij}}{c})^2\right)$
3. minimize with respect to c in the sense of a least-square error the function $\|\vec{e}\|^2$ where $e(1) = min(K) - K_{min}$ and $e(2) = max(K) - K_{max}$ using as an initial guess for c the median of the non-diagonal values of d.

The above procedure can be seen as an automated way to determine the optimal RBF kernel parameter which suits best the data, i.e. allows to extract most information from the data at hand. In practice, some reasonable values are: $K_{min} = 0.3$ and $K_{max} = 0.8$. We thus avoid an almost diagonal kernel matrix or a matrix of ones. Once c is determined, the feature map F is computed and the encoding is known.

A linear empirical kernel map is identical to PCA if the data is centered. Indeed, for PCA we have:

$$DD^T = V\Lambda V^T \quad \text{and} \quad R = V^T D \tag{A.45}$$

with the same notations as in section A.2 and where R the matrix of encodings. We can deduce that $D = VR$ and thus $V\Lambda V^T = DD^T = VRR^TV^T$. This is only possible for $R = \Lambda^{\frac{1}{2}} H^T$ where H is orthogonal and we can then write:

$$D^T D = R^T R \quad \text{for} \quad R = \Lambda^{\frac{1}{2}} H^T \tag{A.46}$$

For the linear empirical kernel map we have, as shown above:

$$D^T D = W\Omega W^T = F^T F \quad \text{for} \quad F = \Omega^{\frac{1}{2}} W^T \tag{A.47}$$

Although the sizes of the matrices for PCA and the kernel map differ, they have same rank and the results are thus identical. The above comparison is similar to what was done for classical PCA versus the Snap Shot Method (linear KPCA).

A.7 Gabor Wavelet Filters

Gabor wavelet filters, or receptive fields, can be seen as one of the most widely-spread models in the neuroscience community for the description of the processing of visual information since the milestone work of [Hubel and Wiesel, 1962]. This type of preprocessing can be regarded as biologically-plausible or inspired and has found application in various studies in computer vision [Mel, 1997] and the design of sparse codes [Olshausen and Field, 1996]. Further, these type of filters have been used to model the response of receptive fields in the visual cortex when confronted with natural scenes [Willmore and Smyth, 2003]. We apply below such receptive fields as filters on the images of the original database. The convolution of such filters at various scales and orientations with the original images yields a high-dimensional highly sparse vector for each image of the database. The dimensionality is subsequently reduced using a linear empirical kernel map.

We define the Gabor wavelet filter following [Liu and Wechsler, 2003] as:

$$\lambda = \frac{\pi 2^{\nu+5}}{d} \quad \text{and} \quad \theta = \frac{\pi\mu}{M} \tag{A.48}$$

where d is the image size in pixels (i.e. its resolution), $\nu = 0, \ldots, N$ indicates the scales and $\mu = 0, \ldots, M$ the orientations. We then define the following:

$$k = \lambda \exp(\theta i) \quad \text{and} \quad \vec{k} = (\Re(k), \Im(k)) \tag{A.49}$$

where $\Re$ and $\Im$ represent respectively the real and the imaginary part. This yields the Gabor filter or receptive filed as:

$$G_{\mu\nu} = \left(\frac{\|\vec{k}\|}{\sigma} \right)^2 \exp\left(-\frac{\|\vec{k}\|^2 s}{2\sigma^2} \right) \left(\exp(i \left\langle \vec{k} | \vec{z} \right\rangle) - \exp(-\frac{\sigma^2}{2}) \right) \tag{A.50}$$

where $\sigma = 2\pi$ and the position vector $\vec{z}$ is centered on the image and the elliptic shape of the filter is defined as:

$$s = (R_{-\theta}\vec{z})^T \begin{pmatrix} 2 & 0 \\ 0 & 1 \end{pmatrix} R_{-\theta}\vec{z} \tag{A.51}$$

where the rotation matrix is defined as:

$$R_\theta = \begin{pmatrix} \cos(\theta) & -\sin(\theta) \\ \sin(\theta) & \cos(\theta) \end{pmatrix} \tag{A.52}$$

We define I and $G_{\mu\nu}$ the image and Gabor filter matrix respectively. We denote their Fourier transforms respectively by $\mathcal{F}(I)$ and $\mathcal{F}(G_{\mu\nu})$. The filtering process is the convolution of the image and the filter matrices, or equivalently the product of their Fourier transforms element by element (and not the matrix product):

$$\mathcal{F}(I_{G_{\mu\nu}}) = \mathcal{F}(I) \cdot \mathcal{F}(G_{\mu\nu}) \tag{A.53}$$

The above filtered image is then downsampled in order to remove unnecessary information. In this procedure, we only consider a centered square of the filtered image, the size of this square being given by the minimal spatial extension of the magnitude of the corresponding filter. We then obtain:

$$\hat{I_{G_{\mu\nu}}} = \mathcal{F}^{-1}\mathcal{F}(\hat{I_{G_{\mu\nu}}}) \tag{A.54}$$

The highly sparse feature vector $\vec{f}_I = \{\hat{I_{G_{\mu\nu}}}\}_{\mu\nu}$ corresponding to the image matrix I is then the concatenation of the filtered image vectors obtained for each scale and orientation. In order to exploit this huge vector for the purpose of machine learning, an additional preprocessing step is required. We choose here a linear empirical kernel map (see previous section) as follows:

$$Q^T Q = V \Lambda V^T \quad \rightarrow \quad F = \Lambda^{\frac{1}{2}} V^T \tag{A.55}$$

where $Q = (\vec{f}_I^1 | \ldots | \vec{f}_I^p)$ is a matrix whose columns are the feature vectors $\vec{f}_I$ and $F = \mathcal{Y}$ the matrix whose columns are the vectors corresponding to each image in a space of dimensionality p.

Appendix B

Hyperplane Classifiers

B.1 Overview

We assume being in the context of supervised machine learning and are given an empirical labeled dataset $\mathcal{D} = \{(\vec{x}_i, y_i)\}_{i=1}^{p}$ where $\vec{x}_i \in \mathbb{R}^n$ are the patterns and $y_i = \pm 1$ the target values. We present in this section four supervised dichotomic linear classification algorithms: the prototype learner (Prot), the K-means learner (Kmean), the Support Vector Machine (SVM) and the Relevance Vector Machine (RVM).

For all these algorithms, the decision function is modeled by a separating hyperplane (SH) defined by its normal vector $\vec{w}$ and offset b. The class of an unlabeled pattern $\vec{x}$ is then given by:

$$y(\vec{x}) = sign(f(\vec{x})) \text{ where } f(\vec{x}) = \langle \vec{w} | \vec{x} \rangle + b \tag{B.1}$$

where $f(\cdot)$ is the decision function. The goal of supervised machine learning is to determine the function f such that $y_i = sign(f(\vec{x}_i)) \quad \forall i = 1, \ldots, p$ while allowing a good generalization to new patterns i.e. avoiding overfitting. The generalization ability is mainly obtained using regularization theory which ensures smoothness and simplicity of the solution [Chen and Haykin, 2002].

Furthermore these algorithms can be expressed as classifiers in dual form. In other words, we can write:

$$\vec{w} = \sum_{i=1}^{p} \alpha_i \vec{x}_i \tag{B.2}$$

where $\vec{\alpha}$ is the vector of dual variables and the $\vec{w}$ the vector of primal variables. In other words, $\vec{w}$ is expressed using a linear combination of the patterns and is thus an element in the same vector space as the data—the primal space.

Finally all these classifiers in dual form proceed to classification using a a small set of patterns $\mathcal{R}$ termed *representations*: $\mathcal{R} = \{\vec{r}_i\}_{i=1}^{l}$ with $l < p$.

We then have two types of representations. For the SVM and the RVM, the representations are a subset of the patterns of $\mathcal{D}$: $\vec{r}_i \in \mathcal{D} \quad \forall i = 1, \ldots, l$. To compute the SH, we only need these elements and we can then write: $\vec{w} = \sum_{i|\vec{x}_i \in \mathcal{R}} \alpha_i \vec{x}_i$. For Prot and Kmean, the representations are not elements of $\mathcal{D}$: $\vec{r}_i \notin \mathcal{D} \quad \forall i = 1, \ldots, l$. They are computed using all patterns of $\mathcal{D}$ and we then have: $\vec{w} = \sum_{i=1}^{p} \alpha_i \vec{x}_i$. The number of representations is closely related to the issue of sparseness of the classifier in dual form.

B.2 Prototype Classifiers

One of the simplest and most basic pattern classification algorithms is the mean-of-class prototype learner. It is argued to be biologically-plausible according to findings in psychology [Reed, 1972, Rosch, Mervis, Gray, Johnson, and Boyes-Braem, 1976, Knapp and Anderson, 1984]. Prototype learning belongs to the class of distance-dependent winner-takes-all learning rules. An unlabeled example $\vec{x}$ is assigned to the class whose prototype is closer to it (one nearest-neighbor approach) i.e classification is done by computing the distance between the unlabeled pattern and the prototype of each class as shown in fig. B.1. We consider the Euclidean norm and the following

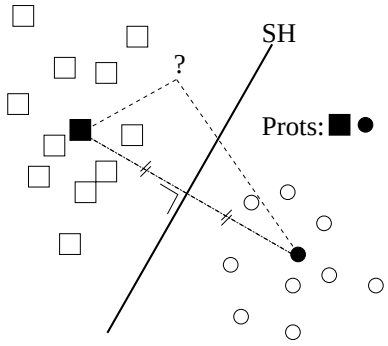


Figure B.1: Schematic view of prototype classification.

decision function can be written:

$$f(\vec{x}) = \frac{\|\vec{x} - \vec{p}_-\|^2 - \|\vec{x} - \vec{p}_+\|^2}{2} = \langle \vec{w} | \vec{x} \rangle + b \tag{B.3}$$

where the prototypes are defined as:

$$\vec{p}_\pm = \frac{\sum_{i=1}^{p} \vec{x}_i (y_i \pm 1)}{\sum_{i=1}^{p} (y_i \pm 1)} = \sum_{i|y_i = \pm 1} \alpha_i \vec{x}_i \tag{B.4}$$

where $\alpha_i = \frac{y_i + 1}{2\sharp\{y_i = 1\}} - \frac{y_i - 1}{2\sharp\{y_i = -1\}}$. The prototype is the center of mass of each class assuming homogeneous punctual mass distributions on each pattern.

Clearly, the prototypes stand for the representations. They can be expressed using a weighted sum of the patterns of the database. The parameters of the SH are then written as:

$$\vec{w} = \vec{p}_+ - \vec{p}_- = \sum_i \alpha_i y_i \vec{x}_i \text{ and } b = \frac{\|\vec{p}_-\|^2 - \|\vec{p}_+\|^2}{2} = -\frac{\langle \vec{w} | \sum_i \alpha_i \vec{x}_i \rangle}{2} \quad \text{(B.5)}$$

Prototype learning can be seen as a powerful "Ansatz" to elaborate novel algorithms which can easily be used and interfaced with (existing) experimental protocols. Moreover, prototype classifiers can be shown to be limit cases of other algorithms: setting $C \to 0$ in the SVM algorithm yields prototype classification and boosting a prototype learner yields a SVM [Graf, Bousquet, and Rätsch, 2004a].

B.3 Kmeans & Nearest-neighbor

We present here a novel algorithm combining the Kmeans clustering algorithm to a nearest-neighbor classifier. This algorithm is a piecewise linear extension of the classical mean-of-class prototype learner. Kmeans[Duda, Hart, and Stork, 2001] is used to compute more than one "prototype" per class (for a conceptually similar approach, see the chorus of prototypes by [Edelman, 1995]). Once the K means for each class are obtained, a new pattern is assigned to the class whose means is nearest. In other words, for a new pattern to be classified, the nearest means of each class are determined and a prototype-like decision rule is applied as shown in fig.B.2. The means

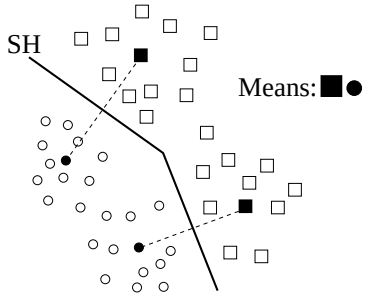


Figure B.2: Schematic view of Kmean clustering combined with Nearest Neighbor classification.

of each class $\{\vec{\mu}_j\}_{j=1}^K$ are determined as follow, assuming the data of this class is given by $\{\vec{z}_i\}_{i=1}^l$:

1. initialization of the means $\vec{\mu}_j$, using for instance the first K principal components of $\{\vec{z}_i\}_{i=1}^l$ i.e. the first K eigenvectors of the covariance matrix of the data

2. update $\vec{\mu}_j$ until convergence, for instance of the norm of the difference of the new and old means, as follows:
   (a) for every pattern $\vec{z}_i$, determine the nearest $\vec{\mu}_j$ according to $u_i = \arg\min_{j=1,\ldots,K} \|\vec{z}_i - \vec{\mu}_j\|$
   (b) recompute the new $\vec{\mu}_j$ as the mean over the patterns nearest to the old $\vec{\mu}_j$ according to $\vec{\mu}_j \leftarrow \langle \vec{z}_{i|u_i=j} \rangle$ where $\langle \cdot \rangle$ is denoting the mean

This algorithm computes the best embedding of the K means in the data and follows similar principles as Locally Linear Embedding [Roweis and Saul, 2000].

Learning of the discriminant function is then done using a nearest-neighbor rule. Assuming the means of each class $\{\vec{\mu}_j^{\pm}\}_{j=1}^{K}$ determined, the mean of each class nearest to the pattern $\vec{x}_i$ is computed as:

$$\vec{k}_{\pm}^{i} = \vec{\mu}_{v_i^{\pm}}^{\pm} \text{ where } v_i^{\pm} = \arg \min_{j=1,\ldots,K} \|\vec{x}_i - \vec{\mu}_j^{\pm}\| \tag{B.6}$$

Learning is done similarly to prototype learning i.e. these means are considered as the prototypes and a linear decision function is computed as follows:

$$f(\vec{x}_i) = \frac{\|\vec{x}_i - \vec{k}_{-}^{i}\|^2 - \|\vec{x}_i - \vec{k}_{+}^{i}\|^2}{2} = \langle \vec{w}_i | \vec{x}_i \rangle + b_i \tag{B.7}$$

where the parameters of th SH are defined as:

$$\vec{w}_i = \vec{k}_{+}^{i} - \vec{k}_{-}^{i} \text{ and } b_i = \frac{\|\vec{k}_{-}^{i}\|^2 - \|\vec{k}_{+}^{i}\|^2}{2} \tag{B.8}$$

where $i = 1, \ldots, K^2$ since there are only K means in each class. Globally, a piecewise linear decision function is obtained with a maximum of K^2 pieces. For each of these pieces, the dual space variable $\vec{\alpha}$ is to be computed using a matrix pseudo-inverse since the Kmeans algorithm does not have a closed-form solution. When considering one piece of the decision function, we have its normal vector $\vec{w}_k$ and an index vector of the elements of the dataset this hyperplane is separating: $\vec{u}_k = \{i | \vec{x}_i \text{ used to compute } \vec{w}_k\} \in \mathbb{R}^q$. We then define the matrix $X_k = (\vec{x}_{u_k^1} | \ldots | \vec{x}_{u_k^l}) \in \mathbb{R}^{n \times q}$ and we can express the normal vector using a weighted sum over the corresponding patterns as:

$$\vec{w}_k = X_k \vec{\alpha}_k \tag{B.9}$$

where $\vec{\alpha}_k \in \mathbb{R}^q$ is the vector of dual variables for this hyperplane. This vector of dual variables can be computed using the matrix pseudo-inverse of the previous expression as

$$\vec{\alpha}_k = (X_k^t X_k)^{-1} X_k^T \vec{w}_k \tag{B.10}$$

Finally the vector of dual variables $\vec{\alpha}$ for the whole set of hyperplanes is written as: $\vec{\alpha}|_{\vec{u}_k} = \vec{\alpha}_k$.

Notice that for $K = 1$ we get the classical mean-of-class prototype learning algorithm and for $K = p$, every pattern is the center of its own cluster and classification amounts then to a nearest-neighbor rule. The Means are the representations and their number K can either be user-defined or determined using cross-validation.

B.4 Support Vector Machines

Support Vector Machines (SVMs) [Vapnik, 2000, Cortes and Vapnik, 1995, Schölkopf and Smola, 2002] arose from statistical learning theory and hinge upon the idea of determining the optimal way to separate both classes i.e. to find a decision hyperplane which divides the dataset such that both classes are as far away as possible from this decision hyperplane. In other words, SVMs maximize the margin between classes (in order to avoid overfitting) while minimizing the number of misclassifications as shown in fig. B.3. It

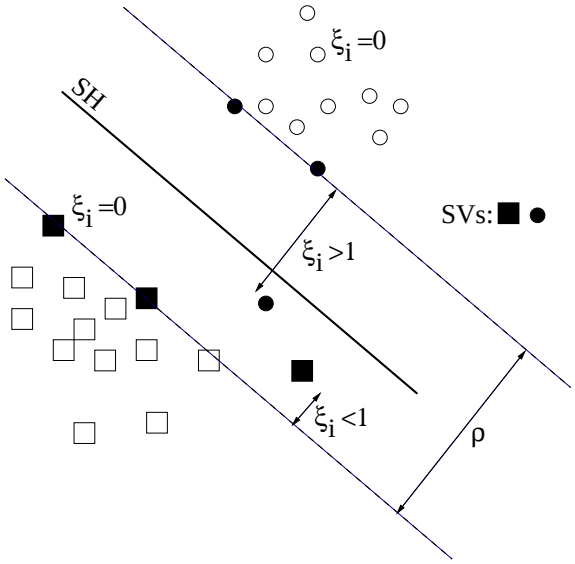


Figure B.3: Schematic view of SVM classification. The two margins are separated by a distance ρ.

can be shown that the distance between both classes is given by $\rho = \frac{2}{\|\vec{w}\|}$ for a canonical representation of the hyperplane i.e. a hyperplane such that $\{|\langle \vec{w}|\vec{x}\rangle + b|\}_{\vec{x}\in \text{margin}} = 1$. This yields a constrained quadratic optimization problem searching the $\vec{w}$, b and $\vec{\xi}$ minimizing the following $1-$Norm soft-

margin cost function:

$$\Phi(\vec{w},\vec{\xi}) = \frac{1}{2}\|\vec{w}\|^2 + C\sum_{i=1}^{p}\xi_i \tag{B.11}$$

such that $\vec{w}$, $\vec{\xi}$ and b satisfy the constraints for $i = 1,\ldots,p$:

$$\begin{cases} y_i(\langle\vec{w}|\vec{x}_i\rangle + b) \geq 1 - \xi_i \\ \xi_i \geq 0 \end{cases} \tag{B.12}$$

The slack variables ξ_i allow a class overlap (wrongly-classified patterns for $\xi_i > 1$ and correctly-classified patterns falling inside of the margin stripe for $0 < \xi_i \leq 1$). The regularization parameter $C > 0$ defined the trade-off parameter between the width of the margin and the number of misclassifications. This parameter is usually determined using cross-validation methods. The resolution of the above problem hinges upon the determination of the saddle points of the following Lagrangian:

$$\mathcal{L}(\vec{w},\vec{\xi},b,\vec{\alpha},\vec{\beta}) = \Phi(\vec{w},\vec{\xi}) - \sum_{i=1}^{p}\alpha_i(y_i(\langle\vec{w}|\vec{x}_i\rangle + b) - 1 + \xi_i) - \sum_{i=1}^{p}\beta_i\xi_i \tag{B.13}$$

where $\vec{\alpha} \geq 0$ and $\vec{\beta} \geq 0$ are the Lagrange variables corresponding to the two constraints of equ.B.12. The saddle points of this Lagrangian yield $\vec{w} = \sum_i \alpha_i y_i \vec{x}_i$ and the constraints of the corresponding dual problem which corresponds to the primal one of equ.B.11. The latter consists of the constrained maximization over $\vec{\alpha}$ of the expression:

$$W(\vec{\alpha}) = \sum_{i=1}^{p}\alpha_i - \frac{1}{2}\sum_{i,j=1}^{p}\alpha_i\alpha_j y_i y_j \langle\vec{x}_i|\vec{x}_j\rangle \tag{B.14}$$

subject to the constraints:

$$\begin{cases} 0 \leq \alpha_i \leq C \qquad i = 1,\ldots,p \\ \sum_{i=1}^{p}\alpha_i y_i = 0 \end{cases} \tag{B.15}$$

The vectors of the dataset corresponding to $\alpha_i \neq 0$ are termed *Support Vectors* (SVs) and correspond to the representations. The SVs can easily be interpreted geometrically. The SVs are elements of the dataset which lie on the margin ($\xi_i = 0$), in the margin stripe ($\xi_i < 2$) or outside of the margin side in the wrong class ($\xi_i > 2$). When $\xi_i = 0$, or equivalently $0 < \alpha_i < C$, the SVs are termed margin SVs since they lie on the margin according to one of the Karush-Kuhn-Tucker complementarity conditions: $\alpha_i\,(y_i(\langle\vec{w}|\vec{x}_i\rangle + b) - 1 + \xi_i) = 0 \quad i = 1,\ldots,p$. The vector $\vec{w}$ and the decision function revert the following form:

$$\vec{w} = \sum_{i\in SV}\alpha_i y_i \vec{x}_i \text{ and } f(\vec{x}) = \sum_{i\in SV} y_i\alpha_i\langle\vec{x}|\vec{x}_i\rangle + b \tag{B.16}$$

where $b = \langle y_i(1-\xi_i) - \sum_{j \in SV} y_j \alpha_j \langle \vec{x}_i | \vec{x}_j \rangle \rangle_{i \in SV}$ is computed as an average over all SVs of the Karush-Kuhn-Tucker complementarity condition mentioned above.

B.5 Relevance Vector Machines

For probabilistic classification, we consider Sparse Bayesian Learning [Tipping, 2000, 2001], and in particular the Relevance Vector Machine (RVM). The RVM is a particular case of Bayesian inference well adapted to our setting. The decision function allows in this case to compute the posterior probability $P(y|\vec{x})$ of membership to a class $y \in \{0, 1\}$ given the input $\vec{x}$:

$$P(y=1|\vec{x}) = s(\vec{x}) \text{ and } P(y=0|\vec{x}) = 1 - s(\vec{x}) \tag{B.17}$$

where we assume the following logistic regression "Ansatz" and decision function:

$$s(\vec{x}) = \frac{1}{1+\exp(-f(\vec{x}))} \text{ and } f(\vec{x}) = \langle \vec{w} | \vec{x} \rangle = \sum_{i=0}^{p} \alpha_i \langle \vec{x}_i | \vec{x} \rangle \tag{B.18}$$

The offset is here included in $\vec{w} = \sum_{i=0}^{p} \alpha_i \vec{x}_i$ using the convention: $w_0 = b$ and $x_{i0} = 1 \quad \forall i = 1, \ldots, p$. The possibly sparse dual space variable $\vec{\alpha}$ (allowing to compute both the normal vector and the offset of the SH) is then to be determined in the learning process. Using the patterns of the dataset, we can write:

$$f(\vec{x}_i) = \sum_{j=0}^{p} \langle \vec{x}_i | \vec{x}_j \rangle \alpha_j = [\Phi \vec{\alpha}]_i \text{ and } s_i = s(\vec{x}_i) = \frac{1}{1+\exp(-[\Phi \vec{\alpha}]_i)} \tag{B.19}$$

where $\Phi_{ij} = [\vec{1} | \langle \vec{x}_i | \vec{x}_j \rangle]$ is the "extended" Gram matrix of the patterns in the dataset. The two classes of the classification task define two possible "states" which can be modeled by a Bernoulli distribution:

$$p(\vec{y}|X, \vec{\alpha}) = \prod_{i=1}^{p} s_i^{y_i} [1-s_i]^{1-y_i} \tag{B.20}$$

where $X = \{\vec{x}_i\}_{i=1}^{p}$ makes the dependency on the data patterns explicit. Gaussian hyperparameters $\vec{\beta}$ are introduced to ensure *sparseness* and *smoothness* of the dual space variable $\vec{\alpha}$:

$$p(\vec{\alpha}|\vec{\beta}) = \prod_{i=1}^{p} \mathcal{N}(\alpha_i | 0, \beta_i^{-1}) \tag{B.21}$$

Learning of $\vec{\alpha}$ then amounts to maximize with respect of $\vec{\beta}$ the probability of get the targets $\vec{y}$ given the patterns X according to:

$$p(\vec{y}|X,\vec{\beta}) = \int p(\vec{y}|X,\vec{\alpha})p(\vec{\alpha}|\vec{\beta})d\vec{\alpha} \tag{B.22}$$

It is not possible to perform this integration analytically and the Laplace approximation can then be used. The latter approximates the integrand locally using a Gaussian around its most probable mode $\vec{\alpha}^{MP}$ according to:

$$p(\vec{y}|X,\vec{\alpha})p(\vec{\alpha}|\vec{\beta}) \simeq \mathcal{N}(\vec{\alpha}|\vec{\alpha}^{MP},\Sigma) \tag{B.23}$$

The log posterior of the integrand of equ.B.22 can be written as using equ.B.20 and B.21:

$$\log\left(p(\vec{y}|X,\vec{\alpha})p(\vec{\alpha}|\vec{\beta}\right) = \sum_{i=1}^{p}[y_i \log s_i + (1-y_i)\log(1-s_i)] - \frac{1}{2}\vec{\alpha}^T B\vec{\alpha} \tag{B.24}$$

where $B = diag(\vec{\beta})$. We can then write an iterative update scheme for $\vec{\beta}$ which maximizes equ.B.22:

1. Initialize $\vec{\beta}$.

2. The vector $\vec{\alpha}^{MP}$ is the value of $\vec{\alpha}$ maximizing equ.B.24 for a fixed $\vec{\beta}$ i.e. the most probable value of $\vec{\alpha}$ given $\vec{\beta}$. We can thus write:

$$\vec{\nabla}_{\vec{\alpha}} \log\left(p(\vec{y}|X,\vec{\alpha})p(\vec{\alpha}|\vec{\beta})\right)|_{\vec{\alpha}^{MP}} = 0 \tag{B.25}$$

   The resolution of the above using equ.B.24 and B.19 yields:

$$B\vec{\alpha}^{MP} = \Phi^T(\vec{y} - \vec{s}(\vec{\alpha}^{MP})) \tag{B.26}$$

   where the dependency of $\vec{s}$ on $\vec{\alpha}$ has been made explicit. The resolution of the above equation, using for example an iterative scheme, then yields $\vec{\alpha}^{MP}$.

3. The variance matrix $\Sigma = -\left(\vec{\nabla}_{\vec{\alpha}}\vec{\nabla}_{\vec{\alpha}} \log \mathcal{N}(\vec{\alpha}|\vec{\alpha}^{MP},\Sigma)\right)^{-1}$ is the variance matrix of the Gaussian approximation of the integrand. We can thus write:

$$\vec{\nabla}_{\vec{\alpha}}\vec{\nabla}_{\vec{\alpha}} \log\left(p(\vec{y}|X,\vec{\alpha})p(\vec{\alpha}|\vec{\beta})\right)|_{\vec{\alpha}^{MP}} = -\Sigma^{-1} \tag{B.27}$$

   which yields using equ.B.24 and B.19:

$$\Sigma = (\Phi^T C\Phi + B)^{-1} \tag{B.28}$$

   where $C = diag(\vec{\gamma})$ and $\gamma_i = s_i(1-s_i)$.

4. The hyperparameter update is then computed as:

$$\beta_i \leftarrow \frac{1 - \beta_i \Sigma_{ii}}{(\alpha_i^{MP})^2}$$

5. Go back to 2. until convergence is reached.

In the update of $\vec{\beta}$, some $\beta_i \to \infty$, implying an infinite peak of $p(\alpha_i|\beta_i))$ around 0. This is then equivalent to setting $\alpha_i = 0$. This feature of the RVM ensures sparseness and defines the *Relevance Vectors* (RVs): $\beta_i < \infty \Leftrightarrow \vec{x}_i \in RV$. The latter are the representations of the RVM.

Since RVMs are based on probabilistic concepts, they do not allow an intuitive geometric explanation as for SVMs for example. Especially, there is no easy way to explain or interpret the RVs. It may just be mentioned that the difficulty of the classification task defines the sharpness of the decision function, this sharpness being proportional to the distance of the RVs to the SH. In easy classification tasks such as in linearly separable datasets, the RVs are far apart. On the other hand, difficult classification tasks such as linearly not separable datasets yield RVs which are close to the SH.

B.6 Comparison of classifiers

The SHs and the representations of the above classifiers are compared in fig. B.4 on a two-dimensional toy dataset. On this dataset, the classification performance is perfect, i.e. there are no misclassifications in any algorithm. The piecewise linear SH of the Kmean algorithm can clearly be seen, as much as the corresponding contour levels of the decision function. The latter also bring to light the regions of each class as selected by the Kmeans clustering algorithm. RVMs and Prots have a low number of representations, whereas SVMs have a high number. The representations of the SVM algorithm are closest to the SH, whereas for the Prot and Kmean algorithms they lie in the middle of the classes. For the RVM, the representations are spread throughout the dataset i.e. they are close of the SH, far of it and in the middle of the dataset. In other words, SVMs deal with the patterns difficult to classify (the SVs) since the decision function has a sum only over the SVs ($\alpha_i \neq 0$), the remaining patterns of the dataset being unused ($\alpha_i = 0$). Prot and Kmean learning consider central elements of each class i.e. very typical patterns. Finally RVM consider elements in a large spectrum of distances from the SH.

Fig.B.5 compares schematically prototype to SV learning and clearly shows the essential difference between both algorithms. SVMs find a SH that separates best the data by taking into account the "geometry" of both classes. On the other hand, prototype classifiers ignore all geometrical information by solely considering the mean of each class for classification purpose.

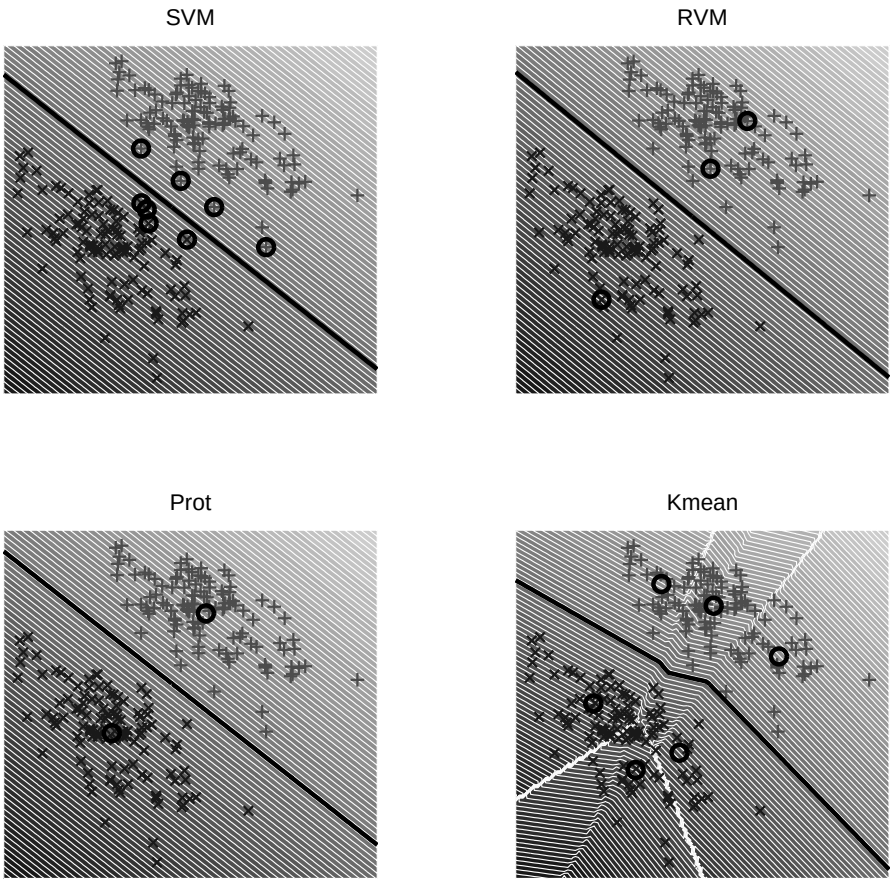


Figure B.4: Comparison of classifiers. The thick line represents the SH, the thick circles the representations and the thin lines the contours of the function $f(\vec{x})$.

Some advantages and limitations of the various algorithms are summarized in the tabular below.

	Prot	Kmean	SVM	RVM
classification perf.	poor	medium	excellent	good
representations	Prots $\notin \mathcal{D}$	Means $\notin \mathcal{D}$	SVs $\in \mathcal{D}$	RVs $\in \mathcal{D}$
$\sharp$(representations)	high	user-defined	med.$\rightarrow$ low	high
user-defined param.	none	K	C	none
multi-class	yes	yes	no	yes
probabilistic output	no	no	no	yes

In the case of RVMs, considering a multi-class problem amounts to consider a multinomial instead of the Bernoulli distribution of equ.B.20. In the case of the Prot, respectively Kmean, algorithms, the multi-class case is accom-

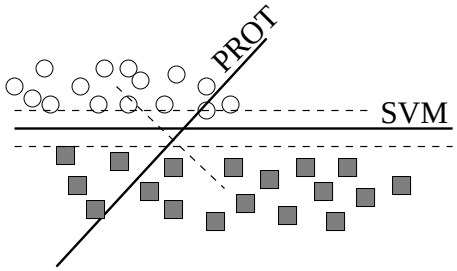


Figure B.5: Schematic difference between prototype and SV classification.

modated by placing one, respectively K, prototypes in each class and using the nearest neighbor algorithm for classification. SVMs, on the other hand, are intrinsically binary classifiers, although some studies related to their extension to multiple cases exist [Platt, Cristianini, and Shawe-Taylor, 2000, Weston and Watkins, 1999, Kressel, 1999]. RVMs are the only algorithm considered here that by construction yield a probabilistic output. Although this is not very well understood in the case of the other three algorithms, the decision function can be put into a logistic regression as $s(\vec{x}) = \frac{1}{1+\exp(-f(\vec{x}))}$ in order to yield a probabilistic output (see [Vapnik, 2000] in the case of SVMs).

The data fed into the SVM needs some "preprocessing", especially for high-dimensional data. One of the most fundamental type of preprocessing is *normalization* of the data i.e. its projection on a unit hypersphere. Although in this process one degree of freedom in the data is lost, it can be shown experimentally that it is most advantage, see [Graf and Borer, 2001, Graf, Smola, and Borer, 2003] in the case of SVMs. Moreover, loosing one degree of freedom can usually be neglected when considering high-dimensional data i.e. $n \gg 1$. Notice that in the case of SVMs and RVMs, normalizing the input patterns creates representations (SVs and RVs) that are also normalized. This is not the case for the Prot and Kmeans algorithms, i.e. neither the Prots nor the Means lie on the unit hypersphere even though the patterns lie on it.

B.7 Nonlinear Extension

The four learning algorithms described above are considered in the input space. It is possible to extend these algorithms to accommodate for nonlinear decision functions using the kernel trick. For this, the elements from the input space $\mathbb{R}^n$ are nonlinearly mapped into a high dimensional feature space $\mathcal{F}$ as:

$$\vec{\varphi} : \mathbb{R}^n \to \mathcal{F} \tag{B.29}$$

Classification in $\mathcal{F}$ is then "easier" according to Cover's theorem i.e a linearly not separable problem may become linearly separable i.e. the data can be separated by a hyperplane. The essential step is to replace the scalar product in $\mathcal{F}$ by a kernel function, yielding the corresponding substitutions:

$$\vec{x} \rightarrow \vec{\varphi}(\vec{x}) \text{ and } \langle \vec{x}|\vec{y}\rangle \rightarrow \langle \vec{\varphi}(\vec{x})|\vec{\varphi}(\vec{y})\rangle = K(\vec{x},\vec{y}) \tag{B.30}$$

The mapping $\vec{\varphi}$ is mostly unknown whereas the kernel function K is known. If an algorithm accepts a dual form, the decision function makes only use of scalar products between patterns and thus only the function $K(\vec{x},\vec{y})$ is needed explicitly.

In the case of SVM and RVM, the extension to a non-linear feature space is immediate by substitution of the scalar product with a kernel function in equ.B.14 and equ.B.16 for the SVM and in equ.B.18 and equ.B.19 for the RVM. However, this procedure is accompanied by a loss in the interpretability of the SH since the weight vector exists then only in the (unknown) feature space.

In the case of the Prot and Kmean algorithms, considering a feature space makes the following question arise: are the representations computed in the input space or in the feature space? Assume a representation can be written as: $\vec{r}_{\pm} = \sum_i \zeta_i^{\pm} \vec{x}_i$ where $\zeta_i^{\pm}$ can be computed in closed form for Prot but may require a matrix pseudo-inverse computation in the case of Kmean. We then have the two following possibilities, where classification is performed in the feature space:

- Representation $\vec{r}_{\pm}$ in input space:

  $$\|\vec{\varphi}(\vec{x}) - \vec{\varphi}(\vec{r}_{\pm})\|^2 = K(\vec{x},\vec{x}) - 2K(\vec{x},\vec{r}_{\pm}) + K(\vec{r}_{\pm},\vec{r}_{\pm}) \tag{B.31}$$

  implying a decision function as follows:

  $$f(\vec{x}) = K(\vec{x},\vec{r}_+) - K(\vec{x},\vec{r}_-) + \frac{K(\vec{r}_-,\vec{r}_-) - K(\vec{r}_+,\vec{r}_+)}{2} \tag{B.32}$$

- Representation $\sum_i \zeta_i^{\pm} \vec{\varphi}(\vec{x}_i)$ in feature space:

  $$\|\vec{\varphi}(\vec{x}) - \sum_i \zeta_i^{\pm} \vec{\varphi}(\vec{x}_i)\|^2 = K(\vec{x},\vec{x}) - 2\sum_i \zeta_i^{\pm} K(\vec{x},\vec{x}_i) + \sum_{i,j} \zeta_i^{\pm}\zeta_j^{\pm} K(\vec{x}_i,\vec{x}_j) \tag{B.33}$$

  with the following decision function:

  $$f(\vec{x}) = \sum_i (\zeta_i^+ - \zeta_i^-) K(\vec{x},\vec{x}_i) + \frac{\sum_{i,j}(\zeta_i^-\zeta_j^- - \zeta_i^+\zeta_j^+) K(\vec{x}_i,\vec{x}_j)}{2} \tag{B.34}$$

  Notice that the representations of Kmean cannot be computed as above since the latter involves averages of vectors in the unknown feature space to determine $\zeta_i^{\pm}$. A prototype learner in feature space is often referred to as a *Parzen Window Estimator* [Schölkopf and Smola, 2002].

In conclusion, using a feature space destroys the interpretability of the parameter $\vec{w}$ of the SH, unless the mapping $\vec{\varphi}$ is known. Moreover, it also makes unclear how to proceed for classification in the case of Prot and Kmean. On the other hand, the main advantage of using a feature space is a better classification performance of the algorithms. Moreover, normalizing the feature space is equivalent to normalizing the Kernel function. This is most advantageous as far as the classification performance is concerned and a modification of the SVM algorithm may also be introduced as shown [Graf, Smola, and Borer, 2003].

Appendix C

Elements from Signal Detection Theory

C.1 Detection of a Signal in Noise

We place ourselves in the context of an experiment where an observer has to decide whether a signal is present (*signal* trial) or absent (*noise* trial). The analysis of the classification performance of the observer is based on signal detection theory [Wickens, 2002]. For this, we assume that both the signal and the noise trials can be assumed to be drawn from a Gaussian distribution, with same unit variance and different means as follows:

$$X_n \sim f_n(x) = \mathcal{N}(x|0,1) \quad \text{and} \quad X_s \sim f_s(x) = \mathcal{N}(x|d',1) \tag{C.1}$$

Fig.C.1 gives a schematic representation and the main parameters. The

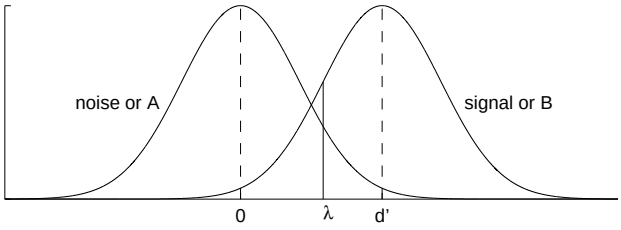


Figure C.1: Noise and signal, respectively stimulus A and B, univariate normal distributions of means 0 and d'. The parameter λ is the decision threshold.

fact that one of the signals is centered at 0 just represents a shift in the x-axis. The fact that both distributions have the same variance is a stronger hypothesis and has to be verified experimentally. Choosing a unit variance then just amounts to rescale the values of x. We want to determine the

position of the signal distribution i.e. d' as much as the decision threshold (or criterion) λ defined as:

$$\begin{cases} x > \lambda \rightarrow \text{answer yes (signal)} \\ x < \lambda \rightarrow \text{answer no (noise)} \end{cases} \tag{C.2}$$

The answers of the observer fall into four categories as shown below:

	no	yes
noise	♯(correct rejection)	♯(false alarm)
signal	♯(miss)	♯(hit)

The values filling the above tabular are given by the observer's responses. The latter are used to define the following rates:

- the *hit* rate: $h = \frac{\sharp(\text{hit})}{\sharp(\text{signal trial})} = \frac{\sharp(\text{hit})}{\sharp(\text{miss})+\sharp(\text{hit})}$
- the *false alarm* rate: $f = \frac{\sharp(\text{false alarm})}{\sharp(\text{noise trial})} = \frac{\sharp(\text{false alarm})}{\sharp(\text{correct rejection})+\sharp(\text{false alarm})}$
- the *miss* rate: $m = 1 - h$ and the *correct rejection* rate: $c = 1 - f$

Since m and c bear no additional information, we shall only work with h and f. The probabilities of making a hit or a false alarm can be computed as follows:

$$\begin{aligned} P_F &= P(yes|noise) = \textstyle\int_\lambda^\infty f_n(x)dx = \Phi(-\lambda) \\ P_H &= P(yes|signal) = \textstyle\int_\lambda^\infty f_s(x)dx = \Phi(d' - \lambda) \end{aligned} \tag{C.3}$$

where Φ is the cumulative normal distribution with zero mean and unit variance. Under the approximation that the samples from the experiment at hand are statistically meaningful ($P_F \simeq f$ and $P_H \simeq h$), we can compute:

$$d' = Z(h) - Z(f) \quad \text{and} \quad \lambda = -Z(f) \tag{C.4}$$

where $Z = \Phi^{-1}$ is the inverse of the cumulative normal distribution with zero mean and unit variance. The above quantities allow to compute the observer's bias as:

$$\log\beta = \log\left(\frac{f_s(\lambda)}{f_n(\lambda)}\right) = \frac{1}{2}\left(Z^2(f) - Z^2(h)\right) \tag{C.5}$$

with:

$$\begin{cases} \log\beta < 0 \rightarrow \text{signal (yes) bias} \\ \log\beta = 0 \rightarrow \text{no bias}, \lambda = \frac{d'}{2} \\ \log\beta > 0 \rightarrow \text{noise (no) bias} \end{cases} \tag{C.6}$$

C.2 Two-alternative Forced-choice Model

We consider here an experiment where an observer has to decide whether a stimulus belongs to class A or class B. Both stimuli are assumed to follow Gaussian univariate distributions with different means as below:

$$X_A \sim \mathcal{N}(x|0,1) \quad \text{and} \quad X_B \sim \mathcal{N}(x|d',1) \tag{C.7}$$

with the same parameters as shown in fig.C.1. We compute d' and the decision criterion λ defined here as:

$$\begin{cases} x > \lambda \rightarrow \text{answer class B} \\ x < \lambda \rightarrow \text{answer class A} \end{cases} \tag{C.8}$$

The following tabular summarizes the different cases:

	$\hat{y} = A$	$\hat{y} = B$
$y = A$	$\sharp$(correct A)	$\sharp$(false A)
$y = B$	$\sharp$(false B)	$\sharp$(correct B)

where y the true class of the stimulus and $\hat{y}$ is the estimated one. We define the correct classification rates for both stimuli as:

- the *correct A* rate: $p_A = \frac{\sharp(\text{correct A})}{\sharp(\text{A trial})} = \frac{\sharp(\text{correct A})}{\sharp(\text{correct A})+\sharp(\text{false A})}$
- the *correct B* rate: $p_B = \frac{\sharp(\text{correct B})}{\sharp(\text{B trial})} = \frac{\sharp(\text{correct B})}{\sharp(\text{false B})+\sharp(\text{correct B})}$

The probability to make a correct classification of A or B is computed as:

$$\begin{aligned} P_A = P(\hat{y} = A|y = A) = \int_{-\infty}^{\lambda} f_A(x)dx = \Phi(\lambda) \\ P_B = P(\hat{y} = B|y = B) = \int_{\lambda}^{\infty} f_B(x)dx = \Phi(d' - \lambda) \end{aligned} \tag{C.9}$$

Under the assumption that $P_A \simeq p_A$ and $P_B \simeq p_B$ (the distributions describe well the samples), the above can be solved, yielding:

$$d' = Z(p_A) + Z(p_B) \quad \text{and} \quad \lambda = Z(p_A) \tag{C.10}$$

The observer's bias to say B can then be written as:

$$\log \beta = \log \left(\frac{f_A(\lambda)}{f_B(\lambda)} \right) = \frac{1}{2} \left(Z^2(p_B) - Z^2(p_A) \right) \tag{C.11}$$

where:

$$\begin{cases} \log \beta > 0 \rightarrow \text{B bias} \\ \log \beta = 0 \rightarrow \text{no bias} \\ \log \beta < 0 \rightarrow \text{A bias} \end{cases} \tag{C.12}$$

Appendix D

Experimental Setup

The psychophysical setup is located in a black chamber. For the presentation of the stimuli, a Cambridge Research Systems VSG 2/5 framebuffer is connected to a fast-phosphor monochrome display by Clinton Monoray. Linearization of the display is performed using a VSG OptiCAL photometer. The responses of the subject are gathered on a four-button VSG response box and on a numerical keypad. The experimenter has his own monitor to program and supervise the experiment. More details can be found below.

The psychophysical laboratory is a black chamber with the subject's side separated from the experimenter's side by a curtain. On the experimenter's side, a keyboard and monitor allow, among others, to program, start, monitor and stop the experiment. The experiment is run on the monitor on the observer's side. The observer's chin is placed on a headrest at $1.48m$ viewing distance from the monitor showing the stimuli. The viewing window is of size $340 \times 250mm$ and is centered at $1.18m$ off the ground.

The PC computer and its peripheral devices are as follows:

- Motherboard: ASUS P4T-E Motherboard with a $1.8GHz$ Pentium IV, running WindowsXP Professional

- Framebuffer: Cambridge Research Systems (`http://www.crsltd.com/`) VSG 2/5 with a 32 bit PCI bus interface to the PC. The VSG card has an embedded 32 bit microprocessor running at $50MHz$ for on-board look-up table animation and pixel operations with 32 MB VRAM and 8 MB DRAM for storing programs, look-up tables and off-screen images. Moreover, it has an extensive external interface capability to support other experimental equipment, a hardware reaction timing and a sophisticated video output circuitry.

- Photometer: Cambridge Research Systems OptiCAL photometer with a $44mm^2$ silicon sensor, 13 degree fixed field-of-view. The full-scale range is $2400\frac{cd}{m^2}$ with a resolution of $0.1\frac{cd}{m^2}$

- Response boxes: Cambridge Research Systems CT3 Response Box (four buttons) and IBM Numerical Keypad

The monitor is a Multisync Clinton Monoray CRT display from Clinton Electronics (see `http://www.clintonelectronics.com/`), which is a modified version of a Richardson Electronics MR2000HB-MED. It has a 20 inch monitor using the phosphor type DP104 and has a visible area of $363 \times 272mm$. The frame rate (vertical frequency) is set to $150Hz$ and the scan rate (horizontal frequency) to $105kHz$. The resolution is 848×636 pixels at $150Hz$ and the luminence varies between 0.2 and $246\frac{cd}{m^2}$.

All of the programs are written in MATLAB, except for the morphing algorithms of the face database and the SVM code which are written in C, the former being compiled under Linux and the latter under Windows. The morphing code, generously provided by Dr. V. Blanz [Blanz and Vetter, 1999], is executed remotely on a Linux PC. The SVM code (LIBSVM version 2.36 by [Chang and Lin, 2001]) is locally compiled under Windows. All the MATLAB code is home-made except for the RVM implementation (SparseBayes version 1.0 by [Tipping, 2002]), the code for ICA (FastICA by [Hyvärinen and Oja, 1997]) and the `psignifit` toolbox by [Wichmann and Hill, 2001b].

Appendix E

Visualization of Parameters of Preprocessors

E.1 Overview

We present in this appendix various visualizations and reconstruction errors corresponding to some of the preprocessors used in this thesis as below:

- **PCA**
  The eigenvalue spectrum is first computed (value of the eigenvalue corresponding to a given PC) as much as the cumulative variance (rescaled to $[0, 1]$), the latter being the sum of all eigenvalues before the eigenvalue corresponding to a given PC. We then plot the mean of the reconstruction error over all elements of the database as function of the number of PCs considered to reconstruct the data. The reconstruction error is defined as the norm of the difference between the vectors corresponding to the original and the reconstructed face: $\|\vec{x}_{original} - \vec{x}_{reconstructed}\|$. We then display 5 original and reconstructed faces to assess visually the quality of the reconstruction. Finally, the 20 first elements from the basis of faces are shown.

- **ICA I, ICA II & NMF**
  The reconstruction error $\|\vec{x}_{original} - \vec{x}_{reconstructed}\|$ for each face in the database is first computed as much as its mean value over the whole database. We then display 5 original and reconstructed faces to assess visually the quality of the reconstruction. Finally, 20 elements from the basis of faces are shown.

In order to be able to plot the basis images, the following is done to reconstruct the faces from:

- the texture data: application of the morphing software on the reconstructed texture data and on the mean of the shape data over the whole dataset

- the shape data: application of the morphing software on the reconstructed shape data and on the mean of the texture data over the whole dataset
- the texture and shape information: direct application of the face morphing software on the texture and on the shape data

In other words, the reconstructed faces for the texture data have all same shape and those for the shape data have all same texture.

E.2 PCA—Image Data

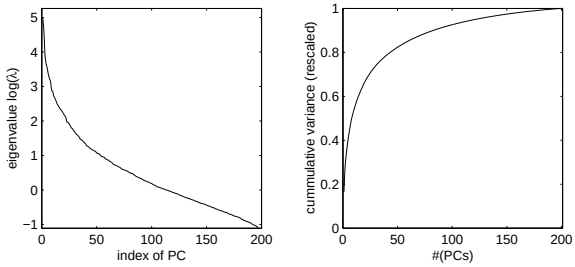


Figure E.1: Eigenvalue spectrum (left) and rescaled cumulative variance (right).

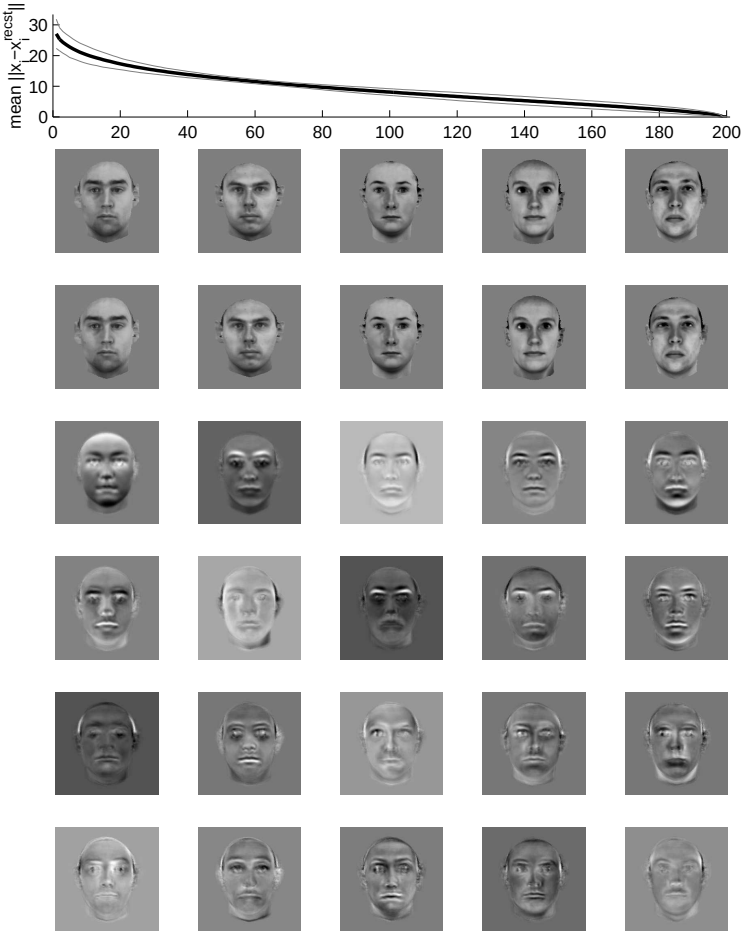


Figure E.2: Reconstruction error (first row), original and reconstructed heads (second and third rows) and the first 20 basis vectors (fourth to seventh row).

E.3 PCA—Texture Data

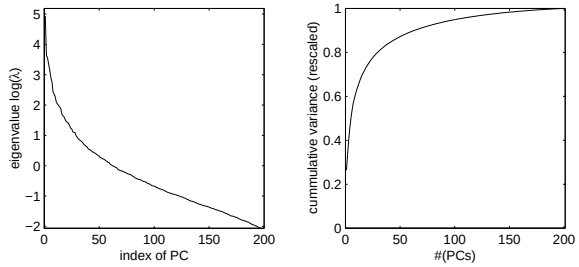


Figure E.3: Eigenvalue spectrum (left) and rescaled cumulative variance (right).

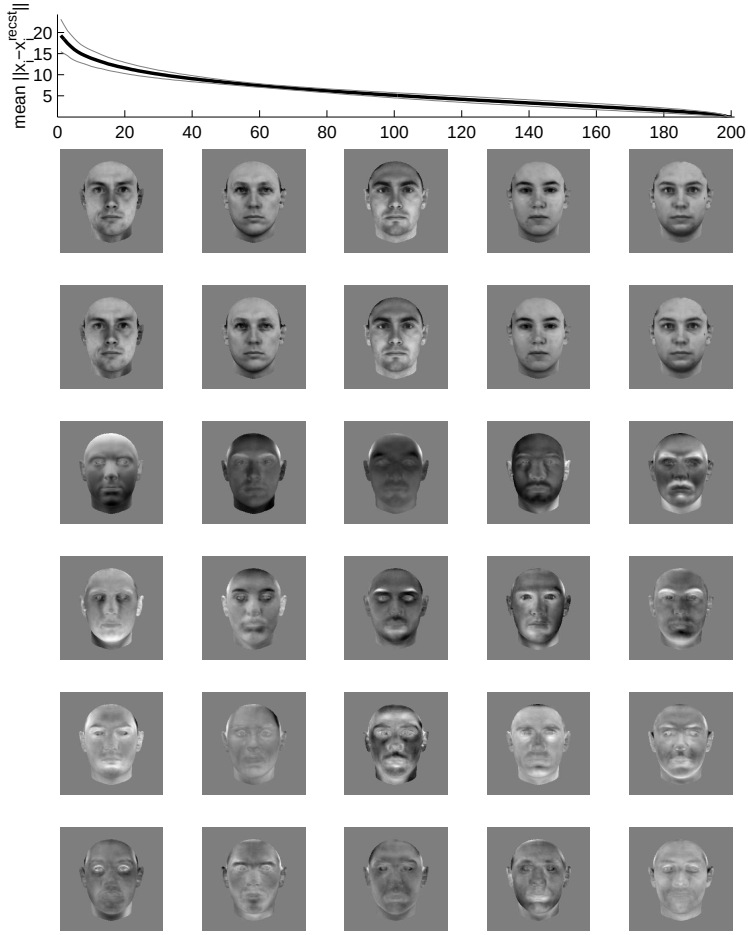


Figure E.4: Reconstruction error (first row), original and reconstructed heads (second and third rows) and the first 20 basis vectors (fourth to seventh row).

E.4 PCA—Shape Data

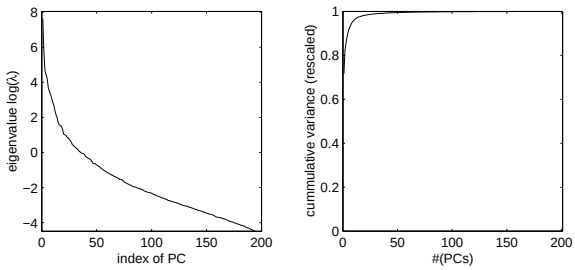


Figure E.5: Eigenvalue spectrum (left) and rescaled cumulative variance (right).

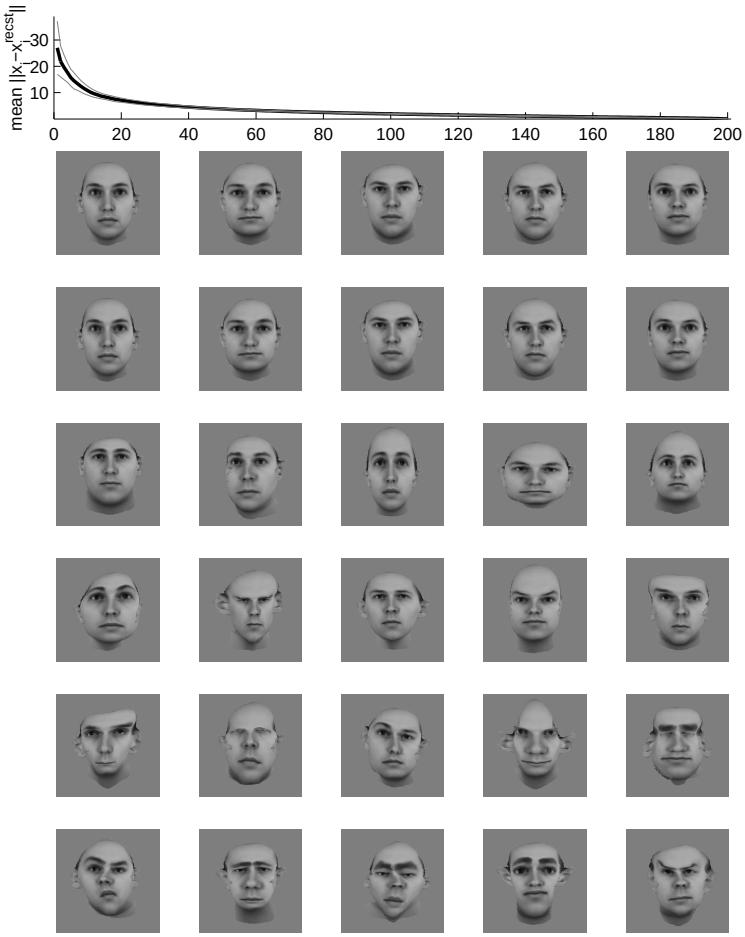


Figure E.6: Reconstruction error (first row), original and reconstructed heads (second and third rows) and the first 20 basis vectors (fourth to seventh row).

E.5 PCA—Texture & Shape Data

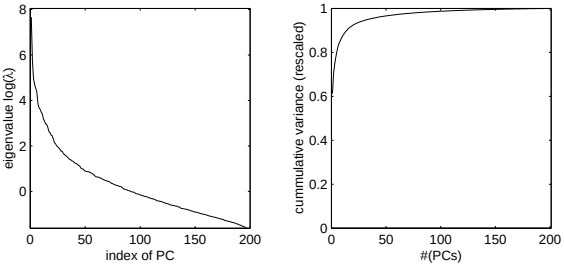


Figure E.7: Eigenvalue spectrum (left) and rescaled cumulative variance (right).

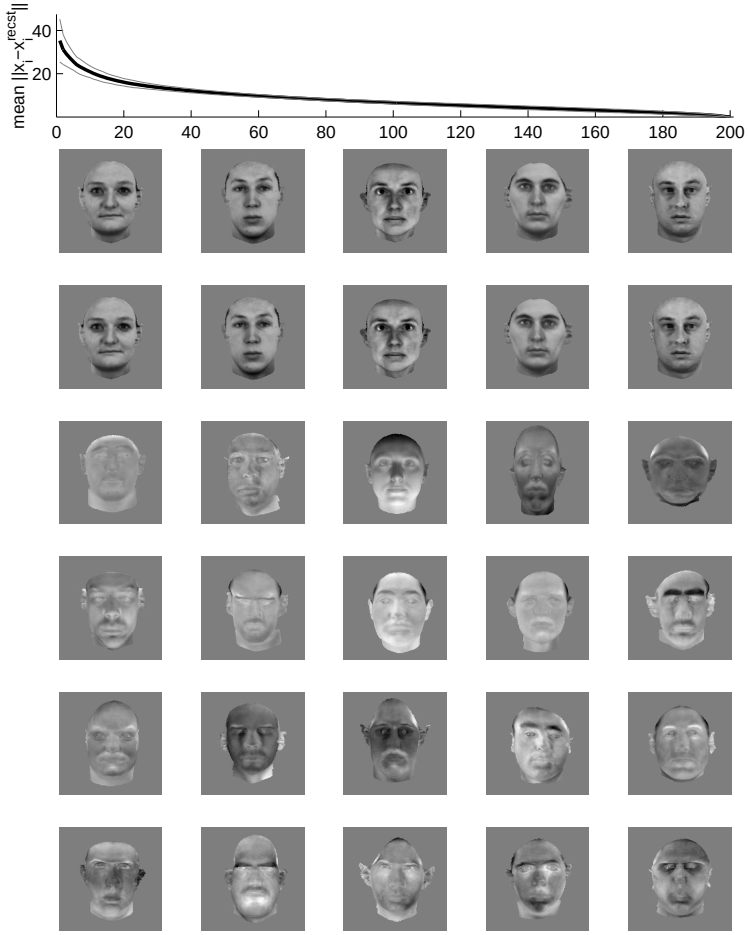


Figure E.8: Reconstruction error (first row), original and reconstructed heads (second and third rows) and the first 20 basis vectors (fourth to seventh row).

E.6 ICA I—Image Data

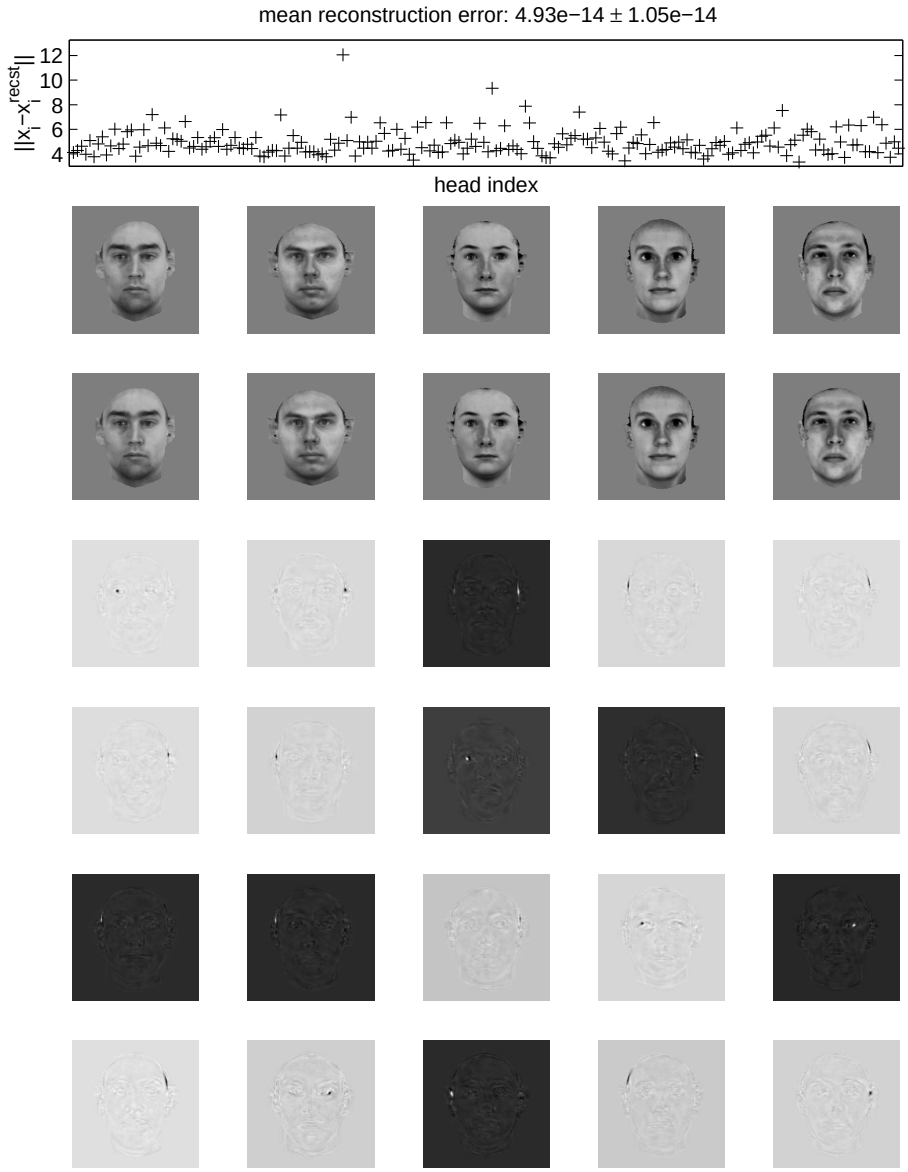


Figure E.9: Reconstruction error (first row), original and reconstructed heads (second and third rows) and the first 20 basis vectors (fourth to seventh row).

E.7 ICA I—Texture Data

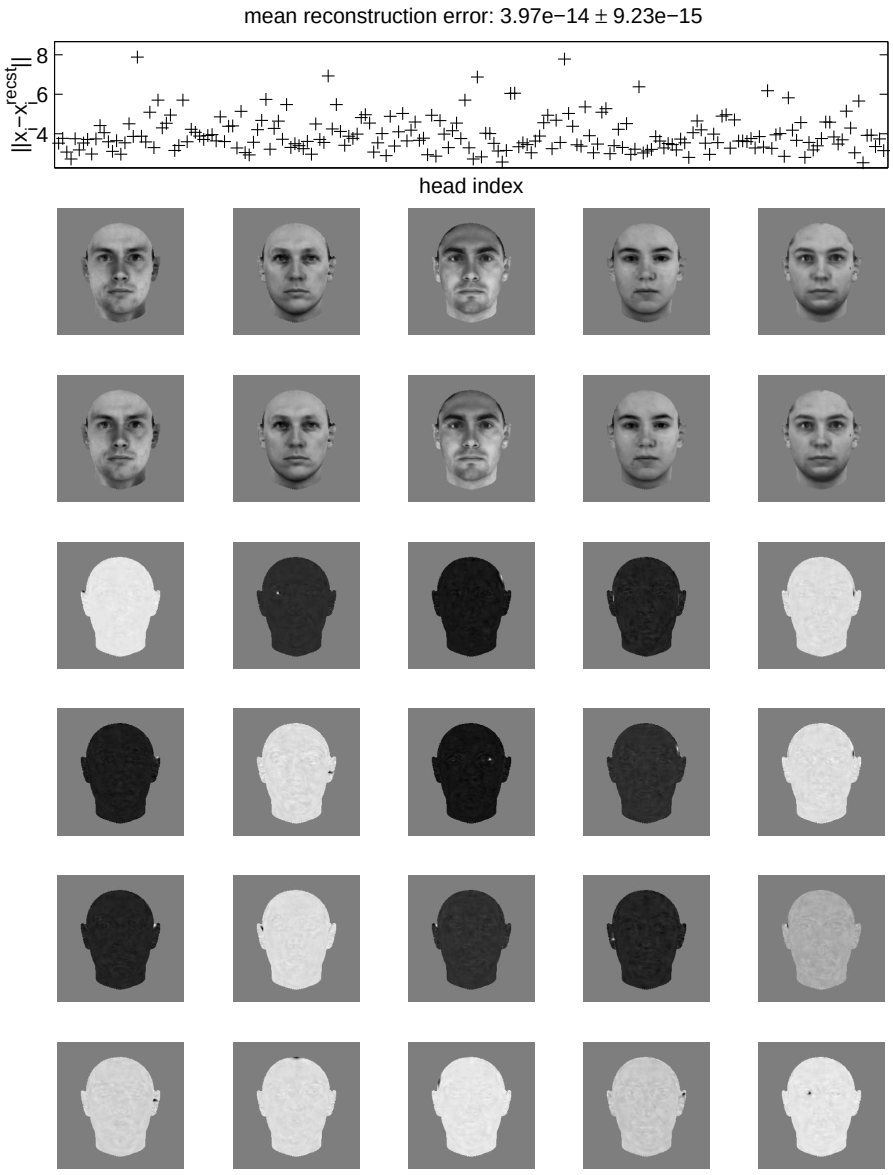


Figure E.10: Reconstruction error (first row), original and reconstructed heads (second and third rows) and the first 20 basis vectors (fourth to seventh row).

E.8 ICA I—Shape Data

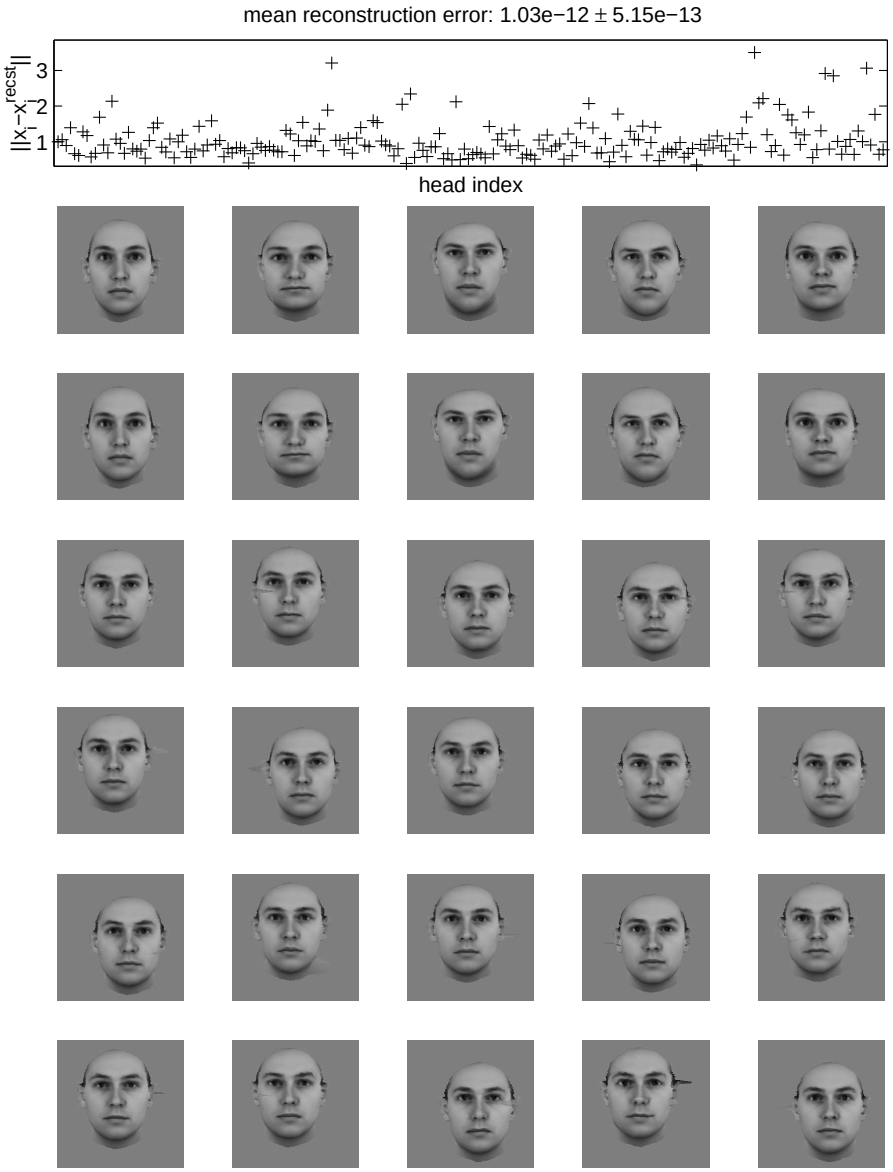


Figure E.11: Reconstruction error (first row), original and reconstructed heads (second and third rows) and the first 20 basis vectors (fourth to seventh row).

E.9 ICA I—Texture & Shape Data

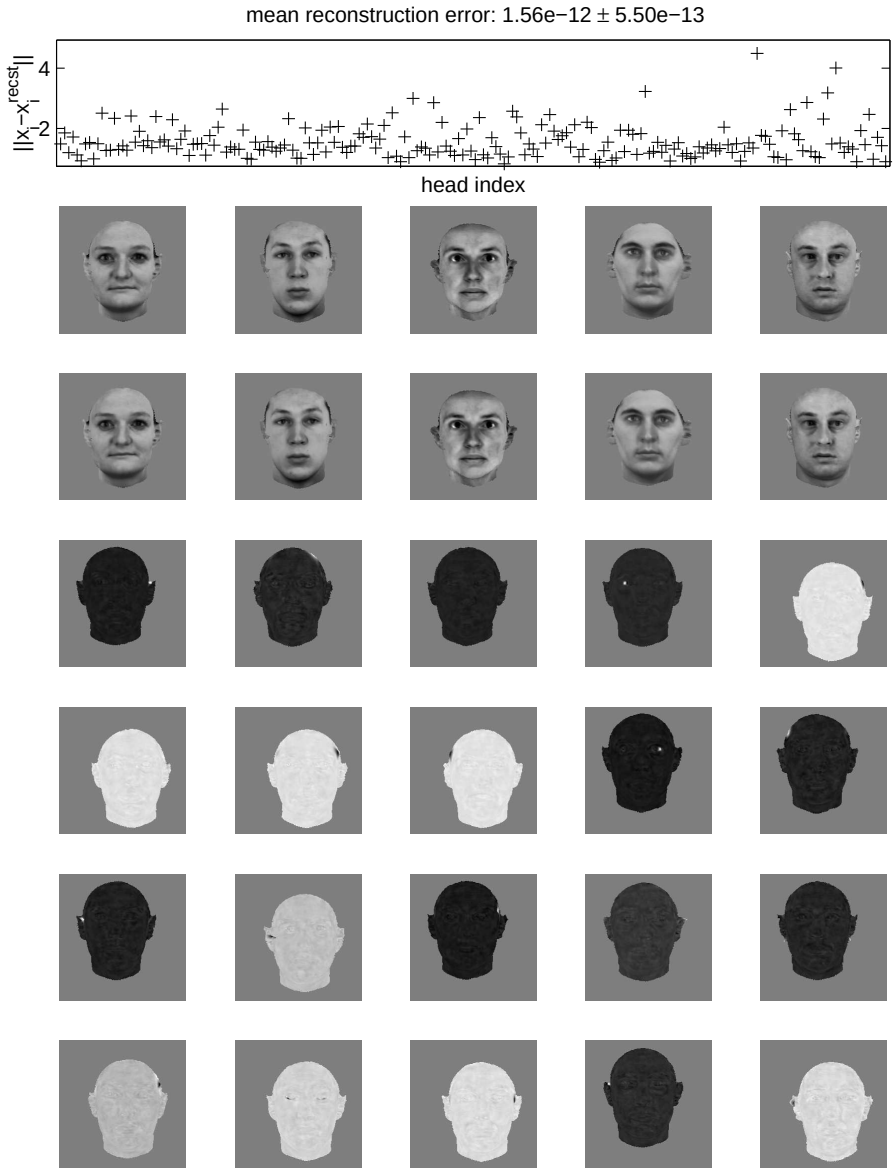


Figure E.12: Reconstruction error (first row), original and reconstructed heads (second and third rows) and the first 20 basis vectors (fourth to seventh row).

E.10 ICA II—Image Data

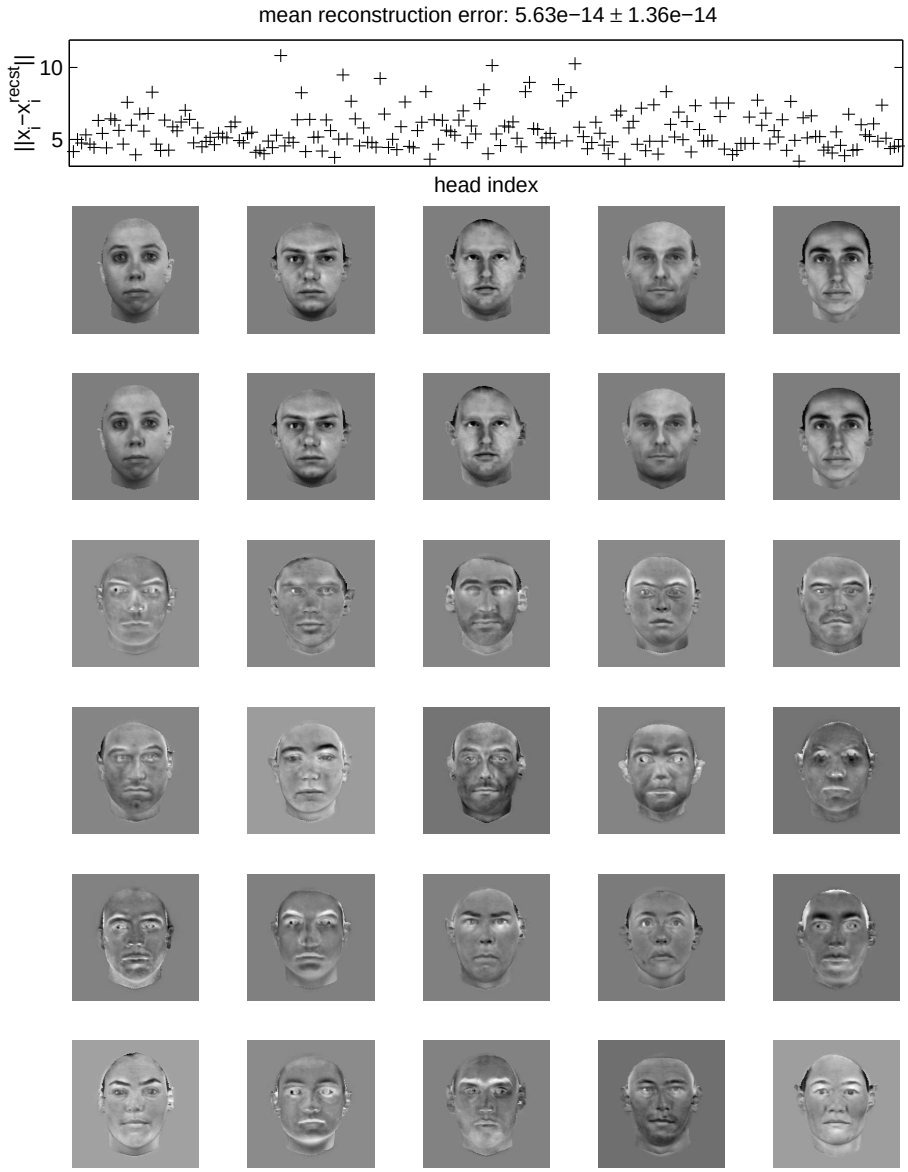


Figure E.13: Reconstruction error (first row), original and reconstructed heads (second and third rows) and the first 20 basis vectors (fourth to seventh row).

E.11 ICA II—Texture Data

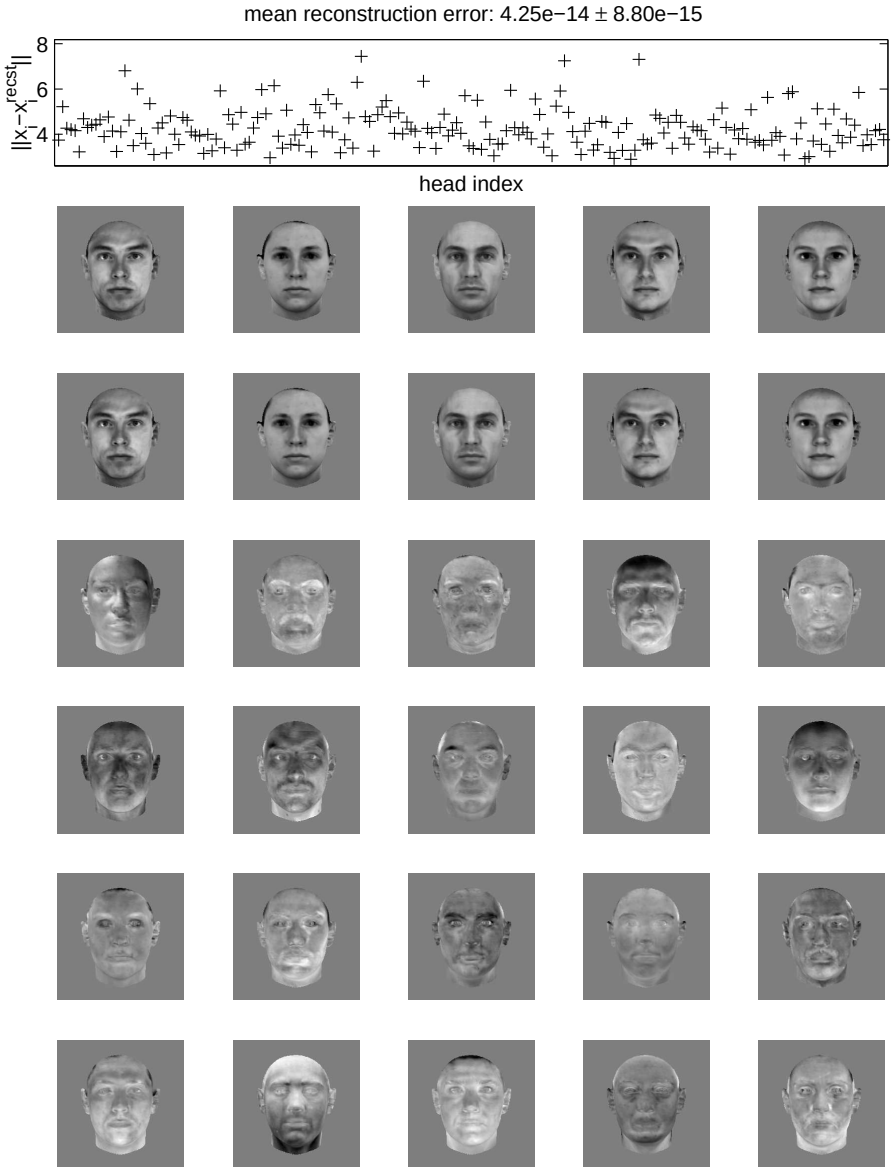


Figure E.14: Reconstruction error (first row), original and reconstructed heads (second and third rows) and the first 20 basis vectors (fourth to seventh row).

E.12 ICA II—Shape Data

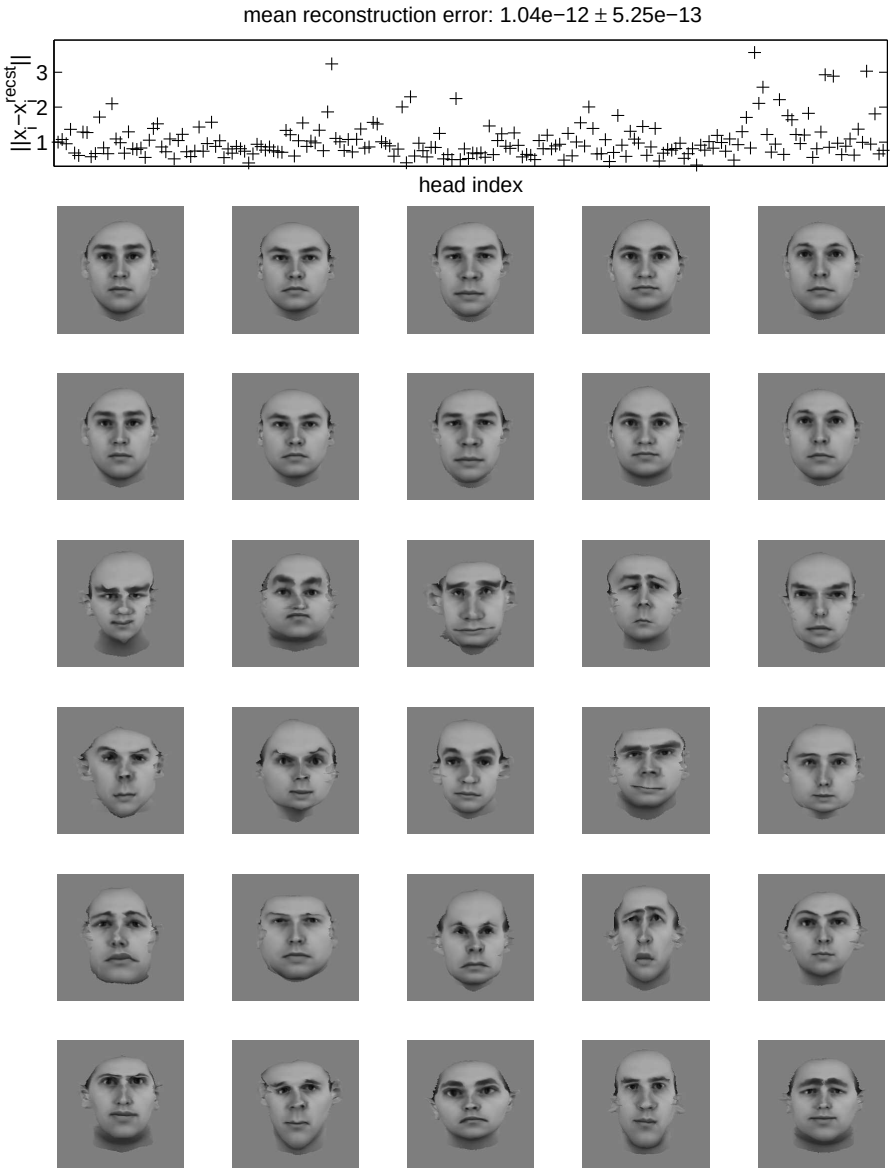


Figure E.15: Reconstruction error (first row), original and reconstructed heads (second and third rows) and the first 20 basis vectors (fourth to seventh row).

E.13 ICA II—Texture & Shape Data

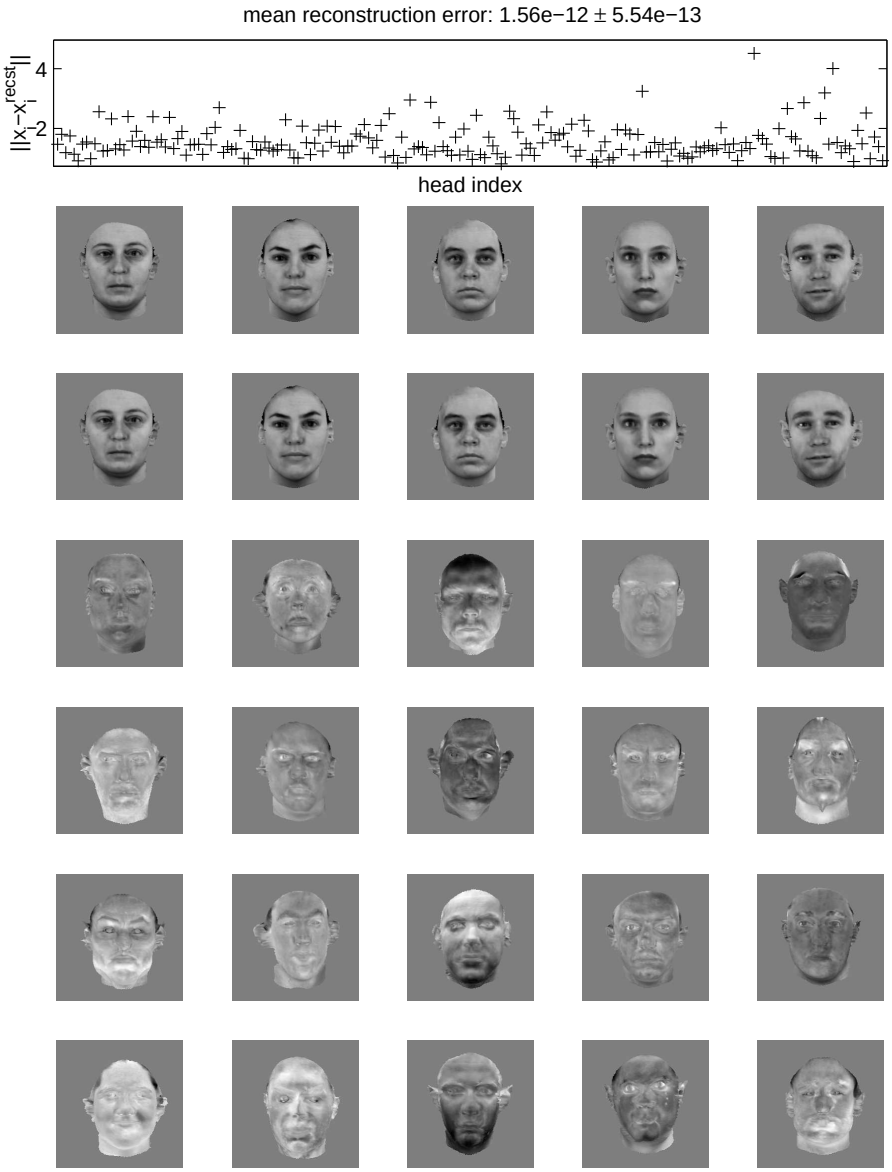


Figure E.16: Reconstruction error (first row), original and reconstructed heads (second and third rows) and the first 20 basis vectors (fourth to seventh row).

E.14 NMF—Image Data

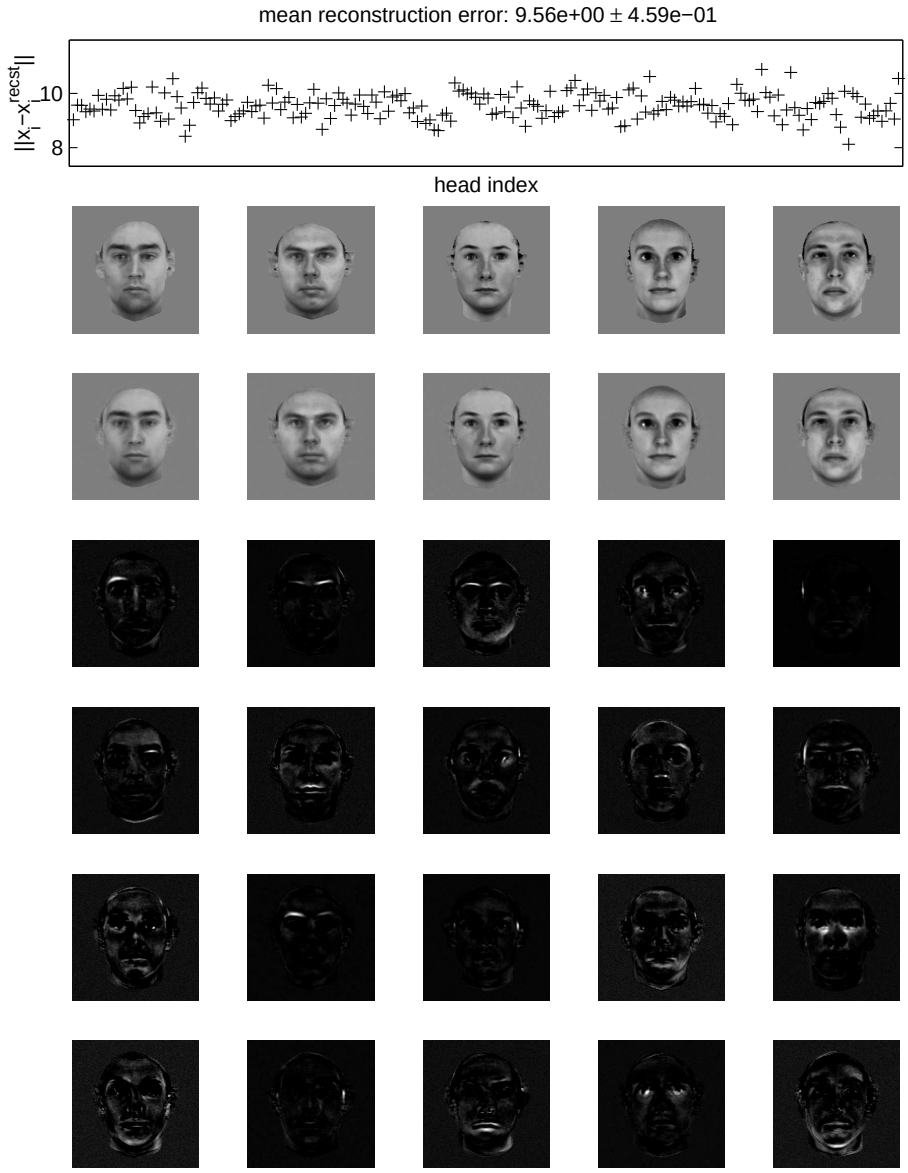


Figure E.17: Reconstruction error (first row), original and reconstructed heads (second and third rows) and the first 20 basis vectors (fourth to seventh row).

E.15 NMF—Texture Data

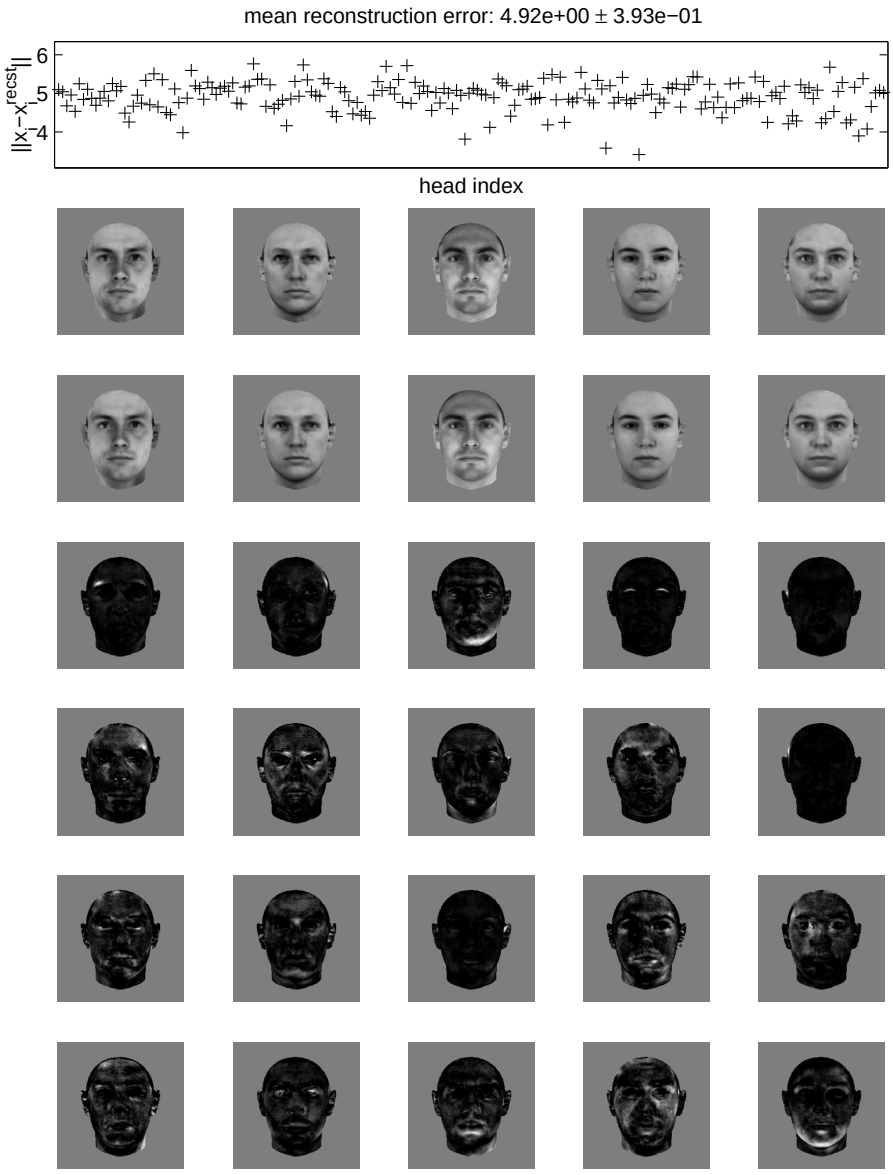


Figure E.18: Reconstruction error (first row), original and reconstructed heads (second and third rows) and the first 20 basis vectors (fourth to seventh row).

E.16 NMF—Shape Data

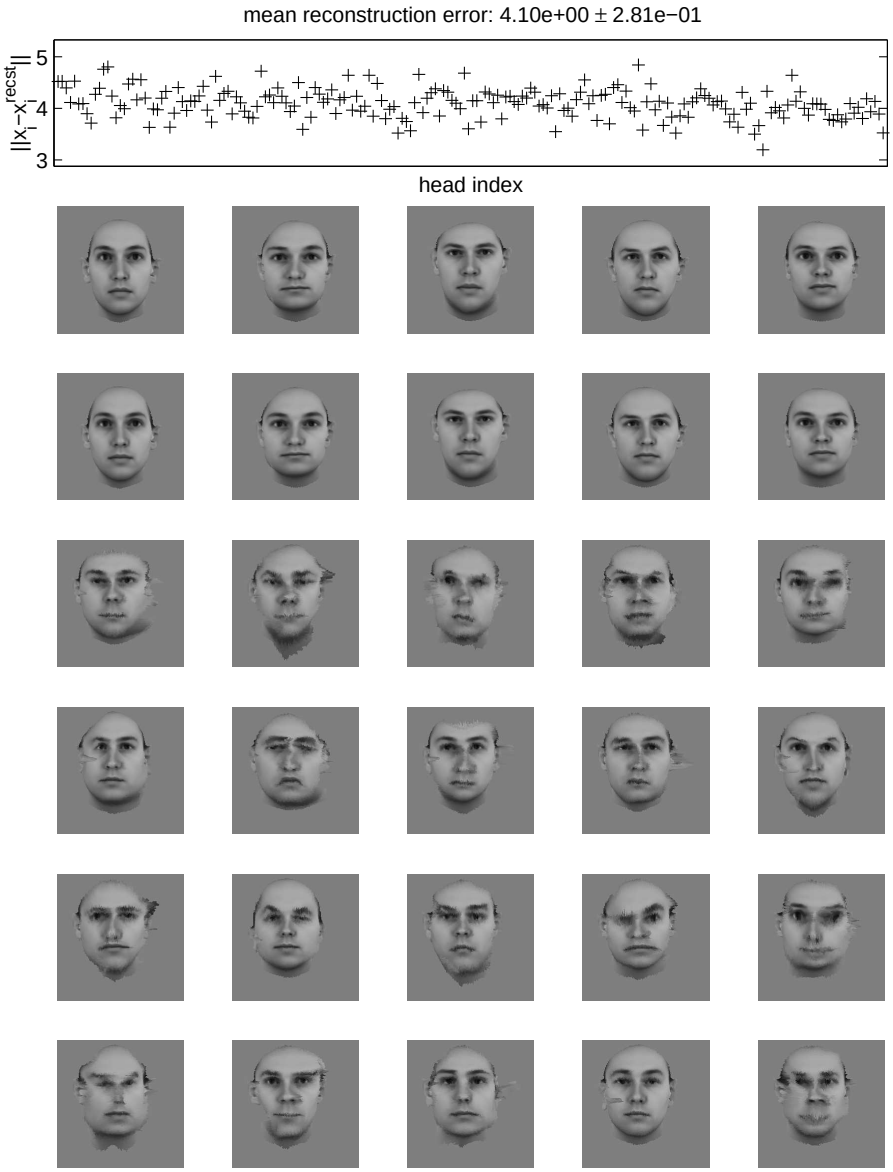


Figure E.19: Reconstruction error (first row), original and reconstructed heads (second and third rows) and the first 20 basis vectors (fourth to seventh row).

E.17 NMF—Texture & Shape Data

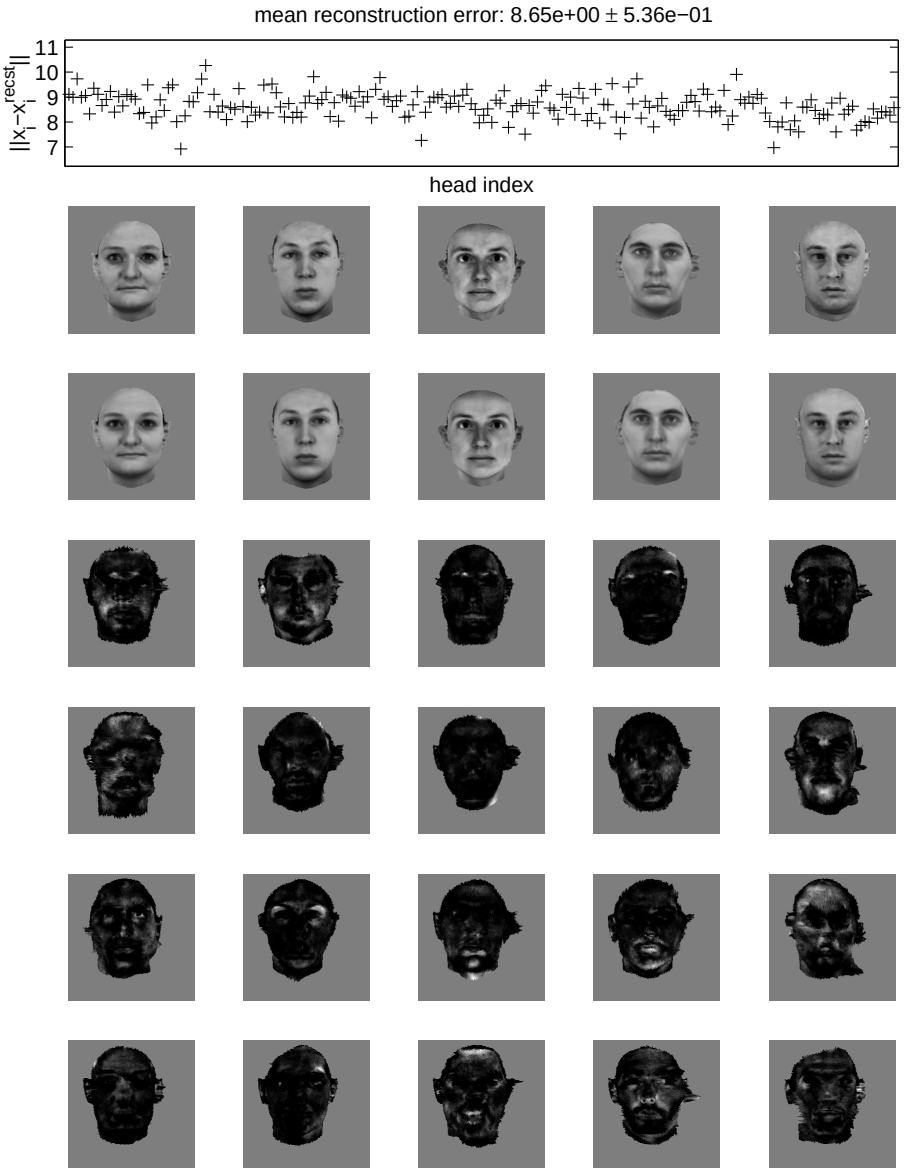


Figure E.20: Reconstruction error (first row), original and reconstructed heads (second and third rows) and the first 20 basis vectors (fourth to seventh row).

Appendix F

Plots Relating Man and Machine

F.1 Overview

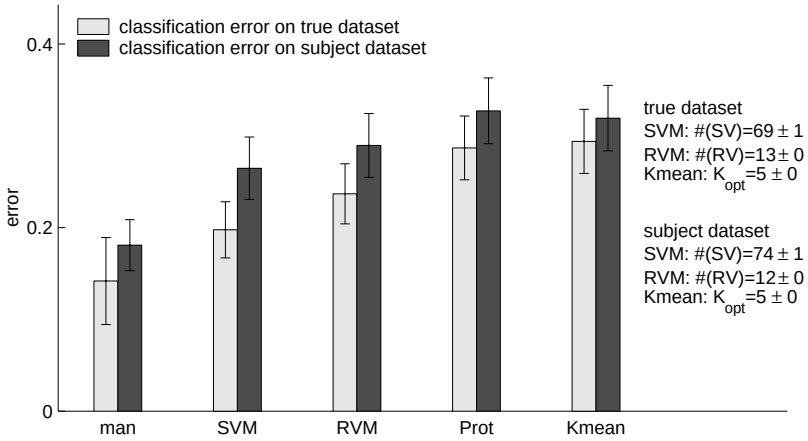
In order to avoid overloading this appendix, we do not put a caption on each of the plots but mention below the captions to be applied.
Left page (descriptions sorted by increasing value of the row)

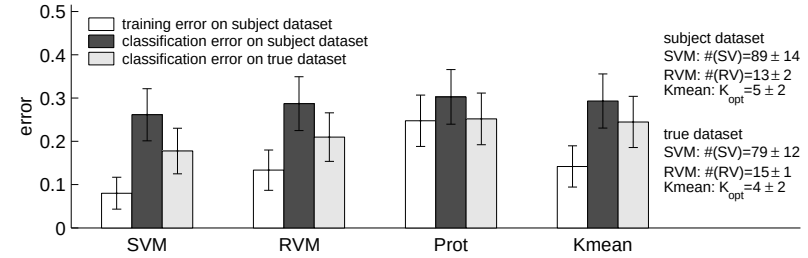
- Comparison of classification performance of man and machine on the true and subject datasets using cross-validation, only for Image or Texture & Shape data.
- Comparison of training and classification errors of machine on the true and subject datasets without cross-validation, and number of representations.
- Histograms of distances of (non-)representations to SH.
- Correlation of classification behavior of man and machine with parameters averaged over subjects and sets of stimuli.

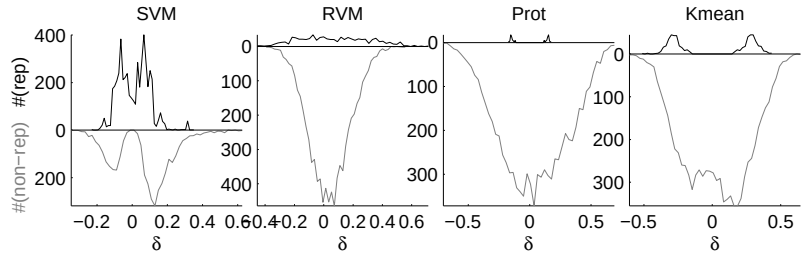
Right page (descriptions sorted by increasing value of the row)

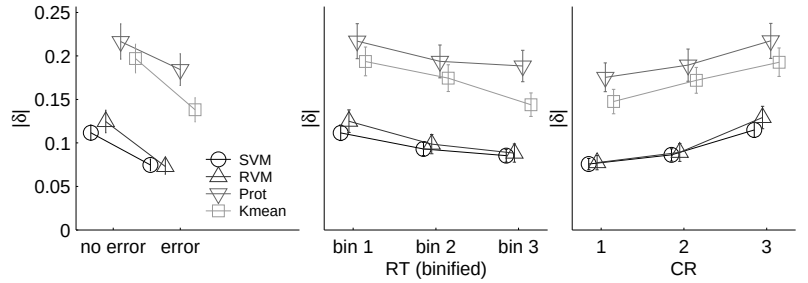
- Correlation of classification behavior of man and machine with parameters averaged over subjects. On the borders: scatter plots and correlation coefficients r relating the classification behavior of man (classification error, RT and CR) and machine (distance $|\delta|$ of the stimuli to the SH). In the center: polar representation of the $|r|$ for each classifier and human response.
- Stability analysis with parameters averaged over subjects. On the right: scatter plots and correlation coefficients r relating the jitter in the subject's classification error and $|\delta|$. On the left: polar representation of $|r|$ for each classifier and human response for the first "+" and second "x" classification experiment, the "$\circ$" on the error axis representing this coefficient for the plots on the right.

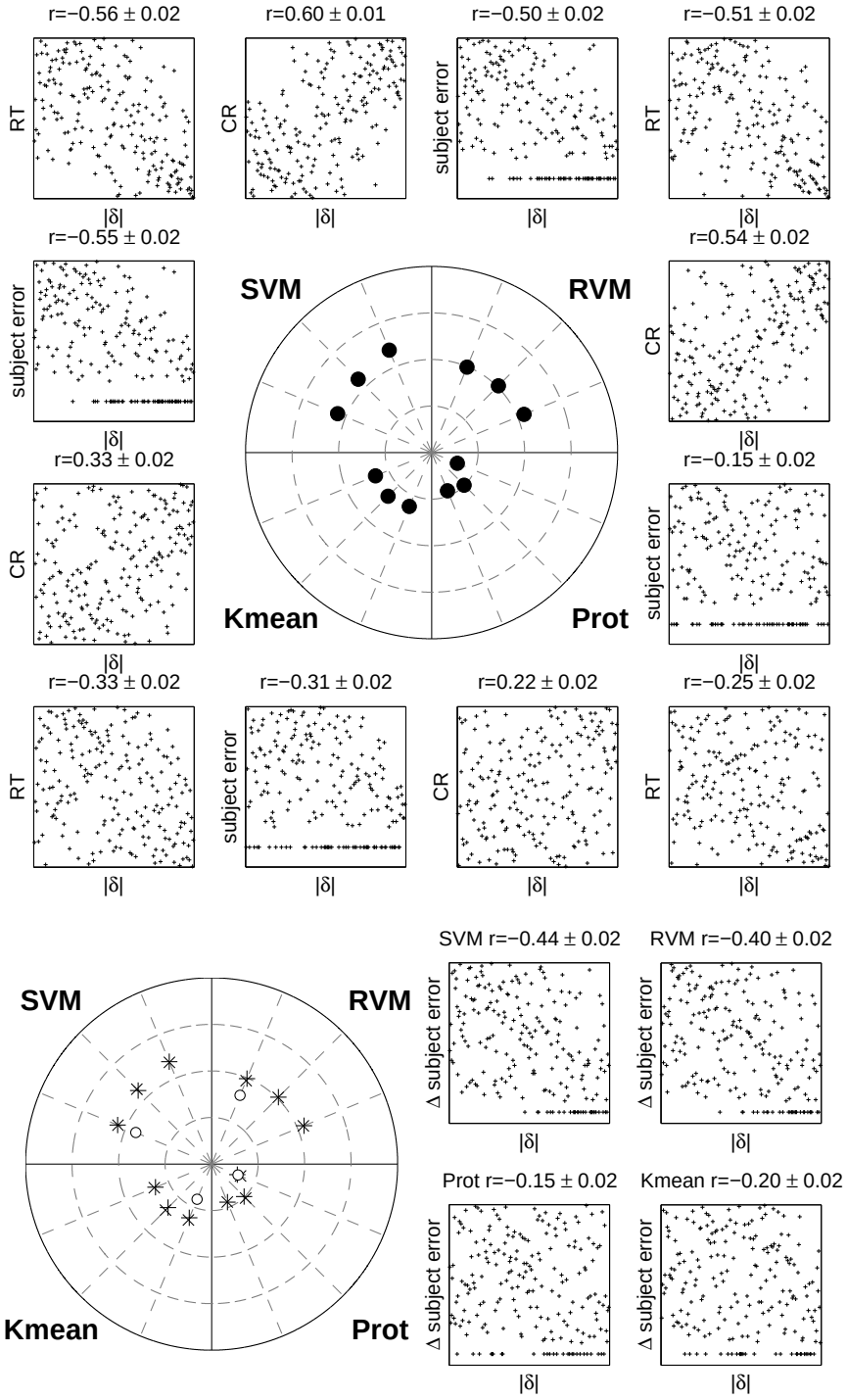
F.2 Image Size Reduction



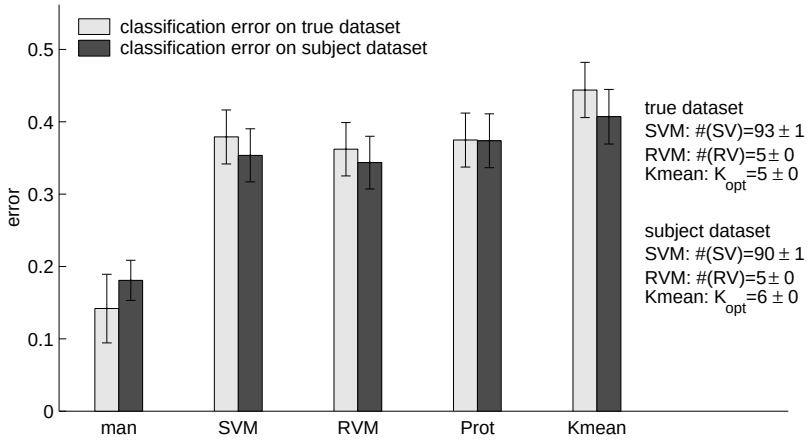


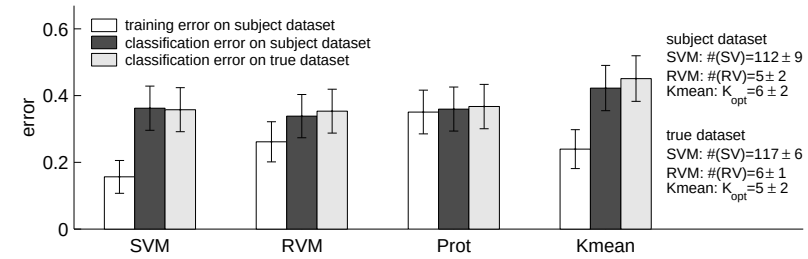


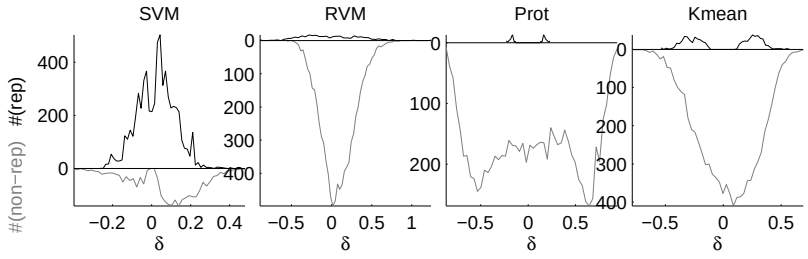


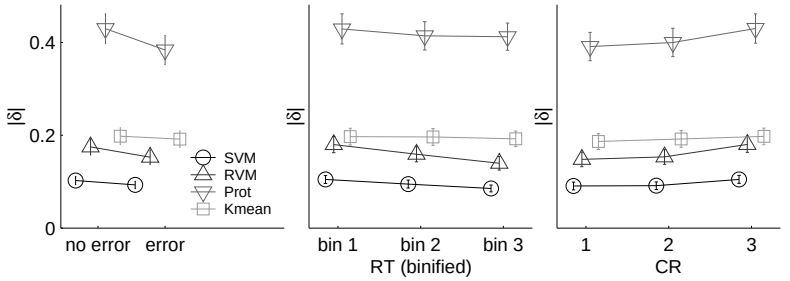

r=−0.56 ± 0.02
RT
|δ|
r=0.60 ± 0.01
CR
|δ|
r=−0.50 ± 0.02
subject error
|δ|
r=−0.51 ± 0.02
RT
|δ|
r=−0.55 ± 0.02
subject error
|δ|
SVM
RVM
Kmean
Prot
r=0.54 ± 0.02
CR
|δ|
r=0.33 ± 0.02
CR
|δ|
r=−0.15 ± 0.02
subject error
|δ|
r=−0.33 ± 0.02
RT
|δ|
r=−0.31 ± 0.02
subject error
|δ|
r=0.22 ± 0.02
CR
|δ|
r=−0.25 ± 0.02
RT
|δ|
SVM
RVM
Kmean
Prot
SVM r=−0.44 ± 0.02
Δ subject error
|δ|
RVM r=−0.40 ± 0.02
Δ subject error
|δ|
Prot r=−0.15 ± 0.02
Δ subject error
|δ|
Kmean r=−0.20 ± 0.02
Δ subject error
|δ|

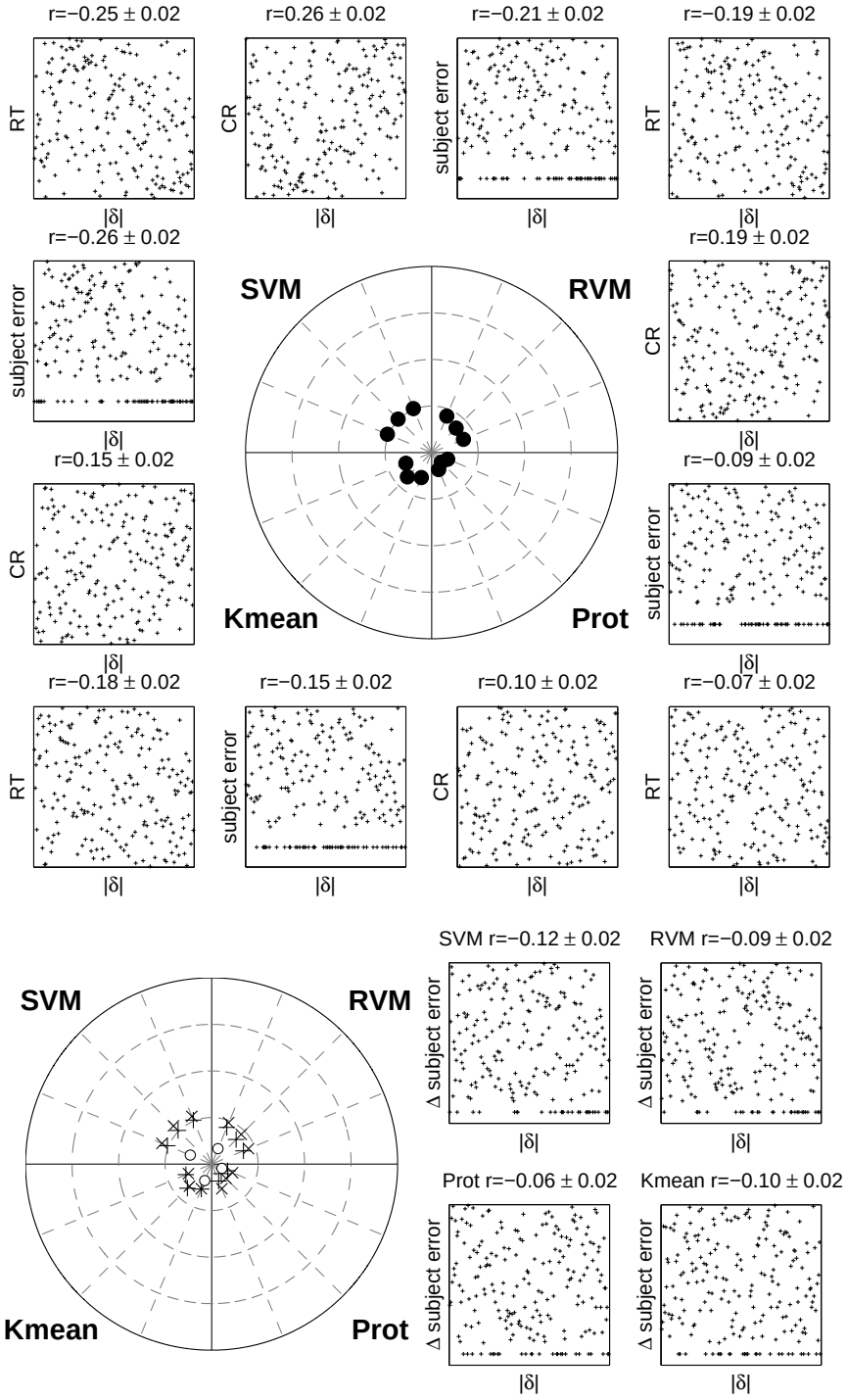
F.3 Histograms



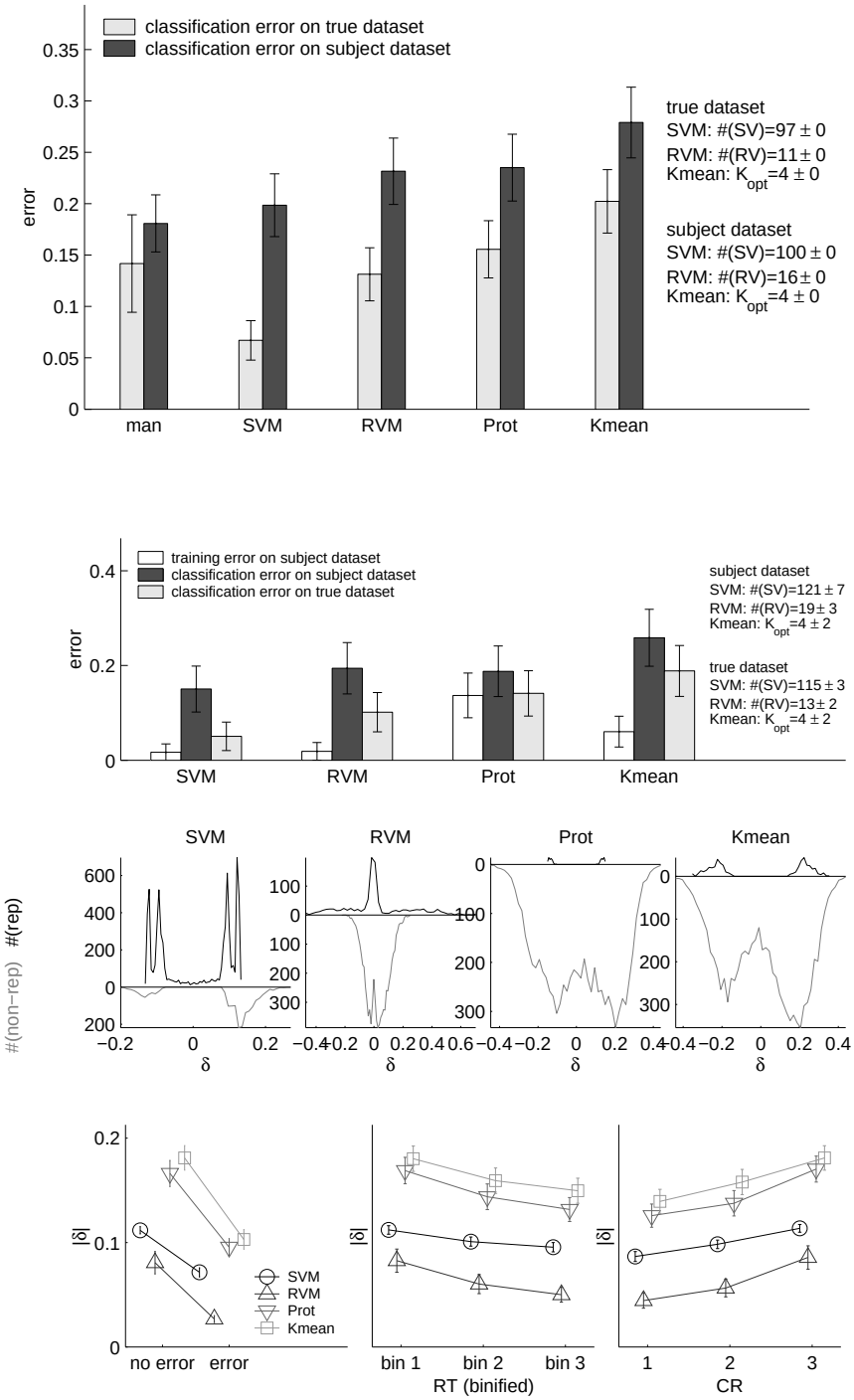


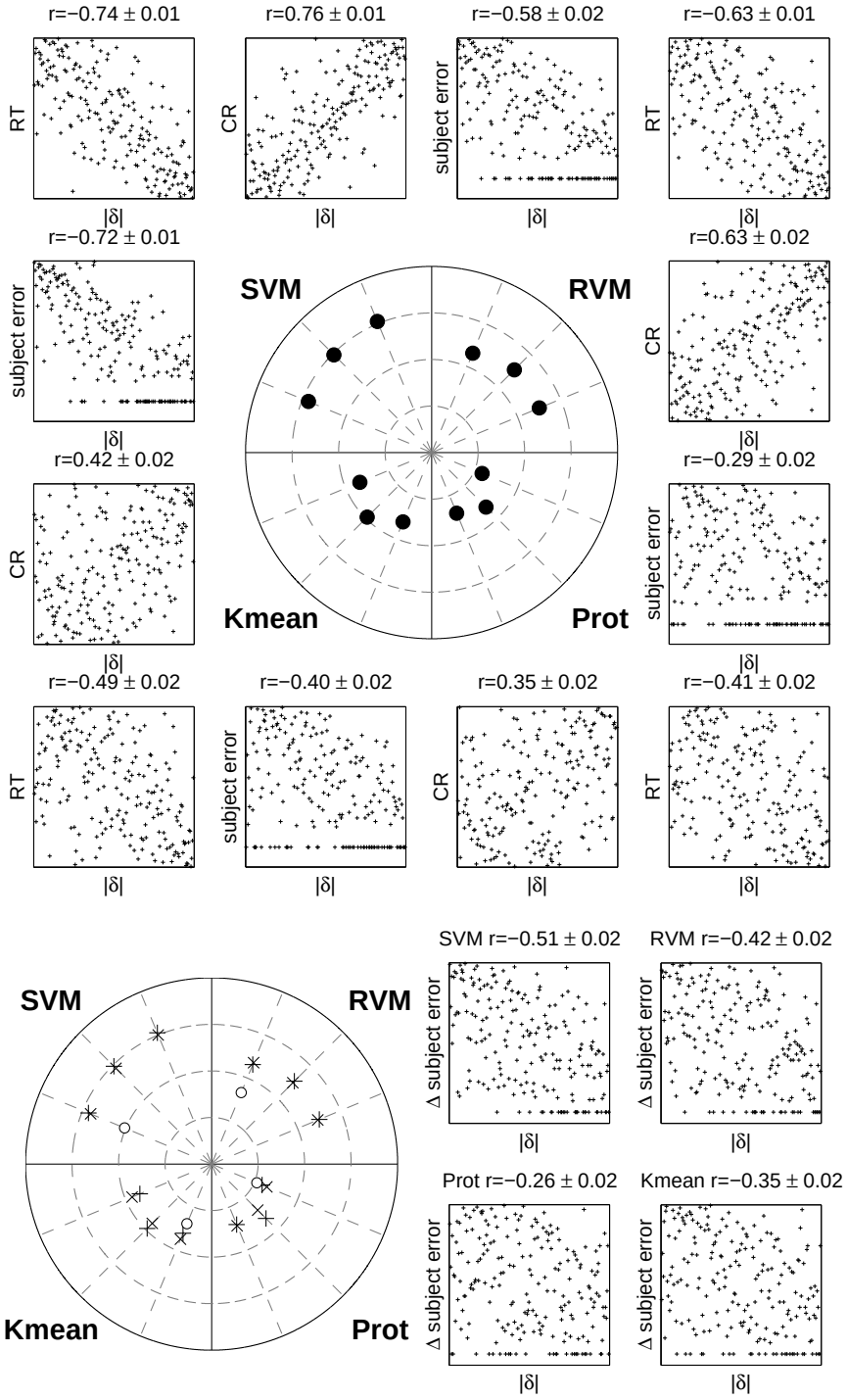




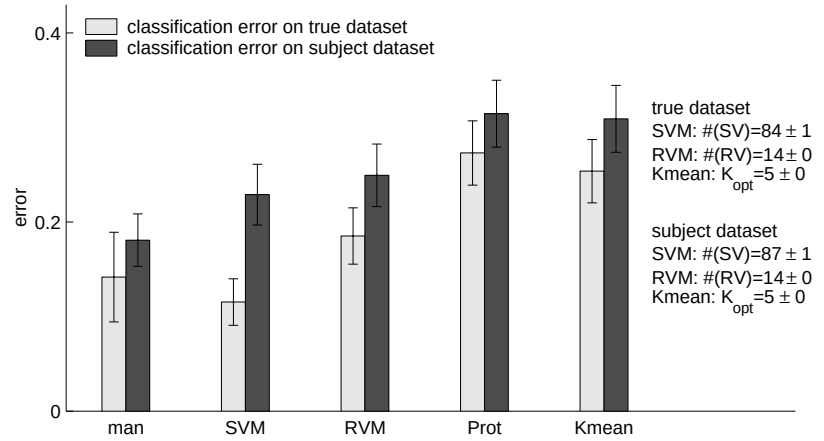

r=−0.25 ± 0.02
RT
|δ|
r=0.26 ± 0.02
CR
|δ|
r=−0.21 ± 0.02
subject error
|δ|
r=−0.19 ± 0.02
RT
|δ|
r=−0.26 ± 0.02
subject error
|δ|
SVM
RVM
r=0.19 ± 0.02
CR
|δ|
r=0.15 ± 0.02
CR
|δ|
Kmean
Prot
r=−0.09 ± 0.02
subject error
|δ|
r=−0.18 ± 0.02
RT
|δ|
r=−0.15 ± 0.02
subject error
|δ|
r=0.10 ± 0.02
CR
|δ|
r=−0.07 ± 0.02
RT
|δ|
SVM
RVM
Kmean
Prot
SVM r=−0.12 ± 0.02
Δ subject error
|δ|
RVM r=−0.09 ± 0.02
Δ subject error
|δ|
Prot r=−0.06 ± 0.02
Δ subject error
|δ|
Kmean r=−0.10 ± 0.02
Δ subject error
|δ|

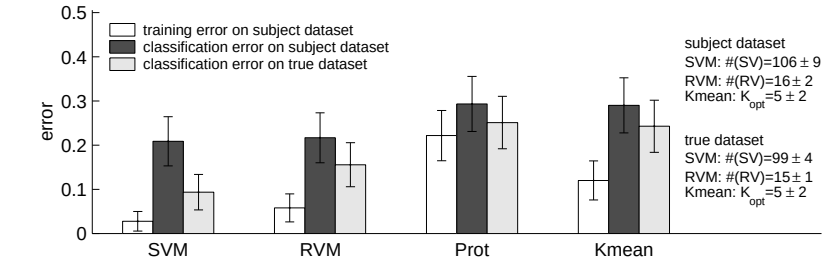
F.4 Gabor Filters

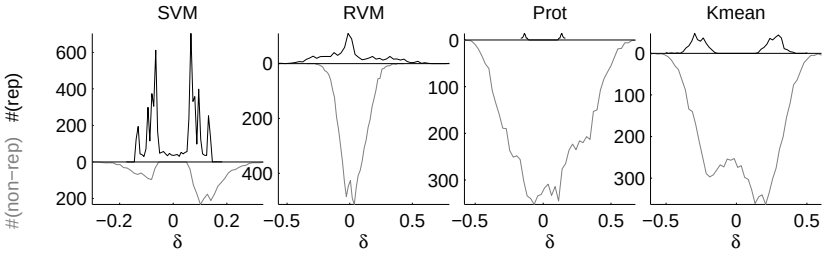


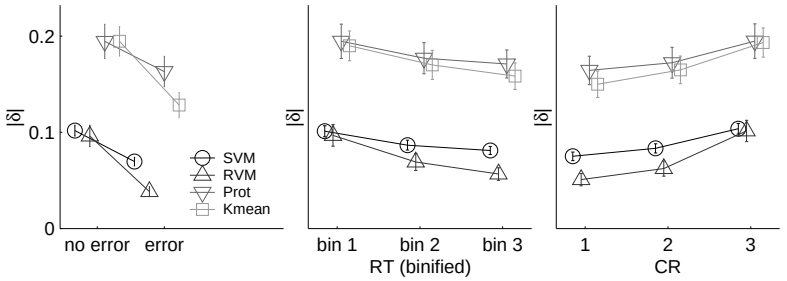

r=−0.74 ± 0.01
RT
|δ|
r=0.76 ± 0.01
CR
|δ|
r=−0.58 ± 0.02
subject error
|δ|
r=−0.63 ± 0.01
RT
|δ|
r=−0.72 ± 0.01
subject error
|δ|
SVM
RVM
Kmean
Prot
r=0.63 ± 0.02
CR
|δ|
r=0.42 ± 0.02
CR
|δ|
r=−0.29 ± 0.02
subject error
|δ|
r=−0.49 ± 0.02
RT
|δ|
r=−0.40 ± 0.02
subject error
|δ|
r=0.35 ± 0.02
CR
|δ|
r=−0.41 ± 0.02
RT
|δ|
SVM
RVM
Kmean
Prot
SVM r=−0.51 ± 0.02
Δ subject error
|δ|
RVM r=−0.42 ± 0.02
Δ subject error
|δ|
Prot r=−0.26 ± 0.02
Δ subject error
|δ|
Kmean r=−0.35 ± 0.02
Δ subject error
|δ|

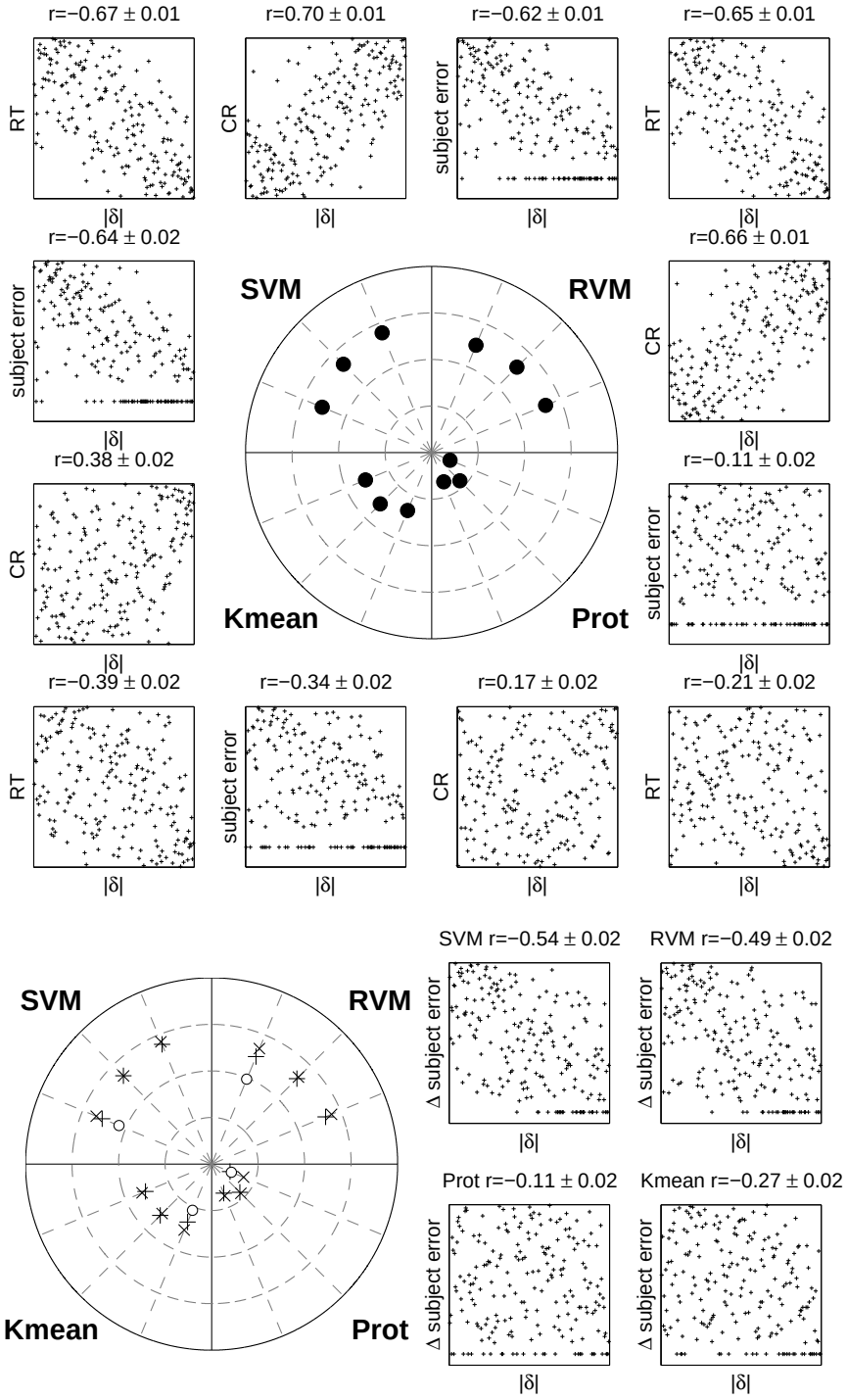
F.5 PCA—Image Data



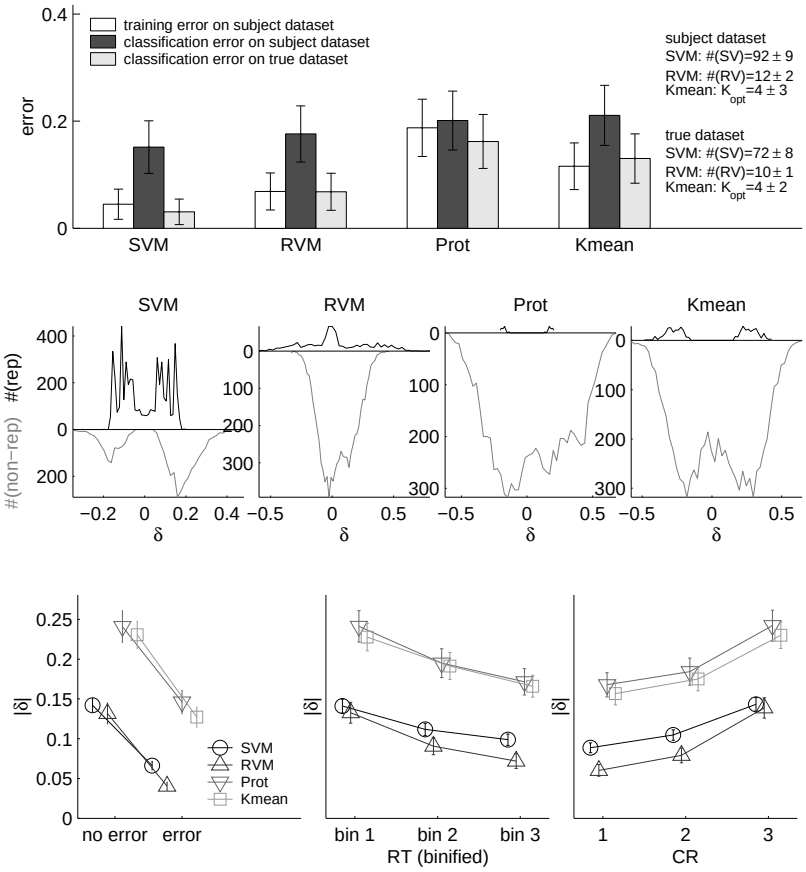


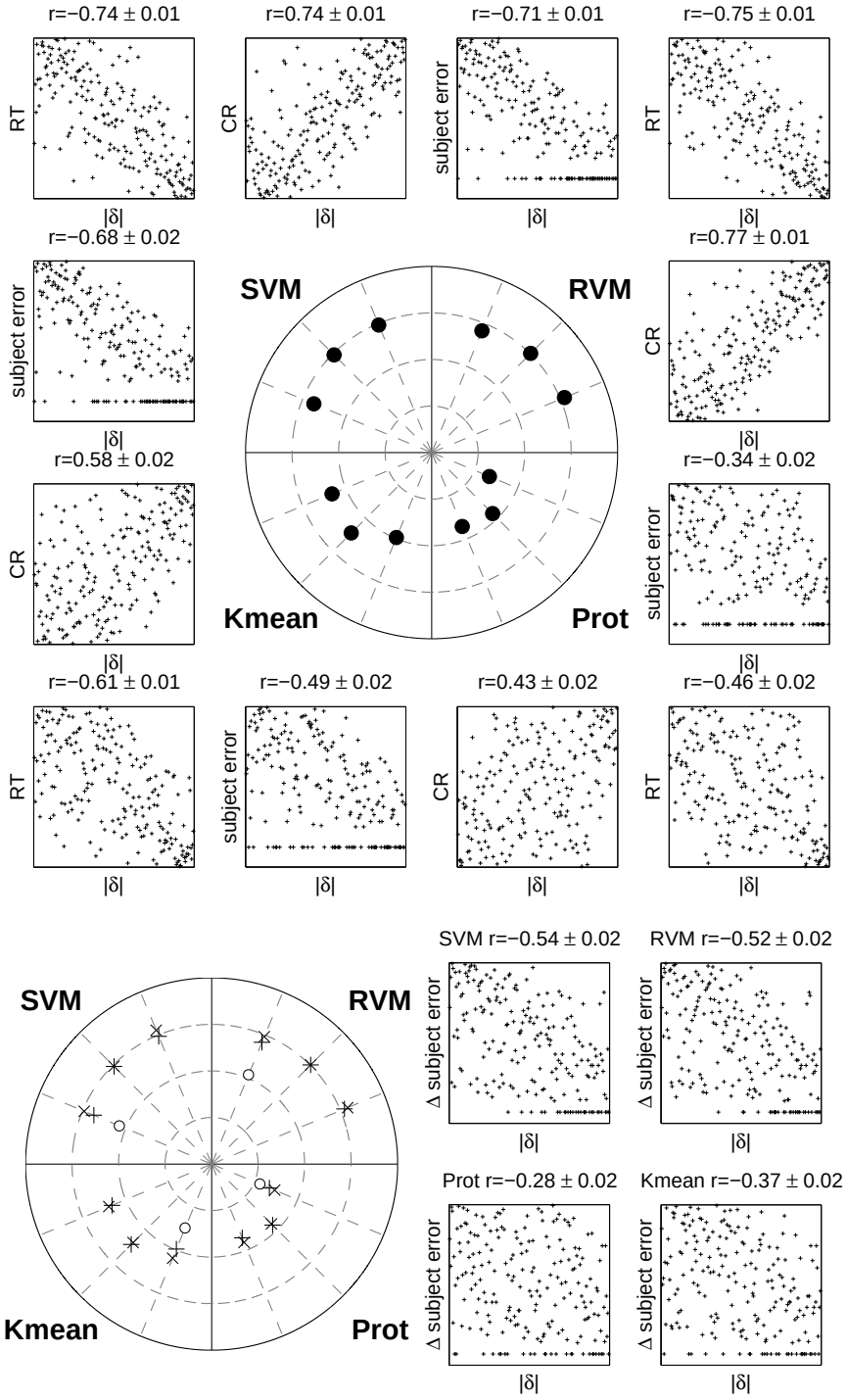




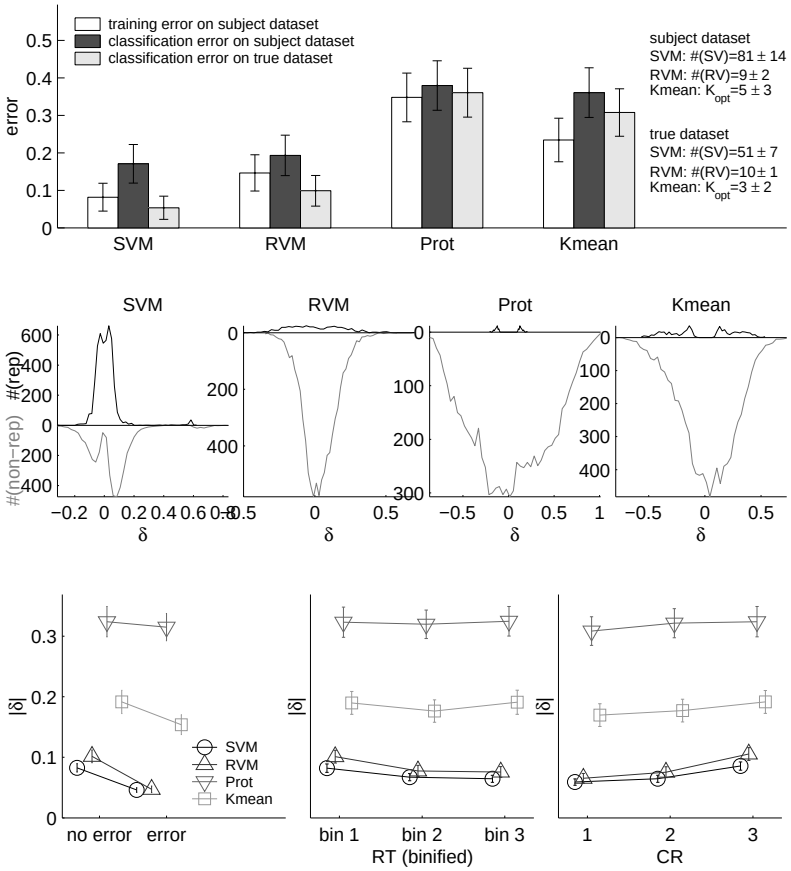

r=−0.67 ± 0.01
RT
|δ|
r=0.70 ± 0.01
CR
|δ|
r=−0.62 ± 0.01
subject error
|δ|
r=−0.65 ± 0.01
RT
|δ|
r=−0.64 ± 0.02
subject error
|δ|
SVM
RVM
Kmean
Prot
r=0.66 ± 0.01
CR
|δ|
r=0.38 ± 0.02
CR
|δ|
r=−0.11 ± 0.02
subject error
|δ|
r=−0.39 ± 0.02
RT
|δ|
r=−0.34 ± 0.02
subject error
|δ|
r=0.17 ± 0.02
CR
|δ|
r=−0.21 ± 0.02
RT
|δ|
SVM
RVM
Kmean
Prot
SVM r=−0.54 ± 0.02
Δ subject error
|δ|
RVM r=−0.49 ± 0.02
Δ subject error
|δ|
Prot r=−0.11 ± 0.02
Δ subject error
|δ|
Kmean r=−0.27 ± 0.02
Δ subject error
|δ|

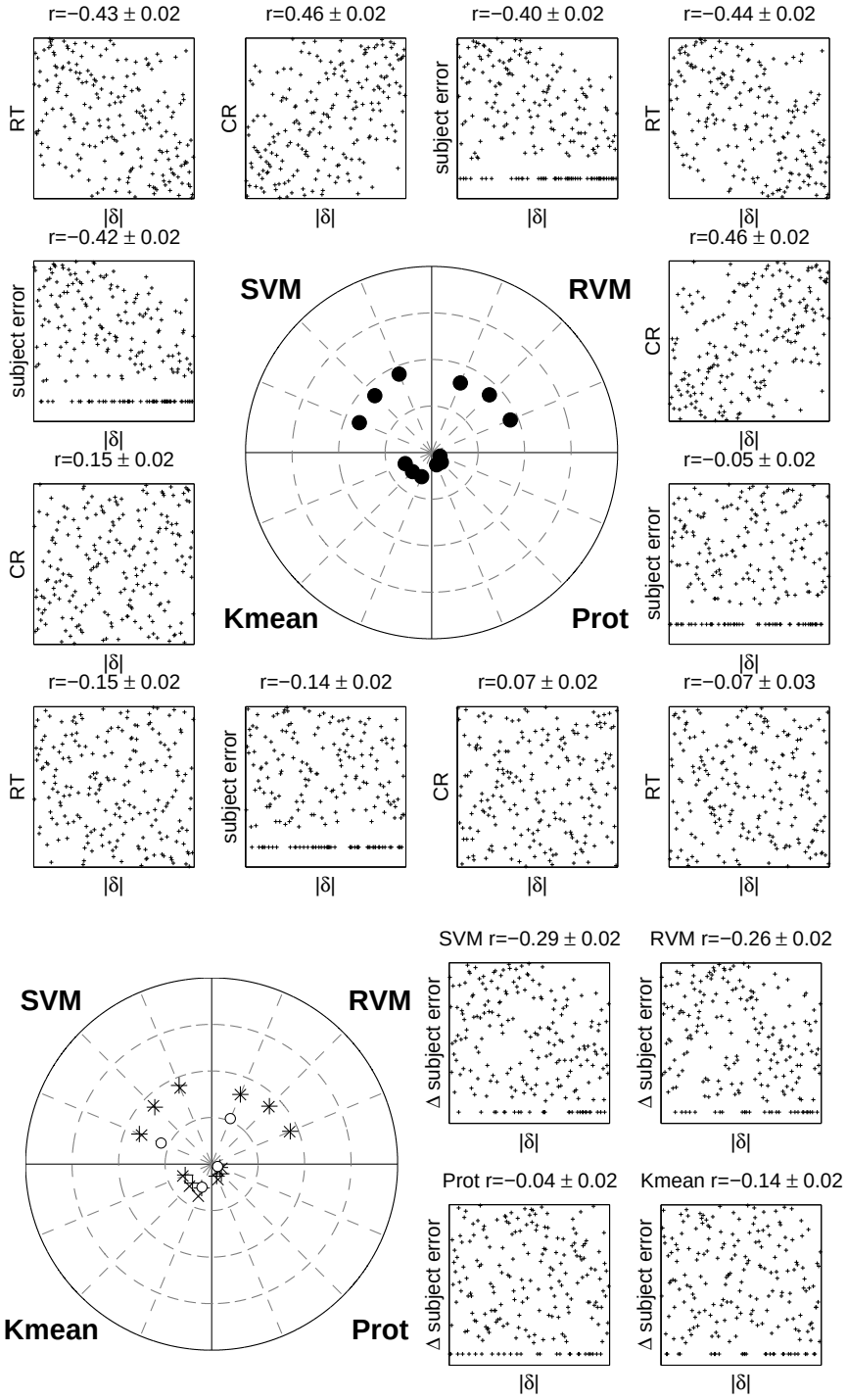
F.6 PCA—Texture Data



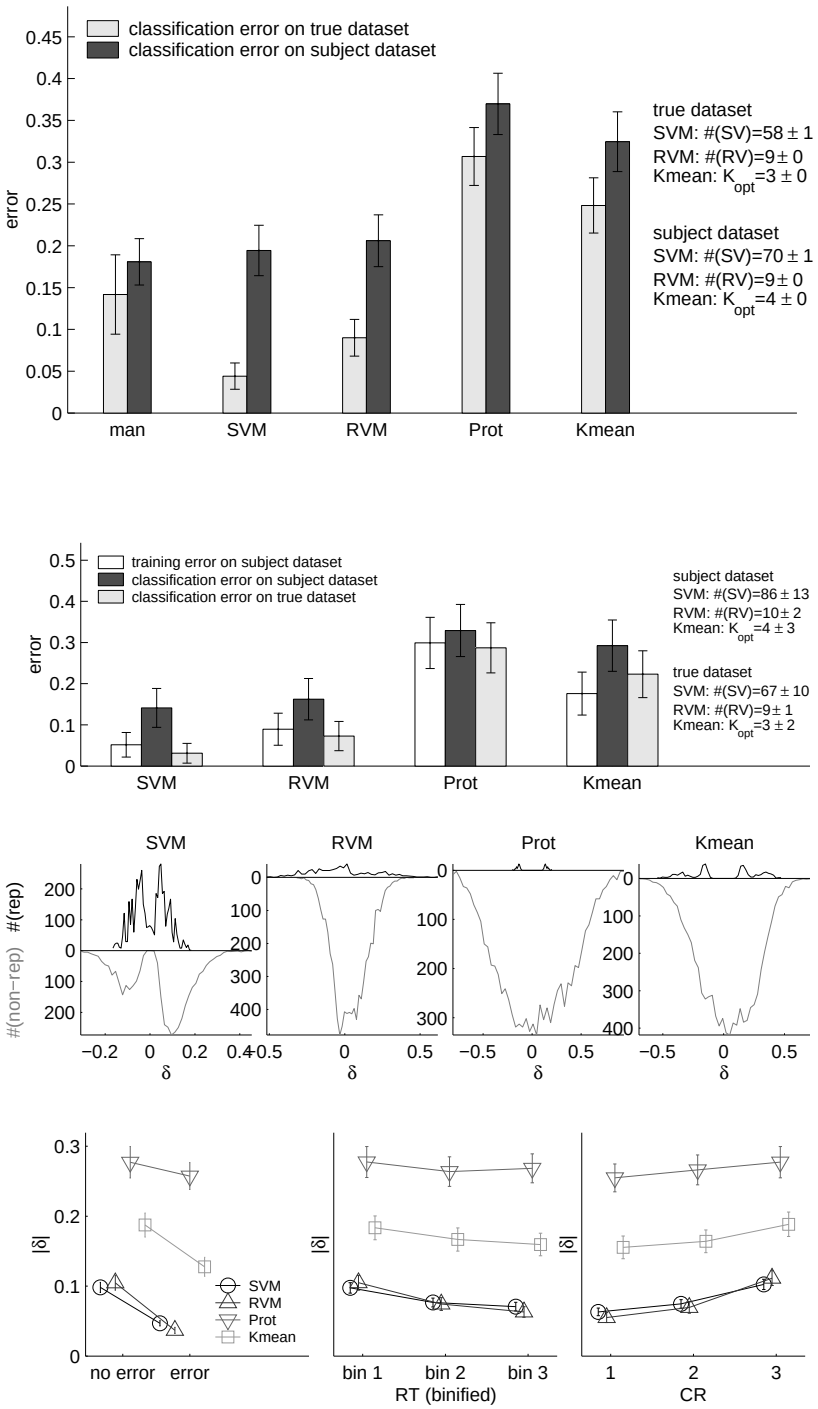

r=−0.74 ± 0.01
RT
|δ|
r=0.74 ± 0.01
CR
|δ|
r=−0.71 ± 0.01
subject error
|δ|
r=−0.75 ± 0.01
RT
|δ|
r=−0.68 ± 0.02
subject error
|δ|
SVM
RVM
r=0.77 ± 0.01
CR
|δ|
r=0.58 ± 0.02
CR
|δ|
Kmean
Prot
r=−0.34 ± 0.02
subject error
|δ|
r=−0.61 ± 0.01
RT
|δ|
r=−0.49 ± 0.02
subject error
|δ|
r=0.43 ± 0.02
CR
|δ|
r=−0.46 ± 0.02
RT
|δ|
SVM
RVM
Kmean
Prot
SVM r=−0.54 ± 0.02
Δ subject error
|δ|
RVM r=−0.52 ± 0.02
Δ subject error
|δ|
Prot r=−0.28 ± 0.02
Δ subject error
|δ|
Kmean r=−0.37 ± 0.02
Δ subject error
|δ|

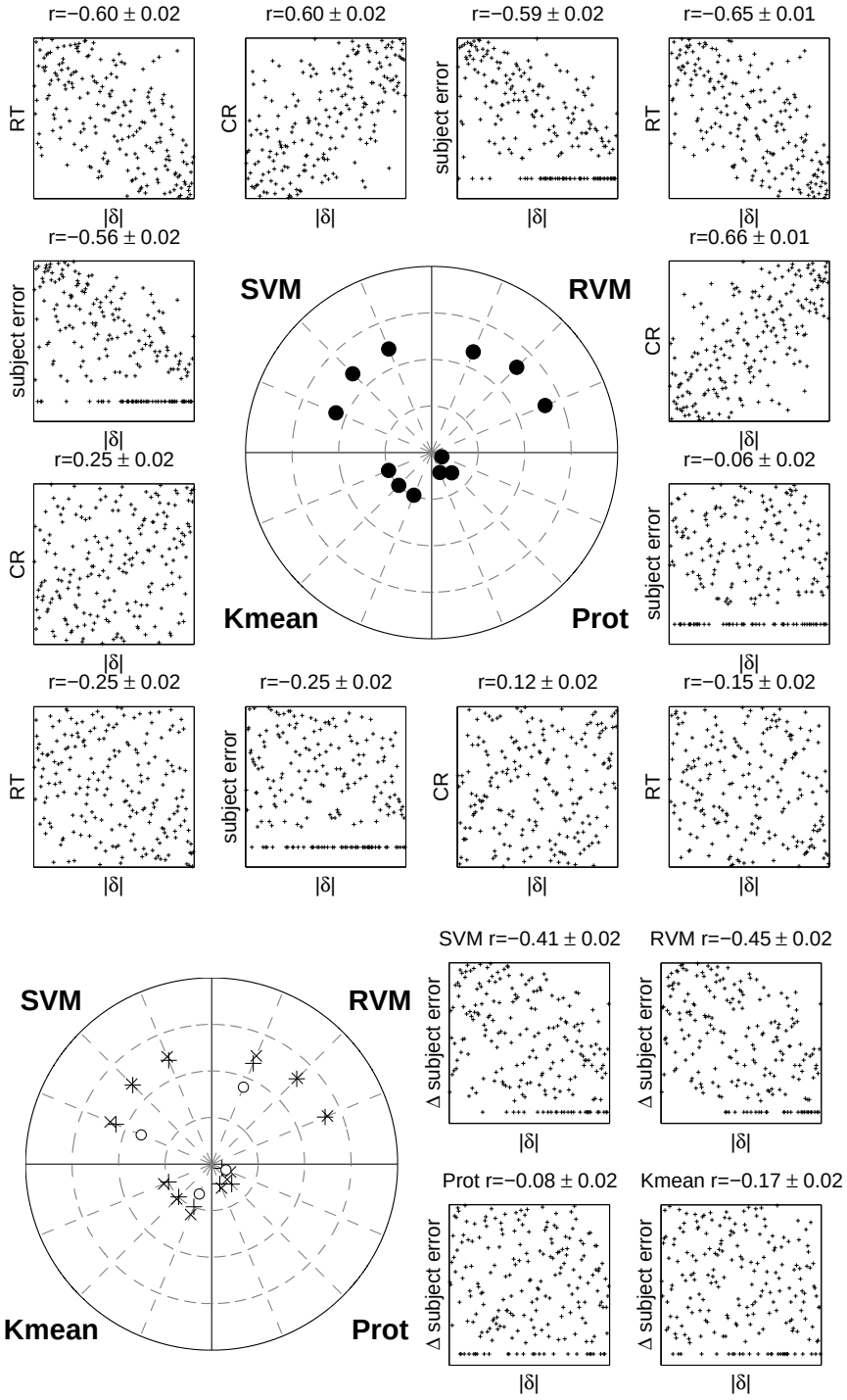
F.7 PCA—Shape Data



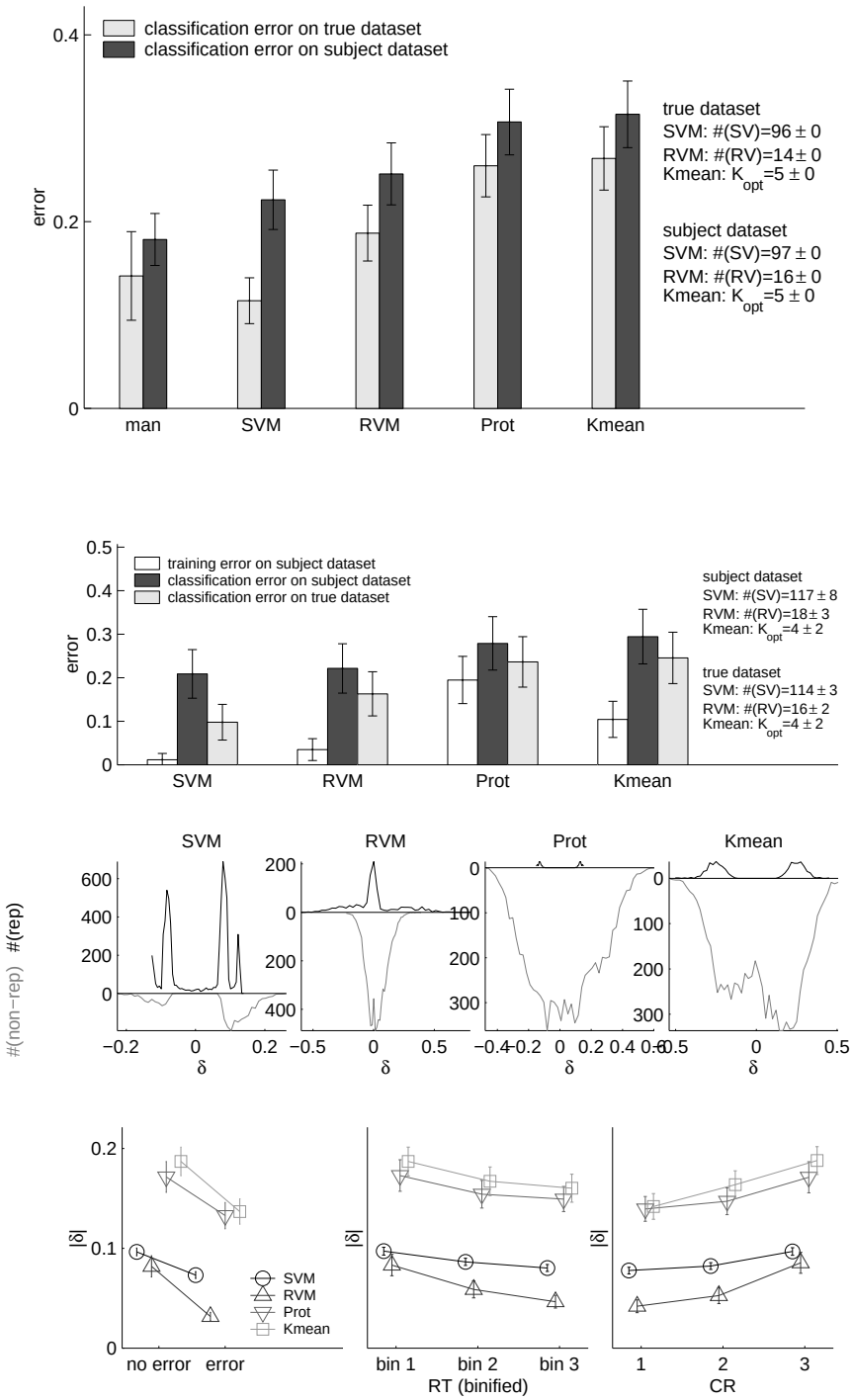

r=−0.43 ± 0.02
RT
|δ|
r=0.46 ± 0.02
CR
|δ|
r=−0.40 ± 0.02
subject error
|δ|
r=−0.44 ± 0.02
RT
|δ|
r=−0.42 ± 0.02
subject error
|δ|
SVM
RVM
Kmean
Prot
r=0.46 ± 0.02
CR
|δ|
r=0.15 ± 0.02
CR
|δ|
r=−0.05 ± 0.02
subject error
|δ|
r=−0.15 ± 0.02
RT
|δ|
r=−0.14 ± 0.02
subject error
|δ|
r=0.07 ± 0.02
CR
|δ|
r=−0.07 ± 0.03
RT
|δ|
SVM
RVM
Kmean
Prot
SVM r=−0.29 ± 0.02
Δ subject error
|δ|
RVM r=−0.26 ± 0.02
Δ subject error
|δ|
Prot r=−0.04 ± 0.02
Δ subject error
|δ|
Kmean r=−0.14 ± 0.02
Δ subject error
|δ|

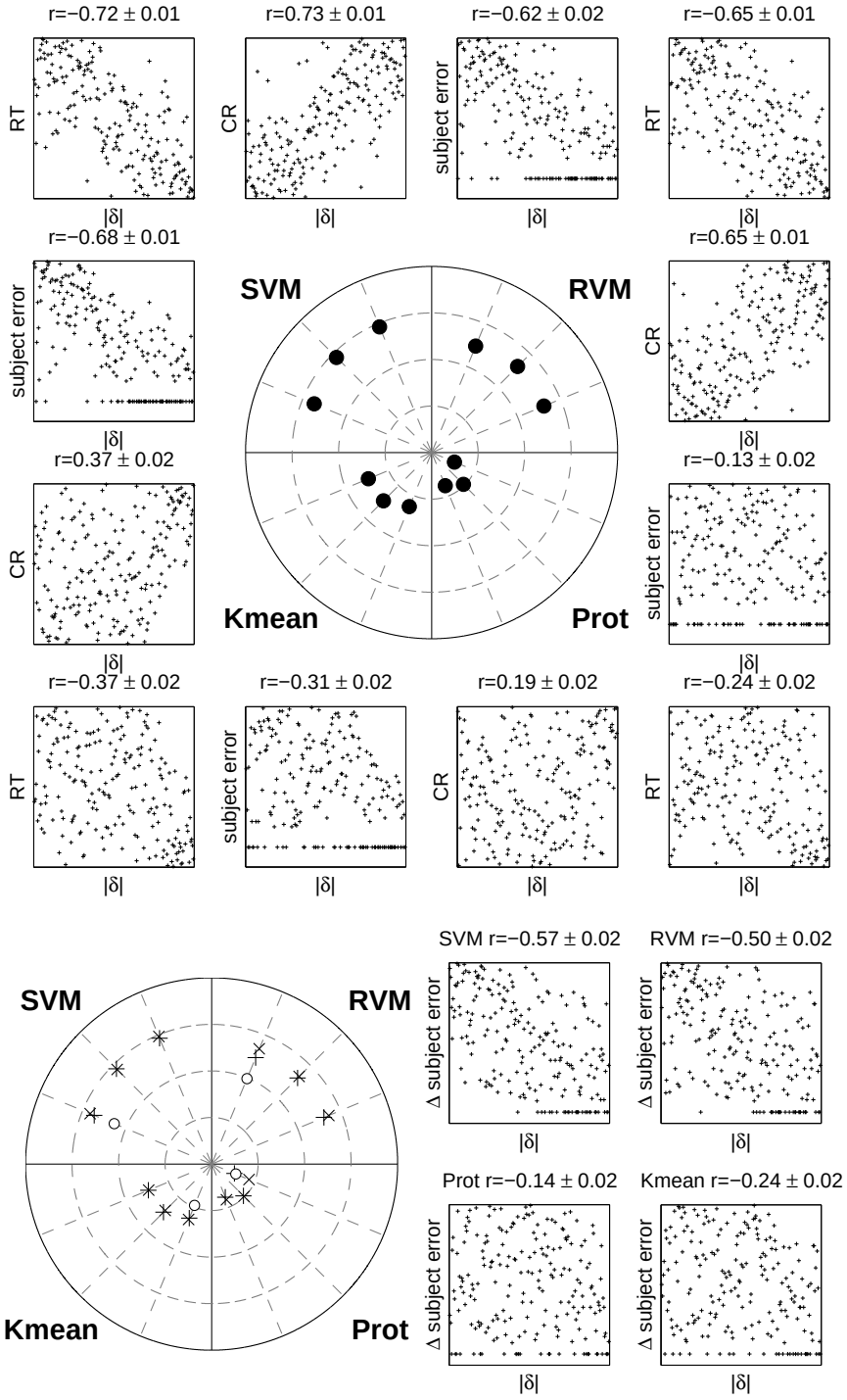
F.8 PCA—Texture & Shape Data



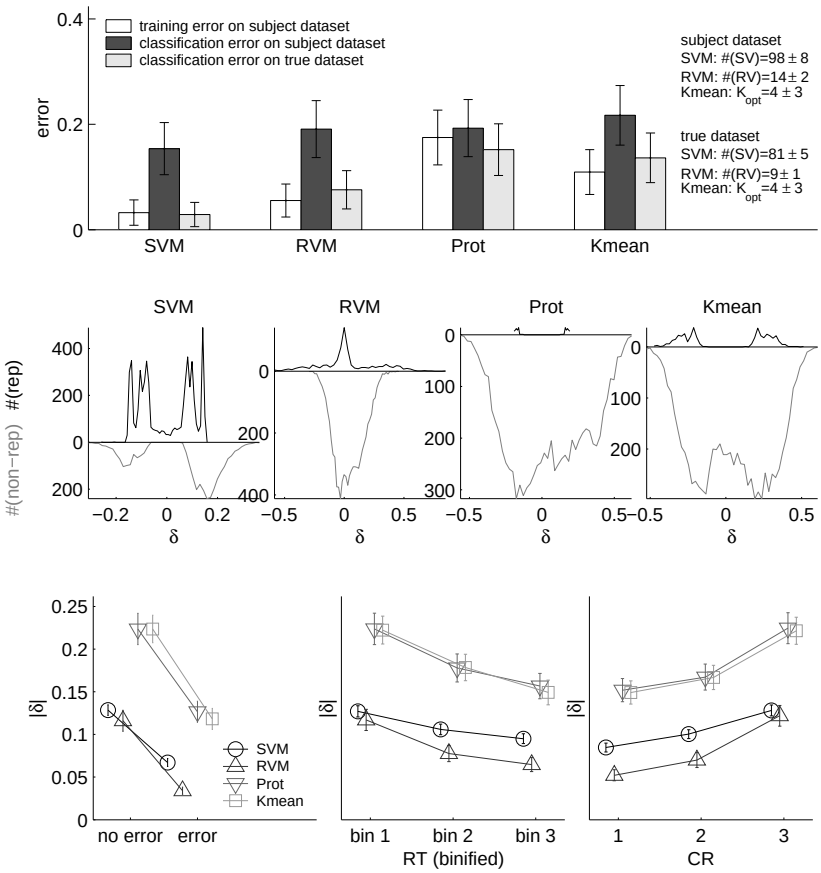

r=−0.60 ± 0.02
RT
|δ|
r=0.60 ± 0.02
CR
|δ|
r=−0.59 ± 0.02
subject error
|δ|
r=−0.65 ± 0.01
RT
|δ|
r=−0.56 ± 0.02
subject error
|δ|
SVM
RVM
Kmean
Prot
r=0.66 ± 0.01
CR
|δ|
r=0.25 ± 0.02
CR
|δ|
r=−0.06 ± 0.02
subject error
|δ|
r=−0.25 ± 0.02
RT
|δ|
r=−0.25 ± 0.02
subject error
|δ|
r=0.12 ± 0.02
CR
|δ|
r=−0.15 ± 0.02
RT
|δ|
SVM
RVM
Kmean
Prot
SVM r=−0.41 ± 0.02
Δ subject error
|δ|
RVM r=−0.45 ± 0.02
Δ subject error
|δ|
Prot r=−0.08 ± 0.02
Δ subject error
|δ|
Kmean r=−0.17 ± 0.02
Δ subject error
|δ|

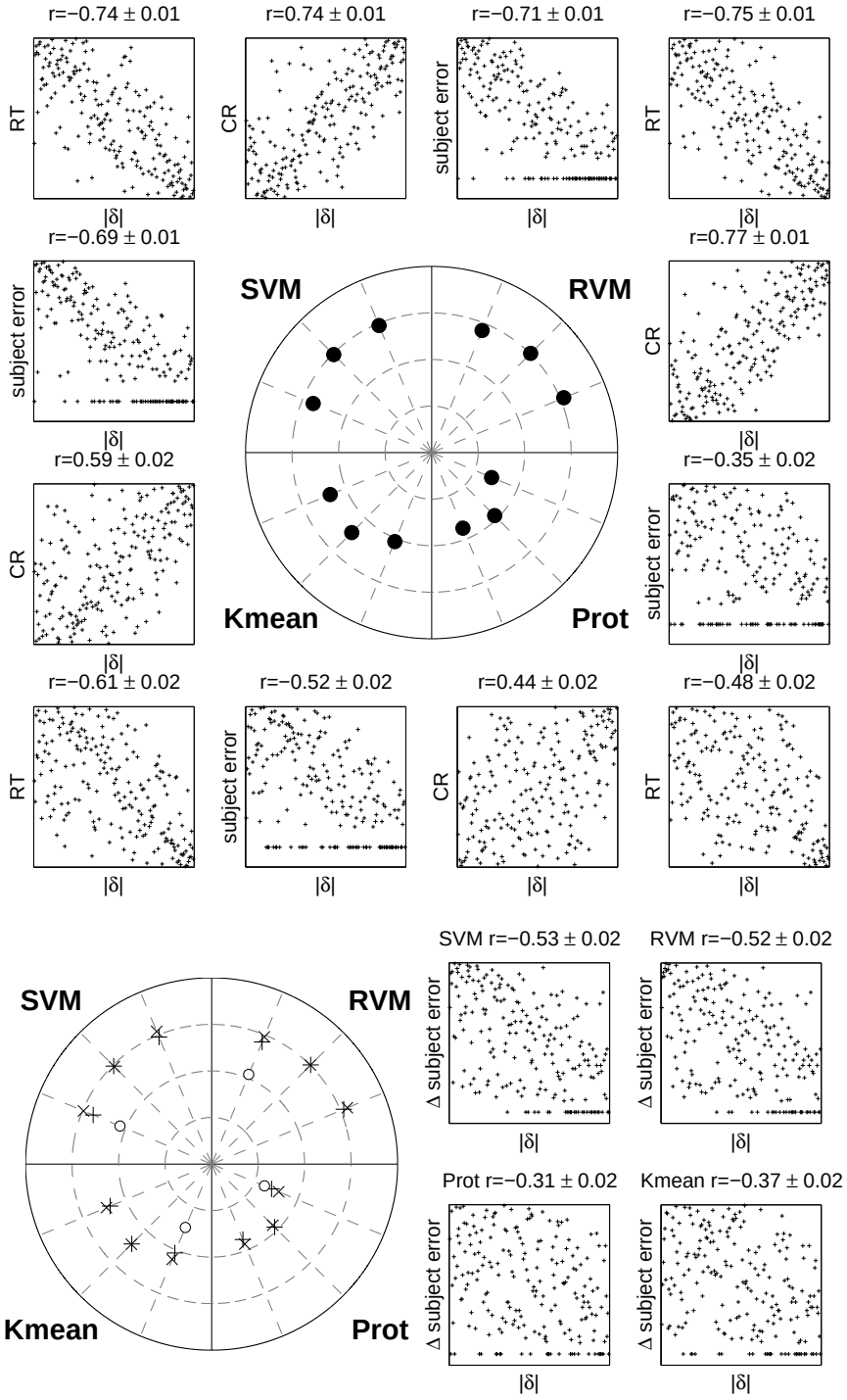
F.9 Kernel Map—Image Data



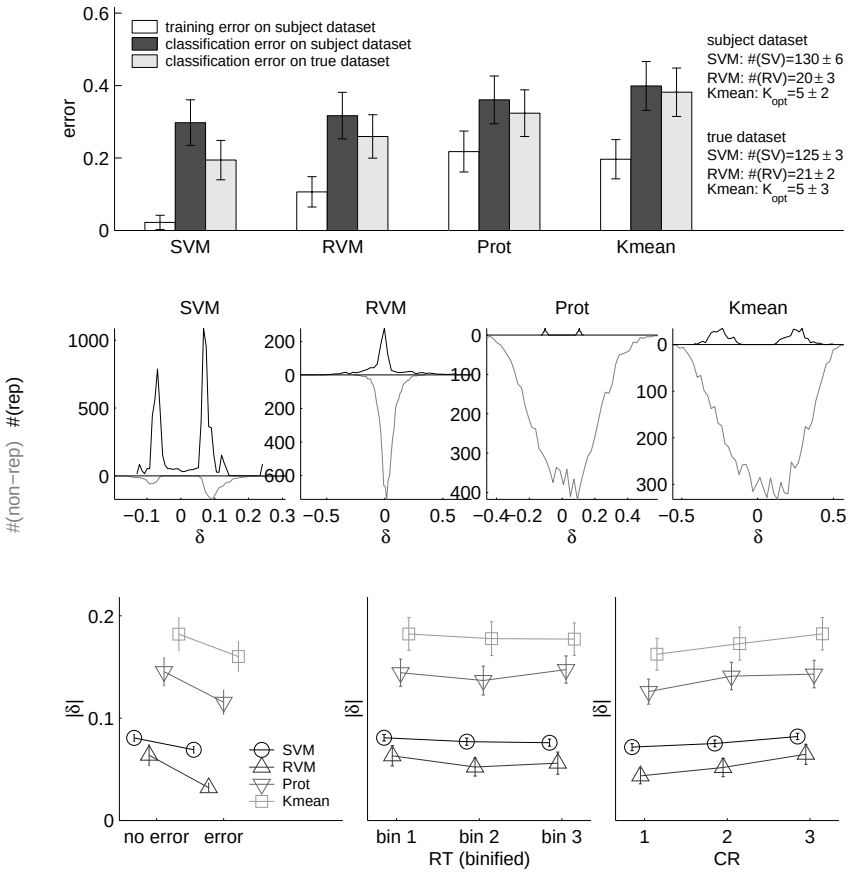

r=−0.72 ± 0.01
RT
|δ|
r=0.73 ± 0.01
CR
|δ|
r=−0.62 ± 0.02
subject error
|δ|
r=−0.65 ± 0.01
RT
|δ|
r=−0.68 ± 0.01
subject error
|δ|
SVM
RVM
Kmean
Prot
r=0.65 ± 0.01
CR
|δ|
r=0.37 ± 0.02
CR
|δ|
r=−0.13 ± 0.02
subject error
|δ|
r=−0.37 ± 0.02
RT
|δ|
r=−0.31 ± 0.02
subject error
|δ|
r=0.19 ± 0.02
CR
|δ|
r=−0.24 ± 0.02
RT
|δ|
SVM
RVM
Kmean
Prot
SVM r=−0.57 ± 0.02
Δ subject error
|δ|
RVM r=−0.50 ± 0.02
Δ subject error
|δ|
Prot r=−0.14 ± 0.02
Δ subject error
|δ|
Kmean r=−0.24 ± 0.02
Δ subject error
|δ|

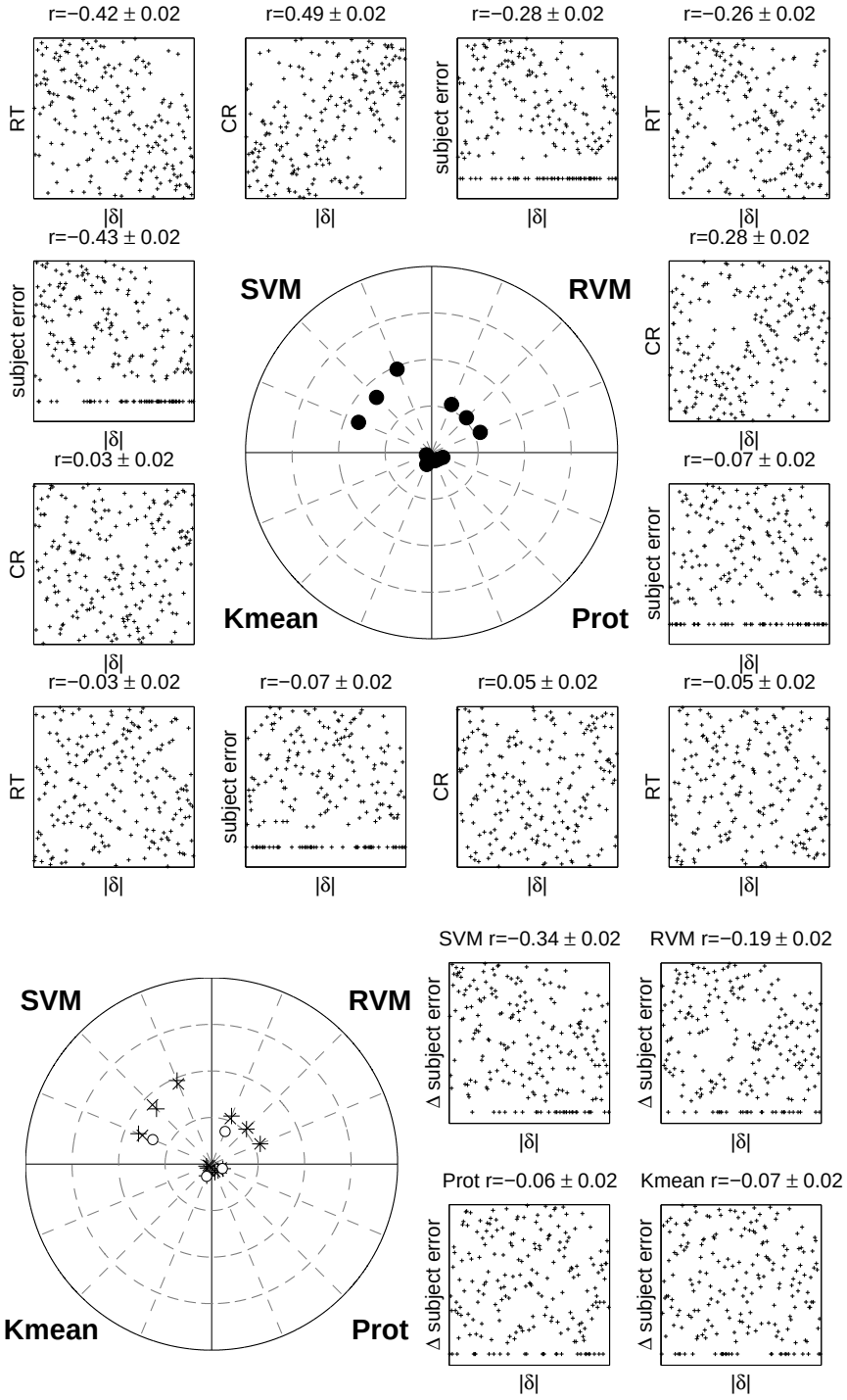
F.10 Kernel Map—Texture Data



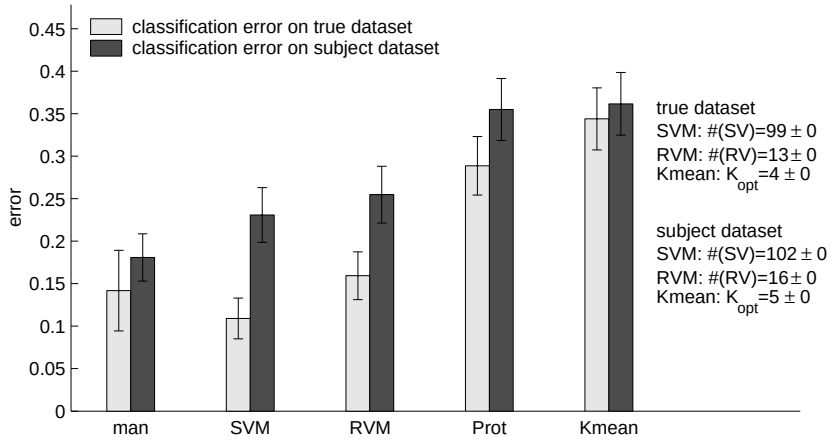

r=−0.74 ± 0.01
RT
|δ|
r=0.74 ± 0.01
CR
|δ|
r=−0.71 ± 0.01
subject error
|δ|
r=−0.75 ± 0.01
RT
|δ|
r=−0.69 ± 0.01
subject error
|δ|
SVM
RVM
Kmean
Prot
r=0.77 ± 0.01
CR
|δ|
r=0.59 ± 0.02
CR
|δ|
r=−0.35 ± 0.02
subject error
|δ|
r=−0.61 ± 0.02
RT
|δ|
r=−0.52 ± 0.02
subject error
|δ|
r=0.44 ± 0.02
CR
|δ|
r=−0.48 ± 0.02
RT
|δ|
SVM r=−0.53 ± 0.02
Δ subject error
|δ|
RVM r=−0.52 ± 0.02
Δ subject error
|δ|
SVM
RVM
Kmean
Prot
Prot r=−0.31 ± 0.02
Δ subject error
|δ|
Kmean r=−0.37 ± 0.02
Δ subject error
|δ|

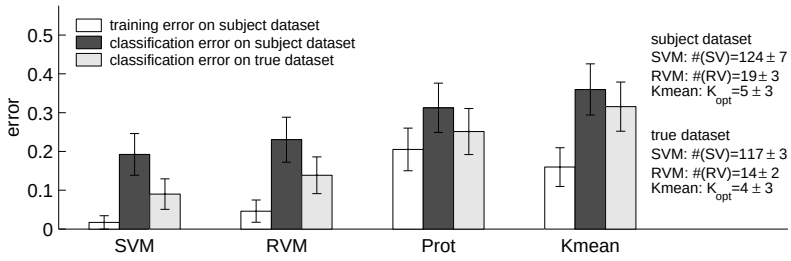
F.11 Kernel Map—Shape Data

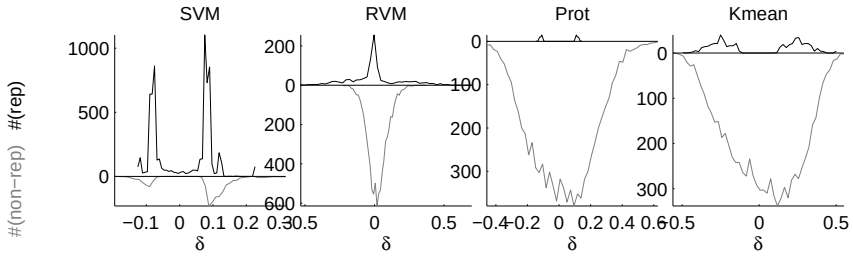


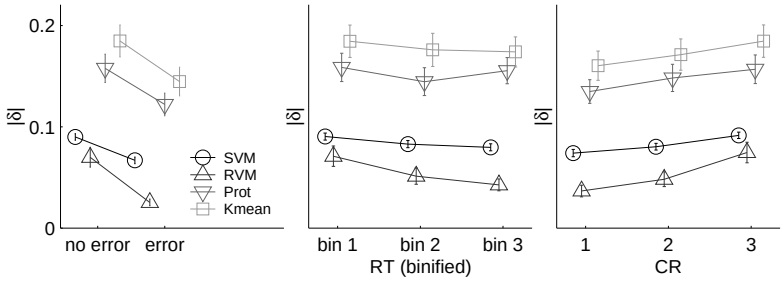

r=−0.42 ± 0.02
RT
|δ|
r=0.49 ± 0.02
CR
|δ|
r=−0.28 ± 0.02
subject error
|δ|
r=−0.26 ± 0.02
RT
|δ|
r=−0.43 ± 0.02
subject error
|δ|
SVM
RVM
Kmean
Prot
r=0.28 ± 0.02
CR
|δ|
r=0.03 ± 0.02
CR
|δ|
r=−0.07 ± 0.02
subject error
|δ|
r=−0.03 ± 0.02
RT
|δ|
r=−0.07 ± 0.02
subject error
|δ|
r=0.05 ± 0.02
CR
|δ|
r=−0.05 ± 0.02
RT
|δ|
SVM
RVM
Kmean
Prot
SVM r=−0.34 ± 0.02
Δ subject error
|δ|
RVM r=−0.19 ± 0.02
Δ subject error
|δ|
Prot r=−0.06 ± 0.02
Δ subject error
|δ|
Kmean r=−0.07 ± 0.02
Δ subject error
|δ|

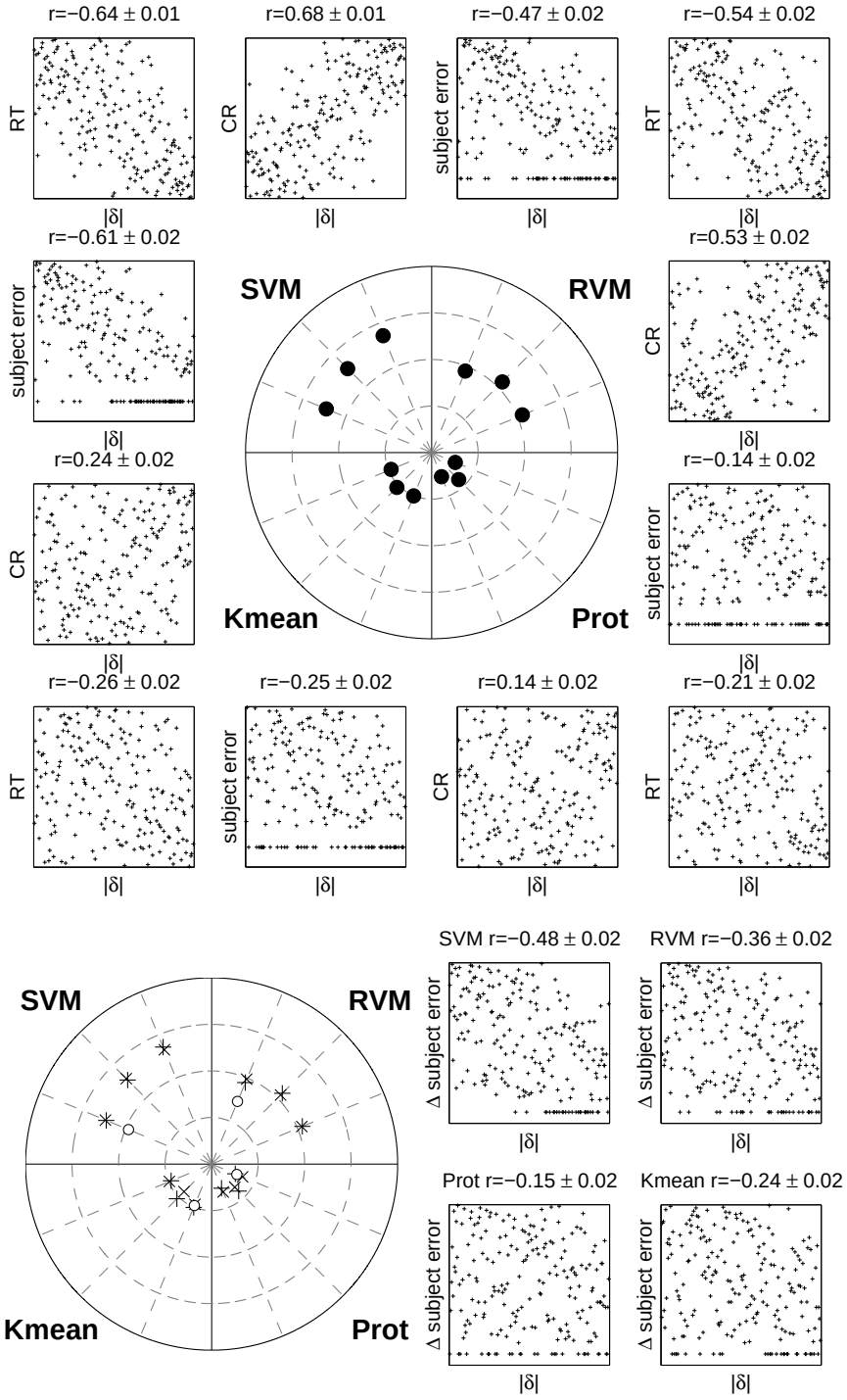
F.12 Kernel Map—Texture & Shape Data



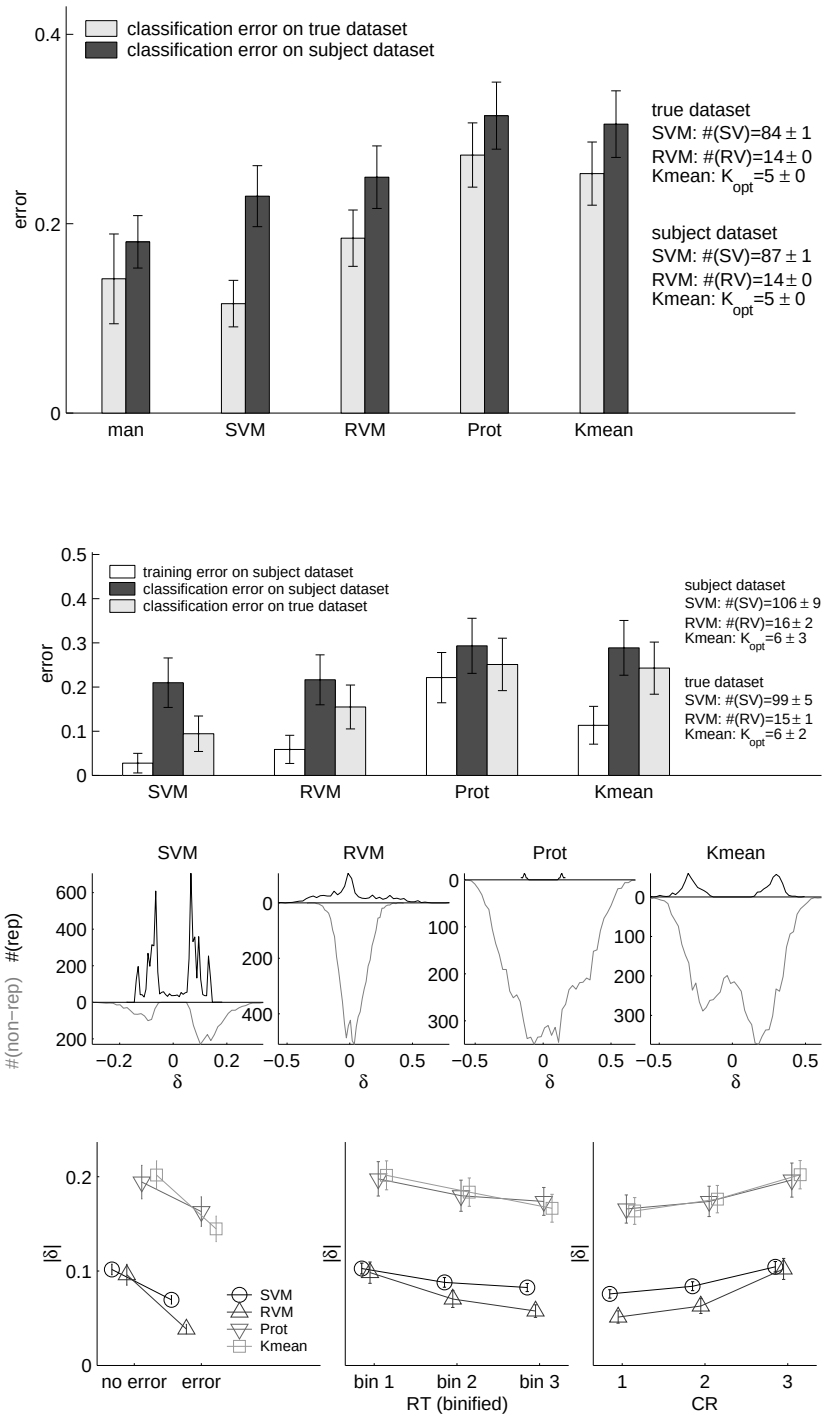


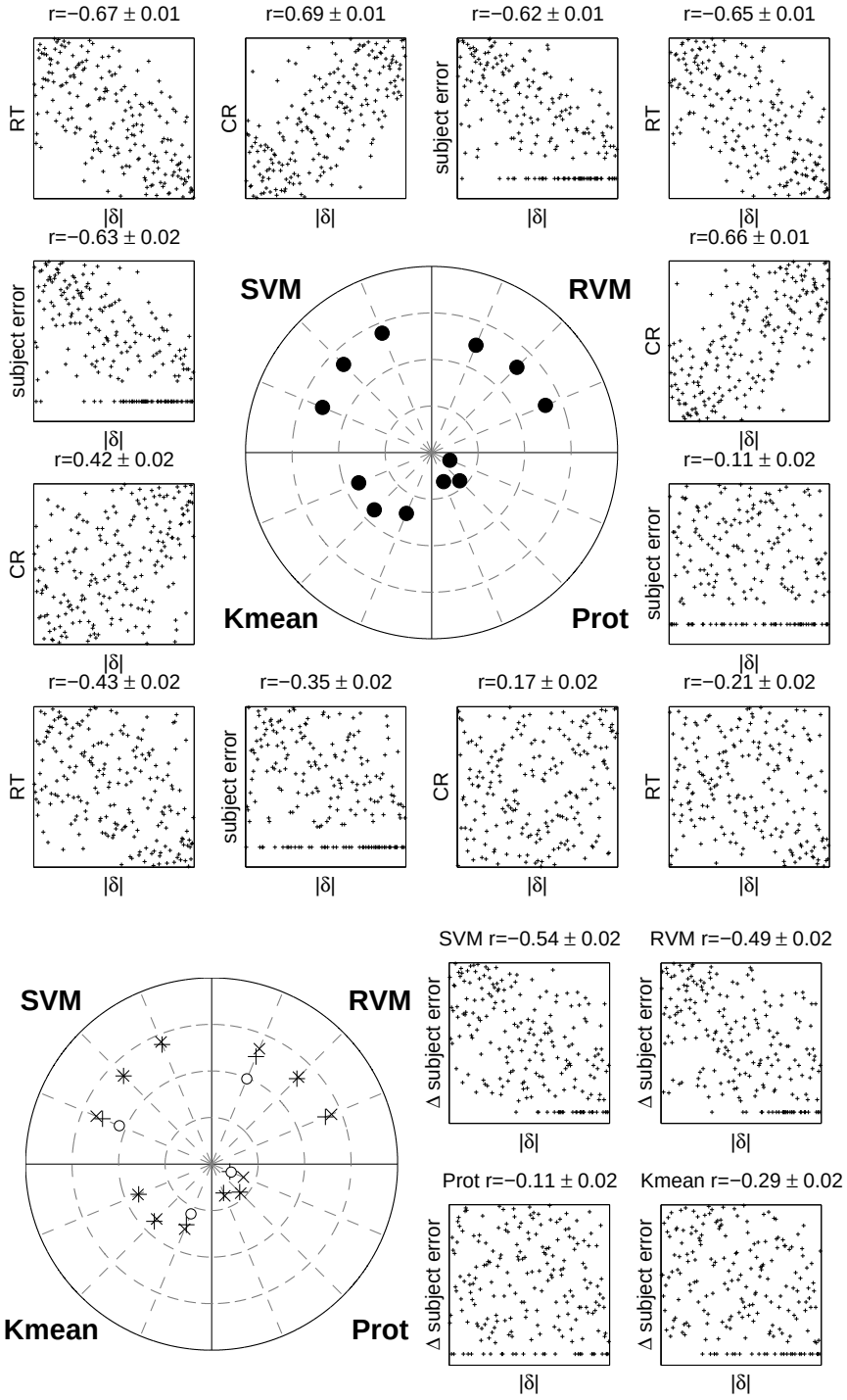




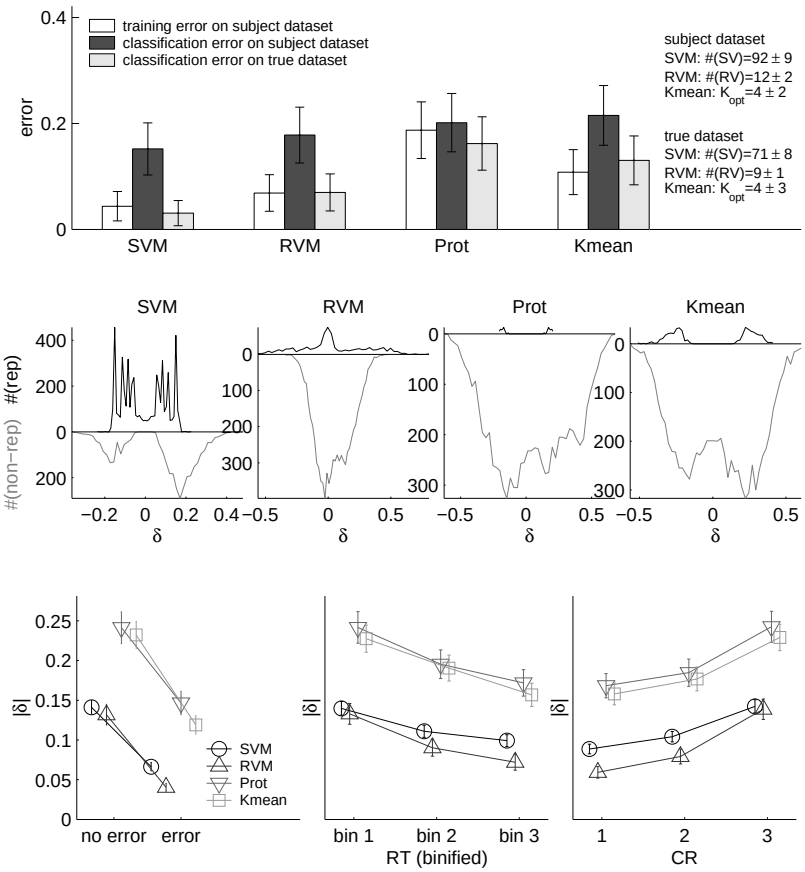

r=−0.64 ± 0.01
RT
|δ|
r=0.68 ± 0.01
CR
|δ|
r=−0.47 ± 0.02
subject error
|δ|
r=−0.54 ± 0.02
RT
|δ|
r=−0.61 ± 0.02
subject error
|δ|
SVM
RVM
Kmean
Prot
r=0.53 ± 0.02
CR
|δ|
r=0.24 ± 0.02
CR
|δ|
r=−0.14 ± 0.02
subject error
|δ|
r=−0.26 ± 0.02
RT
|δ|
r=−0.25 ± 0.02
subject error
|δ|
r=0.14 ± 0.02
CR
|δ|
r=−0.21 ± 0.02
RT
|δ|
SVM
RVM
Kmean
Prot
SVM r=−0.48 ± 0.02
Δ subject error
|δ|
RVM r=−0.36 ± 0.02
Δ subject error
|δ|
Prot r=−0.15 ± 0.02
Δ subject error
|δ|
Kmean r=−0.24 ± 0.02
Δ subject error
|δ|

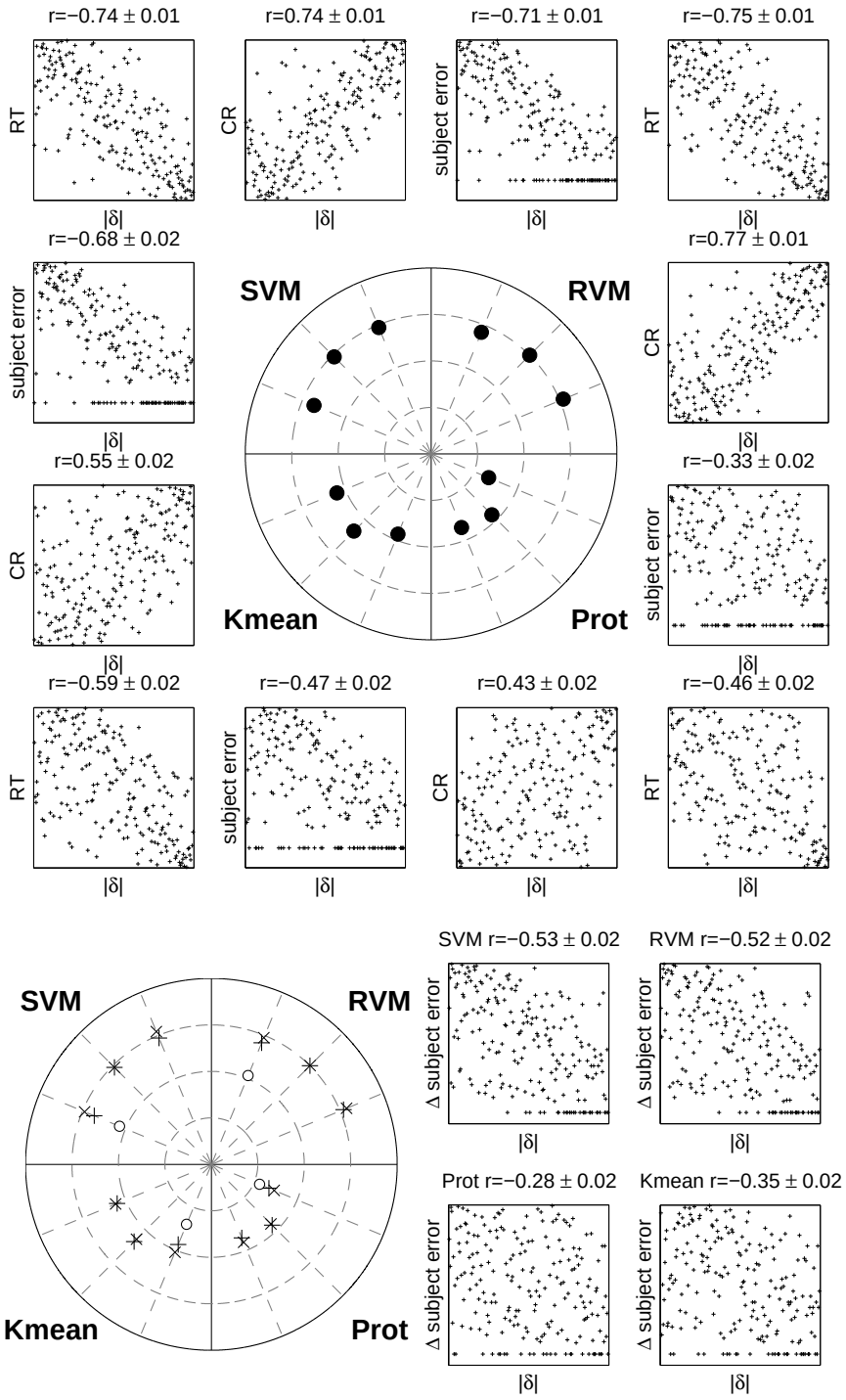
F.13 ICA I—Image Data



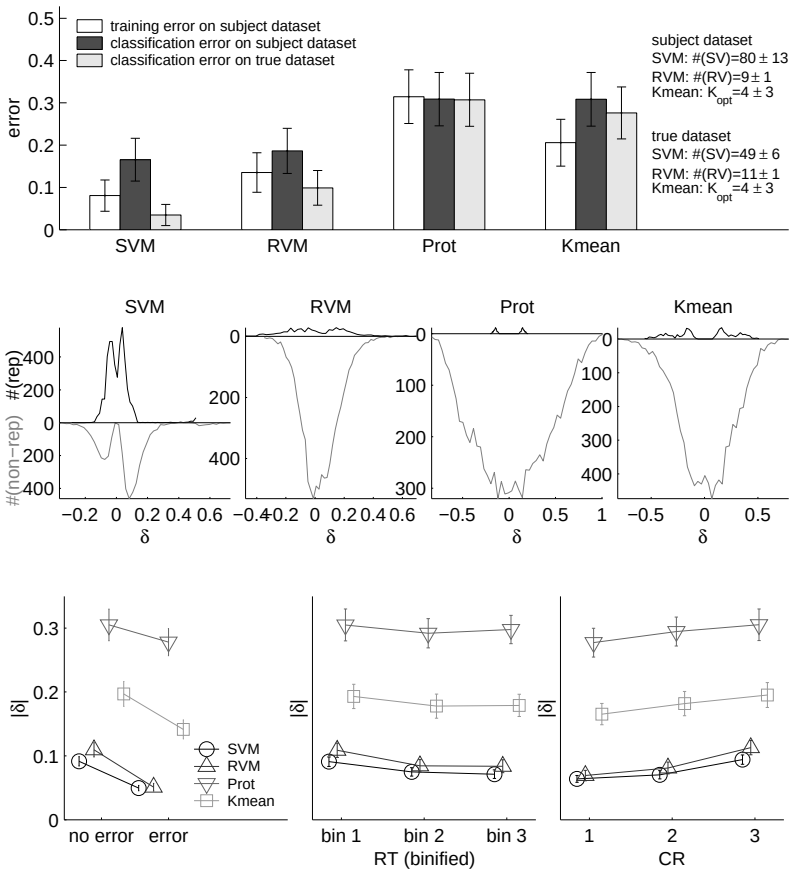

r=−0.67 ± 0.01
RT
|δ|
r=0.69 ± 0.01
CR
|δ|
r=−0.62 ± 0.01
subject error
|δ|
r=−0.65 ± 0.01
RT
|δ|
r=−0.63 ± 0.02
subject error
|δ|
SVM
RVM
Kmean
Prot
r=0.66 ± 0.01
CR
|δ|
r=0.42 ± 0.02
CR
|δ|
r=−0.11 ± 0.02
subject error
|δ|
r=−0.43 ± 0.02
RT
|δ|
r=−0.35 ± 0.02
subject error
|δ|
r=0.17 ± 0.02
CR
|δ|
r=−0.21 ± 0.02
RT
|δ|
SVM
RVM
Kmean
Prot
SVM r=−0.54 ± 0.02
Δ subject error
|δ|
RVM r=−0.49 ± 0.02
Δ subject error
|δ|
Prot r=−0.11 ± 0.02
Δ subject error
|δ|
Kmean r=−0.29 ± 0.02
Δ subject error
|δ|

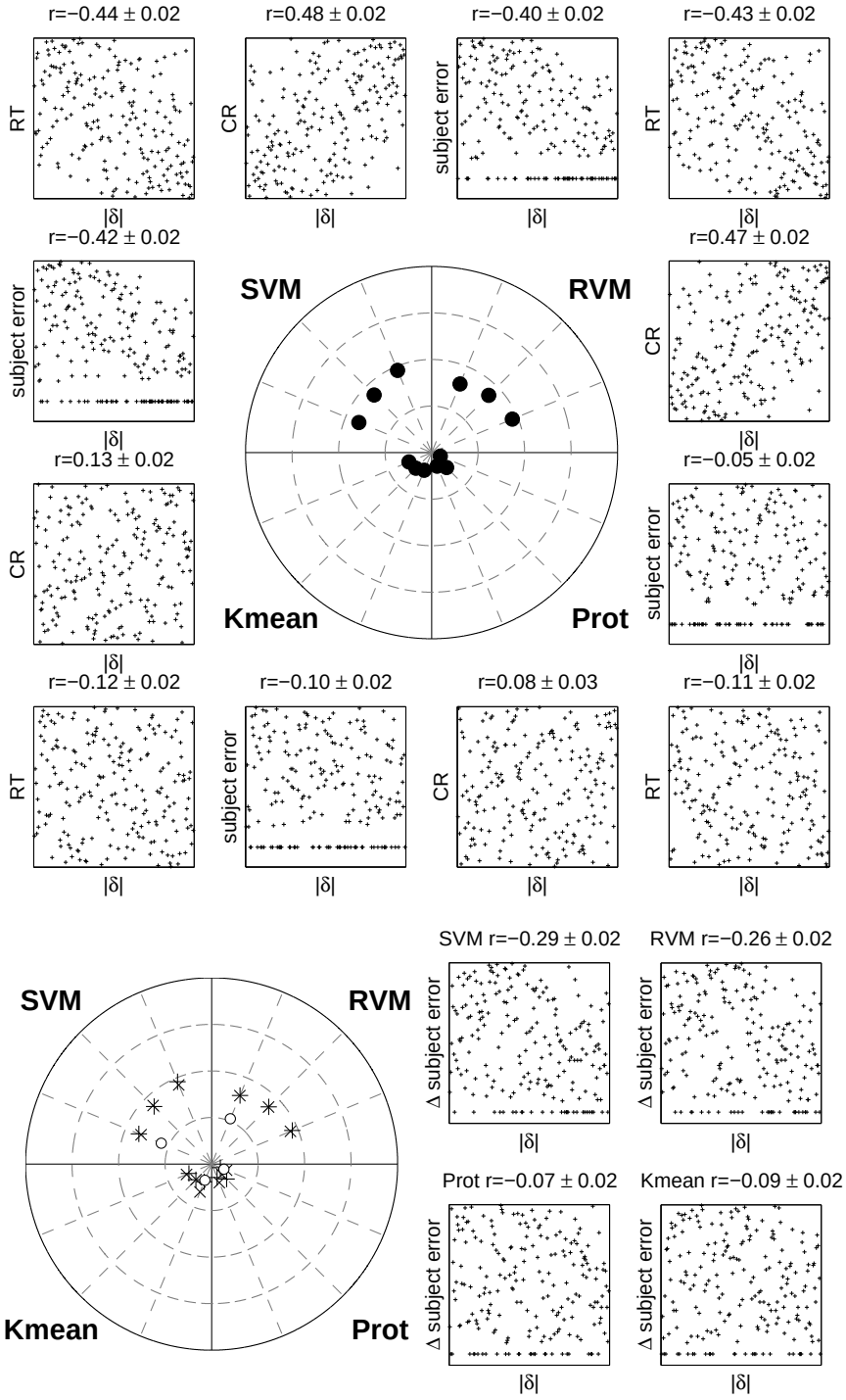
F.14 ICA I—Texture Data



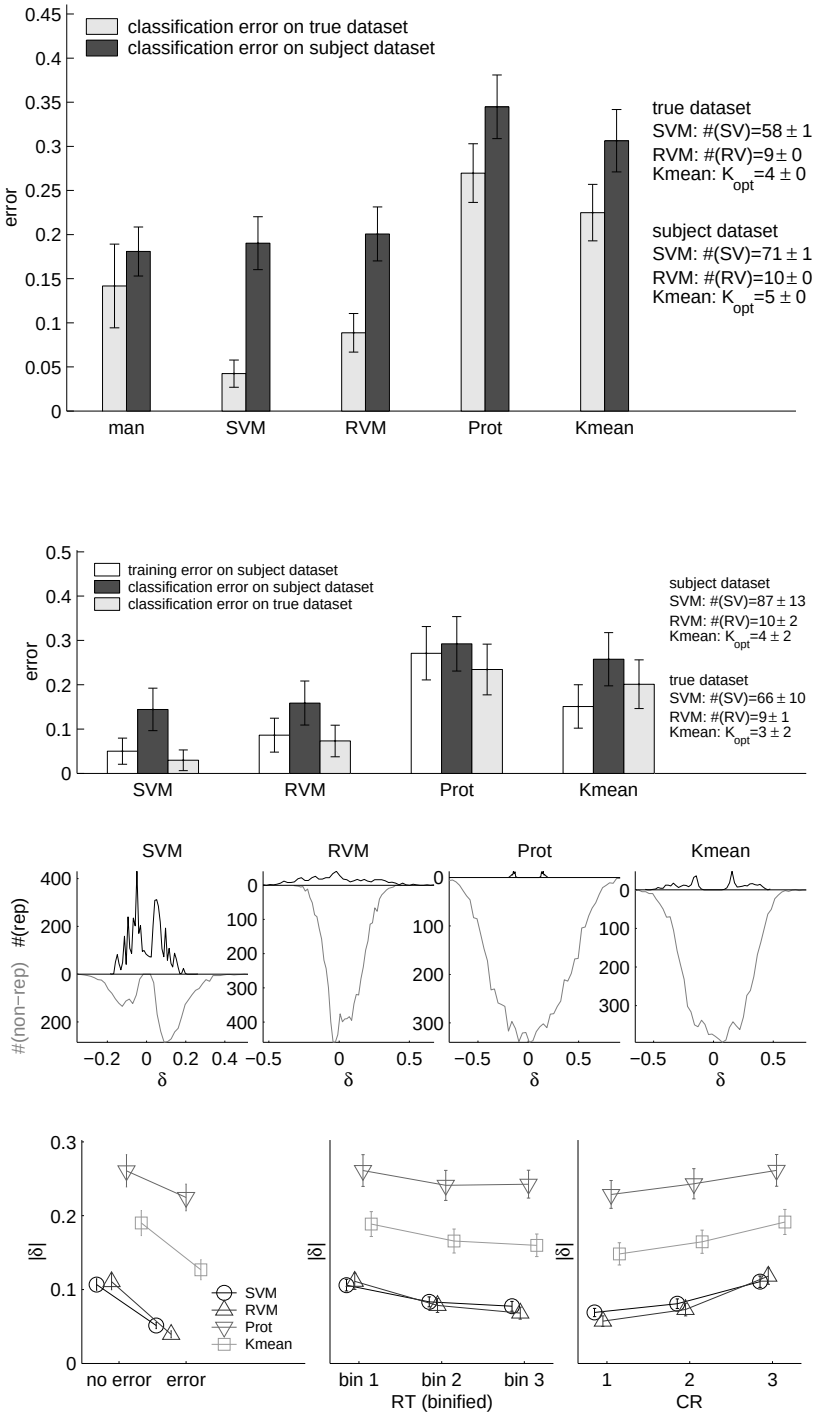

r=−0.74 ± 0.01
RT
|δ|
r=0.74 ± 0.01
CR
|δ|
r=−0.71 ± 0.01
subject error
|δ|
r=−0.75 ± 0.01
RT
|δ|
r=−0.68 ± 0.02
subject error
|δ|
SVM
RVM
r=0.77 ± 0.01
CR
|δ|
r=0.55 ± 0.02
CR
|δ|
Kmean
Prot
r=−0.33 ± 0.02
subject error
|δ|
r=−0.59 ± 0.02
RT
|δ|
r=−0.47 ± 0.02
subject error
|δ|
r=0.43 ± 0.02
CR
|δ|
r=−0.46 ± 0.02
RT
|δ|
SVM
RVM
Kmean
Prot
SVM r=−0.53 ± 0.02
Δ subject error
|δ|
RVM r=−0.52 ± 0.02
Δ subject error
|δ|
Prot r=−0.28 ± 0.02
Δ subject error
|δ|
Kmean r=−0.35 ± 0.02
Δ subject error
|δ|

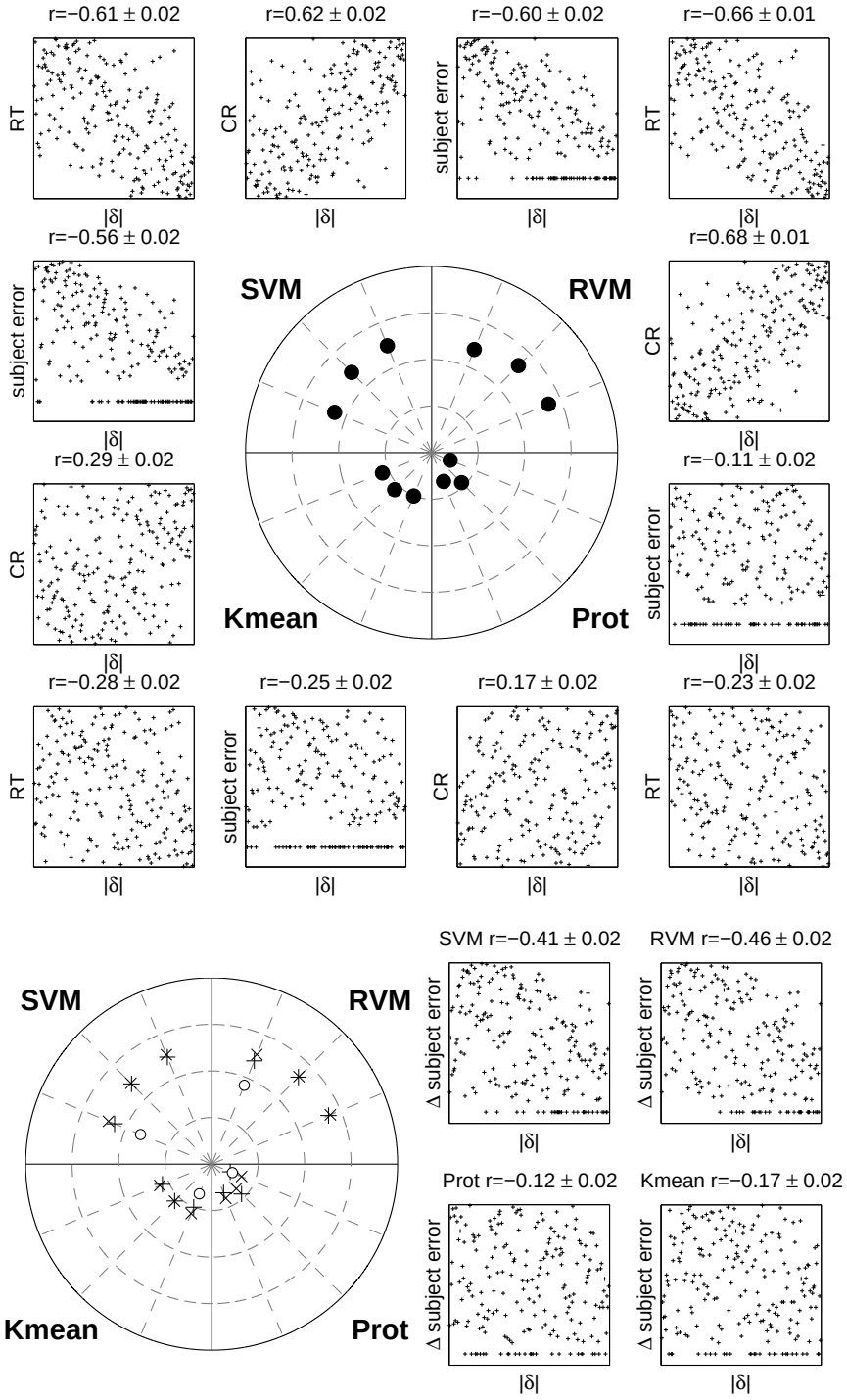
F.15 ICA I—Shape Data



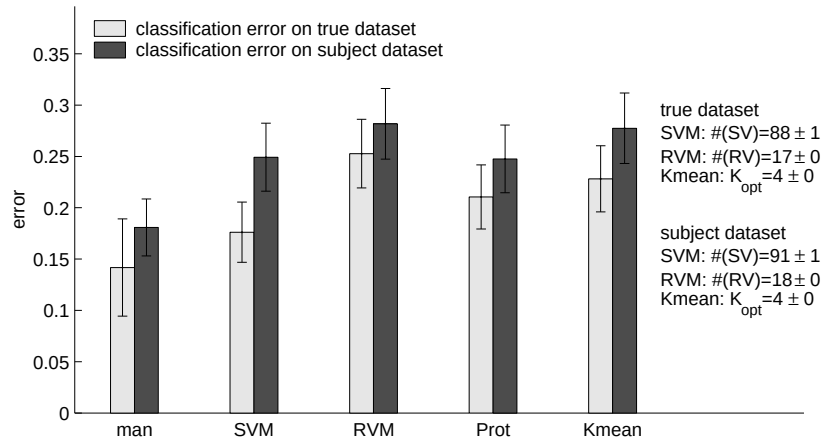

r=−0.44 ± 0.02
RT
|δ|
r=0.48 ± 0.02
CR
|δ|
r=−0.40 ± 0.02
subject error
|δ|
r=−0.43 ± 0.02
RT
|δ|
r=−0.42 ± 0.02
subject error
|δ|
SVM
RVM
Kmean
Prot
r=0.47 ± 0.02
CR
|δ|
r=0.13 ± 0.02
CR
|δ|
r=−0.05 ± 0.02
subject error
|δ|
r=−0.12 ± 0.02
RT
|δ|
r=−0.10 ± 0.02
subject error
|δ|
r=0.08 ± 0.03
CR
|δ|
r=−0.11 ± 0.02
RT
|δ|
SVM
RVM
Kmean
Prot
SVM r=−0.29 ± 0.02
Δ subject error
|δ|
RVM r=−0.26 ± 0.02
Δ subject error
|δ|
Prot r=−0.07 ± 0.02
Δ subject error
|δ|
Kmean r=−0.09 ± 0.02
Δ subject error
|δ|

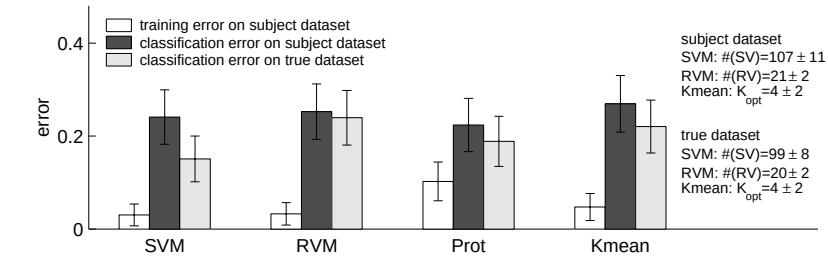
F.16 ICA I—Texture & Shape Data

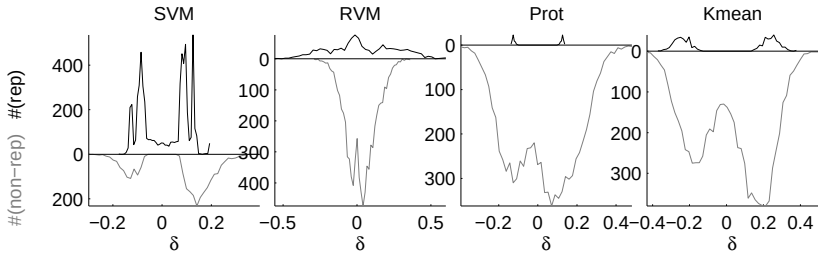


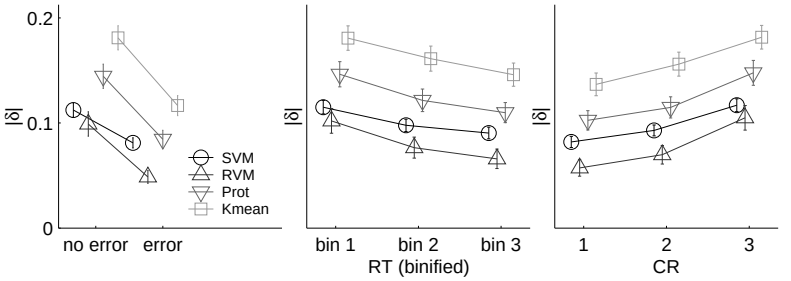

r=−0.61 ± 0.02
RT
|δ|
r=0.62 ± 0.02
CR
|δ|
r=−0.60 ± 0.02
subject error
|δ|
r=−0.66 ± 0.01
RT
|δ|
r=−0.56 ± 0.02
subject error
|δ|
SVM
RVM
Kmean
Prot
r=0.68 ± 0.01
CR
|δ|
r=0.29 ± 0.02
CR
|δ|
r=−0.11 ± 0.02
subject error
|δ|
r=−0.28 ± 0.02
RT
|δ|
r=−0.25 ± 0.02
subject error
|δ|
r=0.17 ± 0.02
CR
|δ|
r=−0.23 ± 0.02
RT
|δ|
SVM
RVM
Kmean
Prot
SVM r=−0.41 ± 0.02
Δ subject error
|δ|
RVM r=−0.46 ± 0.02
Δ subject error
|δ|
Prot r=−0.12 ± 0.02
Δ subject error
|δ|
Kmean r=−0.17 ± 0.02
Δ subject error
|δ|

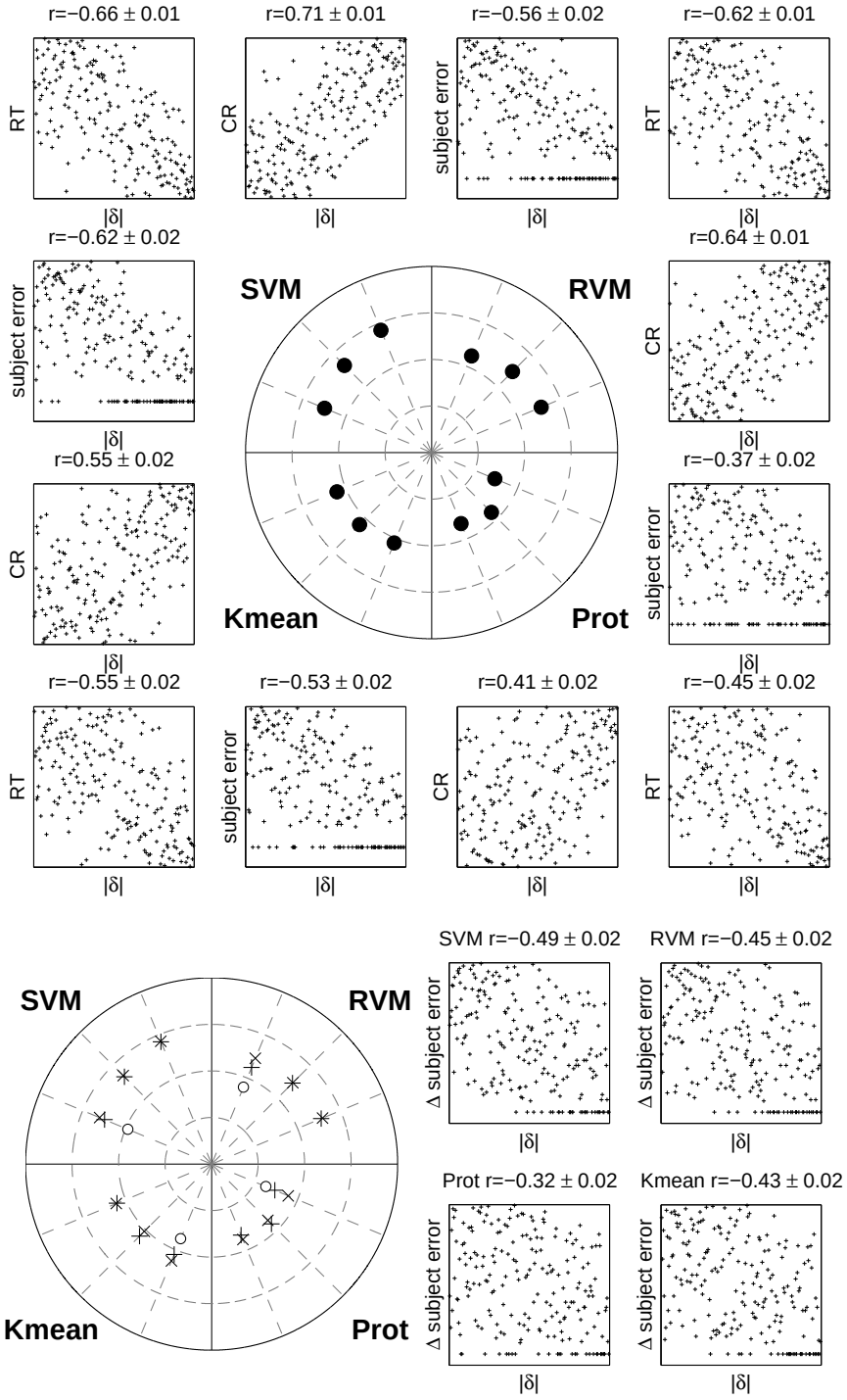
F.17 NMF—Image Data



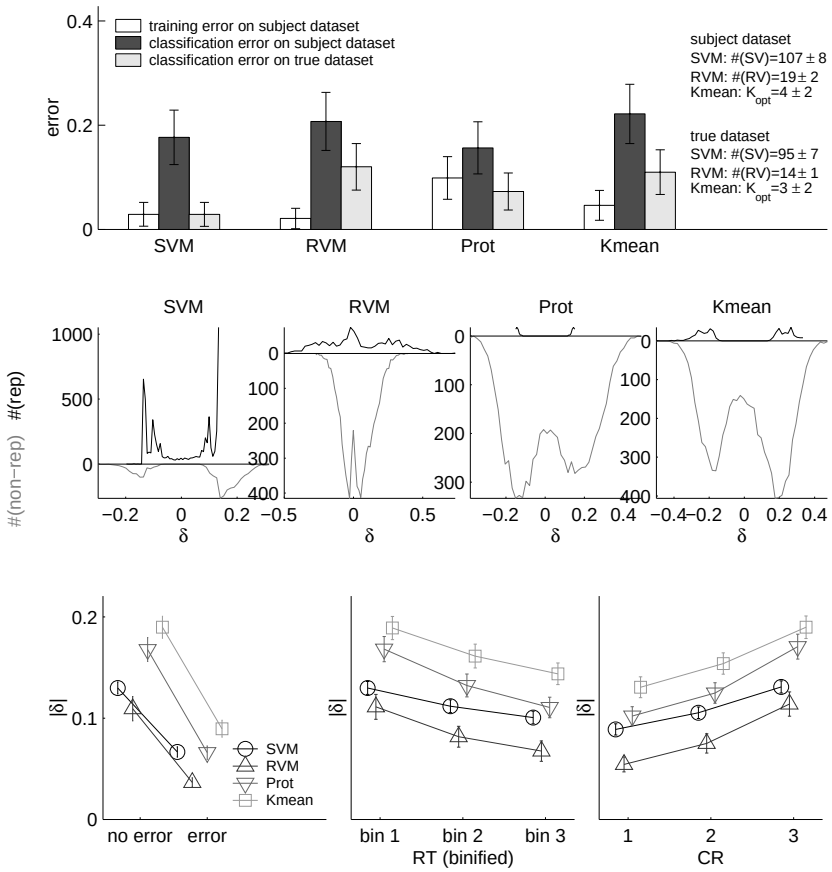


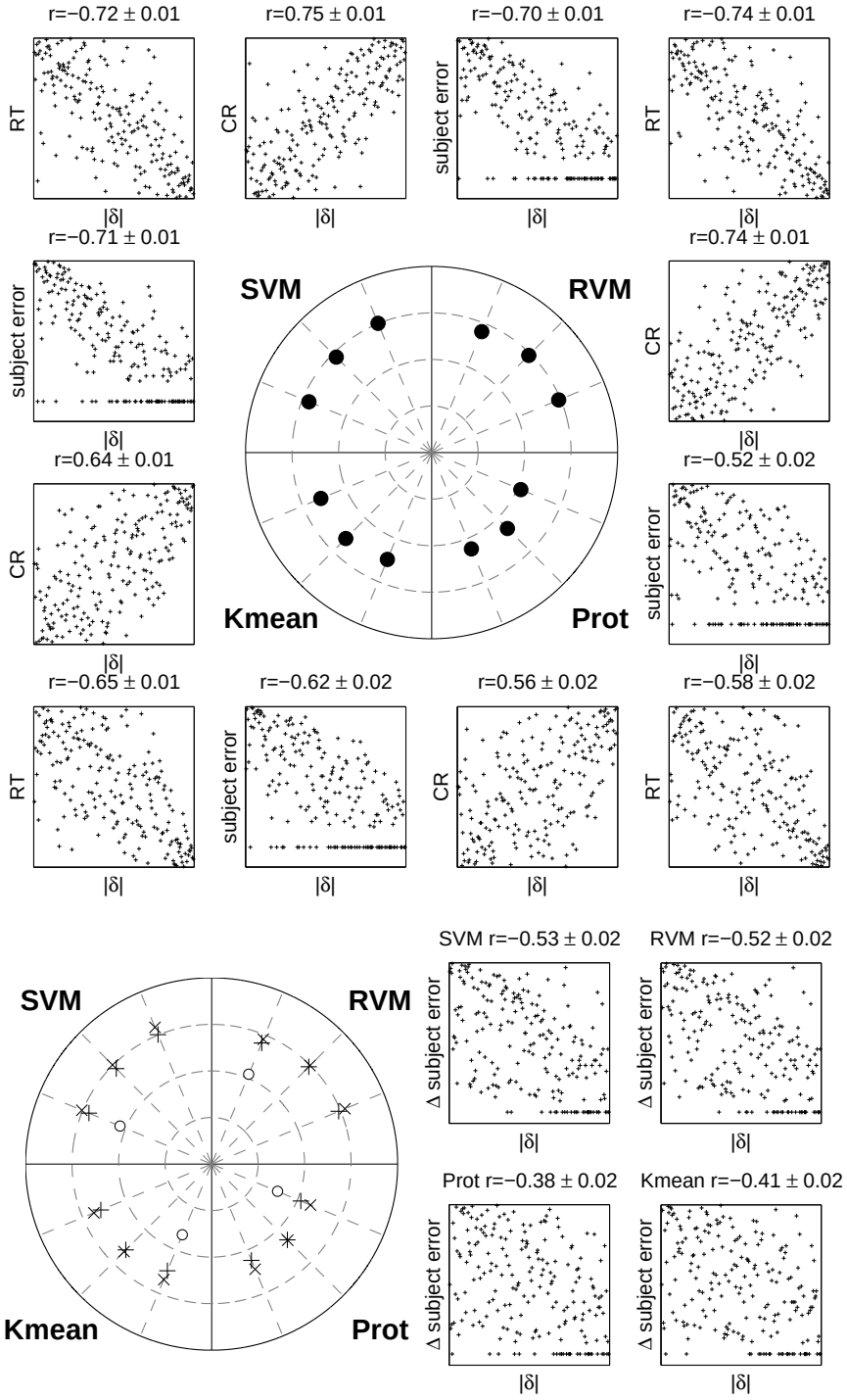




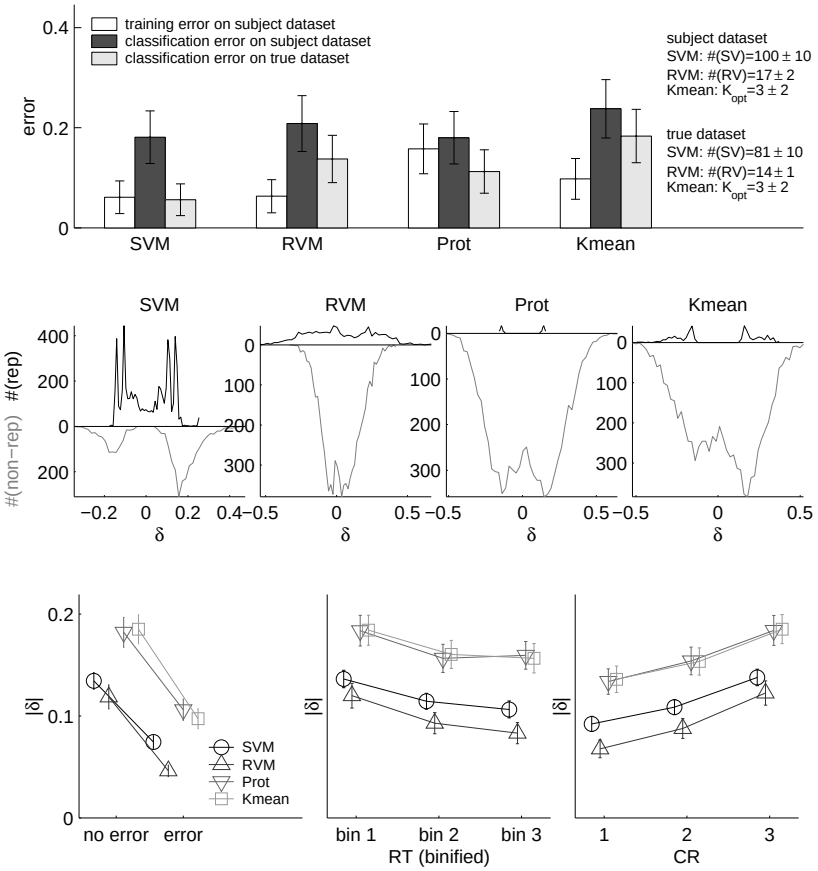

r=−0.66 ± 0.01
RT
|δ|
r=0.71 ± 0.01
CR
|δ|
r=−0.56 ± 0.02
subject error
|δ|
r=−0.62 ± 0.01
RT
|δ|
r=−0.62 ± 0.02
subject error
|δ|
SVM
RVM
Kmean
Prot
r=0.64 ± 0.01
CR
|δ|
r=0.55 ± 0.02
CR
|δ|
r=−0.37 ± 0.02
subject error
|δ|
r=−0.55 ± 0.02
RT
|δ|
r=−0.53 ± 0.02
subject error
|δ|
r=0.41 ± 0.02
CR
|δ|
r=−0.45 ± 0.02
RT
|δ|
SVM r=−0.49 ± 0.02
Δ subject error
|δ|
RVM r=−0.45 ± 0.02
Δ subject error
|δ|
SVM
RVM
Kmean
Prot
Prot r=−0.32 ± 0.02
Δ subject error
|δ|
Kmean r=−0.43 ± 0.02
Δ subject error
|δ|

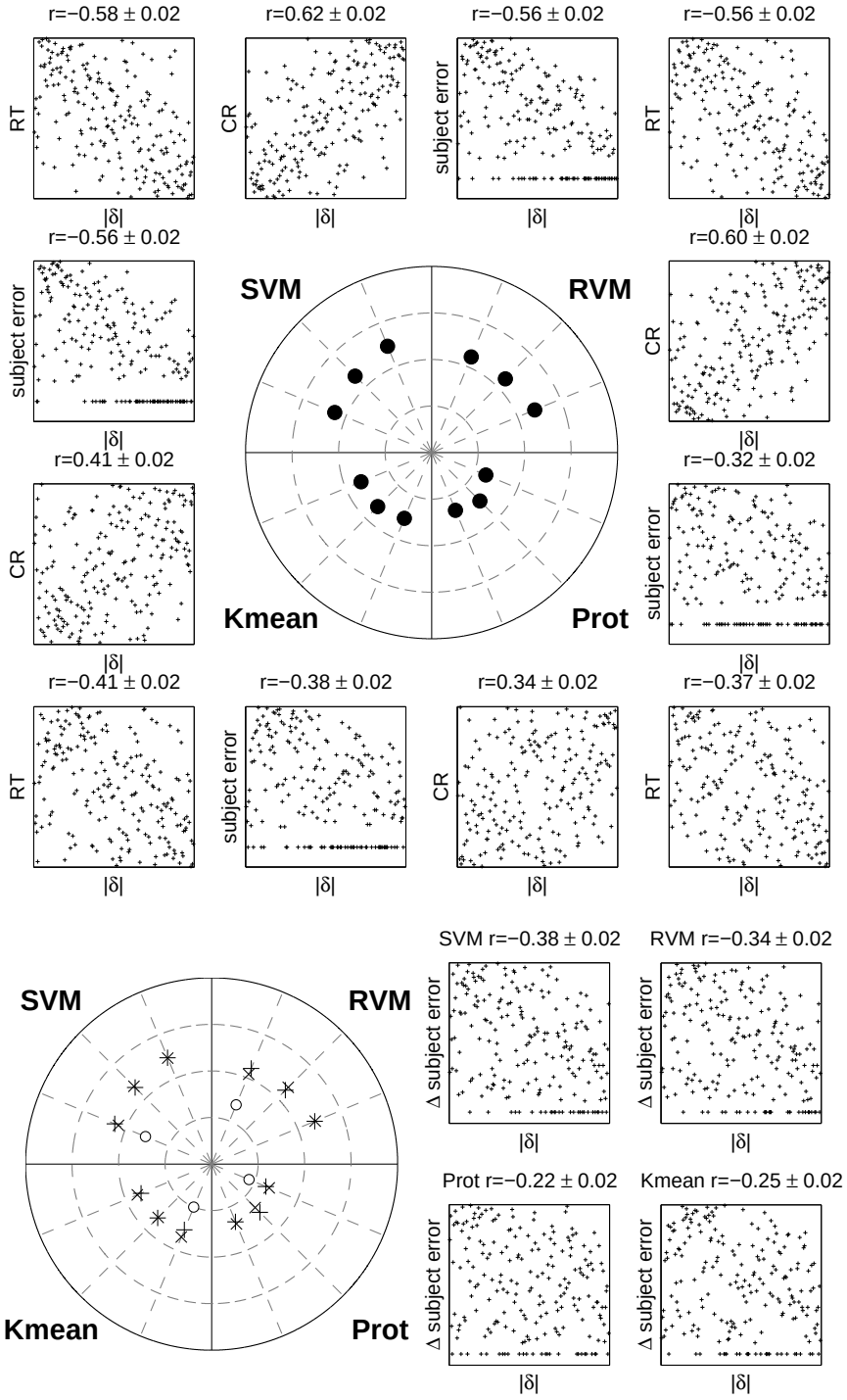
F.18 NMF—Texture Data



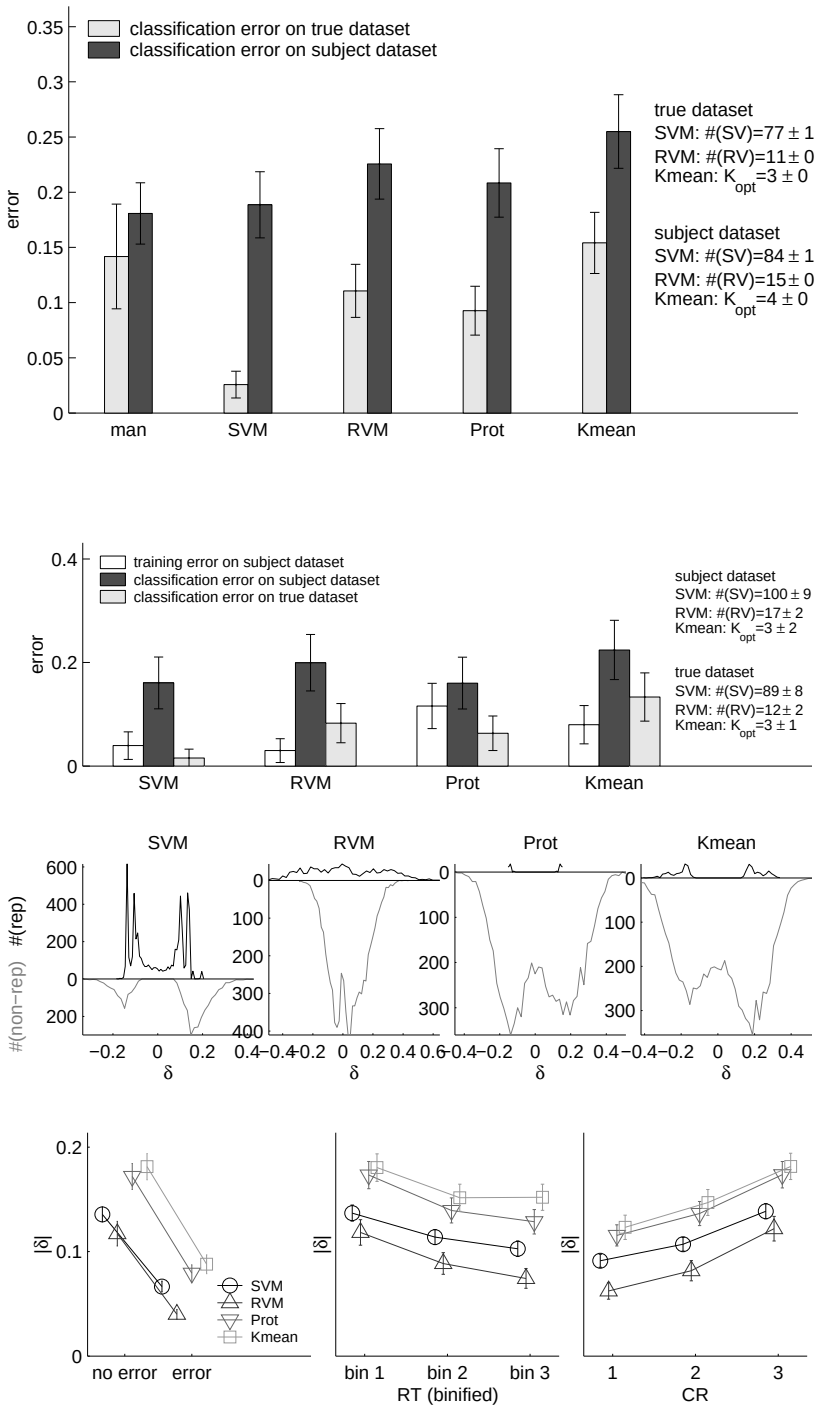

r=−0.72 ± 0.01
RT
|δ|
r=0.75 ± 0.01
CR
|δ|
r=−0.70 ± 0.01
subject error
|δ|
r=−0.74 ± 0.01
RT
|δ|
r=−0.71 ± 0.01
subject error
|δ|
SVM
RVM
Kmean
Prot
r=0.74 ± 0.01
CR
|δ|
r=0.64 ± 0.01
CR
|δ|
r=−0.52 ± 0.02
subject error
|δ|
r=−0.65 ± 0.01
RT
|δ|
r=−0.62 ± 0.02
subject error
|δ|
r=0.56 ± 0.02
CR
|δ|
r=−0.58 ± 0.02
RT
|δ|
SVM r=−0.53 ± 0.02
Δ subject error
|δ|
RVM r=−0.52 ± 0.02
Δ subject error
|δ|
SVM
RVM
Kmean
Prot
Prot r=−0.38 ± 0.02
Δ subject error
|δ|
Kmean r=−0.41 ± 0.02
Δ subject error
|δ|

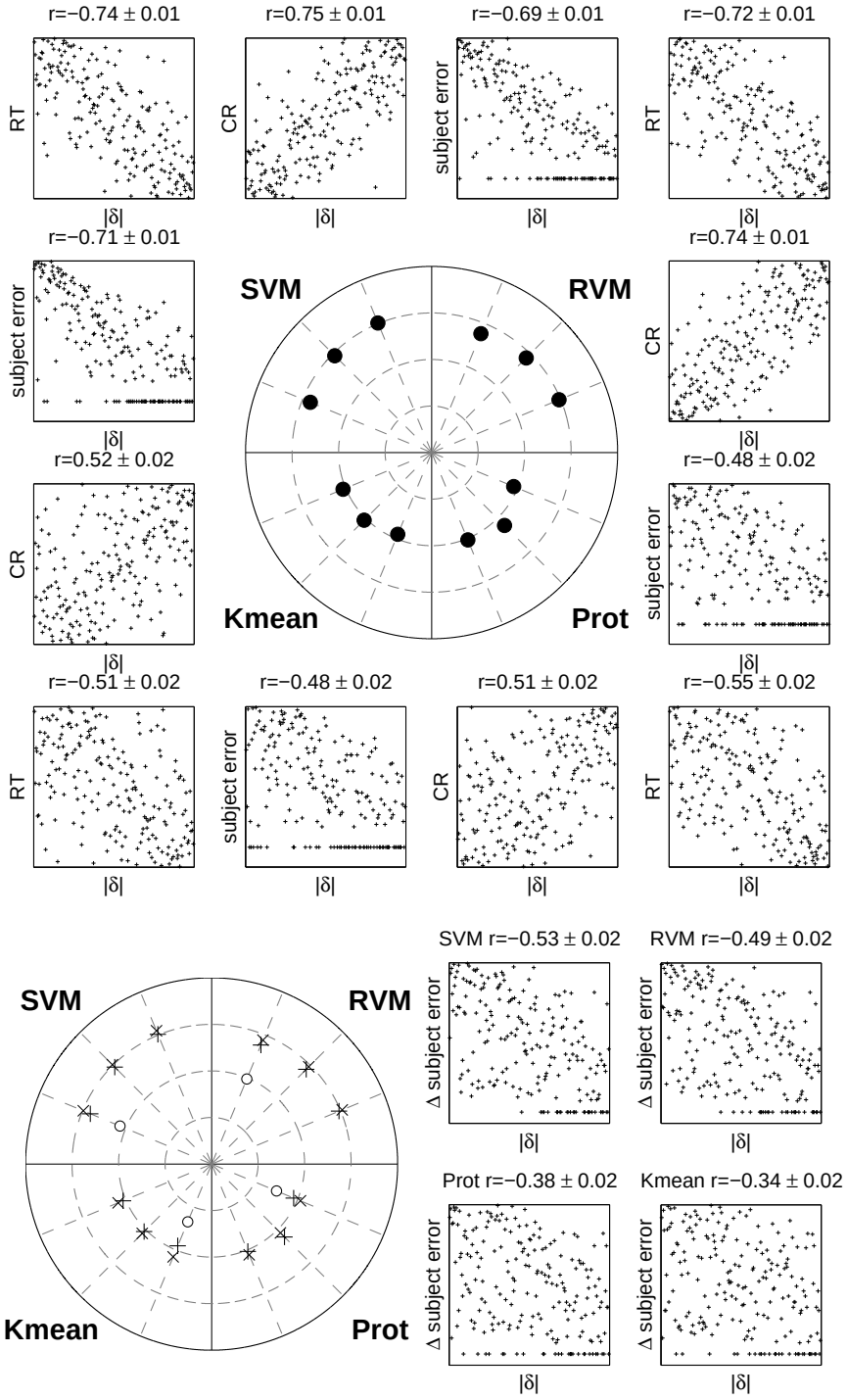
F.19 NMF—Shape Data




r=−0.58 ± 0.02
RT
|δ|
r=0.62 ± 0.02
CR
|δ|
r=−0.56 ± 0.02
subject error
|δ|
r=−0.56 ± 0.02
RT
|δ|
r=−0.56 ± 0.02
subject error
|δ|
SVM
RVM
Kmean
Prot
r=0.60 ± 0.02
CR
|δ|
r=0.41 ± 0.02
CR
|δ|
r=−0.32 ± 0.02
subject error
|δ|
r=−0.41 ± 0.02
RT
|δ|
r=−0.38 ± 0.02
subject error
|δ|
r=0.34 ± 0.02
CR
|δ|
r=−0.37 ± 0.02
RT
|δ|
SVM
RVM
Kmean
Prot
SVM r=−0.38 ± 0.02
Δ subject error
|δ|
RVM r=−0.34 ± 0.02
Δ subject error
|δ|
Prot r=−0.22 ± 0.02
Δ subject error
|δ|
Kmean r=−0.25 ± 0.02
Δ subject error
|δ|

F.20 NMF—Texture & Shape Data




r=−0.74 ± 0.01
RT
|δ|
r=0.75 ± 0.01
CR
|δ|
r=−0.69 ± 0.01
subject error
|δ|
r=−0.72 ± 0.01
RT
|δ|
r=−0.71 ± 0.01
subject error
|δ|
SVM
RVM
Kmean
Prot
r=0.74 ± 0.01
CR
|δ|
r=0.52 ± 0.02
CR
|δ|
r=−0.48 ± 0.02
subject error
|δ|
r=−0.51 ± 0.02
RT
|δ|
r=−0.48 ± 0.02
subject error
|δ|
r=0.51 ± 0.02
CR
|δ|
r=−0.55 ± 0.02
RT
|δ|
SVM r=−0.53 ± 0.02
Δ subject error
|δ|
RVM r=−0.49 ± 0.02
Δ subject error
|δ|
SVM
RVM
Kmean
Prot
Prot r=−0.38 ± 0.02
Δ subject error
|δ|
Kmean r=−0.34 ± 0.02
Δ subject error
|δ|