Semi-Parametric Reduction of Dimensionality

Statistical Detection of Rare Events in Molecular Dynamics

by

Elmar Diederichs, Dipl. Physics, M.A. Philosophy

Thesis

Submitted by Elmar Diederichs

for fulfillment of the Requirements for the Degree of

Doctor of Natural Sciences

as approved dissertation

Supervisors: Prof. Dr. Christof Schütte

Prof. Dr. Vladimir Spokoiny

Department of Mathematics and Computer Science
Freie Universität Berlin

June, 2009

Semi-Parametric Reduction of Dimensionality

Statistical Detection of Rare Events in Molecular Dynamics

Elmar Diederichs, Dipl. Physics, M.A. Philosophy
diederic@math.fu-berlin.de
Freie Universität Berlin, 2009

Supervisor: Prof. Dr. Christof Schütte
schuette@mi.fu-berlin.de
Supervisor: Prof. Dr. Vladimir Spokoiny
spokoiny@wias-berlin.de

date of disputation: 31.7.2009

Abstract

Concerning the analysis of large molecular systems increasing amounts of simulation data and growing dimensionality have led to the demand of data-driven approaches to extract physically interpretable information from large data sets. Hence a mapping to a low dimensional manifold, representing the essential degrees of freedom of a molecular system is sought. A general obstacle to such an analysis is the curse of dimensionality. This thesis is motivated by the fact that most dimension reduction methods are either not reliable in dimensionality regimes of realistic biomolecular systems or restricted to data sets with special features. On the one hand the aim is to develop an unsupervised linear feature extraction method, that allows to extract any multimodal distributed component to a given high dimensional data density. On the other hand the development of a geometric approach to the analysis of the large scale dynamical behavior of biological active molecules is intended. To this end a very general semi-parametric framework for unsupervised feature extraction based on weak structural assumptions on the data density is introduced. We discuss and develop different iterative and non-iterative approaches to semi-parametric dimension reduction allowing for identifying a low-dimensional non-Gaussian component of the whole distribution in a structure adaptive way. The main difference between the approaches discussed consist in the reconstruction of the low dimensional, non-Gaussian target space of the method on focus. We discuss methods based on Principle Component Analysis (PCA), convex projection and semi-definite programming. It turns out that the choice of the optimization problem to be solved in order to reconstruct the target space from some estimators is decisive for the statistical sensitivity of the method to a variety of non-Gaussian components. Currently the best alternative is Sparse NonGaussian Component Analysis based on semidefinite programming. Combining this linear projective method with the so called dip index specialized on the detection of multimodality, we come up with NonGaussian Cluster Analysis (NCA). It is demonstrated that NCA used as a preprocessing step to the metastablility analysis of biomolecules is superior to comparable dimension reduction methods. Combining NCA with the state-of-the-art approach of Hidden Markov Models to metastablility analysis, results in an almost geometrical approach to high dimensional analysis of metastablility as requested.

We have to know.

We will know.

David Hilbert

Contents

Abstract . . . . . **ii**

List of Tables . . . . . **vii**

List of Figures . . . . . **viii**

Acknowledgments . . . . . **xiii**

1 Introduction **1**

2 Complexity **7**
2.1 The Curse of Dimensionality 7
2.1.1 Strange Geometric Phenomena in L^p-Spaces 7
2.1.2 Consequences in Data Analysis 9
2.2 Information-Based Complexity Theory 12
2.3 Reduction of Dimensionality in Data 14
2.3.1 The Continuous Latent Variable Model 15
2.3.2 Semi-Parametric Framework for Dimension Reduction 17

3 Nonparametric Methods For Highdimensional Data **21**
3.1 Taxonomy of Dimension Reduction Methods 21
3.1.1 Geometric Methods 22
3.1.2 Feature selection vs. Feature Extraction 24
3.2 Unsupervised Linear Methods of Feature Extraction 25
3.2.1 Pure Gaussian Analysis 25
3.2.2 Multidimensional Scaling 27
3.2.3 Probabilistic Approaches 28
3.3 Unsupervised Nonlinear Methods of Feature Extraction 31
3.3.1 Nonlinear PCA 31
3.3.2 Pure NonGaussian Analysis 33
3.3.3 Self-Organizing Maps 34

4 Convex Projection in Structural Data Analysis **37**
4.1 The Setup of the Method 37
4.2 Estimation of the Elements from the Target Space 39
4.3 Reduction of Dimensionality and Structural Adaptation 44
4.4 Algorithms 49
4.5 Statistical and Numerical Performance 53

5 Structural Analysis by Semidefinite Programming . . . 61
5.1 Semidefinite Relaxation . . . 64
5.2 Objectives with Convex Structure . . . 69
5.2.1 Variational Inequalities . . . 69
5.2.2 Extragradient Methods . . . 70
5.2.3 Application to Structural Data Analysis . . . 74
5.3 Algorithmic Procedures . . . 74
5.4 Numerical Simulations . . . 79

6 A Geometric Approach to Metastability Analysis . . . 85
6.1 Conformational Dynamics of Biomolecular Systems . . . 87
6.2 Hidden Markov Models with Gaussian Densities . . . 89
6.3 Reduced Conformational Dynamics . . . 92
6.3.1 Clustering of Highdimensional Data . . . 92
6.3.2 Metastability of Polypeptides . . . 97
6.3.3 Penta-Alanine . . . 98
6.3.4 Octa-Alanine . . . 107

7 Summary and Conclusion . . . 115

8 Zusammenfassung . . . 117

Appendix A Proofs . . . 119
A.1 Proof of Theorem 1 . . . 119
A.2 Proof of Theorem 3 . . . 120
A.3 Proof of Theorem 4 . . . 121
A.4 Proof of Theorem 6 . . . 123
A.5 Proof of Theorem 7 . . . 125
A.6 Proof of Theorem 9 . . . 126

Appendix B Statistical Tests . . . 129

List of Tables

4.1 Progress of SNGCA for the test models from above in 10 dimensions with increasing number j of iterations. The empirical mean of $\mathcal{E}(\widehat{\mathcal{I}}, \mathcal{I})$ is denoted by μ_ϵ and σ_ϵ^2 is its empirical variance. 56

List of Figures

1.1 Basic idea of clustering using non-Gaussian projections 2
1.2 general scheme of the geometrical approach to metastability analysis 5

2.1 Illustration of the geometric shape of a d-dimensional hypercube projected on a plane where d is very large. 7
2.2 Figure a) shows the volume of a unit-radius sphere with respect to d. Figure 2 b) shows the ratio between the volume of a unit-radius sphere and the volume of a cube with edge lengths equal to 2. Figure 2 c) shows the ratio between the volumes of two embedded spheres, with radii equal to 1 and 0.9 respectively. Figure 2 d) shows the percentage of the volume of the Gaussian function that falls inside a radius equal to 1.65. For $d = 1$ this percentage 90% but decreases rapidly up to almost 0 for $d \geq 10$, such that almost all the volume of a Gaussian function is contained in its tails. . . . 8
2.3 Probability of a point chosen from $\mathcal{N}(0, I)$ to be at distance $r = 2, 3, 5, 10, 20$ of the center increasing dimensions. 9
2.4 64 data points are simulated from $\mathcal{U}_{[0,1]}$. For $d = 1$ all the data points are clustered together and with increasing dimension the data become more sparse. 10
2.5 Kernel values as a function of the distance to their centers different dimensions $d = 2, 5, 10, 100$, along with the distribution of distances for normally distributed data. Vertical lines correspond to 5 and 95 percentile respectively. 11
2.6 Gneral idea of the curse of dimensionality. 13
2.7 General idea of dimension reduction to a reduced space using a structural assumption on the density where p_x and p_y denote the distribution of the observed and the latent variables respectively. 17

3.1 Taxonomy of dimensionality reduction techniques. 21
3.2 Illustration of lower bound for m for random projections as a number of data points. The upper curve corresponds to $\epsilon = 0.1$, the middle to $\epsilon = 0.2$ and the lower to $\epsilon = 0.5$ 30
3.3 Different configurations of the SOM in the data space as the learning progresses on data depicted as dots goes on. 35

4.1 Dotted line: Gaussian density with zero mean and variance 1/22. Solid line: the same density projected uniformly from distributed data over the 20-sphere, to an arbitrary selected line passing through the origin. 38
4.2 Geometry of ℓ_1-constraint: In a regression or optimization problem the use of the 1-norm as a constraint results in vanishing weights due to the geometry of the feasible set of the 1-norm, since the first touch point of square and ellipsoid containing the solution of the quadratic problem (4.16) is the vertex. 44

4.3 Illustration of the MVEE "rounding ellipsoid" of estimated elements "close" to the target space consisting in the (x, y)-plane. 45
4.4 Illustrative plots of SNGCA applied to toy 20 dimensional data of type **(C)** (see section 4.5): We show $\|\widehat{\beta}\|$ vs. $\cos(\measuredangle(\widehat{\beta}, \mathcal{I}))$ for different iterations of the algorithm where $\mathcal{I}$ is the apriori known target space. 49
4.5 Densities of the non-Gaussian components. From upper left to lower right: $2d$ independent Gaussian mixtures, $2d$ isotropic super-Gaussian, $2d$ isotropic uniform, $2d$ isotropic sub-Gaussian and $2d$ isotropic uniform and dependent $1d$ Laplacian with additive $1d$ uniform. 54
4.6 Performance comparison using toy examples in 10 dimensions of PP and NGCA versus SNGCA (with respect to $\mathcal{E}(\widehat{\mathcal{I}}, \mathcal{I})$) using the index 'tanh(x)'. The doted line denotes the mean, the solid lines the variance of (5.49). . . 55
4.7 From upper left to lower right: $2d$ independent Gaussian mixtures, $2d$ isotropic super-Gaussian, $2d$ isotropic uniform, dependent $1d$ Laplacian with additive $1d$ uniform and $2d$ isotropic sub-Gaussian: Results obtained from the toy densities with respect to $\mathcal{E}(\widehat{\mathcal{I}}, \mathcal{I})$ with deviations of Gaussian components with respect to a geometrical progression on $[10^{-r}, 10^{r}]$ where r is written on the abscissa) 57
4.8 From upper left to lower right: $2d$ independent Gaussian mixtures, $2d$ isotropic super-Gaussian, $2d$ isotropic uniform, dependent $1d$ Laplacian with additive $1d$ uniform and $2d$ isotropic sub-Gaussian: Results obtained from the toy densities with respect to $\mathcal{E}(\widehat{\mathcal{I}}, \mathcal{I})$ with increasing dimension of embedding Gaussian component. 57
4.9 Failure modes of SNGCA obtained from the toy densities - upper figure: model (A) - lower figure: model(B). We show boxplots of the aperture of dimensions where the failures occur. 58

5.1 The convex set $\mathcal{S}$ is separated from the points not in the set by half-spaces. The dashed line separates the plane into two halves, one containing x and the other $\mathcal{S}$. 64
5.2 The set of possible pairs of $g(x)$ and $f(x)$ are shown as the blue region. Left: Any hyperplane which has normal $(w, 1)$ intersects the y-axis at the point $f(x^*) + w^\top > g(x^*)$ where x^* minimizes $\mathcal{L}(x, w)$ with respect to x. Middle: A hyperplane whose y intercept is equal to the minimum of $f(x)$ on the feasible set. The dual optimal value is equal to that of the primal. Right: No hyperplane can achieve the primal optimal value. The discrepancy between the primal and dual optima is called a *duality gap*. The dual optimum value is always a lower bound for the primal. 65
5.3 FLT for a convex (left) and a nonconvex function (right). The hypeplanes $\langle s, x\rangle - d(x)$ are always below $epi(f)$. By f^{**} we denote the biconjugate. Note that from $f(x) \geq f^{**}(x)$ it follows that $f^{**}(x)$ is the convex hull of f and $f^*(x)$ is a supporting hyperplane in x. 71
5.4 Performance comparison in 10 dimensions of PP and SNGCA(1) versus SNGCA(2) (with respect to the error criterion $\mathcal{E}(\widehat{\mathcal{I}}, \mathcal{I})$) using the index 'tanh'. The doted line denotes the mean, the solid lines the variance of (5.49). 80
5.5 Results with respect to the test densities from section 4.5 in terms of $\mathcal{E}(\widehat{\mathcal{I}}, \mathcal{I})$ with deviations of Gaussian components following a geometrical progression on $[10^{-r}, 10^{r}]$ where r is the parameter on the abscissa) 81
5.6 Results with respect to the test densities from section 4.5 in terms of $\mathcal{E}(\widehat{\mathcal{I}}, \mathcal{I})$ with increasing number of gaussian components. 82

6.1 Changes of geometric large scale configurations of a biological active molecule with life times much longer than the time scale of the internal interactions between the atoms and the random perturbations of the molecules with the solvent visualized by AMIRA [214]. 85
6.2 Since some of the bonds in $\mathfrak{M}$ are a planar and thus stiff peptid bond, the rotational degrees of freedom (Φ, Ψ) allow to describe the macroscopic folding process of a biomolecule as a change of the geometric configuration of the backbone of $\mathfrak{M}$. 86
6.3 Illustration of non-parametric mean-shift mode finding process for component-wise normalized data in $\mathbb{R}^2$. The blue circles are the windows of the algorithm. The black stars are the centers od the windows. 91
6.4 Illustration of the important points of a density: modes, bumps, dip and shoulder. Points A and B are modes, shaded areas C and D are bumps, area E is a dip and F is a shoulder point. 93
6.5 Illustration of the estimated values of the dip index significant for multimodality computed for a mixture of Gaussians with increasing distance of their mean values. 95
6.6 Original data consisting of 3 non-overlapping clusters. The colored axis are the basis provided by the concurrent methods: Red color indicates PCA, blue and black color ICA and NCA respectively. 96
6.7 Illustration of general differences of comparable projective feature extraction methods from $\mathbb{R}^3$ to $\mathbb{R}^2$. On the upper left the data projected on the first two eigenvectors obtained from PCA are shown, the upper right and the lower left figure show the analogous result for ICA and NCA respectively. Obviously only NCA gives a sufficient separation of the cluster. 97
6.8 The figure shows the ten peptide angles of 5-alanine determining the secondary structure of 5-alanine, marked by $\Phi_1, \Psi_2, \ldots \Phi_9, \Psi_{10}$. 98
6.9 Comparison of feature extraction methods by means of the dip index and the estimated entropy of data, projected on the basis of $\mathcal{I}$: (A) shows the normed eigenvalues from PCA against the dip index, (B) the results from ICA and (C) the results from NCA. 99
6.10 Densities, estimated by adaptive kernel methods [210] of the data from simulations of 5-alanine projected on the basis vectors of the PCA target space. 100
6.11 Estimated densities of the data from simulations of 5-alanine projected on the basis vectors of the NCA target space. 100
6.12 Plot of first 30 PCCA-eigenvalues from Viterbi-Path-clustering of 5-alanine after dimension reduction with SNGCA. 101
6.13 Schematic Ramachandran plot of penta-alanine. 102
6.14 Empirical Ramachandran-plots of the first 2 conformations with dominating life time in descending order characterizing the effective dynamics of 5-alanine. 103
6.15 Empirical Ramachandran-plots of the conformations 3 and 4 with dominating life time in descending order characterizing the effective dynamics of 5-alanine. 104
6.16 Empirical Ramachandran-plots of the conformations 5, 6 and 7 with dominating life time in descending order characterizing the effective dynamics of 5-alanine. 105
6.17 Empirical Ramachandran-plots of the conformations 8 and 9 with dominating life time in descending order characterizing the effective dynamics of 5-alanine. 106

6.18 8-alanine in $\alpha - helix$ (left) and $hair - pin$ configuration (right) representation with dihedral angles. 107
6.19 Comparison of feature extraction methods by means of the dip index and the estimated entropy of data, projected on the basis of $\mathcal{I}$: (A) shows the normed eigenvalues from PCA against the dip index, (B) the results from ICA and (C) the results from NCA. 108
6.20 Plot of first 30 PCCA-eigenvalues from Viterbi-Path-clustering of 8-alanine after dimension reduction with SNGCA. 109
6.21 Empirical Ramachandran-plots of the first conformation with dominating life time in descending order characterizing the effective dynamics of 8-alanine. 109
6.22 Empirical Ramachandran-plots of the conformations 2 and 3 with dominating life time in descending order characterizing the effective dynamics of 8-alanine. 110
6.23 Empirical Ramachandran-plots of the conformations 4, 5 and 6 with dominating life time in descending order characterizing the effective dynamics of 8-alanine. 111
6.24 Empirical Ramachandran-plots of the conformations 7, 8 and 9 with dominating life time in descending order characterizing the effective dynamics of 8-alanine. 112

Semi-Parametric Reduction of Dimensionality

Statistical Detection of Rare Events in Molecular Dynamics

Declaration

I declare that this thesis is my own work and has not been submitted in any form for another degree or diploma at any university or other institute of tertiary education. Information derived from the published and unpublished work of others has been acknowledged in the text and a list of references is given.

Elmar Diederichs
18.6.2009

Acknowledgments

I would like to express my gratitude to all those who gave me the possibility to complete this thesis. In particular, I would like to thank my supervisors for their help and support during my graduate studies. The idea to the project documented in this thesis goes back to Christof Schütte. His patience and insight into the interdisciplinary field of complex systems makes it a pleasure to work with him. Moreover he keeps the science business running as no one else therefore making this thesis possible. Vladimir Spokoiny has accompanied my statistical studies. He managed to strike the perfect balance between providing direction and encouraging independence while continuously supporting me during my time at the Weierstraß Institute Berlin. I have learned much from him observing the way he lives his scientific life. Anatoli Juditsky from the Université J. Fourier in Grenoble deserve special thanks for providing central ideas to this project. I wish I would have had him as my academical teacher. Moreover I'm grateful to Gilles Blanchard from the Fraunhofer Institute FIRST (IDA) Berlin for helpful discussions on semi-parametric dimension reduction and Illia Horenko from the Free University Berlin for many hints on sideline mathematics. Eike Meerbach took me under his wing and introduced me to the wide world of metastability analysis. The author would like to thank all members of the *Biocomputing Group* at the Free University Berlin and all members of the research group *Stochastic Algorithms and Nonparametric Statistics* at the Weierstraß Institute Berlin for a stimulating and interdisciplinary research atmosphere and last not least for a good time. Furthermore I deeply acknowledge the members of the *Philosophical Colloquium* at the Georg-August-University Göttingen leaving such a big impact on my graduate life during my time at the Free University as well as for long and heated discussions about language, physics, life and the universe. Thanks to all my friends for their support, encouragement and humor. Finally I'm indebted to Sönke, George and Thomas for never failing to provide me with distractions when I needed them and to Jana for reminding me sometimes that there is more in life than a thesis during the last months.

Thankfully this work was funded by the DFG Research Center MATHEON "Mathematics for Key Technologies" (FZT86) in Berlin.

Elmar Diederichs

Freie Universität Berlin
June 2009

Notations

$\mathbb{R}^d$	set of d dimensional real numbers
$\mathcal{S}_d$	cone of symmetric matrices from $\mathbb{R}^d \times \mathbb{R}^d$
$A \succeq 0$	positive semi-definite matrix A
$A \succeq B$	has the same meaning as $A - B \succeq 0$
I	identity matrix from $\mathbb{R}^d$
$\mathrm{Tr}(A)$	trace of a matrix A
$\det(A)$	determinant of a matrix A
$\mathrm{diag}(A)$	main diagonal of a matrix A
$\mathcal{B}_d$	unit ball $\mathcal{B}_d = \{x \in \mathbb{R}^d \mid : \|x\|_2 \leq 1\}$ in $\mathbb{R}^d$
E	vector space over $\mathbb{R}^d$
E^*	dual space to E
$\widehat{f}$	estimator of the function f
$\langle \cdot, \cdot \rangle$	inner product of E
$dist(\cdot)$	metric on E
$\|\cdot\|$	generic vector or operator norm of a space
$\|\cdot\|_*$	dual norm $\|\cdot\|_* \stackrel{\text{def}}{=} \max_{x \in E}\{\langle \cdot, x \rangle : \|x\| \leq 1\}$
$\|\cdot\|_F$	Frobenius norm $(\mathrm{Tr}[A^\top A])^{\frac{1}{2}}$
$\|\cdot\|_p$	canonical norm on L^p
$\mathcal{X}, \mathcal{Y}, \mathcal{Z}, \mathcal{W}$	nonempty, convex, compact sets, domains of a problem
$L^p_\mu(\mathbb{R})$	$\{f : E \to \mathbb{R} \mid f$ measurable wrt. $\mu, \|f\|_p \leq \infty\}$
$\mathcal{C}^2(\mathbb{R}^d, \mathbb{R}^m)$	set of twice continuously differentiable functions $f : \mathbb{R}^d \to \mathbb{R}^m$
$\mathcal{E}_d$	ellipsoid in $\mathbb{R}^d$
∇f	gradient of f
f'	gradient or subgradient of f
$\partial_x f(x)$	partial derivative of $f(x)$ with respect to x
f_*	optimal value of f
$L_{\|\cdot\|}(f)$	Lipschitz constant of f w.r.t. $\|\cdot\|$
$domf$	domain of a function f
$\rho(x)$	function from $\mathbb{R}^d \to \mathbb{R}$ with $\int_{\mathbb{R}^d} \rho(x)\mathrm{d}x = 1$, data density
$\mathbb{P}(\cdot)$	probability measure
X, Y, Z	random variables
$\mathbb{E}[X]$	expectation of a random variable X
$\mathbb{E}_N[X] = N^{-1}\sum_i X_i$	empirical expectation of a random variable X
$Var[X]$	variance of a random variable X
$Cov[X, Y]$	covariance of random variables X and Y
$\mathbb{L}$	log likelihood
Θ	parameter of a model
Σ_{XY}	covariance matrix of random variables X and Y
$\mathcal{N}(\mu, \sigma)$	normal distribution with parameter μ and σ
$\mathcal{U}_{[a,b]}$	uniform distribution in the interval $[a, b]$
χ^2_f	χ^2_f distribution with f degrees of freedom
$\mathbf{P}$	matrix of probabilites
ν	finite measure

$\mathcal{L}$	Lagrange function
$\Pi_{\mathcal{X}}(\cdot)$	orthogonal projector on the set $\mathcal{X}$
Δ_d^+	(full) d-dimensional simplex
$T_\beta(x,s)$	prox mapping
$cpl_\delta(\cdot)$	analytical complexity
$\mathcal{O}(\cdot)$	arithmetical complexity
$\mathcal{O}(1)$	some positive constant
$\lfloor x \rfloor$	The closest integer smaller or equal than x.
$\lceil x \rceil$	The closest integer greater or equal than x.
ϵ	estimation error
δ	numerical error
δ_{ij}	Konnecker symbol
α	constant of strong convexity
α_k	k^{th} step size of a gradient type method
$\lambda_i(A)$	i^{th} eigenvalue of the symmetric matrix A
$\ker[A]$	kernel or null space of a linear mapping
$\mathbf{1}_{\mathcal{X}}$	indicator function with respect to $\mathcal{X}$.

List of Abbrevations

EDR	Effektive Dimension Reduction
EM	Expectation-Maximization
FA	Factor Analysis
FLT	Fenchel-Legendre-Transformation
GCM	Greatest Convex Minorant
HMM	Hidden Markov Model
IBCT	Information Based Complexity Theory
ICA	Independent Component Analysis
IPM	Interior Point Methods
LDA	Linear Discrimination Analysis
LCM	Lowest Concave Majorant
LS	Least Square
MD	Molecular Dynamics
MDS	Multidimensional Scaling
ML	Maximum Likelihood
MVEE	Minimum Volume Enclosing Ellipsoid
MVIE	Maximum Volume Inscribed Ellipsoid
NCA	NonGaussian Clustering Analysis
NGCA	NonGaussian Component Analysis
NLPCA	Nonlinear Principle Component Analysis
ONB	orthonormal basis
PCA	Principle Component Analysis
PCCA	Perron Cluster Cluster Analysis
PCR	Principle Component Regression
PDF	probability density function
PP	Projection Pursuit
PPCA	Probabilistic Principle Component Analysis
QCP	Quadratic Constraint Programming
RP	Random Projections
RRR	Rank Reduced Regression
SDE	Stochastic Differential Equation
SDP	Semi-definite Programming
SIR	Sliced Inverse Regression
SOCP	Second Order Conic Programming
SNGCA	Sparse NonGaussian Component Analysis
SOM	Self-Organizing Maps
SVD	Singular Value Decomposition
VIP	Variational Inequality Problem

Chapter 1

Introduction

In recent years the availability of massive data and challenges from frontiers of research and development have reshaped statistical thinking and data analysis. Nowadays we collect massive amounts of data with relatively low cost. Therefore many mathematical applications in science are confronted with high dimensional data. Such data sets present new challenges in data analysis, since often the data have dimensionality ranging from hundreds to hundreds of thousands of components. Due to high-dimensionality there are general limits to any kind of data analysis, usually referred to as *curse of dimensionality* [15]. Dimensionality is an unwelcome issue arising in many applied scientific fields, ranging from computational biology especially in proteomics [42], structure prediction e.g. metastability analysis of biomolecules [97] to financial engineering or climate research [156]. Generally speaking, the difficulty lies on how to analyze and how to visualize a high dimensional function or data set: It is well known and widely accepted that classifiers and estimators perform poorly in a high dimensional space with a limited number of samples. In fact these limits have their origin in the convex geometry of metric spaces [245].

The burden of responsibility for the mathematical difficulties in analyzing high-dimensional data has to be put on two phenomena. On the one hand high-dimensional spaces have several counter-intuitive geometrical properties far from the properties that can be observed in spaces up to $d = 3$ dimensions. On the other hand statistical data analysis methods are most often designed having in mind intuitive examples in low-dimensional spaces. However to infer algorithmically knowledge or information successfully from data sets primarily depends on two key ingredients:

- Generalization of knowledge on data that are much different from the learning points as yet available is not advisable: Any feasible generalization comes from interpolation but not from extrapolation.
- Even in high dimensions successful learning requires enough data for learning so that they fill the space or parts of the space, where the hypothesis on focus should be valid.

We demonstrate in section 2.1.1 and 2.1.2 that the violation of both assumptions starts not later than dimension $d = 5$ and thus can be considered as completely misleading in dimensions higher than $d = 10$.

Moreover towards higher dimensions other well known problems such as collinearity easily occur, meaning the number of samples available for learning is less than the dimension of the data space. Such problems are even worse when using nonlinear models since most nonlinear methods involve more parameters than input variables e.g. resulting in lack of model identifiability, numerical instability of the methods and overfitting i.e. in a too

efficient modelling where only a special feature of the given samples is represented by the model generalization ability [76]. Due to these facts dimensionality in data analysis should in general be viewed as an independent parameter of mathematical methods.

Structural Analysis and Rare Events: Although these facts provide sufficient reasons to attempt to break or at least to circumvent the curse of dimensionality in data analysis [61], it is far from clear which method will serve as appropriate preprocessing step for information other e.g. statistical or dynamical analysis methods. The data itself as well as the characteristics of the subsequent mathematical analysis heavily influences the choice of the dimension reduction method. Consequently it is always a good advice to have first a close look at the application of interest.

Once in a while the detection of statistically rare events coincides with an extraction of the structured parts in a data set, that are build up by non-normal contributions to a distribution sampled by the given data: If cluster in high dimensional data stem from sudden changes of e.g a physical or biological system in their observation space, then detecting the non-Gaussian components of a data distribution is a suitable preprocessing step to a dynamical analysis of important but rare events in time series reporting the geometrical large scale changes of complex systems. The effect of a data mapping onto an appropriate reduced subspace is illustrated in figure 1.1.

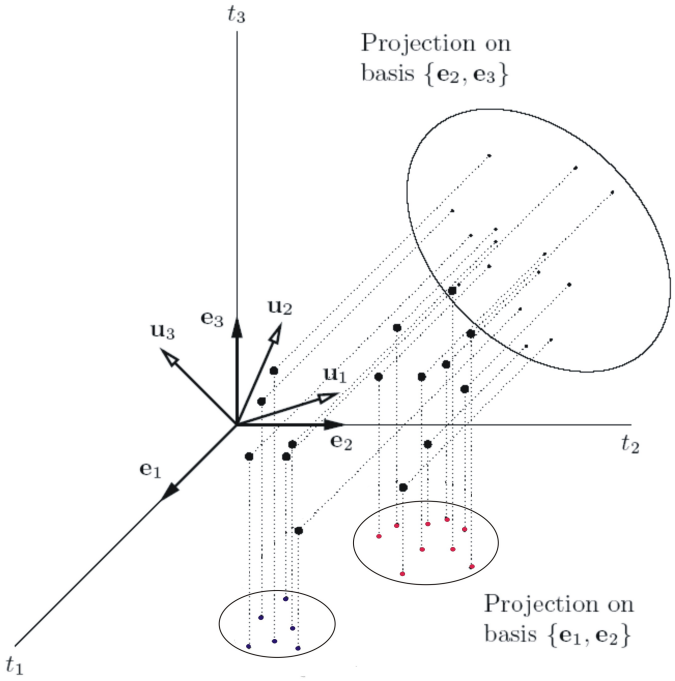


Figure 1.1: Basic idea of clustering using non-Gaussian projections

In this thesis we only concern biomolecular systems as examples for real complex systems. The number of atoms in these biomolecular system typically ranges from at least 33 to hundreds of thousands.

Geometrically large scale transitions in biomolecular systems have been studied for a long time as characteristic features of transition processes using set oriented methods [196; 36; 52; 54; 53; 173; 199; 202] as well as so called *Hidden Markov Models* [70; 100; 96; 95; 201]. In the context of biomolecular systems rare events between so called *metastable* states can

be observed: It is well known that e.g. for proteins the transition from an unfolded in a folded macroscopic state is in the range up to milliseconds. In comparison the small-amplitude motions of the atoms of the amino acid side chains took place at the femtosecond scale [200]. From a physical point of view transitions between metastable states are driven by the mechanical and electrostatic random interactions of the atoms in the molecule or in the solvent constraint by the given electronic bond structure of the biomolecule. Intensity and amplitude of these interactions reflect the temperature at which the rare events occur. From a statistical point of view the dynamics of the transition process is dominated by noise associated with an energy scale with respect to a thermodynamical (canonical) ensemble, that serves as a model for a single molecule. Transitions are rare events, since they depend on the height of the energy barrier separating the metastable states from each other: the time scale of escape from a given metastable state depends exponentially on the ratio of both energies. In spite of the fact that the formal description of a biomolecular system required to observe such transitions can often be reduced using e.g. the torsion or backbone angles of the biomolecule [6], time series from numerical *molecular dynamics* [5] simulation are in almost all cases high dimensional.

Outline of the Thesis: The motivation for this thesis is twofold. On the one hand a survey of existing unsupervised feature extraction methods will show that a data driven method with low complexity and sufficient statistical sensitivity even in high dimensions, currently do not exist. Moreover justified convergence rates for methods of dimension reduction are hardly available. In particular as far as we know up to know there is no possibility to control if the assumptions, providing reasons for the method on focus, are fulfilled, such that interpretable results of the methods are more stroke of luck than quotidian. On the other hand we aim to develop a fully geometric approach to the analysis of metastability of biomolecular systems.

In chapter 2 we start in section 2.1.1 with a survey of the geometrical phenomena occurring in the L^p-spaces, that are responsible for the notorious mathematical limitations in high dimensional data analysis. We summarize in section 2.1.2 some consequences in statistics and machine learning that are frequently discussed but more often ignored in the applied sciences. Then we introduce in section 2.3.1 the well known *Continuous Latent Variable Model* [13], that is used in a variety of popular methods of dimension reduction. In section 2.3.2 we describe an alternative semiparametric framework [212]. In this thesis we show that this framework, based on a very weak structural assumption about the data density, allows to design some new efficient methods of dimension reduction. Finally we give in chapter 3.2 a concise report of existing feature extraction methods, that may be comparable to the here favored approach to dimension reduction, called *Sparse NonGaussian Component Analysis* (SNGCA).

The role of SNGCA as a tool for high dimensional structural analysis can be sketched as follwows: In statistics projections are a common tool for extracting useful information from high dimensional data. Almost all projection methods for feature extraction like e.g. *Principle Component Analysis* [117], *Projection Pursuit* [74; 82], *Partial Least Square Regression* [236; 237], *Conditional Minimum Average Variance Estimation* [239] or *Sliced Inverse Regression Methods* [140; 43; 32] decompose the problem of dimension reduction into two more or less independent tasks: First one has to determine elements from the reduced data space, also referred to as target space. Second, one has to construct a basis of the target space from these elements. For the latter task one has to know the number of required basis elements.

SNGCA, described in this thesis, is a linear unsupervised projection method for feature extraction, that links pure Gaussian (PCA) and pure non-Gaussian *Independent Component Analysis* (ICA) [110]. In this thesis the method comes in two approaches. And as usual there are some good and some bad news about both of them:

- The first approach comes as an iterative and structure adaptive method, that repeats the established two-stage strategy of element estimation and basis construction. The good news is, that the dimension m of the target space can be determined from the estimation procedure. The bad news is that this approach is computationally expensive. We will call this the "convex projection"-approach to SNGCA. The use of convex programming for estimating elements from the target space $\mathcal{I}$ by convex projection is new and allows to realize a uniform bound for the estimation error ϵ based on a well known result from empirical process theory [226].

- The second approach, developed in this thesis, "shortcuts" the intermediary stages described above, and moreover makes the best use of the available information for computing estimatior from the target space. This can be archived by a direct estimation of the so called subprojector onto the target space. However to this end the reduced dimension m must be apriori given to the algorithm as a tuning parameter. We will call this the "semidefinite programming"-approach to SNGCA. We will see that the use of semidefinite programming [238] in structural data analysis instead of Interior Point Methods [184] results in a linear matrix game with bounded convex domains of its arguments. The good news is that this new approach in statistical data analysis is responsible for improving the statistical sensitivity of the complete method while decreasing the required computational effort.

In comparison to ICA we will see that the SNGCA allows for cross-dependence of the non-Gaussian components and for presence of a full dimensional Gaussian part. The only important assumption for the SNGCA approach is that the non-Gaussian part is low dimensional, otherwise no dimensionality reduction will be produced. Correspondingly, the target of the SNGCA is to "kill the noise" rather than to describe the whole distribution. Projecting the data onto a low-dimensional subspace means that the orthogonal complement to this subspace only contains a non-informative noise. SNGCA do not depend on a special difference in the magnitude of the second moments of Gaussian noise and informative data components as e.g. PCA.

In chapter 4 we introduce the "convex projection"-approach to SNGCA as a realization of several desired properties of linear unsupervised feature extraction methods. The label "convex projection" indicates that this approach uses an aggregation strategy to estimate vectors β from the dimension reduced target space $\mathcal{I}$ as convex combinations resulting in outstanding statistical properties: On the one hand as an upper bound for the "prize of aggregation" there is a value $\epsilon = \sqrt{C/N}$ for a fixed positive constant C and a random set A of a dominating probability such that for the projector $I - \Pi_{\mathcal{I}}$ onto the complementary space $\mathcal{I}^c$ it holds $\|(I - \Pi_{\mathcal{I}})\widehat{\beta}\|_2 \leq \epsilon$ for all such constructed vectors $\widehat{\beta}$. Here N denotes the number of samples. In addition the dimension m of the target space $\mathcal{I}$ can be estimated correctly from the data. On the other hand numerical simulations in section 4.5 show that the detection of $\mathcal{I}$ is nearly independent of the variance of the normal noise, that embeds the information representing components to the data density. In section 4.4 we present the algorithmic details of the "convex projection"-approach .

In chapter 5 we describe the "semidefinite programming"-approach to SNGCA that aims to improve the statistical sensitivity of SNGCA. Having introduced the "recipe" of semidefinite relaxation we demonstrate that the arising nonconvex, nonsmooth optimization problem can be efficiently solved in the very general setting of variational inequality problems. The error with respect to original problem made by solving the relaxed problem can be bounded by $\mathcal{O}(1)(\delta\lambda_{\min}(\Sigma))^2$ where δ is the desired numerical accuracy and $\lambda_{\min}(\Sigma)$ the minimal eigenvalue of the covariance matrix Σ of the data distribution. In order to make the "semidefinite programming"-approach viable we "hide" the linear constraints in the geometry of the feasible sets associated with the objective (see section 5.2.3). Numerical simulations with respect to the statistical sensitivity demonstrate that this new approach to SNGCA archives to exploit the information obtained from the data space sampling better than comparing methods.

In chapter 6 we first give a sketch of the metastability analysis by means of Hidden Markov Models, described in section 6.2. Since conventional clustering algorithms have serious drawbacks in high dimensions, we apply the "semidefinite programming"-approach to SNGCA to simulated biomolecules in order to make a dynamical analysis reliable (c.f. [99]).

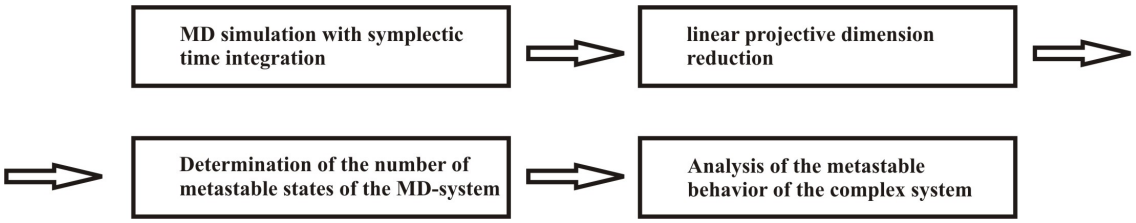


Figure 1.2: general scheme of the geometrical approach to metastability analysis

Using SNGCA as a preprocessing step allows to apply a geometric mean-shift clustering algorithm [35] on the dimension reduced data that provides an necessary initialization for a state-of-the-art approach to metastability described in [153]. The combination of these algorithms results in a completely geometric approach to high dimensional metastability analysis illustrated in figure 1.2. In section 6.3 we present examples of the metastability analysis of oligo-peptides that illustrate the progress made by this strategy.

However with respect to the first motive of this thesis the types of data sets concerned here are subject to some important restrictions:

1. Here we only deal with metric and unlabeled data. Hence in this thesis we put the focus on unsupervised methods for dimension reduction.

2. We assume that all data is a sample from a stationary density. Consequently we are only interested in global methods for dimension reduction.

3. We only account for situations where the number N of the data is much bigger than the dimensionality d of the data space $\mathbb{R}^d$.

The MATLAB-toolbox that implements both approaches to Sparse NonGaussian Component Analysis described in the chapter 4 and 5 is available on the web either here
`http://www.wias-berlin.de/people/spokoiny/` or here
`http://www.math.fu-berlin.de/groups/biocomputing/projects/projekt_A10/index.html`.

Chapter 2

Complexity

2.1 The Curse of Dimensionality

Aside from differences underlying various scientific contexts, requests for dimensionality reduction do have a common geometric root provided by the so-called *curse of dimensionality*. This expression is coined by to Bellman [15] to describe a set of problems caused by the exponential increase in volume associated with adding extra dimensions to e.g. an Euclidean space E. Some of them are relevant to the topics of this thesis and will be described in the following.

2.1.1 Strange Geometric Phenomena in L^p-Spaces

Space is measure linearly and volumetrically. But the relation between linear and volumetric measures itself is not linear causing a breakdown of geometric intuition in high dimensions. In order to illustrate this we consider first of all the volume $Vol(\mathbf{S}_d(r))$ of the d-dimensional hypersphere $\mathbf{S}_d(r)$ of radius r and the volume $Vol(\mathbf{C}_d(r))$ of d-dimensional unit hypercube $\mathbf{C}_d(r)$. Then we get asymptotically:

$$\lim_{d\to\infty} \frac{Vol(\mathbf{S}_d(r))}{Vol(\mathbf{C}_d(r))} = \lim_{d\to\infty} \frac{1}{2^d} \frac{\pi^{\frac{d}{2}}}{\Gamma(\frac{d}{2}+1)} = 0 \tag{2.1}$$

where $\Gamma(\cdot)$ is the gamma function. Obviously towards higher dimensions the volume of the hypercube concentrates on its corners and its central parts shrink to zero [203]. Also, the length of the diagonals goes to infinity. Consequently the hypercube has to be imagined as an anisotropical body: As illustrated in figure 2.1 the inner ball-like part with $r << 1$ is covered with 2^d "spikes" with a length going to infinity for large d [91]. This is shown in figure 2.1.

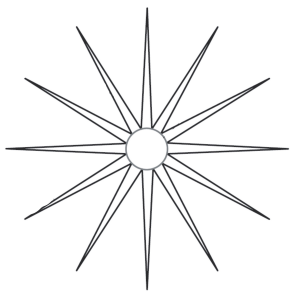


Figure 2.1: Illustration of the geometric shape of a d-dimensional hypercube projected on a plane where d is very large.

This phenomenon is called the *empty space phenomenon*. The analogously argument [235] for the volume between two concentric spheric shells with radii r and $r(1-\epsilon)$ we get asymptotically:

$$\lim_{d\to\infty} \frac{Vol(\mathbf{S}_d(r)) - Vol(\mathbf{S}_d(r(1-\epsilon)))}{Vol(\mathbf{S}_d(r))} = \lim_{d\to\infty} 1 - \left(1 - \frac{\epsilon}{r}\right)^d = 1 \tag{2.2}$$

Hence the content of a hypersphere is in a way concentrated close to its surface, which is only a $(d-1)$-dimensional hypersphere. In terms of a data density in the space, this means that if there are uniformly distributes points over the complete cube, the probability that they fall near the corners is almost one. Moreover even if the two radii only differ by 10%, the ratio between both volumes is almost 0 for $d = 10$. For uniformly distributed data, this means that almost all of them will fall in its skull, and will therefore have a norm equal to 1. These effects of increasing dimensionality is illustrated in figure 2.2.

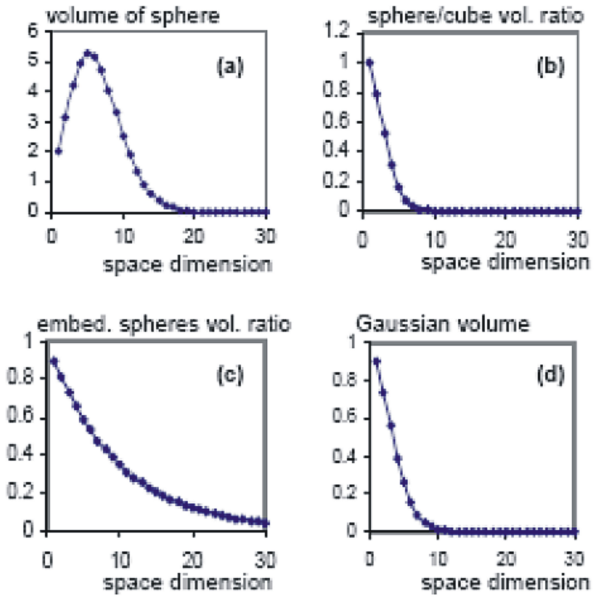


Figure 2.2: Figure a) shows the volume of a unit-radius sphere with respect to d. Figure 2 b) shows the ratio between the volume of a unit-radius sphere and the volume of a cube with edge lengths equal to 2. Figure 2 c) shows the ratio between the volumes of two embedded spheres, with radii equal to 1 and 0.9 respectively. Figure 2 d) shows the percentage of the volume of the Gaussian function that falls inside a radius equal to 1.65. For $d = 1$ this percentage 90% but decreases rapidly up to almost 0 for $d \geq 10$, such that almost all the volume of a Gaussian function is contained in its tails.

Consequently in some sense, almost all of the high-dimensional space is "far away" from its origin. Considering a data set uniformly distributed over $\mathbf{S}_d(r)$ and $\mathbf{C}_d(r)$ most of the data fall near the boundary and edges of the cube and lead to a significant *sparsity* of the data.

Now consider a standard Gaussian in different dimensions and compute the probability density function (pdf) to find a point arbitrarily chosen from $\mathcal{N}(0, I)$ at distance r from the center of the distribution. Figure 2.3 shows that when the dimension increases, the bell shape more or less remains, but is shifted to the right.

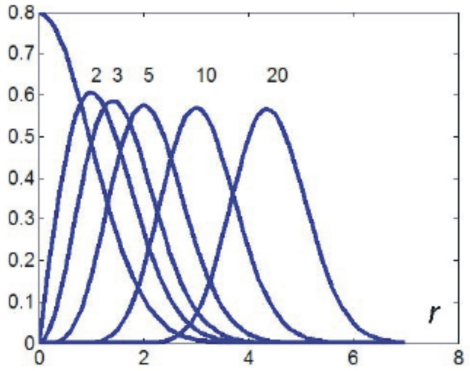


Figure 2.3: Probability of a point chosen from $\mathcal{N}(0, I)$ to be at distance $r = 2, 3, 5, 10, 20$ of the center increasing dimensions.

This means that the distances between all points and the center of the distribution are concentrated in a small interval and relative differences between these distances become less and less discriminative. The phenomenon is called the *concentration of norms*. It has be shown [175] that for every random vector with independent and identically distributed (i.i.d.) components, the mean of their Euclidean norm increases as the square root as the dimension of the space, while the variance of their norm does not increase. Consequently in high dimensions, all vectors are normalized, as the error resulting from taking the mean of their norm instead of their actual norm becomes negligible. In other words towards higher dimensions distance measures become increasingly meaningless: Additional dimensions spread out the points until in very high dimensions, they are almost equidistant from each other for arbitrary distance measures and a wide variety of data distributions [19]. For example this makes a proximity query meaningless and unstable because there is only poor discrimination between the nearest and furthest neighbour. Moreover the concentration of the norms phenomenon results in the fact that Gaussian kernels become a inappropriate tool in high-dimensional spaces [66].

Finally we consider a set of centered diagonal vectors v in $[-1, 1]^d$ from the origin to a corner and let e_d be an arbitrary coordinate axis. Then it holds [203]

$$\lim_{d\to\infty} \cos(\measuredangle(v, e_d)) = \lim_{d\to\infty} \frac{ve_d}{\|v\|\|e_d\|} = \lim_{d\to\infty} \frac{\pm 1}{\sqrt{d}} = 0 \tag{2.3}$$

Thus, all v are nearly orthogonal to all coordinate axes e_d for larger dimensions. This phenomenon is called the *distortion of space*, having consequences for data clustering: A cluster lying near an arbitrary v of the cube will be mapped almost completely into the origin of the coordinate system, while a cluster positioned along a coordinate axis will be visible in some projection. Thus the choice of coordinate systems is critical in some data analysis. Further cheerful facts about the curse of dimensionality can be found in [11; 121].

2.1.2 Consequences in Data Analysis

We restrict ourselves to some striking examples instead of making a complete list of unpleasant effects typically associated with the term *curse of dimensionality*.

Statistics: When data in high-dimensional spaces become inherently sparse the work for a statistican gets harder. We will illustrate this using the uniform and the normal distribution: Recall that the probability that a point from $\mathcal{U}_{[-1;1]^d}$ in the $d = 10$-dimensional sphere falls at a distance of 0.9 or less from the center is only 0.35. This causes severe problems for brute-force implementations of Monte Carlo methods. In particular the sampling of non-Gaussian pdfs requires sample sizes exponentially growing with d [139]. Moreover the analogous problems occur in density estimation methods, as regions of relatively low density can contain a considerable part of the distribution, whereas regions of apparently high density may be completely devoid of observations in a sample of moderate size [210]. In the case of the standard d-dimensional normal distribution, equiprobable contours are hyperspheres. The probability $I\!P$ that a point is within a contour of density ϵ or, equivalently, inside a hypersphere of radius $r = \sqrt{-2\ln\epsilon}$, is:

$$I\!P(\|x\|^2 \leq -2\ln\epsilon) = I\!P(\chi^2_d \leq -2\ln\epsilon) \tag{2.4}$$

since $x \in \mathbb{R}^d \sim \mathcal{N}(0, I)$. Equation (2.4) gives the probability that a random point will not fall in the tails, i.e., that it will fall in the medium- to high-density region. Therefore, and contrarily to our statistical intuition, in high-dimensional spaces the entire sample will be in the tails of a distributions and are much more important than in one-dimensional ones. We give an example of this fact in figure 2.6.

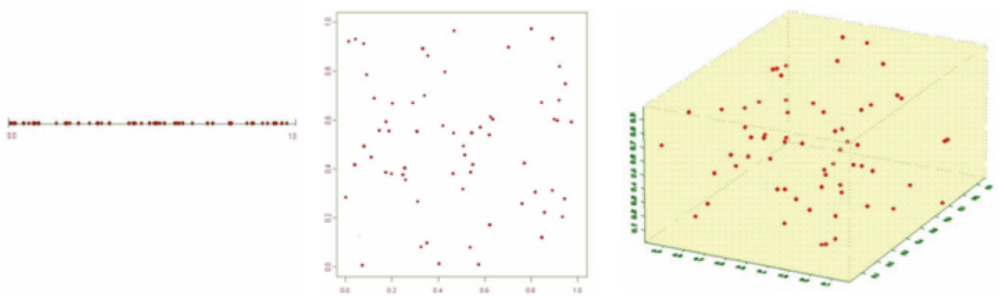


Figure 2.4: 64 data points are simulated from $\mathcal{U}_{[0,1]}$. For $d = 1$ all the data points are clustered together and with increasing dimension the data become more sparse.

The empty space phenomenon implies also that most (spherical) local neighborhoods of e.g. data uniformly distributed over a hypercube in high dimensions are empty. Hence since many estimation methods are based on some local average of the neighboring observations [210], in order to find enough neighbors in high-dimensional spaces, the neighborhood has to reach out farther, such that the locality is lost. This implies that, in the absence of simplifying structural assumptions, the amount of training data, needed to get reasonably low variance estimators is really high [232] and its convergence to the estimated function becomes very slow [23]. Moreover such large data sets may not be available in practical situation.

As an example we consider density based clustering. In high dimensions the method will only consider those regions of the density landscape as clusters that rise above the noise threshold. Another example is given by nonparametric regression. It is shown [216] that under certain regularity assumptions, the optimal rate of convergence varies as $N - p/(2p + d)$, where $N \in \mathbb{N}$ is the number of samples from $\mathbb{R}^d$ with $d \in \mathbb{N}$, and where the regression function is assumed to be $p \in \mathbb{N}$ times differentiable. Consequently in the case of $N = 10^4$, $p = 2$ and $d = 10$ the number of sample points must be increased to approximately 10 million in order to achieve the same optimal rate of convergence compared the case $d = 10$.

Furthermore we consider a typical method of parameter estimation: In order to archive a prescribed mean squared error when estimating the data density needs about $(\epsilon^{-2})^{d+2}$ observations [61]. The required sample size increases if there are linear correlations in the data, that is very likely in high dimensions [203]. In particular the rates of convergence of nonparametric estimates rapidly slows down with the dimensionality for e.g. Lipschitz continuous functions of d variables at rate $\mathcal{O}(N^{\frac{-1}{2+d}})$. Other problems occurring in in non-parametrical testing are reported in [213; 101; 17].

Machine Learning: In order to provide a formal framework for learning problems as inference, classification, model construction, prediction or gaining knowledge statistical learning makes assumption on the statistical nature of the hitherto existing observations of the phenomena on focus [59]. For example it is assumed that the past training data and the future data are from the same general probabilistic model and are independently sampled. Consequently one can construct consistent algorithms, which means that the predictions of the algorithm come closer to optimal predictions the more data are available. In particular many algorithms in statistics or machine learning make use of kernel functions [229; 24; 25] due to the following assumptions:

- Using local functions avoids illegitimate generalization in almost empty regions. Gaussian kernels are deemed to be local.
- Gaussian functions evaluated on the norm of a distance between two points results in a high value if the points are close, and in a low one if the points are far one from each other. Hence Gaussian kernels are used as a smooth measure of similarity between two points.

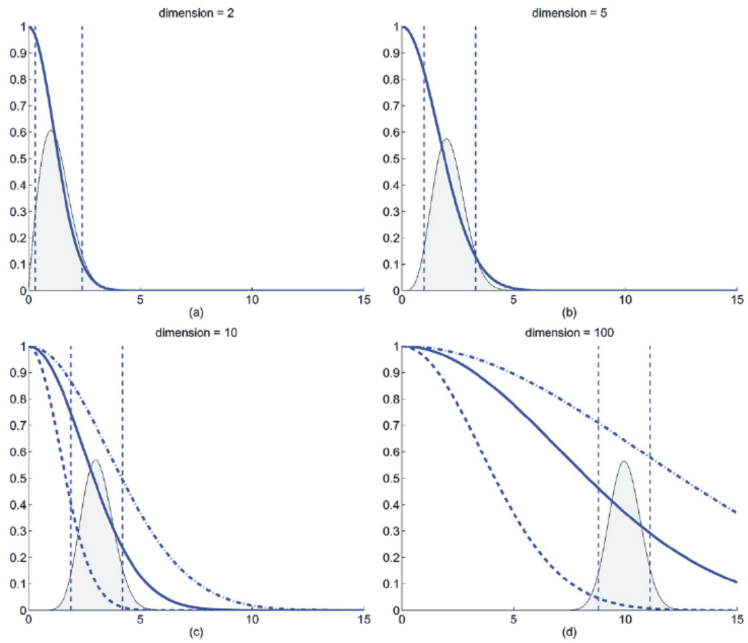


Figure 2.5: Kernel values as a function of the distance to their centers different dimensions $d = 2, 5, 10, 100$, along with the distribution of distances for normally distributed data. Vertical lines correspond to 5 and 95 percentile respectively.

However figure 2.5 illustrates a different statement for highdimensional spaces. The bell-shaped curves (thin lines) show, in dimensions 2, 5, 10 and 100, the distribution of distances

between each sample and the center of a multivariate normal distribution: The vertical lines correspond to the 5% and 95% percentiles respectively. Gaussian kernels are superimposed on the graphs (thick lines). Obviously in dimensions 2 and 5, the values taken by the Gaussian kernels are different for small and large distances found in the distribution (see the dotted vertical lines). Towards higher dimensions this does not remain true. Even by adjusting the standard deviation of the Gaussian kernels (see the dotted kernels), they remain flat in the range of effective distances in the distribution.

Moreover in multivariate data sets obtained in technical or financial applications measured variables are often highly correlated or provide redundant information. Time series often contain significant noise, i.e. subsequences that are the result of random fluctuations. Even if the class densities in these cases are completely known, an increase in the number of features, represented as random variables, will not result in an increase in the probability of misclassification. It has been often observed [185] that the added features may degrade the performance of a classifier if the number of training samples that are used to design the classifier is small relative to the number of features: Noisy or irrelevant features can have the same influence e.g. on classification as predictive features so they will impact negatively on accuracy. On the other hand, a reduction in the number of features may lead to a loss in the discrimination power and thereby lower the accuracy of the resulting recognition system. Moreover classification methods emerging from statistics cannot be accurately modelled when the amount of available samples is small compared with its dimension and the convergence of estimators to the value of an estimated smooth function defined on a space of high dimension is unacceptable slow [131].

2.2 Information-Based Complexity Theory

Complexity Theory aims at understanding the nature of efficient computation [225] and provides a simple way of formalizing what is meant by the curse of dimensionality: In this setting we say that a problem is subject to a curse of dimensionality if the lower bound on computational complexity grows exponentially fast as the dimension d increases. Problems with infinite complexity are called unsolvable or non-computable. A problem which is subject to the curse of dimensionality is said to be intractable. If the computational complexity of a problem is bounded above by a polynomial function of d and ϵ then it is not subject to a curse of dimensionality.

Independently from the occurrence of the geometric phenomena from above the size of a sample from a distribution needed to estimate a function of a high dimensional random variable X with reasonably low variance, grows exponentially with N [188]. This can easily illustrated as follows: Consider 100 evenly-spaced points suffice to sample the interval $[-1, 1]$ with no more than 0.01 distance between every two points. Then an equivalent sampling of a 10-dimensional unit hypercube with a spacing of 0.01 between adjacent points requires 10^{20} sample points. Consequently in some sense, the 10-dimensional hypercube can be said to be a factor of 10^{18} "larger" than $[-1, 1]$. This means an exponential increase of the computational burden for many methods [89]. Hence even if the classification process leads to satisfactory results, the procedure can be prohibitively time consuming.

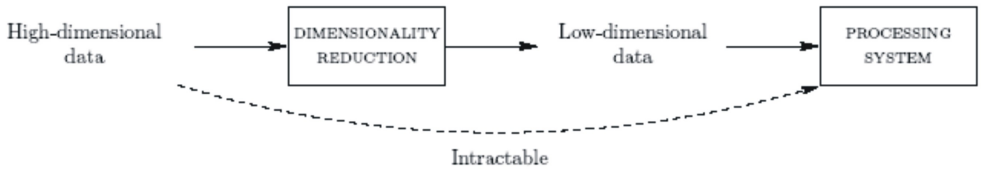


Figure 2.6: Gneral idea of the curse of dimensionality.

Here the curse of dimensionality means the intractability of *accurately approximation* of a high-dimensional Lipschitz-continuous function f: If one wants to approximate f in d variables, one needs evaluations on a grid in order $\mathcal{O}(1/(\epsilon^d))$ to obtain an integration scheme with approximation error ϵ. Hence we also found an completely analogue intractability of *integrating f in d variables*. The same fact can be observed in *optimization by exhaustive search* [61].

The sense of continuous complexity we encounter in the second part of this thesis deals with the task of solving extremely large-scale nonlinear optimization problems of the form

$$\min_{x \in \mathcal{X}} f(x) \qquad f \in \mathbb{C}^1 \tag{2.5}$$

where the nonempty, convex and compact set $\mathcal{X}$ is the feasible domain of the problem and $f : \mathbb{R}^d \to \mathbb{R}$ is a Lipschitz-continuous function with Lipschitz constant $L_{\|\cdot\|}(f)$. By means of these properties (2.5) belongs to a certain class $\mathbf{C}$ such that numerical methods are designed to use the class characteristics in order find an approximate solution with numerical error δ. There are two measures of the complexity of (2.5) [163]:

- The analytical complexity $cpl_\delta(\mathbf{C})$ is the lower bound on the number of calls of a black-box routine, called oracle, which returns on an input point $x \in \mathcal{X}$, the value of f and a subgradient f' of f at x, required to solve the problem up to the given accuracy δ.
- The arithmetical complexity $\mathcal{O}(\cdot)$ is the total number of the arithmetic operations necessary to find an approximation solution to (2.5).

In a sense the information that is accumulated up to the k^{th} step of the optimization process can be given by some kind of information base $\mathcal{D} \stackrel{\text{def}}{=} \{f(x_1), f'(x_2), \ldots, f(x_k), f'(x_k)\}$ and is associated with a characteristic complexity.

In [242; 243] the prove of the negative result that a static nonlinear optimization problem is subject to an inherent curse of dimensionality irregardless of whether deterministic or random algorithms are used was given the first time. However it is possible to break the curse of dimensionality for certain subclasses of problems such as convex optimization problems [241] since the arithmetical complexity for d-dimensional for that class of problems is only $\mathcal{O}(d \log d)$. Some main results on information-based complexity of convex programming provided in [163] are listed below. As usual $\mathcal{X}$ denotes a convex compact set with a nonempty interior and $\mathcal{F}$ is the family of all convex, but not necessarily smooth functions on $\mathbb{R}^d$ normalized by

$$\max_{x \in \mathcal{X}} f(x) - \min_{x \in \mathcal{X}} f(x) \leq 1 \tag{2.6}$$

In the following $\mathcal{O}(1)$ denotes an appropriately chosen positive constant.

Suppose that we ignore the geometry of $\mathcal{X}$. Then it holds for every $f \in \mathcal{F}$, that the analytical complexity $cpl_\delta(\mathcal{F})$ in fixed dimension is nearly independent of the geometry of $\mathcal{X}$. In particular, we get

$$\forall\, \delta \leq \delta_{\mathcal{X}} \quad : \quad \mathcal{O}(1) d \ln\left(2 + \frac{1}{\delta}\right) \leq cpl_\delta(\mathcal{F})$$

where $\delta_{\mathcal{X}}$ has the upper bound $\frac{1}{d^2}$ and depends on the geometry of $\mathcal{X}$. Definitely these are bad news: There is no hope to guarantee an accuracy e.g. of order 10^{-6}, when solving large-scale problems with black-box-oriented methods. With $\mathcal{O}(d)$ steps per accuracy digit and at least $\mathcal{O}(d)$ operations per step to introduce a new search point to the black box routine, the arithmetic cost per accuracy digit has at least a lower bound of $\mathcal{O}(d^2)$, which is prohibitively costly for high values in d.

However if we pay attention to the geometry of the feasible set $\mathcal{X}$ of the problem, there are some good news:

- Consider the d-dimensional box $\mathcal{X} := \{x \in \mathbb{R}^d \,:\, \|x\|_\infty \leq 1\}$. Then we get

$$O(1) d \ln \frac{1}{\delta} \leq cpl_\delta(\mathcal{F}) \tag{2.7}$$

- For a d-dimensional ball $\mathcal{B}_d := \{x \in \mathbb{R}^d \,:\, \|x\|_2 \leq 1\}$ it holds

$$\frac{O(1)}{\delta^2} \leq cpl_\delta(\mathcal{F}) \tag{2.8}$$

- Finally we consider the d-dimensional hyperoctahedron $\mathcal{X} := \{x \in \mathbb{R}^d \,:\, \|x\|_1 \leq 1\}$. In this case, it holds

$$\frac{O(\ln d)}{\delta^2} \leq cpl_\delta(\mathcal{F}) \tag{2.9}$$

In other words there are circumstances where the analytical complexity of minimizing a convex function to a fixed accuracy δ depends only weak on the dimension d of the data space. In spite of the fact that the complexity bounds from above are not polynomial in $\ln(\delta^{-1})$, this might be tolerable when a medium accuracy of the numerical solution is sufficient for the application on focus.

2.3 Reduction of Dimensionality in Data

A long-standing problem in statistics is how not to deduce by analysis but to infer from observations a suitable low-dimensional representation of high-dimensional multivariate data [89], that conveyed by fewer dimensions the same information if the variables are wisely combined. Representation here means that we would like to transform the data, so that the resulting set of low dimensional variables shows its essential structure and captures according to some criterion the content in the original data. Dimensionality reduction is the formal process by which we represent a system that appears as having several degrees of freedom using a smaller number of degrees of freedom instead [67].

In order to illustrate the task of dimension reduction in more detail we consider a certain phenomenon governed by $m \in \mathbb{N}$ stochastically independent variables. In measurements the observable of this phenomenon will actually appear in $\mathbb{R}^d$. The additional degrees of freedom may stem from the influence of a variety of uncontrolled components e.g. noise,

imperfection in the measurement system or the addition of irrelevant observable. In that sense m can be considered as the intrinsic dimension of a phenomenon. From a geometrical point of view m is the dimension of a manifold that approximately contains the structure of the sample data. Alternatively we can consider this structure as a low dimensional signal embedded in high dimensional noise.

More formally the problem of dimension reduction can be stated as follows. Let the m-dimensional manifold $\mathcal{I}$ embedded in a d-dimensional input data space and consider a smooth function $f : \mathbb{R}^d \to \mathbb{R}^m$ having the rank m. Then, the function f is called an embedding. Usually $\mathcal{I}$ is called the feature or latent space. Given a set of N independently identically distributed observations of the d-dimensional random vector $X_1, X_2, \ldots, X_N$, the dimensionality reduction refers to the estimation of the unknown lower (intrinsic) m-dimensional vector $Y = (Y_1, Y_2, \ldots, Y_m)$, such that we get a data model with additive noise

$$X = f(Y) + Z \tag{2.10}$$

with Z denoting the uninteresting noise. As usual the vector Y is the latent or *hidden variable* or *component* in the linear case. If the perturbations do not mask the original model, dimension reduction techniques may be appropriate for understanding the underlying phenomena of interest [67]. The determination of the intrinsic dimensionality m of a process contained in the target space $\mathcal{I}$ is central to the problem of dimensionality reduction, because knowing it would exclude over- or underfitting [89]. Intuitively reduction of dimensionality may be sound if e.g. any of the variables have a variation smaller than the measurement noise and are thus irrelevant or are correlated with each other trough linear or functional dependence. In the latter case a new set of uncorrelated variables should be found.

Fortunately in some applications the underlying variability in observed data is known to result from only a handful of physical interpretable variables, providing a strong motivation for seeking low-dimension representations of the data before attempting a statistical analysis. However, dimension reduction without loss of information is only possible if the data fall exactly on a smooth, locally flat subspace. But real data are typically noisy. Therefore in most cases an exact mapping do not exist. We can only hope to find a mapping approximately preserving some properties of the original data. There will be hardly a single tool that can outperform all the others in every practical situation.

The standard approach to deal with the high dimensional data is based on *structural assumptions* which allows to reduce the complexity or intrinsic dimension of the data without significant loss of statistical information [188; 156]. Since this thesis is interested in linear unsupervised methods we will briefly report the associated underlying frameworks in the next section.

2.3.1 The Continuous Latent Variable Model

Latent variable models are frequently used probabilistic, but not necessarily parametric models that aim to explain a high-dimensional stochastic process in terms of only a small number of continuous, so called *latent variables*. To this end a low-dimensional manifold where the data would live if there is no noise, is sought [67].

A *continuous latent variable model* has three main parts: a prior distribution in the latent space with dimensionality $m < d$ spanned by the latent variables, a smooth, i.e. continuously differentiable mapping from latent to original data space and last not least a probabilistic noise model [12]. To be more precise we define a distribution $I\!P(x) = I\!P(x|g(y))$ with observables $x \in \mathbb{R}^d$ and pairwise stochastically independent latent variables $y \in \mathcal{I}$. The function $g : \mathcal{I} \to \mathbb{R}^d$ is a smooth non-singular mapping. Then the joint data probability is given by

$$I\!P(x) = \int_{\mathbb{R}^d} I\!P(x, y)\, \mathrm{d}x = \int_{\mathbb{R}^d} I\!P(x|y) I\!P(y)\, \mathrm{d}x \tag{2.11}$$

Typically the effect of g is absorbed by $I\!P(x|y)$. Thus, a special latent variable model is essentially a specification of $I\!P(x)$ and $I\!P(x|y)$. The only empirical evidence available concerns $I\!P(x)$ through the sample data and so the only on the probabilities $I\!P(x)$ and $I\!P(t|x)$ is given by (2.11).

In general choosing functional forms for these parts gives different latent variable models. For example, a linear mapping with normal prior and noise give the *Factor Analysis* model (FA) [13]. Other models from this class are *Principal Component Analysis*, *Independent Component Analysis* and *Independent Factor Analysis* (IFA) [117; 111; 10]. In a broad sense many probabilistic models in statistics and machine learning can be considered as latent variable models inasmuch as they include probability distributions for variables which are not observed. Prominent examples are mixture models, where the variable, which indexes the components, is a latent variable, the *Hidden Markov Models* (HMM) assuming that the state sequence is unobserved, Helmholtz machines [50] and *elastic nets* [63].

Due to the use of a prior in latent variable models a dimensionality reduction mapping can be defined from observed to latent variables via Bayes' theorem in many cases. The objective then is to learn the low dimensional generating process defined in terms of latent variables using the noise model, rather than to learn the dimensionality reducing mapping itself. A natural choice of $I\!P(x|g(y))$ should fulfill the following properties: $\forall y \in \mathcal{I} : \; I\!E[X|Y] = g(y)$, where $I\!E[X]$ denotes the expectation and $I\!P(x|g(y))$ should decay gradually as the distance of x to $g(y)$ increases according to some parameter related to the noise covariance. Moreover $I\!P(x|g(y))$ assigns nonzero probability to every point in the observed space. Finally $I\!P(x|g(y))$ should have a diagonal covariance matrix to account for different scales in the different observed variables. A justification of the last request comes from the central limit theorem: if the noise is due to the combined additive action of a number of uncontrolled variables of finite variance, then its distribution will be asymptotically normal. In all the specific latent variable models like IFA or PCA this is realized using the normal distribution. Disadvantages of this choice are its unrealistic symmetry and that its tails decay very rapidly, which reduces robustness against outliers [105].

The latent variable framework is very general, accommodating arbitrary mappings and probability distributions. But the traditional treatment of latent variable models in statistic literature is restricted to the linear normal model since the rigorous analysis of nonlinear models is difficult [134]. However due to the exponential increase of the computational burden to integrate high dimensional functions numerically, this presents severe mathematical and computational difficulties, particularly in the evaluation of integral (2.11) or when maximizing the log-likelihood. In fact, the only tractable case in arbitrary dimensions seems to be when both the prior $I\!P(y)$ in the latent space and the noise model are mixtures of Gaussians or Dirac deltas and when the mapping f is linear. Hence not every choice for the components lead to models, that can be handled, and to convenient

algorithms for parameter estimation. Consequently an alternative framework for the case of additive noise realized in given high dimensional data is sought.

2.3.2 Semi-Parametric Framework for Dimension Reduction

According to the popular parametric approach, that includes some of the latent variable models, the data density function $\rho(x)$ belongs to a parametric family $\mathcal{F} \stackrel{\text{def}}{=} \{\rho_\theta(x)\}$ where θ is an finite dimensional parameter which uniquely identifies the data density $\rho(x)$. Then an algorithmic procedure to find a reliable value for an estimator $\widehat{\theta}$ of θ has to be applied to the data. However parameter estimation has to tackle with the curse of dimensionality and the execution time to evaluate a multivariate density typically has an exponential growth in the number of dimensions [203]. Another drawback of parametric modelling is the requirement that both the structural model and the error distribution have to be correctly specified. In order to avoid these drawbacks, we apply a more flexible semi-parametric approach. The semi-parametric framework for dimension reduction used in this thesis is firstly introduced in [212] and more general compared to the well known latent variable model and only concerned with the case (2.10). We combine a parametric form for most of components of the data generating process with weak non-parametric restrictions on the remainder of the data density.

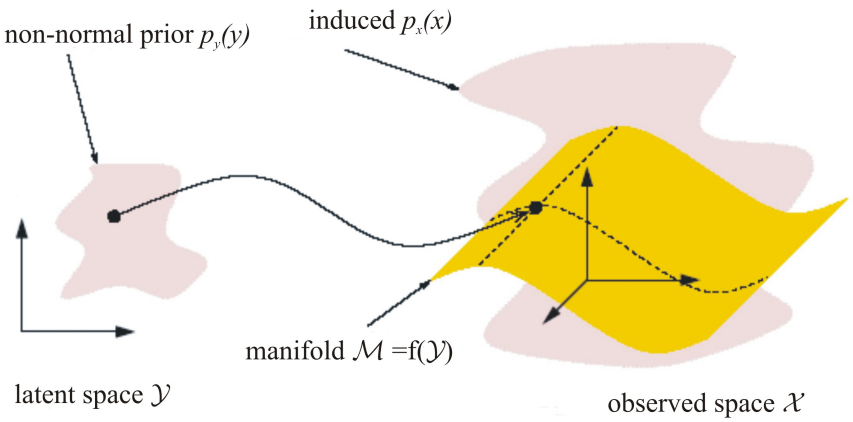


Figure 2.7: General idea of dimension reduction to a reduced space using a structural assumption on the density where p_x and p_y denote the distribution of the observed and the latent variables respectively.

Typically the problem of dimension reduction decomposes into two tasks: First one has to determine elements from the target space. Second, one has to construct a basis of the target space from these elements. Considering the last task we assume in the semi-parametric framework the following stationary data model. Let $X_1, ..., X_N$ be i.i.d. random observable from a distribution $I\!P$ in $\mathbb{R}^d$ describing the random phenomenon of interest. We suppose that $I\!P$ possesses a density ρ w.r.t. the Lebesgue measure on $\mathbb{R}^d$, which can be decomposed as follows:

$$\rho(x) = \phi_{\mu,\Sigma}(x) q(Tx). \tag{2.12}$$

In the sequel we will call this the *semi-parametric assumption* that is illustrated in figure 2.7. In (2.12) $\phi_{\mu,\Sigma}$ denotes the density of the multivariate normal distribution $\mathcal{N}(\mu, \Sigma)$ with expectation $\mu \in \mathbb{R}^d$ and covariance matrix $\Sigma \in \mathbb{R}^{d\times d}$. The function $q : \mathbb{R}^m \to \mathbb{R}$ with $m \leq d$ has to be nonlinear and smooth. $T : \mathbb{R}^d \to \mathbb{R}^m$ is an unknown

linear operator with

$$\mathcal{I} = \ker(T)^{\perp} = range(T^{\top}). \tag{2.13}$$

(2.12) is not a unique representation. However the m-dimensional linear subspace $\mathcal{I} \subset \mathbb{R}^d$ is uniquely defined by (2.13). $\mathcal{I}$ contains the non-Gaussian distributed data and is called the *non-Gaussian subspace*. Loosely speaking, we would like to project X linearly so as to eliminate as much of the noise as possible while preserving the signal information. By analogy with the regression case [43; 141; 140], we may call $\mathcal{I}$ the effective dimension reduction space (EDR-space) alternatively. We call m the *effective* or *intrinsic* dimension of the data. In many applications m is unknown and has to be recovered from the data. Furthermore the semi-parametric assumption can be regarded as the distribution of the low dimensional *signal* Y corrupted by a full dimensional Gaussian noise Z:

$$X = Y + Z. \tag{2.14}$$

This is due to the following theorem [212]:

Theorem 1. *The density $\rho(x)$ for the model (2.14) with the m-dimensional signal Y and an independent Gaussian noise Z can be represented as*

$$\rho(x) = \phi_{\mu,\Sigma}(x) q(Tx).$$

where T is a linear operator from $T: \mathbb{R}^d \to \mathbb{R}^m$, $q(\cdot)$ is some function on $\mathbb{R}^m$ and $\phi_{\mu,\Sigma}$ is the density of the Gaussian component.

The formal proof of this theorem is given in the Appendix.

From the point of view of the interpretation of the new semi-parametric framework, the semi-parametric assumption is motivated by the well known theorem:

Theorem 2. *(maximum entropy property)*
Let $X \in \mathbb{R}^d$ be a random vector with density ρ, $\mathbb{E}[X] = 0$ and $\Sigma = \mathbb{E}[XX^{\top}]$. Then it holds

$$h(x) \leq \frac{1}{2} \log(2\pi e)^d \det \Sigma \tag{2.15}$$

where $h(X) \stackrel{\text{def}}{=} -\int \rho(x) \log \rho(x) dx$ denotes the differential entropy of X. In particular equality is attained if and only if $X \sim \mathcal{N}(0, \Sigma)$.

A nice proof of this theorem can be found in [44]. Consequently among all distributions with the same variance, the normal distribution has the least information in the sense of Fisher information as well as in the sense of negative entropy. Hence in this thesis the Gaussian components of $\rho(x)$ are considered as entropy maximizing and consequently as non-informative noise such that only the non-Gaussian components contained in the target space $\mathcal{I}$ represent the structure realized in the data.

In order to detect non-Gaussian components general contrast functions can be used. Such contrast functions J_g are formulated to have good statistical properties without requiring knowledge of their distributions and to allow simple interpretation and algorithmic implementation. They measure non-Gaussianity of the standardized random variable X compared to a standard Gaussian variable Y via a smooth non-quadratic even function $g(\cdot)$ by

$$J_g(x) = \|\mathbb{E}[g(x)] - \mathbb{E}[g(y)]\|^p \tag{2.16}$$

where $1 \leq p \leq 2$. Estimators based on optimizing generalized contrast functions have superior statistical properties than cumulant-based estimators.

Chapter 3

Nonparametric Methods For Highdimensional Data

There is an overwhelmingly number of dimension reduction methods such that in this thesis we can only give an impression of the ideas leading to certain methods that can be combined to new algorithms. In this thesis we only discuss *geometric methods* for feature extraction and dimensional reduction. In particular in a complex field like dimension reduction, it would not be advisable to anticipate the existence of any single method that can outperform all other methods for all data sets.

3.1 Taxonomy of Dimension Reduction Methods

Generally speaking, dimensionality reduction techniques aim to extract the low-dimensional information about a single one or a collection of signals from a high-dimensional data space. Figure 3.1 gives an (incomplete) impression of the diversity of methods.

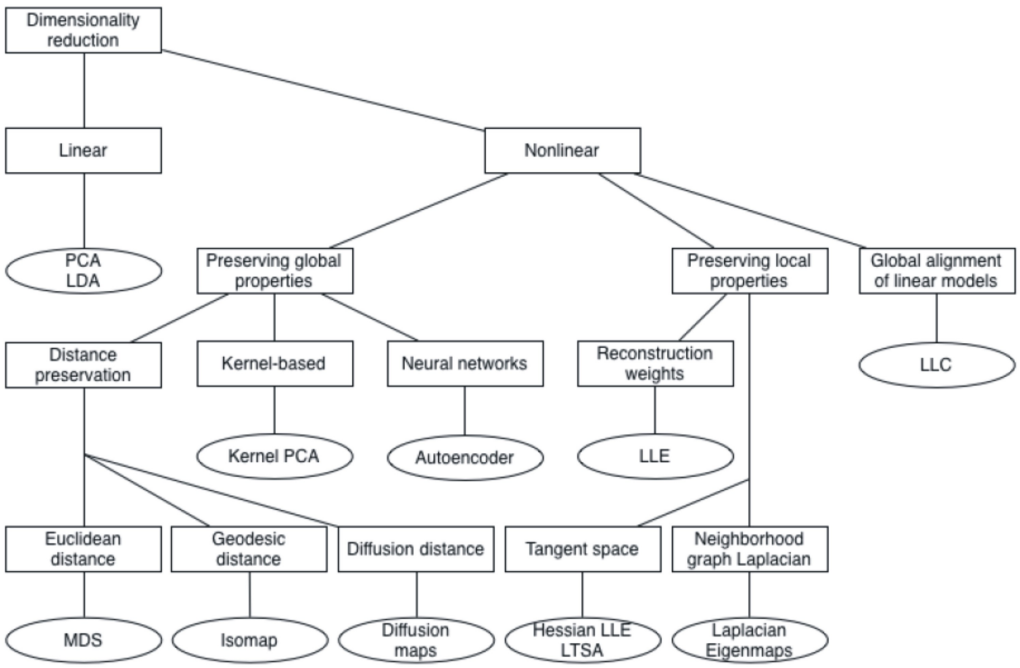


Figure 3.1: Taxonomy of dimensionality reduction techniques.

Common tasks include approximation and compression, in which the goal is to maintain a low-dimensional representation of a signal from which a faithful approximation to the original signal can be recovered.

3.1.1 Geometric Methods

Geometric methods of dimension reduction can be devided into projective methods as e.g. PCA, Linear Discriminant Analysis (LDA), Singular Spectrum Analysis [78] or ICA and methods that model the manifold on which the data lies e.g. multidimensional scaling (MDS), Isomaps [113], local-linear embedding [188], Laplacian eigenmaps [146] are some of the approaches that estimate an underlying nonlinear manifold. In this thesis we are only interested in completely data driven models that disclaim to model the assumed manifold. New techniques that allow only non-negative values in their decomposition factors such as *Non-negative Matrix Factorization* [37], *Local Non-negative Matrix Factorization* [219] and *Discriminant Non-negative Matrix Factorization* [130] were designed only for a special task and hence are not discussed also. In the age of information overload computational cost and memory requirements should be as important as intuitively appeal and theoretical reliability for a comparison of existing methods.

Linear Methods: The simplicity and efficiency of linear transformations are the main reason for their popularity for extracting features represented by some data. Examples of their applications include image compression and reconstruction, discriminant analysis, pattern classification and image retrieval. Assuming that the data are stored in the columns of $\mathbf{X} \in \mathbb{R}^{d \times N}$, the linear dimension reduction techniques attempt to decompose the data according to

$$\mathbf{X} = B\mathbf{Y} \tag{3.1}$$

where $B \in \mathbb{R}^{d \times m}$ is a matrix with each column of the input data $\mathcal{X}$ is a linear combination of the elements in B. $\mathbf{Y} \in \mathbb{R}^{m \times N}$ convey in its columns the new m-dimensional (hidden) variables. The linear combinations of random variables can be viewed as linear *projections*. Nonlinear methods reduce dimension through the use of nonlinear functions of random variables.

The linear dimensionality reduction techniques can only retrieve the linear structure of the target space. In spite of the fact that approximating a nonlinear manifold globally by a low-dimensional hyperplane as e.g. in PCA or ICA will fail in general, linear approaches will be instead useful in cases with close proximity of important data points to a linear or nonlinear underlying manifold with negligible curvature. However in some cases, the assumption that the input space can be represented as linear combination of the feature subspaces does not always account for expected results from the real-world scenarios. Therefore, undertaken a nonlinear decomposition of the input space can lead to more appropriate subspace representation [31]. However the roles of these more general methods in statistical learning remains to be fully investigated.

Sparse Models: Sparse signal models arise commonly in audio and image processing [220]. In a sparse signal model, every signal can be at least approximately represented, where the relevant set of basis functions may change from signal to signal. A simple example is the wavelet representation of piecewise smooth signals: Discontinuities in each signal can be sparsely represented in the wavelet domain and as the locations of the discontinuities vary from signal to signal, the required set of wavelets varies from signal to signal [60].

Supervised Learning Methods: Supervised techniques perform dimension reduction in a manner that allows for optimal prediction of a variable of interest, e.g. class membership or some other response variable. They are concerned with problems as inference, classification, model construction, prediction or gaining knowledge. The focus is often on e.g. measurement cost, classification errors, returns and risks minimization rather than on the accuracy of estimated probabilistic model parameters θ. Selecting some random variables e.g. as reliable predictors is fundamental [25]. Most representative methods of this type are clustering, discriminant analysis and regression methods. *Linear Discriminant Analysis* (LDA) [217] attempts to maximize the linear separability between labeled data points belonging to different classes. The aim of regression methods is to estimate the regression function, describing the relationship between a dependent (response) variable $X \in \mathbb{R}^d$ and the so called explanatory variable $Y \in \mathbb{R}^d$ in order to predict X. The relationship on focus can be, without prior knowledge and with full generality, modelled non-parametrically [69]. Regression methods can be used for dimension reduction when the goal is to model a response variable X in terms of Y. Currently there are two main approaches to deal with the curse of dimensionality in this setting: to assume a simpler form of the regression function or to reduce the dimension of the space of explanatory variable [140; 239]. More formally a dimension reduction regression model has the form

$$x = a + f(B^T y) + \epsilon \tag{3.2}$$

where is a vector of intercepts, and $B \in \mathbb{R}^{d \times m}$ is an unknown orthogonal matrix of regression coefficients. The smooth regression function f can be linear or non-linear imposing a specific structure on the regression curve $I\!E(Y|X)$ [32]. Typical examples for such function approximation methods are *Sliced Inverse Regression* [140; 43], *Principal Component Regression* [117] and *Partial Least Squares Regression* [236; 237]. In the regression context, it is generally assumed that the Y_i's were carefully selected, uncorrelated, and relevant to explaining the variation in X. In current data mining applications however, those assumptions rarely hold. Variable selection or dimension reduction is therefore needed for such cases (see for example [126]). However traditional variable selection criterions such as *Mallows* C_p [149], the *Akaike Information Criterion* [3] and the *Bayesian Information Criterion* [152] involves a combinatorial optimization problem which is NP-hard, i.e. associated with computational time increasing exponentially with increasing dimensionality. The expensive computational cost makes traditional procedures infeasible for high-dimensional data analysis.

Examples for further supervised methods related to regression include *Projection Pursuit Regression* [106], generalized linear [61; 151] and additive [88] models, neural network models, *Principal Hessian Directions* [141; 142], conditional minimum average variance estimation [239] single and multi-index models along with different fitting methods such as average derivatives [103].

Unsupervised Learning Methods: The general characteristic of such methods is to perform dimension reduction in order to optimally predict the given data from the reduced representation while preserving some properties of the original data. It is obvious that in order to develop a pure geometrical approach to metastability analysis of stochastic dynamical systems unsupervised methods are the methods of choice. Most prominent already existing methods are:

- Principal Component Analysis that finds a few orthogonal linear combinations of the X-components with the largest variance.

- Factor Analysis and Principal Factor Analysis that estimate unknown common data generating factors
- Projection Pursuit that defines the "interestingness" of a direction according to a given projection index and than locates directions maximizing that index.
- Independent Component Analysis that finds linear projections that are as nearly statistically independent as possible.
- Multidimensional Scaling, a technique for identifying a m-dimensional representation of X so that the distances among the points in the new space reflect the proximities in the data.

Since we aim to develop an unsupervised method for dimension reduction we will briefly summarize some of these methods in a subsequent subsection.

3.1.2 Feature selection vs. Feature Extraction

Currently there are essentially two approaches to *unsupervised* high dimensional data analysis: function approximation using e.g. additive models [88] or projection pursuit regression [74] and dimension reduction. Reducing the dimensionality is an effect typically involved in one of three very general mathematical tasks: feature subset selection, feature extraction and feature transformations. Although feature transformation methods [75] are actually not designed to reduce the data space they are frequently used to obtain a subspace representing some interesting features of the data [144].

Feature transformation: Feature transformations techniques are commonly used on high dimensional data sets. The transformations generally preserve the original, relative distances between the data points [233]. In this sense they summarize the data set by creating linear combinations of the features, and hopefully uncover a latent structure. For dimension reduction a feature transformation method must be combined with a subset selection criterion [180]. Feature transformations like PCA, Nonnegative Matrix Factorization, *spectral clustering* [171] or Factor Analysis are often a preprocessing step, allowing e.g. a clustering algorithm to use just a few of the newly created features [55]. In spite of the fact that they attempt to summarize a data set in fewer dimensions by creating combinations of the original features, these techniques do not actually remove any of the original features from consideration. Consequently information from irrelevant dimensions such as noise is preserved, making these techniques ineffective when there are large numbers of irrelevant features that mask the data structure. Moreover using combinations of features are difficult to interpret. Hence feature transformations techniques are currently best suited to data sets where most of the dimensions are relevant to the task to be performed, but many are either highly correlated or even redundant.

Feature selection: Concerning unsupervised learning methods it is important to make another distinction between feature selection and feature extraction methods. The term *feature selection* refers primarily to algorithms that select the best subset of the input feature set according to some given criterion, while neglecting variables that do not contribute to a given classification task. The general aim is to seek m features out of the available, but not necessarily labeled N data [75]. Feature selection algorithms fall into two broad categories, called the filter model and the wrapper model [81]. The filter modelling relies on general characteristics of some given training data to select features without involving any learning step. The wrapper model however requires at least one learning step. This can be a hypothesis or a classifier. Then the result is used to evaluate and to determine

which features have to be selected [115]. The wrapper model tends to find features better suited to the learning step resulting in a superior learning performance. However it also tends to be more computationally expensive than the filter model [25]. Since the number of possible subsets grows combinatorically, feature selection approaches are impractical for more than moderate values of d and m [159].

Feature extraction: In this case data features are extracted algorithmically as linear or non-linear functions of the original set of features. According to some predefined error criterion, such methods are designed to map the original data space into a lower dimensional subspace of non-redundant features, preserving as much as possible of the local structure in the original data. Supervised feature extraction techniques usually relate to the discriminant analysis technique [75], which typically uses the within and between-class scatter matrices. Currently there is a growing number of methods [7] assuming that the data points form low-dimensional nonlinear and apriori given manifolds: Nonlinear mappings e.g. *Isomaps* [113], *Multidimensional Scaling* [133] and *Sammon's Mapping* [193], *Local-Linear Embedding* [188], *Laplacian Eigenmaps* [146] are some of the approaches that estimate such underlying manifolds. However the roles of these more general methods in statistical learning remains to be fully investigated. In comparison linear representations are well understood and attractive due to their simplicity and computational efficiency.

We will now discuss some methods of the last type in more detail including some very popular feature transformation methods in order to point out that there is a need for new linear, completely data driven feature extraction method with sufficiently low complexity.

3.2 Unsupervised Linear Methods of Feature Extraction

Non-probabilistic methods for dimensionality reduction, are methods that do not assume a probabilistic model for the data. These include linear methods as for example PCA, projection pursuit, kernel methods, principal curves, vector quantization methods as elastic net or self-organizing maps or multidimensional scaling methods. Since we are here interested to present the ideas behind the methods on focus, numerical and algorithmic aspects are discarded.

3.2.1 Pure Gaussian Analysis

Principal Component Analysis also known as the singular value decomposition (SVD), Karhunen-Loeve transform, Hotelling transform, or empirical orthogonal function method, is the most simple method for dimension reduction. Moreover most of the other methods that were formulated up to now in the continuous latent variable framework described in section 2.3.1 have analytical relations to PCA.

The purpose of PCA is to identify the dependence structure behind multivariate stochastic observations in order to obtain a compact description of it, such that projecting the data loses as little information represented by data as possible. When the observed variables have a non-zero correlations, the dimensionality d of the data space does not represent the number m of independent variables, that are really needed to fully describe the data: The more correlated the observed variables are, the smaller is the number of independent variables that can adequately describe them. Due to three important properties PCA is a widely used dimensionality reduction technique in data analysis: First, it is a linear distribution-free scheme which is optimal in the sense of the 2-norm for compressing a set of high dimensional vectors into a set of lower dimensional vectors. Second, the model

parameters can be computed directly from the data - e.g. by diagonalizing the data covariance matrix. Third given the model parameters compression and decompression require only matrix multiplications.

The basic idea PCA is to project the given data onto a hyperplane spanned by the eigenvectors of the estimator of the data covariance matrix. To be more precise assume that i.i.d. random variable $\{X_i\}_{i=1}^N \in \mathbb{R}^d$ are given and $Y_j = \Pi^\top (X_i - I\!E_N[X])$ with $\Pi \in \mathbb{R}^{d\times d}$ denote the linear transformation to the so called principle components or features $Y_j \in \mathbb{R}^d$. Then the sought orthogonal transformation fulfills [13]

$$\widehat{\Pi} = \arg\max_{\Pi\,:\,\Pi\Pi^\top = I}[-I\!E\|X - \Pi\Pi^\top X\|^2] = \arg\min_{\Pi\,:\,\Pi\Pi^\top = I} \mathrm{Tr}[\Pi^\top I\!E[XX^\top]\Pi]$$

As a consequence we get for all i, j:

$$I\!E[Y_j] = 0 \qquad I\!E[Y_j Y_j] = \lambda_j \qquad I\!E[Y_i, Y_j] = \delta_{ij}$$

where $\{\lambda_j\}$ are the eigenvalues of $I\!E[XX^\top]$ and δ_{ij} denotes the Konnecker symbol. Typically for the purpose of dimensionality reduction principal components with small variance are discarded. There is a variety of proposals how to do this in a reasonable way (see e.g. [13]).

It is well known that PCA attains the best dimensionality reduction map $\Pi_m^\top \in \mathbb{R}^{m\times d}$ in the sense of maximal variance [117] in the projected space

$$\max_{\Pi\Pi^\top = \mathbf{1}} \left\{ \frac{1}{N} \sum_{i=1}^N X_i X_i^\top \right\} = \sum_{n=1}^d I\!E[Y_n^2] = \sum_{n=1}^d \lambda_n$$

and in the sense of least squared sum of errors of the reconstructed data

$$\min_{\Pi^\top} \left\{ \frac{1}{N} \sum_{i=1}^N \|X_i - \Pi\Pi^\top X_i\|^2 \right\} = \sum_{n=m+1}^d \lambda_n$$

Thus the maximization of the projection variance is equivalent to the minimization of the mean squared reconstruction error. Moreover assuming the data vectors are normally distributed, the mutual information $I(X, Y)$ between the original vectors and their projections on the hyperplane is maximal [44]:

$$\max_{\Pi^\top}\{I(X, Y)\} = \frac{1}{2} \ln \left[\prod_{n=1}^m 2\pi e \lambda_n \right]$$

where $\lambda_1, \ldots, \lambda_m$ are the first m eigenvalues of $I\!E[XX^\top]$. Due to these Gaussian assumptions PCA can be viewed as a limiting case of a particular class of linear-Gaussian models

$$Y = AX + \epsilon \quad X \sim \mathcal{N}(0, \mathbf{1}) \quad \epsilon \sim \mathcal{N}(0, \Sigma)$$

where X_i and ϵ_i are assumed to be i.i.d. as the covariance of the noise becomes infinitesimally small and equal in all directions of the data space. The PCA-projector can be computed by the SVD of the empirical covariance matrix of the data with computational complexity $\mathcal{O}(d^3)$. In case of Gaussian data, the dimension of the reduced linear space might be obtained from a gap in the spectrum of the empirical covariance matrix. Since the variance depends on the scale of the variables, it is customary to first standardize each

variable to have mean zero and standard deviation one.

There are some statistical limitations for the application of PCA: On the one hand for many cases the demand of orthogonality of the components is too strict. On the other hand if the geometrical data distribution is far from being Gaussian PCA will introduce artificial correlations in the reduced data set and hence can not be combined with other statistical method as ICA. Moreover PCA is obviously not designed to find a clustering structure and for non-Gaussian data the result of PCA may be difficult to interpret or even misleading. Furthermore the sample covariance matrix is strongly influenced by extreme outliers. Consequently all methods relying on it will not be robust against outliers.

3.2.2 Multidimensional Scaling

The term "multidimensional scaling" (MDS) [45; 185] covers a variety of techniques that analyze a matrix of dissimilarities. In general MDS searches for a linear mapping of the dissimilarities to a low dimensional Euclidean space such that the pair-wise dissimilarities become squared distances under the constraint that the original d-dimensional feature space is preserved as faithfully as possible in the target space [133; 205]. The dissimilarities are typically measured by a distance measure: The more similar two items are, the smaller is their distance. The methods differ by their used error measure. Various stress or objective functions are used for measuring the performance of the mapping [26]. The most popular criterion is the so called stress function [193]. The complexity of MDS is at least $\mathcal{O}(d^3N^2)$.

The two basis types are metric and non-metric MDS. The metric MDS assumes that the data are quantitative, so that there exists a functional relationship between the inter-point distances and the given dissimilarities [193]. Non-metric MDS assumes the data to be qualitative, having ordinal significance and procedures produce configurations that attempt to maintain the rank order of the dissimilarities.

Metric Multidimensional Scaling: In the classical version of MDS the first step is to compute the matrix A of squared point proximity measures e.g. the empirical covariance matrix. Then one has to conduct the so called double centering by $B = -J^\top AJ$ using the centering matrix $J = I - N^{-1}ee^\top$ where $e^\top$ is the column of N ones. Obviously J has rank $N-1$ and projects onto the subspace $\mathbb{R}^{N-1}$ orthogonal to e. It is shown [26] that B is positive semidefinite if and only if A is the distance matrix with embedding space $\mathbb{R}^m$. Moreover the minimal value for m is the rank of B and the embedding vectors are any set of Gram vectors scaled by $\frac{1}{\sqrt{2}}$. Consequently in order to find the embedding vectors for a given distance matrix, it is sufficient to extract the m largest positive eigenvalues of B and the corresponding m eigenvectors. Therefore a m-dimensional spatial configuration of the N objects is derived from $V_m\Lambda_m^{\frac{1}{2}}$, where V_m is the matrix of m eigenvectors and Λ_m is the diagonal matrix of m scaling eigenvalues of B.

In other words MDS typically [193] uses a stress function of the form

$$\sum_{i=1}^{N}\sum_{j=i+1}^{N}\frac{(f(x_{ij})-d_{ij})^2}{a_{ij}} \tag{3.3}$$

in order to measure the discrepancy between the given dissimilarities and the derived distances. Here $0 \leq a_{ij}$ refer to scaling factors and d_{ij} to some geometric distances measure. Metric scaling uses $f(x_{ij}) = x_{ij}$ and the 2-norm, i.e. the data is compared

directly to the Euclidean distances. The minimization of that stress function yields a projection onto the first m principal components if d_{ij} are the Euclidean distances [75]. Based on (3.3) several non-linear versions [88; 151] of MDS have been developed.

3.2.3 Probabilistic Approaches

Factor Analysis: Factor analysis (FA) [67] as an other example for the continuous latent variable model uses a Gaussian distributed latent space prior $y \sim \mathcal{N}(0, I)$, an additive model of uncorrelated noise and a linear mapping from the data space to the latent space

$$x = Ay + \mu + z \tag{3.4}$$

where $z \in \mathbb{R}^m$ is the noise term. The latent variables y are often referred to as factors. The columns of the $d \times m$ matrix A are referred to as the factor loadings. FA explains the observed covariance structure in the data. We assume $rank(A) = m$, i.e. linear independent factors although there exist varieties of factor analysis where these factors y are correlated. In addition the factors are assumed to be standardized with variance one. The noise model is normal centered at μ with diagonal covariance matrix Ψ. In general A and Ψ must be estimated and the conditional distribution of the observed random variable is

$$I\!P(x|y) \;=\; \mathcal{N}(Ay + \mu, \Psi). \tag{3.5}$$

Consequently the marginal distribution is given by

$$I\!P(x) \;=\; \mathcal{N}(\mu, AA^\top + \Psi). \tag{3.6}$$

where $\sum_{j=1} a_{ij}^2$ represents the variance of x_i common to all hidden variables. If several variables x_i have high loadings a_{ij} on a given factor y_j, then these variables measure the same unobservable quantity and are therefore redundant. Unlike PCA the factor model does not depend on the scale of the observed variables.

For the posterior distribution in the latent space we get [13]:

$$I\!P(y|x) \;=\; \mathcal{N}(B^{-1}A^\top(y - \mu), \Psi^{-1}B) \tag{3.7}$$

where $B = AA^\top + \Psi$. The reduced-dimension representative is taken as the posterior mean (coinciding with the mode) and is usually referred to as Thomson scores:

$$I\!E[Y|X] = B(x - \mu)$$

Since (3.4) is invariant to rotations, we can apply an invertible linear transformation g with matrix R to the factors y in order to obtain a new set of factors $y' = Ry$. The prior distribution $p(y')$ is still normal and the new factor loadings are given by $A' = AR^{-1}$. Only if R is an orthogonal matrix the new factors y' will still be independent and $\Psi' = \Psi$. Thus from all the factor loading matrices A, we are free to choose that one which is easiest to interpret according to some application-dependent criterion.

The parameters (A, Ψ, μ) are obtained as the log-likelihood of a normal distribution $\mathcal{N}(\mu, \Sigma)$ with covariance $\Sigma = AA^T + \Psi$:

$$\mathbb{L}(A, \Psi) = -\frac{N}{2}\Big(d\ln(2\pi) + \ln(|\Sigma|) + \mathrm{Tr}(\widehat{\Sigma}\Sigma^{-1})\Big)$$

where $\widehat{\Sigma}$ is the sample covariance matrix. In the case of FA the log-likelihood gradient is given by:

$$\begin{aligned}\frac{\partial \mathbb{L}}{\partial A} &= -N(\Sigma^{-1}(I - \widehat{\Sigma}\Sigma^{-1})A) \\ \frac{\partial \mathbb{L}}{\partial \Psi} &= \frac{N}{2}\mathrm{diag}(\Sigma^{-1}(I - \widehat{\Sigma}\Sigma^{-1}))\end{aligned}$$

In theory different local maxima are possible and should be due to an underdetermined model. The parameters of a factor analysis model may be estimated using e.g. an expectation-maximization algorithm [191; 192] with arithmetical complexity $\mathcal{O}(N^2d^2)$.

Probabilistic PCA: PCA can also be expressed as the maximum likelihood solution of a probabilistic latent variable model. In this form *Probabilistic PCA* (PPCA) answers the question, how to construct a mixture of PCA models [222] and is an extension of factor analysis as it assumes a model of the form (3.4) with $\Psi = \sigma^2 I$ without assuming the model and sample covariances being equal. Hence only σ^2 must be estimated from the data. Furthermore it is assumed that the smallest $d - m$ eigenvalues of the model are all equal to σ^2 and that the sample covariance $\widehat{\Psi}$ is equal to the model covariance. Considering the eigenvalue decomposition of $\widehat{\Psi}$ it is shown [223], that fortunately the resulting maximum likelihood estimates of A and σ^2 have a closed form:

$$\begin{aligned}A &= V\sqrt{(\Lambda - \sigma^2 I)}R \\ \sigma^2 &= \frac{1}{d-m}\sum_{i=m+1}^{d}\lambda_i\end{aligned}$$

where V is the matrix of the m principal column eigenvectors of the sample covariance matrix. Λ is the corresponding diagonal matrix of principal eigenvalues λ_i and R is an arbitrary orthogonal matrix in the latent space. Thus σ^2 captures the variance lost in the discarded FA projections and the PCA directions appear in the maximum likelihood estimate of A. In the special case $\sigma \to 0$ and $R = I$ the factors y become the PCA projections of the x. The advantages of a probabilistic model are obvious: for example the weight that each mixture component gives to (3.7) of a given data point can be computed. Moreover PPCA allows to perform PCA in case of partially missing components and m can be estimated using Bayesian techniques [22]. However PPCA is limited to second order statistics as well as PCA. The arithmetical complexity of this method is $\mathcal{O}(d^3)$.

Random Projections: The method of random projections is a technique that uses random matrices to map the data into lower dimensional spaces preserving distances nicely. To this end the original i.i.d. data obtained from $X \in \mathbb{R}^d$ is transformed to the lower dimensional representation $Y \in \mathbb{R}^m$ with complexity $\mathcal{O}(mdN)$ via $Y = \Pi X$ where the columns of $\Pi \in \mathbb{R}^{m\times d}$ are typically but not necessarily realizations of i.i.d. zero-mean normal variables scaled to unit length. Strictly speaking, there is no such projection since Π is not orthogonal, since otherwise the linear mapping can cause significant distortions in the data set. However orthogonalizing Π is computationally expensive. Instead, the method relies on the proposition [92] that in a high-dimensional space there exist a much larger number of almost orthogonal than non-orthogonal directions. Thus vectors having random directions might be sufficiently close to orthogonal, and equivalently $\Pi^T\Pi$ would nicely approximate an identity matrix.

The key idea of random mapping arises from the Johnson-Lindenstrauss lemma: If points are projected onto a randomly selected subspace of suitably high dimension, then the distances between the points are approximately preserved [47].

Theorem 3. *(Johnson-Lindenstrauss Lemma)*
Let $\mathbf{X} \in \mathbb{R}^{d \times N}$ *be a data matrix and* $\beta > 0$ *be the failure probability. Then there is a linear but non-sparse mapping* $\Pi \in \mathbb{R}^{d \times m}$ *such that all pairwise distances are preserved up to a uniform distortion* ϵ*:*

$$\forall i, j: \quad I\!P\Big(\frac{1}{1+\epsilon} \leq \frac{\|\Pi x_i - \Pi x_j\|^2}{\|x_i - x_j\|^2} \leq 1+\epsilon\Big) \geq 1 - N^2\beta \tag{3.8}$$

if $m \geq 9\ln(N)(\epsilon^2 - \epsilon^3)^{-1}$.

A proof of this theorem can be found in [116]. Furthermore it is shown, that the choice of at least $\mathcal{O}(N)$ directions will result in a projector having arbitrarily high probability of preserving distances. An improved approach to this method is given in [1], where the projection matrix Π is more easier constructed in one of the following ways paying the price of a slight loss in accuracy:

- $\Pi_{ij} = \pm 1$ with probability 0.5.
- $\Pi_{ij} = \sqrt{3} \pm 1$ with probability 1/6 or 0 with probability 2/3 [128].

Then on the average the distortion of the inner products is zero and its variance is at most $2/d$ [120]. It has been shown empirically that results obtained by random projections are comparable with results obtained with Principle Component Analysis [120].

However, there are several drawbacks of this method: Random projections are highly unstable meaning that different random projections may lead to radically different projection results.

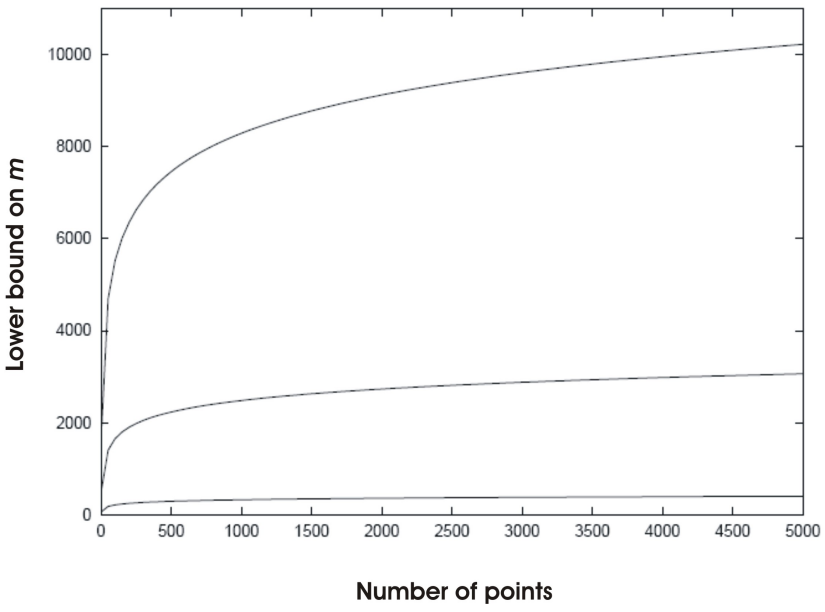


Figure 3.2: Illustration of lower bound for m for random projections as a number of data points. The upper curve corresponds to $\epsilon = 0.1$, the middle to $\epsilon = 0.2$ and the lower to $\epsilon = 0.5$

Moreover even if the elements of x are mutually independent the same does not hold for the elements of Πx when conducting such a projection. This is due to the fact that each element of $y = \Pi x$ is a linear combination of x with a column of Π. Last not least figure 3.2 illustrates that the lower bound of the reduced dimensionality of the embedding is not dominated by the data properties but depends on N. In particular the last fact is not a recommendation for the use of that method with large data sets.

3.3 Unsupervised Nonlinear Methods of Feature Extraction

3.3.1 Nonlinear PCA

First of all nonlinear versions of PCA like e.g. *Principle Curves* [87] have to be distinguished from *generalized PCA*, which is not a method to find nonlinear dependencies between random variables, but an algebraic tool for subspace clustering based on fitting data with a set of polynomials. The intuitive idea of generalized PCA is to rewrite the target space bases in terms of factors of the polynomials [230].

It is well known that the purpose of PCA is to identify linear correlations between random variables [117]. This may be an appropriate assumption for many data sets. In some cases however, it may be more appropriate to assume that the hidden factors are nonlinear functions of the variables. The problem in *Nonlinear PCA* (NLPCA) is then to minimize the mean squared reconstruction error

$$\arg\min_{\Pi,\Pi\Pi^\top=I} I\!E[\|x - \Pi g(\Pi^\top x)\|_2^2] \tag{3.9}$$

where $y = g(x)$ are the non-linear principal components [176] and $g \in \mathcal{C}^1$. Since the optimal choice of g is not unique, non-linear hidden factors depend on g. Commonly used non-linear and smooth functions $g(\cdot)$ are odd functions like $g(y) = \tanh(y)$ or $g(y) = y^3$. Nonlinear principal component analysis is commonly seen as a nonlinear generalization of PCA, not only generalizing the principal components from straight lines to curves, but including also higher order statistics. Several versions of nonlinear PCA have been proposed [57]. For illustration we be briefly summarize the kernel PCA method proposed in [229].

Kernel PCA: Standard PCA formalizes the intuition that the structure in the data only depends only on first and second moments of the data in linear subspaces, whereas kernel PCA does not have such a limitation. Kernel PCA can find non-linear subspaces with high variance. The basic idea is to extend the original features with a large number of non-linear features and then to apply linear PCA in the new feature space. Increasing the number of features means, that it becomes unfeasible to compute PCA via an eigenvalue decomposition of the data covariance matrix since its size grows quadratically with the number of features. Moreover if we use the inner product data matrix the time needed to compute the inner products will increase linearly with the increasing number of features. However it turns out that for appropriate choices of new features, we can rewrite the resulting inner product as a function of the original features that can be evaluated with much smaller arithmetical complexity than the number of new features.

The function used to compute the inner products is called "kernel function" corresponding to these features. Conversely, Mercer's theorem provides the conditions under which a kernel function computes the inner product in some associated feature space. Using a kernel allows a representation of data in extremely high dimensional spaces without explicitly mapping the data to this feature space and thus avoiding the computational

burden of using such rich representations [195]. Now recall that the so called kernel trick assumes an algorithm depending only on dot products of the data. Using the nonlinear transformation $x \mapsto \Phi(x) \in E$, where E is a Hilbert space, the algorithm depends only on the dot products $\langle \Phi(x_i), \Phi(x_j) \rangle$. Then suppose there exists a continuous, positive and symmetric kernel function $k(x_i, x_j)$ such that for all $x_i, x_j \in \mathbb{R}^d$, $k(x_i, x_j) = \langle \Phi(x_i), \Phi(x_j) \rangle$ where $\Phi(x_i)$ denotes a basis function of the feature space E. Then $\Phi(x)$ never has to be computed explicitly. Instead the kernel form can always be used.

In [195] it is shown that PCA can be written entirely in terms of dot products transforming thereby into a nonlinear feature extraction method due to some key observations: The eigenvectors of the covariance matrix in E lie in the span of the centered and mapped data. Therefore no information in the eigenvalue equation is lost if the equation is replaced by m equations, formed by taking the dot product of each side of the eigenvalue equation with each centered and mapped data point. To be more precise we have the covariance matrix

$$\Sigma = \frac{1}{m} \sum_{i=1}^{m} (\Phi(x_i) - \mu)(\Phi(x_i) - \mu)^\top \tag{3.10}$$

where $\mu = \frac{1}{N} \sum_i \Phi(x_i)$. Then the eigenvectors v of Σ lying in the span of $\Phi(x_i) - \mu$ are sought, i.e.

$$v = \sum_{i=1}^{m} a_i (\Phi(x_i) - \mu) \tag{3.11}$$

where the $a_i \in \mathbb{R}$ are suitable constants. Consequently we get m eigenvalue equations

$$(\Phi(x_i) - \mu)^\top \Sigma v = \lambda (\Phi(x_i) - \mu)^\top v \tag{3.12}$$

Introducing the so called kernel matrix with elements $K_{ij} = (\Phi(x_i) - \mu) \cdot (\Phi(x_j) - \mu)$ (3.12) can be written as

$$K^\top K u = N \lambda K u \tag{3.13}$$

where $u \in \mathbb{R}^N$. Obviously any solution of $Ku = N\lambda u$ is a solution of (3.13) also. Moreover any solution of (3.13) is a solution of $Ku = N\lambda u$ plus a vector $w \perp u$ where w fulfils $\sum_{i=1}^{N} w_i (\Phi(x_i) - \mu) = 0$ and hence do not contribute to (3.11). Consequently we only have to ask for the equality

$$1 = v \cdot v = N \lambda u \cdot u \tag{3.14}$$

Therefore u computed by an eigenvalue decomposition of the centered kernel matrix must be rescaled to have length $(N\lambda)^{-1/2}$. For dimensionality reduction purposes, the projections on the N principal components can be taken as features.

Unfortunately the complete procedure is sensitive to the used kernel, but we do not know a priori what kernel is advisable to use. Moreover in spite of the fact that with Kernel PCA nonlinear components are obtained without any nonlinear optimization, Kernel PCA obviously has a higher computational effort than classical PCA for large data sets.

3.3.2 Pure NonGaussian Analysis

Independent Component Analysis, as a special case of blind source separation, is a computational linear method for separating a multivariate, linear mixture of a number of signals from variate mathematical or physical models into additive, unknown, zero mean and time-dependent subcomponents supposing the mutual statistical independence of the univariate components of the given data density [111]. More precisely the noisy variant of ICA assumes the linear mixing model

$$X = A(Y - I\!E[Y]) + \epsilon \quad \text{with} \quad \epsilon \sim \mathcal{N}(0, \Sigma)$$

where $A \in \mathbb{R}^{d \times d}$ is the mixing matrix and Y are the independent components. The mixing matrix A needs to be of full rank, but not to be orthogonal. Since A and Y are unknowns, the variances of Y are not detectable such that a whitening pre-procedure of the data is required. ICA aims to learn unsupervised the linear transformation $W^\top \in \mathbb{R}^{d \times d}$ such that

$$y = W^\top x \qquad \text{subject to} \qquad W^\top A = \mathbf{1}_d$$

from the data where the hidden and data generating, non-Gaussian distributed components Y_i, Y_j are mutually independent. Hence the functional to be numerically maximized is

$$\widehat{W}^\top = \arg\max_{W^\top} - KL(I\!P(x_1, \ldots, x_d) | I\!P_1(x_1) \cdot I\!P_2(x_2) \cdot \ldots \cdot I\!P_d(x_d))$$

where $I\!P(\cdot)$ denotes the joint probability and $I\!P_i(\cdot)$ its marginal probabilities. KL denotes the Kullback-Leibler divergence

$$KL(I\!P_1, I\!P_2) \stackrel{\text{def}}{=} \int I\!P_1(x) \log \left(\frac{I\!P_1(x)}{I\!P_2(x)} \right) \mathrm{d}x.$$

Since it is difficult to estimate mutual information using observations of random variables, several approximations of this functional have been applied to obtain ICA in the literature. Typically ICA uses Projection Pursuit [74] as a method. In this case the basic algorithmic idea is to locate successively one-dimensional, linear projections from the high- to a low-dimensional space that reveal the most details about the structure of the data maximizing a so called index

$$\iota \in L^2(\mathbb{R}^d) \to \mathbb{R} \tag{3.15}$$

that is sensitive to the data structure on focus. Popular examples are curtosis or negentropy [110]. Typical projection indices include indices based on higher-order cumulants and on the Fisher information [112], that are minimized by Gaussian distributions. Consequently ICA is a parametric pure non-Gaussian analysis and incorporates higher than second-order information as used in PCA. However for Gaussian data PCA is a special case of ICA. Moreover ICA can be considered as another factor rotation method, where the goal is to find rotations that maximize certain independence criteria [89]. For the purpose of dimension reduction the number of independent components to be extracted from the data must be given apriori.

As an implementation of ICA we use symmetric FastICA in this thesis. This variant of ICA based on Newton's method is implemented by Hyvarinnen and Oja [112] with numerical complexity $\mathcal{O}(d^2 + N \log N + m^2)$. For convenience we briefly summarized the algorithm here. To this end let $G(*)$ be a suitable measure of stochastic independence and

n the current number of FastICA iterations. Moreover $\epsilon \in \mathbb{R}$ is an numerical error and err a bound for ϵ used as a tuning parameter in the following algorithm.

Algorithm 1: FastICA

Data: $\{X_i\}_{i=1}^N$, d, m, ϵ
Result: m-dimensional basis of target space of independent components
for *j=1* **to** d **do**
  while $\epsilon \leq err$ **do**
    Choose an initial vector $w_j \in \mathbb{R}^d$ of unit norm.
    Set: $w_j^{(n)} = I\!E_N\big[G'(w_j^{(n-1)}x)x\big] - I\!E_N\big[G''(w_j^{(n-1)}x)\big]w_j^{(n-1)}$
    where G' denotes the first and G'' the second derivative of G.
    orthogonalize : $w_j^{(n)} = w_j^{(n)} - \sum_{k\neq j}(w_j^{(n)}w_k)w_k$.
    normalize: $w_j^{(n)} = w_j^{(n)}\|w_j^{(n)}\|^{-1}$
    stopping rule: compute $\|w_j^{(n)} - w_j^{(n-1)}\| = err$
  end
end

There are well known drawbacks of ICA: The projection pursuit procedure tends to identify outliers since the presence of the latter gives it the sample appearance of non-normality. This may obscure the clusters or other interesting structure being sought. The methodological problem is the unrealistic demand on a product structure of the whole distribution and the requirement of NonGaussianity for all the components. Moreover Comon [41] showed that W is identifiable up to scaling and permutation of its columns if and only if at most one of the source signals is normally distributed. Furthermore heavy tailed densities are difficult to detect for ICA and the optimization algorithms may not work effectively if the data contains both super- and sub-gaussian contributions to the whole density [38].

3.3.3 Self-Organizing Maps

A severe problem with almost all methods to find a dimensionality reducing mapping is that one can not know apriori that the method or its parameterization is appropriate for the given data. However with the development of neural networks, some new possibilities finding non-linear dimension reducing mappings were created. There are several methods based on unsupervised finding a continuous map to transform nonlinear statistical relationships from high-dimensional data into a lower-dimensional lattice $\mathcal{L}$ of apriori given (reduced) dimension. Amongst them *Self-Organizing Maps* (SOM) are probably the most well known [23]. A self-organizing map consists of a discrete map of lattice points that will represent the topology preserving mapping [129; 127] in the sense that data clusters that are nearby in the latent space will typically contain similar data. The mapping consists of a lattice of reference points that are fitted to the data space in order to approximate its density function in an ordered way.

In order just to sketch the of SOMs idea let $d_{\mathcal{L}}$ and $d_{\mathcal{D}}$ denote a distance measure in the lattice and in the data space respectively. Further define a symmetric and unimodal neighborhood $h_{ij}\ \mathcal{L} \to [0,1]$ with $h_{ii} = 1$ for any node in the lattice $\mathcal{L}$. The further the node j is from node i the smaller is h_{ij}. By means of a threshold τ for h_{ij} the width of the neighborhood is determined. A typical neighborhood is given by

$$h_{ij} = e^{-\frac{d_{\mathcal{L}}(i,j)}{2\tau^2}}$$

Kohonen's rule uses an initially random set of reference vectors $\{r_i \in \mathbb{R}^m\}_{i=1}^n$, then updates them iteratively according to the data distribution such that the final reference vectors will be dense in regions of $\mathbb{R}^d$ where the data is clustered.

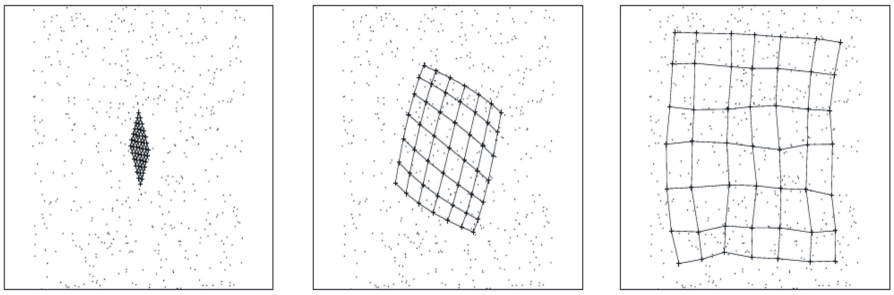


Figure 3.3: Different configurations of the SOM in the data space as the learning progresses on data depicted as dots goes on.

Until convergence the following procedure over all data is iterated:

- For a given data point x_i find the closest r_j to it in the data space by

$$i^* = \arg\max_{\{1,\dots,J\}} d_{\mathcal{D}}(x_i, r_j)$$

- Then at iteration k and with k monotonically decreasing learning rate $\alpha^{(k)} \in [0,1]$ make a gradient-type update of r_j by

$$r_j^{(k+1)} = r_j^{(k)} + \alpha^{(k)} h_{i^*j}^{(k)} (x_{i^*} - r_j^{(k)})$$

Figure 3.3 illustrates the development of the lattice as a function of the iteration count k: If the initial values for $\{r_i \in \mathbb{R}^m\}_{i=1}^n$ are not randomly selected but as a regular array lying on the subspace e.g. spanned by the largest principal components of input data, computation of the SOM may be much faster, since from the beginning the SOM is already approximately organized such that one can start with a narrower neighborhood function and smaller learning-rate factor $\alpha^{(k)}$.

Since neighboring lattice points will be neighbors in the input space, the mapping preserves topology. When maintaining an equal sampling of the lattice's space when d increases, the total complexity of SOM becomes $\mathcal{O}(dN\exp(m))$. However there are well known drawbacks of SOMs: There is no implicit criteria that SOMs try to optimize, no explicit rules to optimally update $\alpha^{(k)}$ and $h_{ij}^{(k)}$ and hence there is no proof of convergence in general [65].

In the following chapters with $\mathbb{E}_N[\cdot]$ we refer to the empirical mean of the expectation $\mathbb{E}[X]$, i.e. for any function $f(x)$ on $\mathbb{R}^d$ we set

$$\mathbb{E}_N\big[f(X)\big] \stackrel{\text{def}}{=} \frac{1}{N}\sum_{i=1}^{N} f(X_i).$$

Chapter 4

Convex Projection in Structural Data Analysis

In the last chapter we have seen that existing unsupervised projective feature extraction methods for dimension reduction as ICA, PCA, PPCA, FA, MDS, SOM, kernel PCA or RP have at most two drawbacks: On the one hand their inherent optimality criteria are often unrelated to the application or based on unrealistic assumptions about the geometric or distribution properties of the data. On the other hand concise statements about the convergence rate of the methods or the loss of information represented by the structure in the data are hardly available.

Thus in this section we introduce an alternative linear projective semi-parametrical method of dimension reduction [58] as a general preprocessing tool for other e.g. statistical or dynamical analysis methods.

4.1 The Setup of the Method

To this end we will first give an outline of SNGCA and summarize its properties in order to motivate the design of the method.

Framework of Dimension Reduction: For SNGCA we use the semi-parametric framework already described in section 2.3.2, i.e. by using the semi-parametric assumption

$$\rho(x) = \phi_{\mu,\Sigma}(x)q(Tx). \tag{4.1}$$

for a given sample from i.i.d. random variable $X_1, \ldots, X_N$. Thus we combine a parametric form for most of the components of the data generating process with weak non-parametric restrictions on the remainder of the data density: Obviously (4.1) links pure Gaussian (PCA) and pure non-Gaussian (ICA) modelling. For the sake of simplicity we assume from now on that the expectation of X vanishes: $\mathbb{E}[X] = 0$. This is easily achieved by removing the empirical mean $\mathbb{E}_N[X]$ from the data.

Non-informative Gaussian Components: As usual in the statistical literature [44] we assume that the Gaussian components in (4.1) are uninformative noise and that the structure of a data set is represented by non-Gaussian components of the data density $\rho(x)$. Note that the suggested way of treating the Gaussian distribution as a noise in general exclude the use of the classical PCA for searching the informative density components because PCA heavily relies on the Gaussian distribution of the data and looks at the directions with the largest variance.

Linear Projective Method: Obviously a linear method seems to be attractive due to its simplicity, since it is identified by a m-dimensional projector from the data space $\mathbb{R}^d$ onto the target space $\mathcal{I}$ containing the non-Gaussian structure representing component of the data density. Maybe the inquiry of data projections to some lower dimensional subset is the most simple approach to get information from the data. One realization of this strategy is already given by PP described in section 3.2.1. Methods of that type have been proven to be robust to noisy or irrelevant features [106] when applied to regression problems [74], where the regression is using a sum of ridge functions.

However any projective approach is limited in general since for many high-dimensional non-Gaussian distributed clouds of points, most low-dimensional projections are approximately Gaussian: For some given data let $d = d(\nu)$ and $N = N(\nu)$ such that $\lim_{\nu\to\infty} d(\nu) \to \infty$ and $\lim_{\nu\to\infty} N(\nu) \to \infty$. Further suppose that with the parameter $\nu \to \infty$ the fraction of vectors in the data space not approximately orthogonal to each other tends to zero as the distortion of space phenomenon in section 2.1.1 tells us. Then theorem 1.1 in [56] states that the empirical distribution of the projections onto an arbitrary vector ω from the unit ball $\mathcal{B}_d \subset \mathbb{R}^d$ converges weakly in probability to $\mathcal{N}(0, \Sigma)$. This can be illustrated as follows.

As an example consider a set of points in the unit ball $\mathcal{B}_d$ distributed according to the uniform distribution $\mathcal{U}_{[-1,1]}$ and compute the density of that points projected to an arbitrary $\omega \in \mathcal{B}_d$ passing through the origin and parameterized by $\theta = \cos(\measuredangle(x, \omega))$. Then the density along ω is proportional to the volume of an d-sphere of radius $\sin(\measuredangle(x, \omega))$. Hence we conclude for that density $\rho(x) = C(1 - \theta^2)^{\frac{d-1}{2}}$ with a normalizing constant $C = 2^{-\frac{d+1}{2}} d![(0.5(d-1))!]^{-2}$. If we now plot the projected uniform density against an one dimensional Gaussian density with variance $\sigma^2 = (d + 2)^{-1}$ in the case of $d = 20$ we get figure 4.1:

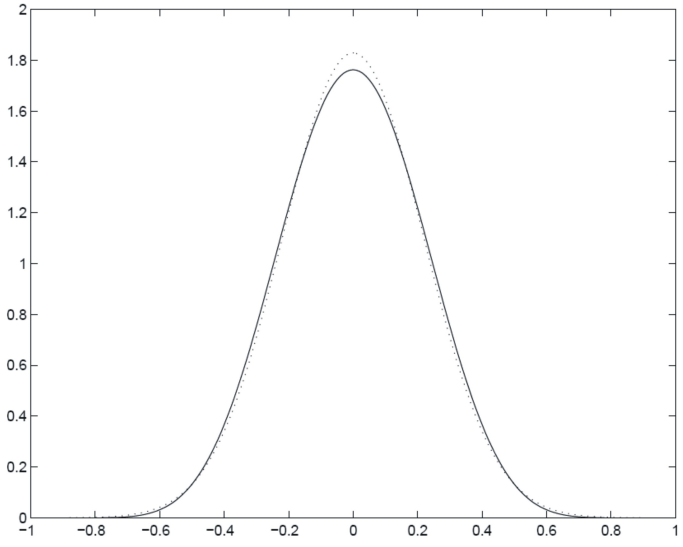


Figure 4.1: Dotted line: Gaussian density with zero mean and variance 1/22. Solid line: the same density projected uniformly from distributed data over the 20-sphere, to an arbitrary selected line passing through the origin.

Thus we cannot hope to uncover every such data structure using one dimensional projections. For projections on higher dimensional subspaces an analogous result is give in [104]. However if the data have long-tailed distributions e.g. for Cauchy distributed data, most

of all projections are not normal [56]. Hence finding projections along which the projected density departs from normality, seems always to be a good idea.

Unsupervised and Structural Adaptive Method: The range of application of a dimension reduction method would become as wide as possible, if we could get rid of all model assumptions and tuning parameter. Consequently we should not assume any apriori knowledge about the density $\rho(x)$ of the original data or about the spatial distribution of the informative data part lying on the manifold in focus or the parameter T, q and Σ in (4.1). Moreover using an iterative method gives the chance to incorporate the result of former estimations from the data set into the current estimation as a "good" initial guess.

Design of the Structural Data Analysis: If we combine all desirable properties of our method we come up to the following iterative and structure adaptive approach to an unsupervised, completely data driven, linear projective method.

i) *Whitening*: The data is re-centered by subtracting the empirical mean and then re-scaled according to $Y_i = \mathrm{diag}(\sigma)X_i$ where $\sigma \stackrel{\mathrm{def}}{=} (\sigma_1, \ldots, \sigma_d)$ are the standard deviations of the components of the random variables X_i.

ii) *Directional sampling*: Let $1 \leq j \leq J$ and $1 \leq l \leq L$. Choose the components of ω_{jl} according to $\mathcal{U}_{[-1,1]}$ from $\mathcal{B}_d$ as directions to project the data and evaluate the projected data using a general contrast function (2.16).

iii) *Estimation*: Use the j^{th} set $\{\omega_{jl}\}_{l=1}^{L}$ in order to estimate an element $\beta_j \in \mathcal{I}$ such that $dist(\widehat{\beta}_j, \mathcal{I}) \leq \epsilon$ where $0 < \epsilon$ should be an uniform bound to the estimation error. Here $\widehat{\beta}_j$ denotes the estimate of β_j.

iv) *Building an ONB*: Reconstruct an orthonormal basis for $\mathcal{I}$ using a numerical method with low total complexity.

v) *Dimension reduction*: Chose m vectors from the reconstructed ONB in order to define the linear mapping from the data space to $\mathcal{I}$.

vi) *Structural adaptation*: Combine the idea of structural adaptation with the projective approach by using some estimators $\widehat{\beta}_j^k$ to get a "better" initial guess for the directional sampling in the $(k+1)^{th}$ computation of the SNGCA-procedure.

Note that this is a two-stage procedure since the estimation of elements from the target space and the reconstruction of a (reduced) basis are separated from each other. Traditionally the detection of the reduced dimensionality $m = rank(\mathcal{I})$ is the most challenging part for unsupervised methods of dimension reduction.

Next we will explain how elements $\beta \in \mathcal{I}$ can be estimated from the data without estimating the parameters μ, Σ and q of ρ in (4.1).

4.2 Estimation of the Elements from the Target Space

The whole approach of SNGCA is essentially based on the following theorem.

Theorem 4. *Let X follow the distribution with the density $\rho(x)$ according to (4.1) and let $I\!E[X] = \mu = 0$. Suppose that $\psi \in \mathcal{C}^1(\mathbb{R}^d, \mathbb{R})$ is a function[1] fulfilling the condition*

$$\gamma(\psi) \stackrel{\mathrm{def}}{=} I\!E[X\psi(X)] = 0, \tag{4.2}$$

[1] We assume here, that $\mathcal{C}^p(\mathbb{R}^n, \mathbb{R}^m)$ is the normed space of functions $f : \mathbb{R}^n \to \mathbb{R}^m$ which are p -times continuously differentiable.

Define

$$\beta(\psi) \stackrel{\text{def}}{=} I\!E\big[\nabla_x \psi(X)\big] = \int \nabla_x \psi(x)\, \rho(x)\, dx, \tag{4.3}$$

where $\nabla_x \psi$ *denotes the gradient of* ψ *in* x*. Then* $\beta(\psi)$ *belongs to* $\mathcal{I}$*. In particular if (4.2) is not fulfilled, then there is a* $\beta \in \mathcal{I}$ *such that*

$$\|\beta(\psi) - \beta\|_2 \leq \epsilon$$

where ϵ *is the uniform error bound:*

$$\epsilon = \Big\| \Sigma^{-1} \int x \psi(x) \rho(x)\, dx \Big\|_2 . \tag{4.4}$$

Hence the distance between $\beta(\psi)$ *and the non-Gaussian subspace* $\mathcal{I}$ *is uniformly bounded as given by (4.4).*

Equivalently one can state the result (4.4) in the form

$$\|(I - \Pi_{\mathcal{I}})\beta(\psi)\|_2 \leq \epsilon$$

where I is the unit operator and $\Pi_{\mathcal{I}}$ is the orthogonal projector on $\mathcal{I}$ in $\mathbb{R}^d$. The proof of this theorem is given in the appendix.

The basic strategy of every approach to SNGCA is the algorithmic realization of (4.3) and (4.2) in Theorem 4, that relies on the vectors $\gamma(\psi)$ and $\beta(\psi)$ which in turn depend on the unknown density ρ. However, both vectors are integrals with respect to ρ. Therefore, they can be easily estimated from the data by using their empirical counterparts:

$$\begin{aligned} \widehat{\gamma}(\psi) &= I\!E_N[X\psi(X)] = N^{-1} \sum_{i=1}^{N} X_i \psi(X_i), \\ \widehat{\beta}(\psi) &= I\!E_N \nabla\psi(X) = N^{-1} \sum_{i=1}^{N} \nabla\psi(X_i). \end{aligned}$$

The important point of the semi-parametric framework from section 2.3.2 at this point is, that the mathematical manner of using theorem 4 within this framework is not unique. In this thesis we will describe some, but not all of them.

Early NonGaussian Component Analysis: In the following let $\eta(\psi) \stackrel{\text{def}}{=} I\!E[\nabla\psi(x)]$ and let $\widehat{\eta}(\psi)$ denote its empirical counterpart. In [212] is was suggested to construct "approximating vectors" $\{\widehat{\beta}_l\}_{l=1}^{L}$ according to

$$\widehat{\beta}_l = \widehat{\eta}_l - \widehat{\Sigma}^{-1}\widehat{\gamma}_l \tag{4.5}$$

with $l = 1, \ldots, L$ and $\widehat{\Sigma}$ as estimator of the data covariance matrix Σ. In the first stage $\phi = h_\omega$ was chosen with $h_\omega \in \mathcal{C}^{1,1}(\mathbb{R}^d \times \mathbb{R}^d, \mathbb{R})$ of the form

$$h_\omega(x) = h(\omega^\top x) e^{-\lambda\|x\|^2/2}. \tag{4.6}$$

where $\omega \in \mathcal{B}^d$ and $\lambda > 0$. Then due to (4.4) the vector $\widehat{\beta}$ is informative if its length is larger in order than the accuracy of approximation ϵ.

By means of the choice of ω the directional sampling was realized. The choice of these directions is crucial to the algorithm because the estimated vectors $\widehat{\beta}$ are searched as "aggregations" of the vectors ω_l. In all approaches a Monte-Carlo sampling from the uniform distribution $\mathcal{U}_{[-1,1]}$ for the coefficients of $\omega_l \in \mathcal{B}_d$ was used. Due to the sparsity of high dimensional data this step becomes more and more computationally expensive with increasing number of dimensions.

In general the function $h \in \mathcal{C}^{1,1}(\mathbb{R}, \mathbb{R})$ should be informative with respect to non-Gaussian components. One may consider different parameter-dependent families of symmetric and non-symmetric functions h. For the numerical simulations later shown in this thesis, we use the families

$$\begin{array}{rcll} h(t) & = & t^3 \exp(-0.5t^2) & \text{(Gauss-Pow3)} \\ h(t) & = & t^4 \exp(-0.5t^2) & \text{(Gauss-Pow4)} \\ h(t) & = & \tanh(t) & \text{(hyperbolic tangent)} \\ h(t) & = & (1+t^2)^{-1} \exp(t) & \text{(asymmetric Gauss).} \end{array}$$

of so called test functions, playing the role of indices ι in section 3.3.2. The multiplier $e^{-\lambda\|x\|^2/2}$ ensures that $h_\omega(x)$ is bounded and integrable with respect to the data density ρ over the whole space $\mathbb{R}^d$. Then at the second stage the projector was estimated by $\widehat{\Pi} = \sum_{j=1}^m v_j v_j^T$, where v_j, $j = 1, ..., m$, are m principal eigenvectors of the matrix $\sum_{l=1}^L \widehat{\beta}_l \widehat{\beta}_l^T$. Note that in this early approach m must be apriori given.

However this implementation of has some serious drawbacks: The way in (4.5) focus on the major non-Gaussian directions and discards the less pronounced directions. Moreover it relies upon the estimation of the covariance matrix Σ of the Gaussian component, which can be hard when d increases. Moreover $\widehat{\Sigma}$ is bad conditioned in high dimensions. Poor estimation of Σ then will result in badly estimated vectors $\widehat{\beta}_l$. This - of course - limits the accuracy of the estimation of the reduced target space $\mathcal{I}$. Furthermore using the eigenvalue decomposition of $\sum_{l=1}^L \widehat{\beta}_l \widehat{\beta}_l^T$ entails that the variance of the estimation $\widehat{\Pi}$ of the projector Π on $\mathcal{I}$ is proportional to L. Consequently only relatively small families $\{h_l\}$ can be used to recover the target subspace, since the choice of h_l becomes dependent of the unknown data properties of the application.

Sparse NonGaussian Component Analysis: In order to circumvent the above limitations of the approach in [212] we propose here a different procedure to obtain estimates $\widehat{\beta}$ of vectors β from the target space. We refer to that method as *convex projection*. The main difference is, that while copying the idea of the directional sampling, the convex projection approach aims to learn $\psi(x)$ as a convex combination of smooth functions h from the data: Let $\{\omega_l\}_{l=1}^L$ be a given set of vectors $\omega_l \in \mathbf{B}_d$ and let h' denote the derivative of h. Then define

$$\psi_{h,c}(x) \stackrel{\text{def}}{=} \sum_{l=1}^L c_l \, h_{\omega_l}(x) \tag{4.7}$$

Then using definition (4.7) this yields

$$\beta(\psi_{h,c}) = \sum_{l=1}^L c_l I\!E\big[\nabla h_{\omega_l}(X)\big] = \sum_{l=1}^L c_l \, \eta_{\omega_l} \,. \tag{4.8}$$

where we used the definition:

$$\eta_{\omega_l} \stackrel{\text{def}}{=} I\!E[\nabla_x h_{\omega_l}(X)] = I\!E\big[\omega_l h'(\omega_l^\top X)e^{-\lambda\|x\|^2/2} - \lambda X h_{\omega_l}(X)\big]$$

Similarly with $\gamma_{\omega_l} \stackrel{\text{def}}{=} I\!E\big[X h_{\omega_l}(X)\big]$ we get

$$\gamma(\psi_{h,c}) = \sum_{l=1}^{L} c_l I\!E\big[X h_{\omega_l}(X)\big] = \sum_{l=1}^{L} c_l\, \gamma_{\omega_l}\,. \tag{4.9}$$

The data counterparts of these expressions playing the central role in the algorithm of SNGCA are given by

$$\widehat{\gamma}_{\omega_l} = I\!E_N\big[X h_{\omega_l}(X)\big] = \frac{1}{N}\sum_{i=1}^{N} X_i h_{\omega_l}(X_i) \tag{4.10}$$

$$\widehat{\eta}_{\omega_l} = I\!E_N\big[\nabla h_{\omega_l}(X)\big] = \omega_l \frac{1}{N}\sum_{i=1}^{N} h'(\omega_l^\top X_i) - \lambda\widehat{\gamma}_l \tag{4.11}$$

$$\begin{aligned}\widehat{\beta}(\psi_{h,c}) &= \sum_{l=1}^{L} c_l\, \widehat{\eta}_{\omega_l} \\ &= \frac{1}{N}\sum_{l=1}^{L} c_l\omega_l \sum_{i=1}^{N} h'(\omega_l^\top X_i) - \lambda\sum_{l=1}^{L} c_l\,\widehat{\gamma}_l. \end{aligned}\tag{4.12}$$

We will also use the abbrevations $\widehat{\gamma}_l$ and $\widehat{\eta}_l$ instead of $\widehat{\gamma}_{\omega_l}$ and $\widehat{\gamma}_{\omega_l}$ respectively. In sum the decisive task of estimating $\beta \in \mathcal{I}$ reduces to that of finding a "good" corresponding coefficient vector $c \in \mathbb{R}^L$ in (4.7).

Aggregation and Concentration: It is well known [80] that the general problem to find an "aggregated" estimate $\widehat{f}(x) \stackrel{\text{def}}{=} \sum_j \widehat{c}_j f_j(x)$ nearly as good as the closest to $f(x)$ convex combination of a set of given $\|\cdot\|_\infty$-bounded Borel functions f_j is associated with a characteristic mean square estimation error in case of $N \to \infty$ i.i.d observations and $f \in L^2_\mu$: Let c^* be the vector of optimal coefficients and $\widehat{c}$ their estimates. Then the difference between the expected distance from f to the result of the aggregation and the distance from f to the "ideal" aggregate is bounded [20] as

$$\begin{aligned} &\mathcal{O}(1)\frac{\sigma L_{\|\cdot\|_2}(f)\sqrt{\ln J}}{\sqrt{N}} \le \\ I\!E\left[\min_{\|c\|_1\le 1}\left\{\int_{\mathbb{R}^d}[f(x)-\sum_{j=1}^{J} c_j f(x)]^2\mathrm{d}x\right\} - \int_{\mathbb{R}^d}[f(x)-\widehat{f}(x)]^2\mathrm{d}x\right] &\le \\ &\mathcal{O}(1)\frac{[L^2_{\|\cdot\|_2}(f) - \sigma L_{\|\cdot\|_2}(f)]\sqrt{\ln J}}{\sqrt{N}} \end{aligned}\tag{4.13}$$

where $L_{\|\cdot\|_2}(f) = \mathcal{O}(1)\sigma$ and σ^2 is the variance of the homogeneous normal noise in the data. The good news about this bound is, that the loss of accuracy caused by the aggregation is nearly independent of J. However from the perspective of complexity, the bad news is that a procedure using an aggregation step may involve all our J functions f_j where J is very large. Hence we are interested in the consequences for the upper bound of the statistical aggregation error that arise if we set K randomly chosen coefficients c_l to zero. It turns out [161], that the new upper bound for (4.13) ends in the following

"$\sqrt{N}$-concentration result":

$$\mathcal{O}(1)\frac{[L^2_{\|\cdot\|_2}(f)-\sigma L_{\|\cdot\|_2}(f)]\sqrt{\ln J}}{\sqrt{N}}+\frac{L_{\|\cdot\|_2^2}(f)}{K}$$

where $K=\lceil\sqrt{d}/\sqrt{\ln J}\rceil$ and $L_{\|\cdot\|_2}(f)$ denotes the Lipschitz constant of f. Naturally we are interested in a similar result for Sparse NonGaussian Component Analysis.

Uniform error bound: A well known result from the empirical process theory [226] claims that $\widehat{\gamma}_\omega$ approximates the unknown vector γ_ω with the accuracy of order $N^{-1/2}$. Moreover, this result can be stated uniformly over all $\omega\in\mathcal{B}_d$. The same holds for the differences $\widehat{\eta}_\omega-\eta_\omega$. Then the use of convex combinations $\psi_{h,c}(x)=\sum_l h(\omega_l^\top x)$ allows to extend this accuracy of approximation ϵ on the difference $\widehat{\beta}(\psi_{h,c})-\beta(\psi_{h,c})$. The next result justifies to construct $\beta(\psi_{h,c})$ in (4.8) together with the constraint $\|c\|_1\le 1$.

Theorem 5. *Suppose that $f(x,\omega)$ is continuously differentiable in w and for some fixed constant f_1^* and any $\omega\in\mathcal{B}_d$, $x\in\mathbb{R}^d$*

$$\begin{aligned}\operatorname{Var}\bigl[X_j\,f(X,\omega)\bigr]\le f_1^*,\quad \operatorname{Cov}\bigl[X_j\,\nabla_\omega f(X,\omega)\bigr]\le f_1^* I,\\ \operatorname{Var}\biggl[\frac{\partial}{\partial x_j}f(X,\omega)\biggr]\le f_1^*,\quad \operatorname{Cov}\biggl[\nabla_\omega\frac{\partial}{\partial x_j}f(X,\omega)\biggr]\le f_1^* I,\end{aligned}$$

Consider the (random) set

$$\mathscr{C}=\bigl\{c\in\mathbb{R}^L:\|c\|_1\le 1,\,\widehat{\gamma}(c)=0\bigr\}. \tag{4.14}$$

Then for any $\epsilon>0$ there is a set $A\subset\Omega$ of probability at least $1-\epsilon$ such that on A for all $c\in\mathscr{C}$,

$$\bigl\|(I-\Pi^*)\widehat{\beta}(c)\bigr\|_2\le\sqrt{d}\,\delta_N\bigl(1+\|\Sigma^{-1}\|_2\bigr),$$

where

$$\delta_N=N^{-1/2}\inf_{\lambda\le\lambda_1^*N^{1/2}}\bigl\{5\mathfrak{n}_0 f_1^*\lambda+2\lambda^{-1}\bigl[\mathfrak{e}_d+\log(2d/\epsilon)\bigr]\bigr\}$$

and $\mathfrak{e}_d=4d\log 2$.

The proof of this theorem is given in the appendix. In other words theorem 4 shows that the convexity condition $\sum_l|c_l|\le 1$ leads to the claims that there is a value $\epsilon=\sqrt{C/N}$ for a fixed positive constant C and a random set A of a dominating probability such that $\|(I-\Pi_{\mathcal{I}})\widehat{\beta}\|_2\le\epsilon$ for all such constructed vectors $\widehat{\beta}$. Consequently the idea of the convex projection approach is to repeat this for different combinations of $\xi,\omega_1,\ldots,\omega_L$ leading to a family of estimated vectors $\widehat{\beta}$. Then the subspace $\mathcal{I}$ can be recovered from the set of $\widehat{\beta}$'s. Due to this result, any vector $c\in\mathscr{C}$ can be used to estimate a vector $\widehat{\beta}(c)$ which is "close" to $\mathcal{I}$. However, vectors constructed in this way are only informative if its length is significant relative to the estimation error in theorem 4.

We will describe in the next section, how to determine the coefficients $\{c_l\}_{l=1}^L$ with $c=\{c_l\}_{l=1}^L$ fulfills $\|c\|_1\le 1$ by means of solving an optimization problem that can be called *convex projection*.

4.3 Reduction of Dimensionality and Structural Adaptation

Convex Projection: Using the definitions from above, consider for a given element $\beta_j \in \mathcal{I}$ to estimate the non-smooth, non-convex optimization problem

$$\widehat{c}_j = \arg\min_{\|c_j\|_1 \leq 1} \Big\| \xi_j - \sum_l c_{lj}\, \widehat{\eta}_l \Big\| \text{ s.t. } \Big\| \sum_l c_{lj}\, \widehat{\gamma}_l \Big\| = 0 \tag{4.15}$$

where $\sum_l$ means $\sum_{l=1}^{L}$ in the sequel. We are interested to bound the distance of $\widehat{\beta}$ from the target space $\mathcal{I}$. This distance is naturally measured by the value $\|(I - \Pi_{\mathcal{I}})\widehat{\beta}\|_2$. Again $\Pi_{\mathcal{I}}$ means the orthogonal projector on $\mathcal{I}$.

Obviously (4.15) is a non-smooth optimization problem. However an equivalent but smooth [27] and convex version of (4.15) is given by

$$\begin{aligned} &\arg\min_{c^-, c^+} \Big\| \xi - \textstyle\sum_{l=1}^{L} c_l^+ \widehat{\eta}_l + \sum_{l=1}^{L} c_l^- \widehat{\eta}_l \Big\|_2^2 \\ \text{such that} \quad &\textstyle\sum_{l=1}^{L} (c_l^+ - c_l^-)\, \widehat{\gamma}_l = 0, \quad \sum_{l=1}^{L} (c_l^+ - c_l^-) \leq 1, \quad 0 \leq c_l^+, c_l^-. \end{aligned} \tag{4.16}$$

The estimation procedure of elements from the target space described here solves only this smooth problem. In order to solve this problem, we refresh the problem as an equivalent linear second order conic problem (SOCP) [145]. In the latter form the original non-smooth and non-convex projection problem can be solved by a fast self-dual interior point method (IPM) [238] to high accuracy. In the MATLAB-toolbox implemented during this thesis, we use a commercial solver[2].

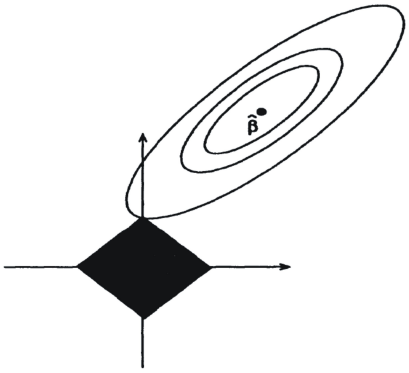


Figure 4.2: Geometry of ℓ_1-constraint: In a regression or optimization problem the use of the 1-norm as a constraint results in vanishing weights due to the geometry of the feasible set of the 1-norm, since the first touch point of square and ellipsoid containing the solution of the quadratic problem (4.16) is the vertex.

Shrinkage: Moreover, it is well known [60; 62; 221; 246] that a ℓ_1-constraint in (4.14) realizes a numerical stable, continuous shrinkage technique and thus leads to a *sparse* solution in only d of L coefficients c_l are different from zero. This holds even in the non-orthogonal design in many cases [86]. Consequently (4.14) suppresses directions ω_l with small weights which are uninformative about $\mathcal{I}$. Hence a welcome side-effect of the ℓ_1-constraint is, that in addition perturbations in the estimation procedure are suppressed. The intuitive idea of the shrinkage in this case is illustrated in figure 4.2.

[2] `http://www.mosek.com`

The next important step of the SNGCA procedure is to recover the subspace $\mathcal{I}$ from the estimated vectors $\widehat{\beta}_j$. At the first glance this problem is a special case of the so called *Reduced Rank Regression* (RRR) problem.

PCA-solution: A standard and popular solution of the RRR problem is given by minimizing the sum of orthogonal complements $\sum_{j=1}^{J} \|(I - \Pi_{\mathcal{I}})\widehat{\beta}_j\|_2^2$ over all projectors $\Pi_{\mathcal{I}}$ of a given rank m, i.e.

$$\widehat{\Pi}_{\mathcal{I}} = \arg\min_{\Pi_{\mathcal{I}}} \sum_{j=1}^{J} \|(I - \Pi_{\mathcal{I}})\widehat{\beta}_j\|_2^2 \quad \text{s.t.} \quad rank(\Pi_{\mathcal{I}}) = m.$$

The solution of this problem is known as PCA solution and it is given by the span $\langle \cdots \rangle$ of the first m eigenvectors of the matrix $\widehat{D} \stackrel{\text{def}}{=} \sum_{j=1}^{J} \widehat{\beta}_j \widehat{\beta}_j^{\top}$:

$$\widehat{\mathcal{I}} = \langle \text{first } m \text{ eigenvectors of } \widehat{D} \rangle.$$

Let β_j be the vectors from $\mathcal{I}$ such that $\|\widehat{\beta}_j - \beta_j\|_2 \leq \epsilon$. The closeness of the subspace $\mathcal{I}$ and its estimate $\widehat{\mathcal{I}}$ can be measured by the error function

$$\mathcal{E}(\widehat{\mathcal{I}}, \mathcal{I}) \stackrel{\text{def}}{=} \|\Pi_{\widehat{\mathcal{I}}} - \Pi_{\mathcal{I}}\|_F^2 \tag{4.17}$$

where $\|\cdot\|_F$ is the Frobenius norm.

However consider the matrix $D = \sum_{j=1}^{J} \beta_j \beta_j^{\top}$. This matrix is of the rank $m(D) \leq m$. Simple algebra yields

$$\|\widehat{D} - D\|_2^2 = \mathrm{Tr}(\widehat{D} - D)^2 \leq J\epsilon^2.$$

Therefore D can be well identified if its first m^{th} eigenvalues fulfill the condition

$$\lambda_m(D) > J\epsilon^2.$$

This condition is verified if some significant fraction of the vectors β_j are significant (informative) in the sense $\|\beta_j\|_2 \geq \kappa$ with some fixed $\kappa > 0$. However, if the most of vectors β_j are non-informative, the PCA solution is very volatile. Moreover the larger is the number of non-informative vectors the worse is the quality of recovering the subspace $\mathcal{I}$. This drawback requires to consider more robust estimates of $\mathcal{I}$.

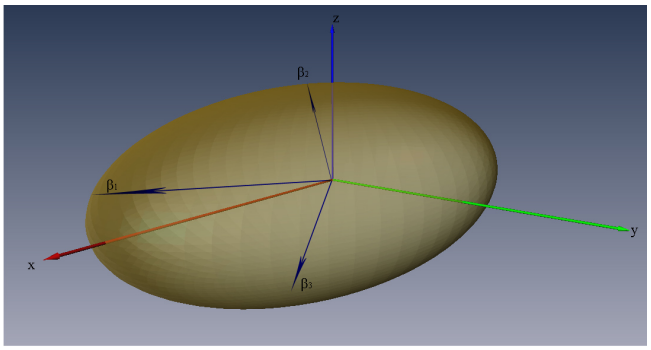


Figure 4.3: Illustration of the MVEE "rounding ellipsoid" of estimated elements "close" to the target space consisting in the (x, y)-plane.

"Rounding ellipsoid" solution: Another way of recovering the subspace $\mathcal{I}$ is given by the "rounding ellipsoid" idea illustrated in figure 4.3. Consider the symmetrized set $\mathcal{S}$ of estimators $\widehat{\beta}_j$ with $j = 1, \dots, J$:

$$\mathcal{S} \stackrel{\text{def}}{=} \{\widehat{\beta}_1, -\widehat{\beta}_1, \widehat{\beta}_2, -\widehat{\beta}_2, \dots\}. \tag{4.18}$$

For any direction orthogonal to the linear subspace $\mathcal{I}$, theorem 4 states that $\mathcal{S}$ expands only with the distance not larger than ϵ while the for the directions within $\mathcal{I}$ we expect at least some informative vectors. This leads to the idea of building an ellipsoid which contains $\mathcal{S}$ and hence its convex hull $conv(\mathcal{S})$ [3]. Then we can take its m largest axes for estimating the subspace $\mathcal{I}$.

The problem of computing a minimum volume enclosing ellipsoid (MVEE) of the symmetrized convex set $conv(\mathcal{S})$ can be considered as the problem of computing the Löwner-John ellipsoid:

Theorem 6. *(Existence and Uniqueness) [114]*
For every convex, bounded, centrally symmetric and non-empty set $\mathcal{C}$ there is a unique ellipsoid $\mathcal{E}$ of minimum volume that covers $\mathcal{C}$ with the center at zero. Moreover, the following Fritz-John-inequality holds:

$$d^{-1/2} MVEE(\mathcal{C}) \subseteq conv(\mathcal{C}) \subseteq MVEE(\mathcal{C}).$$

In the sequel let $\mathcal{E}_{\sqrt{d}}$ denote the $\sqrt{d}$-rounding of the MVEE of $\mathcal{S}$. This ellipsoid is described by a matrix B:

$$\mathcal{E}_{\sqrt{d}}(B) \stackrel{\text{def}}{=} \{x \in \mathbb{R}^d : \ \|B^{-1/2} x\|_2 \le 1\} \tag{4.19}$$

Finding the matrix B is a convex optimization problem. Numerically efficient gradient-type schemes for computing $\sqrt{d}$-rounding ellipsoids are available, see e.g. [169]. We will discuss them in more detail in the next chapter.

We measure the quality of estimation of the subspace $\mathcal{I}$ by the closeness of the estimated projector $\widehat{\Pi}$ to Π^*, where Π^* denotes the "ideal" projector to the target space:

$$\mathcal{E}(\mathcal{I}, \widehat{\mathcal{I}}) = \|\widehat{\Pi} - \Pi^*\|_2^2 = \mathrm{Tr}\big[(\widehat{\Pi} - \Pi^*)^2\big]. \tag{4.20}$$

The property of the spatial information recovery, based on the idea of rounding ellipsoids, is described in the following theorem.

Theorem 7. *(*approximation of rounding ellipsoid*)*
1. Let $\mathscr{S}$ be the convex envelope of the set $\{\pm\widehat{\beta}_j\}$, $j = 1, ..., J$, and let $\mathscr{E}_1(B)$ be an ellipsoid inscribed into $\mathscr{S}$, such that $\mathscr{E}_{\sqrt{d}}(B)$ is a $\sqrt{d}$-rounding ellipsoid for $\mathscr{S}$. Then for any unit vector $v \perp \mathcal{I}$,

$$v^\top B^{-1} v \le \varrho^2.$$

2. If there is $\mu \in \mathbb{R}^J$ with $\mu_j \ge 0$ and $\sum_j \mu_j = 1$ such that

$$\lambda_m\Big(\sum_j \mu_j \beta_j \beta_j^\top\Big) \ge \lambda^* > 2\varrho^2,$$

[3] Recall that the convex hull of S, denoted by $conv(S)$, is the intersection of all convex sets containing S. Alternatively, one can also think of $conv(S)$ as the union of all possible convex combinations of points in S.

where $\lambda_m(A)$ *stands for the* m*-th principal eigenvalue of a symmetric matrix* A*, then*

$$\lambda_m(B^{-1}) \geq \frac{\lambda^* - 2\varrho^2}{2\sqrt{d}}\,. \tag{4.21}$$

3. Moreover, let $\widehat{\Pi} = \widehat{\Gamma}_m \widehat{\Gamma}_m^\top$ *where* Γ_m *is the matrix of* m *principal eigenvectors of* B^{-1}*. Then*

$$\|\widehat{\Pi} - \Pi^*\|_2^2 \leq \frac{4\varrho^2 d\sqrt{d}}{\lambda^* - 2\varrho^2}.$$

The proof of the theorem is presented in the appendix. The results of Theorems 4 and 6 provide a kind of theoretical justification for the algorithms presented in the next section. Indeed, suppose that the test functions $h_1, ..., h_L$ and the vectors $\xi_1, ..., \xi_J$ are chosen in such a way that there are at least m vectors with "significant" projection on $\mathcal{I}$ among $\widehat{\beta}_1, ..., \widehat{\beta}_J$. Then the projector estimate $\widehat{\Pi}$, computed using the ellipsoid $\mathcal{E}(B)$ which is rounding for the set $\{\pm\widehat{\beta}_j\}$, will be with high probability close to Π^*.

However the results about the estimation quality depend critically on the dimension d. Numerical simulations also indicate that with growing dimension, the fraction of non-informative vectors $\widehat{\beta}_j$ increases. Furthermore due to the random choice of the projected directions ξ the length of the informative vectors is no longer correlated with small values of

$$\|(I - \Pi_{\mathcal{I}})\widehat{\beta}(\psi)\|_2.$$

In higher dimensions this leads typically to the situation when some of the longest semi-major axis of $\mathcal{E}_{\sqrt{d}}(B)$ are also non-informative and nearly orthogonal to $\mathcal{I}$. Motivated by this observation we propose to identify the semi-axis of $\mathcal{E}_{\sqrt{d}}(B)$ close to $\mathcal{I}$ using statistical tests on normality. Finally all numerical methods to compute a rounding ellipsoid depend on the computation of the eigenvalue decomposition of the data covariance matrix that typically becomes bad conditioned in high dimensions.

Identifying the non-Gaussian subspace by statistical tests: Currently the estimation procedure of the vectors $\beta(\psi_{h,c})$ itself does not allow the identification of the semi-axis of $\mathcal{E}_{\sqrt{d}}(B)$ within the target space. Hence the basic idea is to apply statistical tests on normality with respect to a significance level α to the original data from $\mathbb{R}^d$ projected on every semi-axis of $\mathcal{E}_{\sqrt{d}}(B)$. In order to avoid misleading results due to large sample sizes N (see e.g. [118]), we chose randomly 1000 points from the projected data for each semi-axis. If the hypothesis of normality is rejected with respect to the projected data, the corresponding semi-axis is used as a basis vector for the reduced target space $\mathcal{I}$. In general this use of statistical tests allows to determine algorithmically the reduced dimension m from the data.

Since statistical tests are specialized to a certain deviations from the normal distribution, are more powerful, we use different tests inside of SNGCA in order to cope with different deviations from normality of the projected data. To be more precise we use the K^2-test according to D'Agostino-Pearson [244] to identify a significant asymmetry in the projected distribution and a EDF-test according to Anderson-Darling [8] with the modification of Stephens [215], which is sensitive to the tails of the projected distribution. In order to confirm these test results from above we use the Shapiro-Wilks test [204] based on a regression strategy in the version given by Royston [189; 190]. Once we have classified the semi-axis of $\mathcal{E}_{\sqrt{d}}(B)$ as being close to the target space we can use the identified subset of axis in the structural adaptation step to be described in the next section.

Structural Adaptation: At the beginning of the algorithm, we have no prior information about $\mathcal{I}$ and therefore we have to sample the directions ξ_j and ω_l randomly from the uniform law. However the SNGCA procedure assumes that the obtained estimated structure $\widehat{\mathcal{I}}$ delivers some information about $\mathcal{I}$ which can be used for improving the sample mechanism and therefore the final quality of estimation. This leads to a *structurally adaptation* in the iterative procedure [102]: the step of estimating the vectors $\{\widehat{\beta}_j\}_{j=1}^J$ and the steps of estimating $\mathcal{I}$ are iterated such that the estimated structural information given by $\widehat{\mathcal{I}}$ can be used to improve the quality of estimating the vectors $\widehat{\beta}_j$ in the next iteration of SNGCA.

Statistically this structural adaptation idea is justified by the following Theorem:

Theorem 8. *Let $\mathscr{A}$ be a random set on which*

$$\max_l \|\gamma_l - \widehat{\gamma}_l\|_2 \leq \epsilon,$$
$$\max_l \|\eta_l - \widehat{\eta}_l\|_2 \leq \epsilon.$$

and let β^* denote the "ideal aggregation" $\beta^* = \sum_l c_l^* \eta_l$. Then it holds:

$$\begin{aligned} \|\xi - \widehat{\beta}\|_2 &\leq \|\xi - \beta^*\|_2 + \epsilon, \\ \|\Pi_{\mathcal{I}}(\xi - \widehat{\beta})\|_2 &\leq \|\Pi_{\mathcal{I}}(\xi - \beta^*)\|_2 + (1 + C_1)\epsilon. \end{aligned}$$

The proof of this theorem can be found in the appendix.

In other words if the sampling directions $\{\xi_j\}_{j=1}^J$ and $\{\omega_l\}_{l=1}^L$ are informative then the corresponding vectors $\eta_l = I\!E \nabla h_{\omega_l}(X)$ are expected to be informative as well. This ensures that the vector $\beta^* = \sum_l c_l^* \eta_l$ coming out of the "ideal" optimization problem:

$$\{c_l^*\} = \arg\min_{\|c\|_1 \leq 1} \Big\|\xi - \sum_{l=1}^L c_l \eta_l\Big\|_2 \quad \text{s.t.} \quad \sum_{l=1}^L c_l \gamma_l = 0$$

is also informative. The message of Theorem 7 is that in this situation the estimated vector $\widehat{\beta}$ delivers as much information as β^* up to a small error of estimation. Therefore we sample a fraction of directions $\{\xi_j\}_{j=1}^J$ and $\{\omega_l\}_{l=1}^L$ due to the previously estimated ellipsoid $\widehat{B}$ and the other part randomly. The fraction of the randomly selected directions decreases during iteration.

One Step Improvement: We will now illustrate the iterative gain of information about the target space. To this end we use the projection of $\widehat{\beta}_j$ to the target space in order to demonstrate, how the algorithm works. The next figure 4.4 shows that $dist(\widehat{\beta}, \widehat{\mathcal{I}})$ with

$$\sin(\measuredangle(\widehat{\beta}, \widehat{\mathcal{I}})) = \frac{dist(\widehat{\beta}, \widehat{\mathcal{I}})}{\|\widehat{\beta}\|} = \frac{\epsilon}{\|\widehat{\beta}\|}$$

decreases with increasing number of iterations. As expected we observe, that estimators $\widehat{\beta}$ with higher norm tend to be close to $\mathcal{I}$. Nevertheless this can not be assured for much higher dimensions. Moreover improvement in each iteration heavily depends on the size of the MC-sampling of the measurement directions.

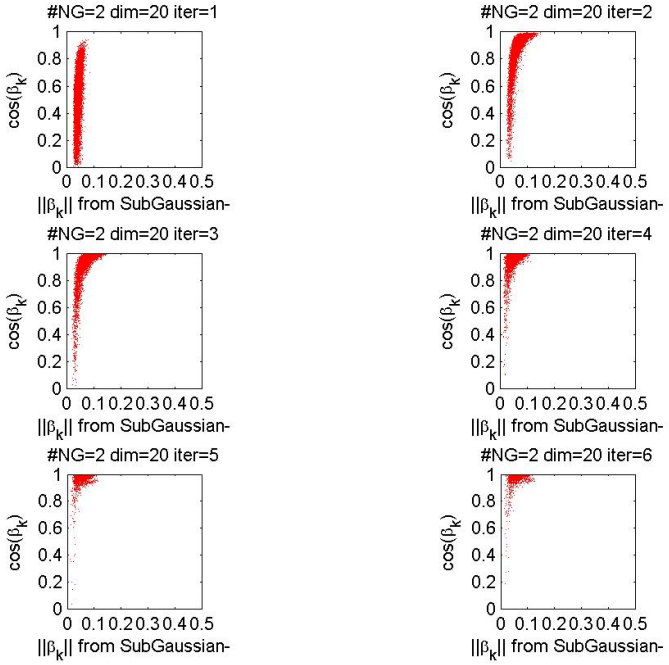


Figure 4.4: Illustrative plots of SNGCA applied to toy 20 dimensional data of type **(C)** (see section 4.5): We show $\|\widehat{\beta}\|$ vs. $\cos(\measuredangle(\widehat{\beta}, \mathcal{I}))$ for different iterations of the algorithm where $\mathcal{I}$ is the apriori known target space.

4.4 Algorithms

Normalization: As a preprocessing step the SNGCA procedure uses a componentwise normalization of the data. To this end let $\sigma = (\sigma_1, \dots \sigma_d)$ be the standard deviations of the data components of $x_1, \dots, x_d$. For $i = 1, \dots, N$ the componentwise normalization of the data is done by $Y_i = \mathrm{diag}(\sigma^{-1})X_i$.

Estimation of the vectors from non-Gaussian subspace: Here we repeat in more detail the estimation procedure already presented in section (4.2). In the sequel we will call the directions ω_l and ξ_j the measurement directions.

Algorithm 2: linear estimation of $\beta(\psi_{h,c})$

Data: Y,L,J
Result: $\{\widehat{\beta}_j\}_{j=1}^J$
Sampling: choice of measurement directions
for *j=1* **to** J **do**
  for *l=1* **to** L **do**
    Compute:
    $\widehat{\eta}_{jl} = \frac{1}{N}\sum_{i=1}^N \nabla h_{\omega_{jl}}(Y_i)$
    $\widehat{\gamma}_{jl} = \frac{1}{N}\sum_{i=1}^N Y_i h_{\omega_{jl}}(Y_i)$
  end
  Compute (4.15) and than $\widehat{\beta}_j = \sum_{l=1}^L \widehat{c}_j \widehat{\eta}_{jl}$.
end

Reduction of dimensionality: Let us consider again the central symmetrized set $\mathcal{S} \stackrel{\text{def}}{=} \{\widehat{\beta}_1, -\widehat{\beta}_1, \widehat{\beta}_2, -\widehat{\beta}_2, \ldots\}$ already defined in (4.18). From theorem 6 we know that there is a minimum volume ellipsoid $\mathcal{E}$, that covers $conv(\mathcal{S})$. For a polytope $conv(x_1, x_2, \ldots)$ of given points $x_1, x_2, \ldots$ the MVEE and the maximum volume inscribed ellipsoid (MVIE) are affine invariant. In this case the computation of the MVEE can be reduced to the computation of the MVIE [122; 123]. Even though the latter problem can be solved using interior-point-methods in $\mathcal{O}(d^3 \log N)$ iterations, they are computationally expensive and restricted to the case of a full-dimensional ellipsoid [224].

Recently a gradient-type method for computing the ellipsoidal rounding for some polytopes was proposed in [169]. The heart of this algorithm is the alternate computation of the MVIE and MVEE taking only $\mathcal{O}(d^2 J \log(J))$ operations. Moreover we expect the computation of an high dimensional MVEE to be numerically bad conditioned. Consequently while approximating the MVEE by a numerical procedure we try to avoid computations of the inverse of $\sum_l \widehat{\beta}_l \widehat{\beta}_l^T$. However this method needs just a single computation of the inverse of $\sum_l \widehat{\beta}_l \widehat{\beta}_l^T$ at the beginning of the procedure. Then due to (4.4), the Fritz-John theorem and the ℓ_1-constraint in (4.16) we know that there are at least d estimated points $\widehat{\beta}$ lying on an ellipsoid bowl that exists in at least m dimensions. Hence we can randomly choose additional points from $\mathbb{R}^d$ that are less informative about $\mathcal{I}$ than every estimated point $\widehat{\beta}_l$. Hence these additional points will not change the shape of the MVEE, but easily lead to a regularized version of the original algorithm to compute an approximation of the MVEE. For convenience we repeat that algorithm here:

Algorithm 3: Compute of the $\sqrt{d}$-rounding of the MVEE

Data: $\{\widehat{\beta}_j\}_{j=1}^J$
Result: $\widehat{B}$,
Let $\delta_i^{k^*} = \max_{1 \le j \le J} \langle \widehat{\beta}_j, \widehat{B}_i \widehat{\beta}_j \rangle$ and set $\nu_i = \delta_i^{k^*} d^{-1}$.
Let $\widehat{B}_0$ be the inverse empirical covariance matrix of the $\widehat{\beta}_j$ and set $t_i = \frac{\nu_i}{(\delta_i^{k^*} d^{-1} - 1)}$. Let i be the loop index.
repeat
  $x_i = \widehat{B}_i \widehat{\beta}_{k^*}$
  $\widehat{B}_{i+1} = \frac{1}{1-t_i}\left(\widehat{B}_i - \frac{t_i}{1+\nu_i} x_i x_i^\top\right)$
  $\delta_{i+1}^{k^*} = \frac{1}{1-t_i}\left(\delta_i^{k^*} - \frac{t_i}{1+\nu_i} \langle \widehat{\beta}_{k^*}, x_i \rangle^2\right)$
until $\delta_i^{k^*} \le C \cdot d$ *where* C *is a tuning parameter.*

The next algorithm (4) reports the pseudocode for constructing a basis of the target space from the estimated elements:

Algorithm 4: Dimension Reduction

Data: $\widehat{B}$
Result: $\langle$first m eigenvectors of $\widehat{B}\rangle$
Let $\widehat{V}$ be the matrix of eigenvectors $\widehat{v}_i$ from $\widehat{B}$ according to algorithm 3.
for *i=1* **to** d **do**
  Project the data orthogonal on $\widehat{v}_i$.
  Compute tests on normality of the projected data.
end
Discard every eigenvector with associated normal distributed projected data.

Structural Adaptation: In algorithm 2 we start with a random initialization of the non-parametric estimator (4.12) by means of a Monte-Carlo sampling of the directions ω_{jl} and ξ_j. However we can use the result of the first iteration $j = 1$ of SNGCA in order to accumulate information about $\mathcal{I}$ in a sequence $\widehat{\mathcal{I}}_1, \widehat{\mathcal{I}}_2, \ldots$ of estimators of the target space. The procedure is described in detail in algorithm 5.

Algorithm 5: structural adaptation of the linear estimation

Data: $\langle$first m eigenvectors of $\widehat{B}\rangle$
Let $\{\widehat{v}_i\}_{i=1}^m$ denote the reduced set of eigenvectors from $\widehat{B}$ and let k iterations be completed. To initialize iteration $k+1$ choose random numbers $z_{j,1}, \ldots, z_{j,m}$ and $u_{l,1}, \ldots, u_{l,m}$ from $\mathcal{U}_{[-1,1]}$ and set

$$\xi_j \stackrel{\text{def}}{=} \sum_{s=1}^m z_{j,s}\widehat{v}_{i_s} \text{ for } 1 \le j \le n_1 < J$$
$$\omega_l \stackrel{\text{def}}{=} \sum_{s=1}^m u_{l,s}\widehat{v}_{i_s} \text{ for } 1 \le l \le n_2 < L$$

Then define $\omega_{L-n_2}, \ldots, \omega_L$ and $\xi_{J-n_1}, \ldots, \xi_J$ analogous to the case $k = 1$. Now compose the sets

$$\{\xi_1^{(k)}, \ldots, \xi_{n_1}^{(k)}, \xi_{n_1+1}^{(k)}, \ldots, \xi_J^{(k)}\}$$
$$\{\omega_1^{(k)}, \ldots, \omega_{n_2}^{(k)}, \omega_{n_2+1}^{(k)}, \ldots, \omega_L^{(k)}\}$$

For the initialization in the case $k = k + 1$. Moreover we choose $n_1 = kd$ and $n_2 = kd$ until $n_1 > J - d$ or $n_2 > L - d$. Otherwise set $n_1 = J - d$ or $n_2 = L - d$.

In the sequel we call that part of measurement directions which are chosen by a Monte-Carlo method the *Monte-Carlo*-part.

Stopping criterion: Suppose that $\mathcal{I}$ is apriori given. Then the convergence of SNGCA can be measured according to the criterion (5.49). More precisely we assume convergence if the improvement of the error measured by (5.49) from one iteration to the next one is less than δ percent of the estimation error in the former iteration.

Suppose now that $\mathcal{I}$ is unknown. Then compute the maximum angle θ between the subspaces specified by the matrix of eigenvectors $V^{(k)} = \left[\widehat{v}_1^{(k)}, \widehat{v}_2^{(k)}, \ldots\right]$ and $V^{(k+1)} = \left[\widehat{v}_1^{(k+1)}, \widehat{v}_2^{(k+1)}, \ldots\right]$ given by

$$\cos(\theta) = \max_{x,y} \frac{|x^\top V^{(k)\top} V^{(k+1)} y|}{\|V^{(k)}x\|_2 \, \|V^{(k+1)}y\|_2}$$

The algorithm stops if the change of the subspace angle is less than κ percent.

We we will now describe, how SNGCA makes use of the algorithms 2, 4 and 5 in order to realize an iterative estimation procedure of $\mathcal{I}$.

Full Description of the Procedure: For convenience we will now give a detailed description of the complete SNGCA algorithm. The choice of the parameters will be explained in the sequel.

Algorithm 6: full procedure of SNGCA

Data: $\{X_i\}_{i=1}^N$,L,J,α
Result: $\widehat{\mathcal{I}}$

Normalization: The data $(X_i)_{i=1}^N$ are recentered. Let $\sigma = (\sigma_1, \dots \sigma_d)$ be the standard deviations of the components of X_i. Then $Y_i = \operatorname{diag}(\sigma^{-1})X_i$ denotes the componentwise empirically normalized data.

Main Procedure:; `// loop on` k

while $\sim$ *StoppingCriterion*$(\mathcal{I}, \widehat{\mathcal{I}})$ **do**

Sampling: The components of the *Monte-Carlo*-parts of $\xi_j^{(k)}$ and $\omega_{jl}^{(k)}$ are randomly chosen from $\mathcal{U}_{[-1,1]}$. The other part of the measurement directions are initialized according to the structural adaptation approach described in algorithm 5. Then $\xi_j^{(k)}$ and $\omega_{jl}^{(k)}$ are normalized to unit length.

Linear Estimation Procedure:
for *j=1* **to** J **do**

for *l=1* **to** L **do**

$$\widehat{\eta}_{jl}^{(k)} = \tfrac{1}{N}\sum_{i=1}^N \nabla h_{\omega_{jl}^{(k)}}(Y_i)$$
$$\widehat{\gamma}_{jl}^{(k)} = \tfrac{1}{N}\sum_{i=1}^N Y_i h_{\omega_{jl}^{(k)}}(Y_i)$$

end

Compute the coefficients $\{c_l\}_{l=1}^L$ by solving the second-order conic optimization problem (4.15):

$$\begin{gathered} \min\ q \qquad \text{s.t.} \\ \tfrac{1}{2}\|z\|_2 \le q \\ \textstyle\sum_{l=1}^L (c_l^+ - c_l^-)\widehat{\eta}_{jl}^{(k)} - z = \xi_j^{(k)} \\ \textstyle\sum_{l=1}^L (c_l^+ - c_l^-)\widehat{\gamma}_{jl}^{(k)} = 0 \\ \textstyle\sum_{l=1}^L (c_l^+ - c_l^-) \le 1, \quad 0 \le c_l^+, c_l^- \quad \forall l \end{gathered}$$

Compute $\widehat{\beta}_j^{(k)} = \sum_{l=1}^L (\widehat{c}_l^+ - \widehat{c}_l^-)\widehat{\eta}_{jl}^{(k)}$

end

Dimension Reduction:
Compute the symmetric matrix $\widehat{B}^{(k)}$ defining the approximation of the Löwner-John ellipsoid $\mathcal{E}$ in (4.19) according to algorithm 3. Reduce the basis of $\mathcal{X}$ according to algorithm 4.

end

Choice of parameters: One of the advantages of the algorithm proposed above is the fact that there are only a few tuning parameters.

i) Suppose now that ω_i is an absolute continuous random variable with $\omega_i \sim \mathcal{U}_{[-1,1]}$. Without loss of generality we set $e = (1, 0, \dots, 0)$. Due to the normalization of $(\omega_1, \dots, \omega_d)$, it holds:

$$P\big(|(\omega_1, \dots, \omega_d)^\top e| \ge 0.5\big) = \big(\sqrt{d}\big)^{-1}$$

However the choice of J and L heavily depends on the non-gaussian components. In the experiments we use $7d \leq J \leq 18d$ and $6d \leq L \leq 16d$.

ii) Set the parameter of the stopping rule to $\delta = 0.05$.

iii) Set the constant in the stopping rule for the computation of the MVEE to $C = 2$.

iv) Set the significance level of the statistical tests to $\alpha = 0.05$.

v) The tuning parameter χ in the dimension reduction step is set to $\chi = 3$.

In the next section we compare the "convex-projection"-approach to SNGCA to the early methods of NonGaussian Component Analysis and to Independent Component Analysis with respect to their statistical sensitivity and stability using some artificial toy examples. Again we use the estimation error defined in (5.49).

Complexity: Let us now estimate the arithmetical complexity of SNGCA. We restrict ourselves to the leading polynomial terms of the complexity of corresponding computations counting only the multiplications.

1. The numerical effort to compute η_{jl} and γ_{jl} in algorithm 2 heavily depends on the choice of $h(\omega^\top x)$. Let $h(\omega^\top x) = \tanh(\omega^\top x)$. Then this step takes $\mathcal{O}(J(\log N)^2 N^2)$ operations.

2. Algorithm 3 takes $\mathcal{O}(d^2 J \log(J))$ operations [169].

3. For the optimization step in 2 we use a commercial solver [4] based on an interior point method. The constrained convex projection solved as an SOCP takes $\mathcal{O}(d^2 n^3)$ operations there n is the number of constraints.

4. The computation of the statistical tests in one dimension: Let N denote the number of samples. D'Agostino-Pearson-test needs $\mathcal{O}(N^3 \log N)$ and the Anderson-Darling-test $\mathcal{O}((\log N)^2 N^2)$ operations. The test of Shapiro-Wilks takes $\mathcal{O}(N^2)$. In order to avoid robustness problems [118] in SNGCA the number of samples is limited to $N \leq 1000$. For larger data sets, $N = 1000$ points are randomly chosen.

5. The computation of the entropy estimator takes only $\mathcal{O}(N \log N)$ operations [135].

Hence the SNGCA procedure computes an estimate $\widehat{\mathcal{I}}$ of $\mathcal{I}$ in $\mathcal{O}(J(\log N)^2 N^2 + d^2 J \log(2J))$ arithmetical operations in each iteration.

4.5 Statistical and Numerical Performance

The numerical comparison of different unsupervised feature extraction methods is based on non-Gaussian densities used as informative component to ρ.

(A) **Gaussian mixture**: 2-dimensional independent Gaussian mixtures with density of each component given by $0.5\, \phi_{-3,1}(x) + 0.5\, \phi_{3,1}(x)$.

(B) **Dependent super-Gaussian**: 2-dimensional isotropic distribution with density proportional to $\exp(-\|x\|)$.

(C) **Dependent sub-Gaussian**: 2-dimensional isotropic uniform with constant positive density for $\|x\|_2 \leq 1$ and 0 otherwise.

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

(D) **Dependent super- and sub-Gaussian**: 1-dimensional Laplacian with density proportional to $\exp(-|x_{Lap}|)$ and 1-dimensional dependent uniform $\mathcal{U}(c, c+1)$, where $c = 0$ for $|x_{Lap}| \leq \log(2)$ and $c = -1$ otherwise.

(E) **Dependent sub-Gaussian**: 2-dimensional isotropic Cauchy distribution with density proportional to $\lambda(\lambda^2 - x^2)^{-1}$ where $\lambda = 1$.

Each of the following test data sets includes 1000 samples in 10 dimensions and each sample consists partly of 8-dimensional independent, standard and homogeneous Gaussian distributions. The other 2 components of each sample are non-Gaussian with variance unity. That means, that the non-normal distributed data are located in a linear subspace. In all simulations the number of non-Gaussian dimensions is apriori given to each algorithm.

Figure 4.5 illustrates the densities of the non-Gaussian components of the test data.

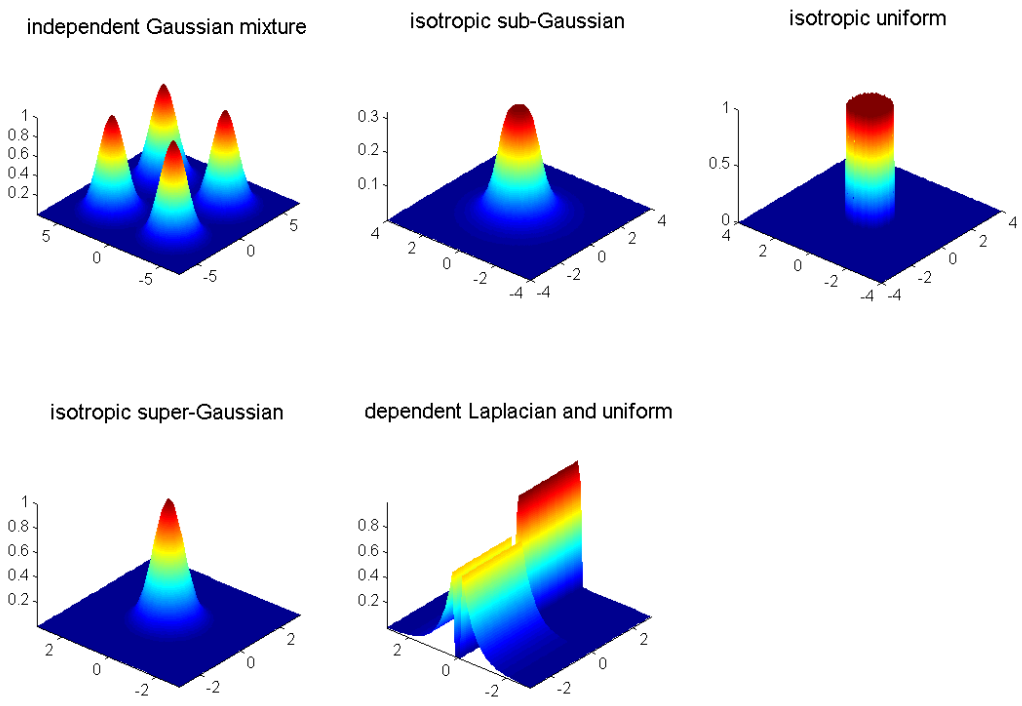


Figure 4.5: Densities of the non-Gaussian components. From upper left to lower right: $2d$ independent Gaussian mixtures, $2d$ isotropic super-Gaussian, $2d$ isotropic uniform, $2d$ isotropic sub-Gaussian and $2d$ isotropic uniform and dependent $1d$ Laplacian with additive $1d$ uniform.

Each of the following simulations is repeated 100 times. All simulations are done with the index 'tanh'. Since the speed of convergence varies with the type of non-Gaussian components we use the maximum number $maxIter = 3\log(d)$ of allowed iterations to stop SNGCA. In the experiments the error measure $\mathcal{E}(\widehat{\mathcal{I}}, \mathcal{I})$ is used only to determine the final estimation error. All simulations other than whose with respect to model (C) are computed with a componentwise pre-normalization.

In the figure 5.4 we present boxplots of the error (5.49) of the methods PP, NGCA and SNGCA. Since the optimizer used in PP tends to trap in local a minimum in each of the 100 simulations, PP is restarted 10 times with random starting points. The best result with respect to (5.49) is reported as the result of each PP-simulation.

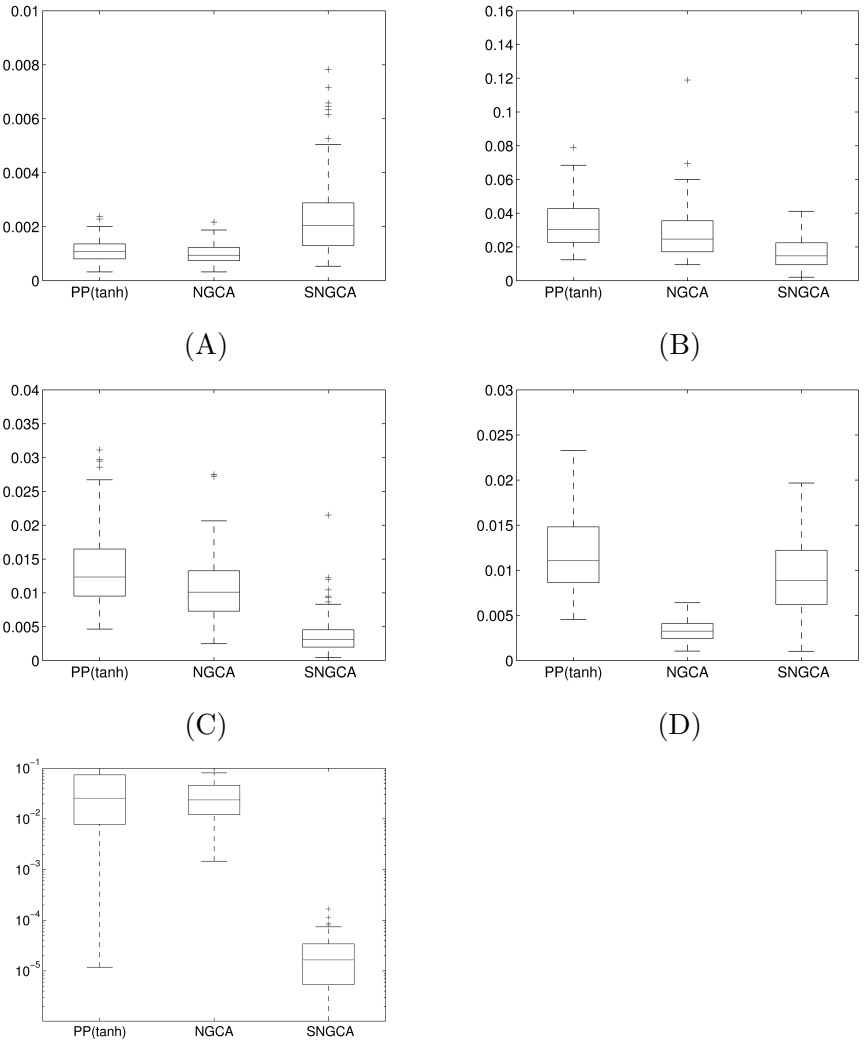


Figure 4.6: Performance comparison using toy examples in 10 dimensions of PP and NGCA versus SNGCA (with respect to $\mathcal{E}(\widehat{\mathcal{I}}, \mathcal{I})$) using the index '$\tanh(x)$'. The doted line denotes the mean, the solid lines the variance of (5.49).

Concerning the results of SNGCA on the data sets (A) and (D) we observe a slightly inferior performance compared to NGCA. In case of model (A) this is due to the fact that most of the data projections have almost a Gaussian density. Consequently the decrease of the estimation error is slow with increasing number of iterations. In case of the model (D) the higher variance of the results indicate that the initial MC-sampling of the data sets gives a poor result. Consequently more iterations are needed to get an estimation error with is comparable to the result of NGCA.

In order to illustrate this interpretation we report in table (4.1) the progress of SNGCA with respect to the estimation error $\mathcal{E}(\mathcal{I}, \widehat{\mathcal{I}})$ in each iteration for every test model. The next table 4.1 reports these results in more detail.

j	μ_ϵ	σ_ϵ^2
1	0.232504	0.045787
2	0.163022	0.072263
3	0.066537	0.032436
4	0.009380	0.021975
5	0.002359	0.000853

(A)

j	μ_ϵ	σ_ϵ^2
1	0.30350	0.175313
2	0.144430	0.057856
3	0.088142	0.015168
4	0.041420	0.008197
5	0.026436	0.000917

(B)

j	μ_ϵ	σ_ϵ^2
1	0.040556	0.004215
2	0.016012	0.002441
3	0.012427	0.001105
4	0.008874	0.000169
5	0.003770	0.000125

(C)

j	μ_ϵ	σ_ϵ^2
1	0.203419	0.044672
2	0.023023	0.000314
3	0.019960	0.000211
4	0.012709	0.000197
5	0.009343	0.000127

(D)

j	μ_ϵ	σ_ϵ^2
1	0.2762e-3	0.1371e-6
2	0.0450e-3	0.0031e-6
3	0.0416e-3	0.0033e-6
4	0.0360e-3	0.0014e-6
5	0.0287e-3	0.0024e-6

(E)

Table 4.1: Progress of SNGCA for the test models from above in 10 dimensions with increasing number j of iterations. The empirical mean of $\mathcal{E}(\widehat{\mathcal{I}}, \mathcal{I})$ is denoted by μ_ϵ and σ_ϵ^2 is its empirical variance.

Now let us switch to the question of robustness of the estimation procedure with respect to a bad conditioning of the covariance matrix Σ of the data. In figure 5.5 we consider the same test data sets as above. The non-Gaussian coordinates always have variance unity, but the standard deviation of the 8 Gaussian dimensions now follow the geometrical progression $10^{-r}, 10^{-r+2r/7}, \ldots, 10^{r}$ where $r = 1, \ldots, 8$. Again we apply a componentwise normalization procedure to the data from the models (A), (B), (D), (E).

In figure (5.5) we observe that the condition of the covariance matrix heavily influences the estimation error for the methods NGCA and PP(tanh). In comparison SNGCA is independent of differences in the noise variance along different directions in most cases. Only the detection of the uniform distribution by SNGCA is influenced by the condition of the data variances in Σ.

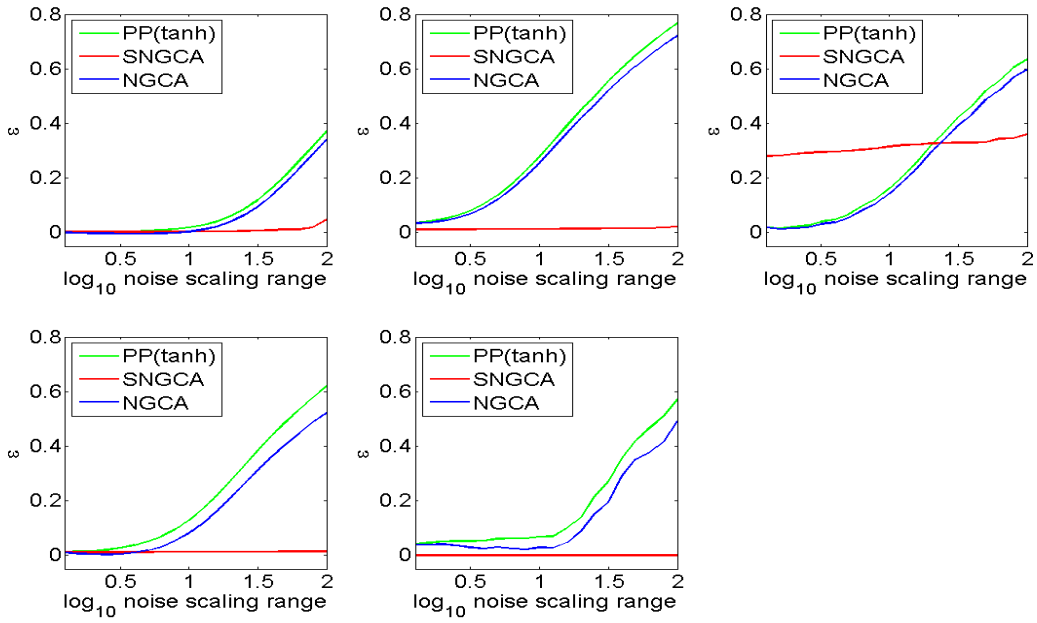


Figure 4.7: From upper left to lower right: $2d$ independent Gaussian mixtures, $2d$ isotropic super-Gaussian, $2d$ isotropic uniform, dependent $1d$ Laplacian with additive $1d$ uniform and $2d$ isotropic sub-Gaussian: Results obtained from the toy densities with respect to $\mathcal{E}(\widehat{\mathcal{I}}, \mathcal{I})$ with deviations of Gaussian components with respect to a geometrical progression on $[10^{-r}, 10^{r}]$ where r is written on the abscissa) .

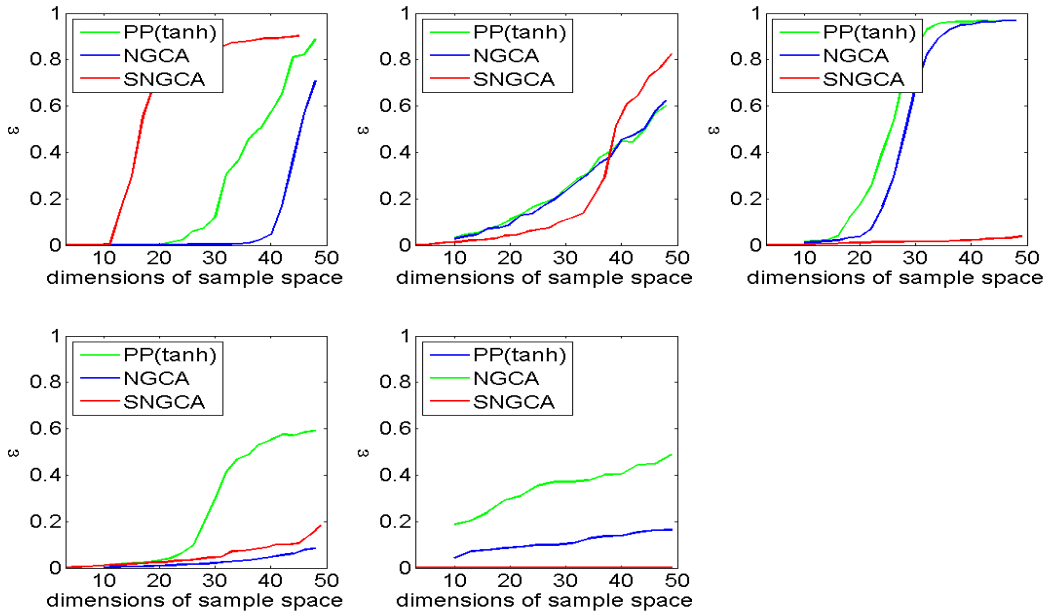


Figure 4.8: From upper left to lower right: $2d$ independent Gaussian mixtures, $2d$ isotropic super-Gaussian, $2d$ isotropic uniform, dependent $1d$ Laplacian with additive $1d$ uniform and $2d$ isotropic sub-Gaussian: Results obtained from the toy densities with respect to $\mathcal{E}(\widehat{\mathcal{I}}, \mathcal{I})$ with increasing dimension of embedding Gaussian component.

Finally we study the ability of the different algorithms to detect the non-Gaussian components in embedding high dimensional noise called statistical sensitivity. Figure 5.6 compares the behavior of SNGCA with PP and NGCA as the number of standard and homogeneous Gaussian dimensions increases. As described above we use the test models with 2-dimensional non-Gaussian components with variance unity. We plot the mean of errors $\mathcal{E}(\widehat{\mathcal{I}}, \mathcal{I})$ over 100 simulations with respect to the test models (A) to (E).

Again concerning the mean of errors $\mathcal{E}(\widehat{\mathcal{I}}, \mathcal{I})$ over 100 simulations of PP and NGCA we find a transition in the error criterion to a failure mode for the test models (A), (C) between $d = 30$ and $d = 40$ and between $d = 20$ and $d = 30$ respectively. For the test models (B),(D) and (E) we found a relative continuous increase in $\mathcal{E}(\widehat{\mathcal{I}}, \mathcal{I})$ for the methods PP and NGCA. In comparison SNGCA fails to analyze test model (A) independently from the size of the MC-sampling, if the dimension increase $d = 12$. Concerning test model (B) there is a sharp transition in the simulation result between $d = 35$ and $d = 40$. Moreover some deviations from normality are much harder to detect as others: For example we expect most of the projected distributions of the Laplacian density to be normal, since it has almost a Gaussian shape. However since the projections of a Cauchy distribution are again non-normal (see section 4.1, the difference in the statistical sensitivity with respect to to the Cauchy distributed data density components and the other non-Gaussians is expected.

Failure modes: In order to provide a better insight into the details of the failure modes we present box plots of the error criterion $\mathcal{E}(\widehat{\mathcal{I}}, \mathcal{I})$ in the transitions phase with respect to the models (A) and (B).

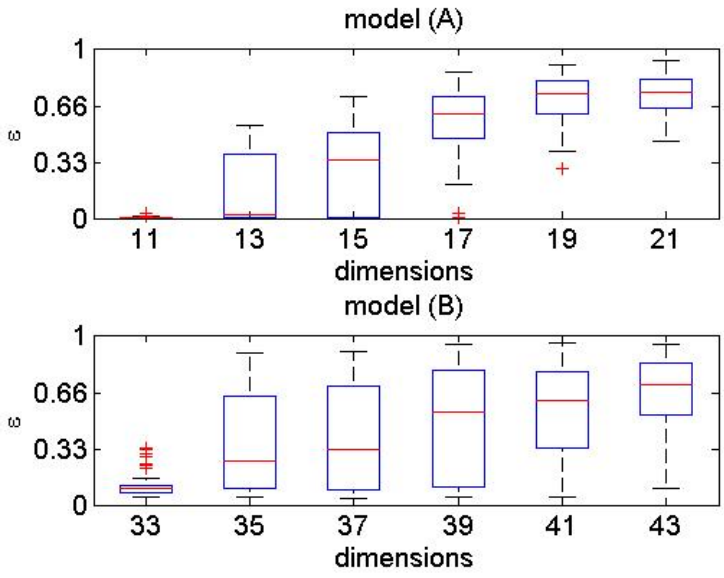


Figure 4.9: Failure modes of SNGCA obtained from the toy densities - upper figure: model (A) - lower figure: model(B). We show boxplots of the aperture of dimensions where the failures occur.

Figure 4.9 demonstrates the differences in the transition phases of model (A) and (B) respectively. The transition phase of SNGCA is characterized by high variance of the estimation error. For model (A) the increase of the variance $\sigma^2_{\mathcal{E}}$ of the error $\mathcal{E}(\widehat{\mathcal{I}}, \mathcal{I})$ beginning at dimensions 13 and its decrease beginning at dimension 15 indicates that

a sharp transition phase is happens in the interval $[13, 15]$. For higher dimensions iterations of SNGCA have a decreasing effect on the estimation result. This indicates that by the MC-sampling of the measurement directions, we can not detect the non-Gaussian components of the data density.

For model (B) the transition phase starts at dimension 35 and ends at dimension 43. Moreover the decrease of $\sigma^2_{\mathcal{E}}$ towards higher dimensions and the increase of the mean of $\mathcal{E}(\widehat{\mathcal{I}}, \mathcal{I})$ is much slower. This indicates that the non-Gaussian components of the data density might be detectable if we would allow much more iterations of SNGCA and an enlarged size of the set of measurement directions. This observation motivates the interpretation that the Monte-Carlo sampling is a very poor strategy which fails to provide sufficient information about the Laplace distribution in high dimensions. Since we can not increase the size of the directional sampling at exponential rate the performance of SNGCA currently is limited.

Chapter 5

Structural Analysis by Semidefinite Programming

In this chapter we will benefit from our numerical simulations from section 4.5 in the sense that they suggest to improve our current strategy of semi-parametric dimension reduction due to the reasons of computational complexity and statistical efficiency.

Complexity: Suppose a set $\{\widehat{\beta}_j\}_{j=1}^J$ of elements from $\widehat{\mathcal{I}}$ according to (4.12) has to be estimated. Then for every $\widehat{\beta}_j$ "good" coefficients $\{c_l\}_{l=1}^L$ can be found by computing the convex projection

$$\min_{\|c\|_1 \leq 1} \left\| \xi_j - \sum_{l=1}^{L} c_{lj}\, \widehat{\eta}_{lj} \right\|_2^2 \qquad \text{subject to} \qquad \sum_{l=1}^{L} c_{lj}\, \widehat{\gamma}_{lj} = 0. \tag{5.1}$$

of an arbitrary vector $\xi \in \mathcal{B}_d$ on the convex hull of $\{\widehat{\eta}_{\omega_l}\}_{l=1}^L$ as sparse [60; 62] solution to a linear constrained, quadratic but non-smooth optimization problem (QCP). Obviously (5.1) suppresses directions ω_l which are less informative about $\mathcal{I}$. Recall that in order to compute the set $\{\widehat{\beta}_j\}_{j=1}^J$ we have to solve J problems of the form (5.1) based on J directional samplings each taking $\mathcal{O}(LN^2)$ operations. Moreover the "convex projection"-approach to SNGCA uses a smooth and convex reformulation of (5.1) as an SOCP that is a linear problem with quadratic constraints.

Hitherto it is widely accepted that the best tool for solving large scale convex optimization problems are Polynomial Time Interior Point methods (IPM) since they enjoy at most superlinear convergence and their computational effort to find an approximate solution is proportional to the number of accuracy digits where the proportionality coefficient growing polynomially with the dimension of the problem [184]. This property means rapid convergence in terms of the number of calculations and provides high-accuracy solutions. However in order to solve a SOCP with L variables an Newton-type IPM iteration requires assembling and solving a $L \times L$ Newton system of linear equations [238] that takes $\mathcal{O}(L^3)$ operations unless the equation system is sparse with favorable patterns. In the context of the "convex projection"-approach $\mathcal{O}(JLN^2 + (16L)^3)$ operations are needed for the k^{th} iteration of SNGCA, if a primal-dual IPM is used. For 10^5 variables or higher this makes the algorithmic cost of the iterative approach to SNGCA prohibitively large and thus limits the dimensionality of the data sets. Hence a non-iterative approach to the structural data analysis of SNGCA with almost linear in d complexity is sought.

Efficiency: Hence let us consider again the development of the estimation error as a function of increasing dimensionality of the embedding Gaussians in case of the failure modes reported in section 4.5. The failure of SNGCA in case of the Laplacian component to the data density is more or less expected: There is a relative slow transition with high variance demonstrating that although the iterative procedure improves significantly the accuracy of the recovery of $\mathcal{I}$ the choice of "informative" probe vectors ξ at the first iteration $k = 0$ remains a challenging task and hitherto is a weak point of the procedure. However the fast transition at $d \leq 15$ in the mixed Gaussian case indicates that in the "convex projection"-approach SNGCA do not make the best use of the available information to estimate the target space.

"Semidefinite Programming"-Approach: In order to improve the former "convex projection"-approach to SNGCA within the semi-parametric framework from section 2.3.2 let $G \in \mathbb{R}^{d\times L}$ be a matrix of averaged gradients of the test functions h_ω with columns γ_l and $U \in \mathbb{R}^{d\times L}$ a matrix of averaged functions xh_ω with columns γ_l. Analogously we build $\widehat{G} \in \mathbb{R}^{d\times L}$ and $\widehat{U} \in \mathbb{R}^{d\times L}$ from the data counterparts respectively such that $\widehat{G}$ and $\widehat{U}$ are estimators of G and U with

$$\|G - \widehat{G}\|_2 \leq \epsilon \quad \text{and} \quad \|U - \widehat{U}\|_2 \leq \epsilon. \tag{5.2}$$

Now observe if $c \in \mathbb{R}^L$ satisfies $Gc = \sum_{l=1}^L c_l\gamma_l = 0$ then $Uc = \sum_{l=1}^L c_l\eta_l$ belongs to $\mathcal{I}$, i.e. $(I - \Pi^*)Uc = 0$ where Π^* denotes the Euclidean projector on $\mathcal{I}$. To be more precise suppose that the set $\{h_l\}$ of test functions is comprehensive in the sense that vectors Uc span $\mathcal{I}$ where c fulfils the constraint $Gc = 0$. Since the projector $\Pi_\mathcal{I}$ is a symmetric $d \times d$ matrix of $rank\Pi = m$ with the eigenvalues $0 \leq \lambda_i(\Pi) \leq 1,\ i = 1, ..., d$ and $\mathrm{Tr}[\Pi] = m$, Π is identified by

$$\Pi^* = \min_\Pi \max_c \left\{ \|(I - \Pi)Uc\|_2^2 \;\middle|\; \begin{array}{c} 0 \preceq \Pi \preceq I,\ \mathrm{Tr}[\Pi] = m,\ rank\Pi = m; \\ c \in \mathbb{R}^L,\ Gc = 0 \end{array} \right\} \tag{5.3}$$

Here $\mathrm{Tr}[\cdot]$ denotes the trace of a matrix. For simplicity we will write Π^* instead of $\Pi^*_\mathcal{I}$ if their is no risk of confusion.

Now we aim to adapt (5.3) to the task of structural analysis. To this end we substitute $\widehat{U}$ and $\widehat{G}$ for U and G into (5.3). Thus we have to exchange $Gc = 0$ by the inequality constraint $\|\widehat{G}c\|_2 \leq \delta$ in order to keep the optimal solution c^* of (5.3) feasible in the "perturbed variant"

$$\min_\Pi \max_c \left\{ \|(I - \Pi)\widehat{U}c\|_2^2 \;\middle|\; \begin{array}{c} 0 \preceq \Pi \preceq I,\ \mathrm{Tr}[\Pi] = m,\ rank\Pi = m; \\ c \in \mathbb{R}^L,\ \|c\|_1 \leq 1,\ \|\widehat{G}c\|_2 \leq \delta \end{array} \right\}. \tag{5.4}$$

of (5.3). Currently nonconvex minmax problems as (5.4) can be solved efficiently only in the case of convex-concave games (c.f. [163]).

Relaxation: Consequently to make the "semidefinite programming"-approach viable we have to "relax" the quadratic and linear constrained problem (5.4) to a semidefinite problem. This is a classical approach called *Semidefinite Relaxation* (or SDP-relaxation) [218]. Concerning the task of structural analysis this means that the new approach to SNGCA is unified in the sense that the intermediary stages of estimating vectors from the target space and constructing an ONB are combined to the estimation of Π^* from only one directional sampling. Otherwise we transfer sampling, estimation procedure and

the dimension reduction concept to the new approach except the strategy of iteration and structural adaptation.

Dual Extrapolation Method: Furthermore we have to rewrite (5.4) such that the resulting problem is equivalent to (5.4) in the sense of optimization and belongs to a class of problems that can be solved efficiently. All known computationally cheap optimization techniques are black box oriented [27]. In general the task of solving semidefinite problems numerically with almost linear complexity in d, the requirement of low complexity currently leaves us with Quasi-Newton or Conjugate Gradient methods and gradient-type methods alternatively. However it is not obvious how to handle the constraints with Quasi-Newton or Conjugate Gradient methods [132]. Hence we focus on first-order methods with computationally cheap iterations.

Gradient-type methods for non-smooth convex optimization originate from the subgradient descent algorithm [208; 178]. The main update of the objective in the algorithm becomes

$$x_{k+1} = \Pi_{\mathcal{X}}(x_k - \alpha_k f'(x_k)) \tag{5.5}$$

where $\Pi_{\mathcal{X}}(y) \stackrel{\text{def}}{=} \arg\min_{y \in \mathcal{X}} \|x - y\|_2$ is the projector onto $\mathcal{X}$ and $\alpha_k > 0$ are stepsizes. Subgradient descent methods were extensively studied in the literature (see, e.g. [125; 124]). It is well known that subgradient descent methods and their extensions are intrinsically related to problems with Euclidean geometry. The non-Euclidean extensions of gradient-type methods allow to adjust, to some extent, a method to the geometry of feasible sets of the optimization problems on focus [16].

However standard first-order methods e.g. subgradient methods [18] are unable to utilize a priori knowledge of the data or structure of the problem such that progress is obtain solely on the basis of local information. Consequently in the large-scale case due to performance limits resulting from IBCT on black-box-oriented models, they exhibit only sublinear convergence and thus are unable to produce high-accuracy solutions on realistic time scales. Their achievable convergence rate depends on the smoothness of the objective, the geometry of the feasible sets and is never better than $\mathcal{O}(1/k^2)$. In the large-scale non-smooth case, the best guaranteed rate of convergence is $\mathcal{O}(1/\sqrt{k})$ [163]. However medium-accuracy solutions are in some sense welcome: If we consider problems of the form (5.4) based on statistical data as in SNGCA, we expect that the first iterations of the optimization method correspond to some progress in the approximation quality as long as the solution is adjusted to the non-Gaussian components of the data. But the more iterations are done the more progress in optimization is due to adjusting the actual solution to some individual, but perhaps statistically meaningless feature of the current sample of the data. Hence in the setting of SNGCA solely medium-accuracy numerical solutions in nonparametric statistics should not be considered as a drawback.

Fortunately motivated by results on IBCT, methods for large scale convex-concave saddle point problems with low analytical complexity of $\mathcal{O}(d \log d)$, linear memory requirements and convergence rate of $\mathcal{O}(1)k^{-1}$ in case of special geometry of the given feasible sets [119; 162; 167; 165] have been recently introduced. They belong to the family of subgradient descent-ascent methods and deal with at least one gradient step in the dual space E^*. Their total complexity $\mathcal{O}(1/\delta)$ is cheap compared with subgradient-type or cutting plane schemes [240] consuming $\mathcal{O}(1/\delta^2)$ evaluations of the objective where $\delta = \|f(x^*) - f(k_k)\|$ is a desired accuracy.

In the next sections we will report on semidefinite relaxation and the dual extrapolation method published in [167], that aims to solve variational inequalities with monotone operators. This type of problem is the most general optimization problem possessing a convex structure [68]. Then we give an equivalent, relaxed reformulation of (5.4) as convex-concave semidefinite problem such that the constraints in (5.1) are represented by the geometry of the feasible sets. The resulting problem is simple enough to apply the ideas of the dual extrapolation algorithm. Finally we illustrate the obtained progress using the toy examples already introduced in section 4.5.

5.1 Semidefinite Relaxation

Basics from Convex Analysis: First of all let us briefly summarize some concepts from convex analysis needed to introduce the "recipe" of relaxation. For this summary we follow [170],[93] and [238].

A set $\mathcal{S} \subset \mathbb{R}^d$ is called *convex* if $\forall\, x, y \in S$ and all $\lambda \in \mathbb{R}$ with $0 \geq \lambda \geq 1$ then $\lambda x + (1-\lambda)y \in S$. A function $f \;:\; \mathbb{R}^d \to \mathbb{R}$ is convex, if $\forall\, \lambda \in \mathbb{R}, 0 \geq \lambda \geq 1$ and all $x, y \in \mathbb{R}^d$ it holds

$$\lambda f(x) + (1-\lambda) f(y) \geq f(\lambda x + (1-\lambda) y)$$

Alternatively a convex function f is a real-valued function whose epigraph is convex. The epigraph of a function $f : \mathbb{R}^d \to \mathbb{R}$ is given by

$$\mathrm{epi}(f) \stackrel{\mathrm{def}}{=} \{(x, y) \;:\; x \in \mathbb{R}^d,\; y \geq f(x)\}.$$

The function $-f$ is concave is f is convex. Moreover any linear function is convex only if the quadratic forms arising from matrices A have nonnegative eigenvalues, i.e. $AA^\top \succeq 0$. Convexity may be checked by inspecting derivatives: $f \in \mathcal{C}^1$ is convex if and only if

$$\forall y \;:\; f(y) \geq f(x) + \nabla f(x)^{\mathcal{T}} (y - x)$$

Alternatively for $f \in \mathcal{C}^2$ convexity is equivalent to $\nabla^2 f \succeq 0$.

An interesting property of convex sets is that they are the intersection of all halfspaces which contain them.

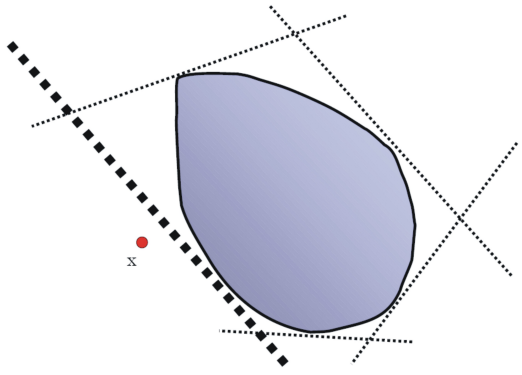


Figure 5.1: The convex set $\mathcal{S}$ is separated from the points not in the set by half-spaces. The dashed line separates the plane into two halves, one containing x and the other $\mathcal{S}$.

That is, if $x \notin \mathcal{S}$, then the Euclidean space can be divided into two halves, one half containing x and the other half containing the convex set. This property suggests that when

trying to find an optimal point x^* in $\mathcal{S}$, one could also search over the set of half-spaces which contain the set, see figure 5.1.

Applying this line of thoughts to optimization, consider the so called *primal problem*

$$\begin{aligned} &\min_{x\in\mathcal{X}} f(x) \\ \text{subject to} \quad &f_i(x) \leq 0 \quad j = 1, \ldots, M \end{aligned}$$

where f is the objective, $\mathcal{X}$ the feasible set and f_i are functional constraints. The *Lagrangian* for this problem is given by

$$\mathcal{L}(x, w) = f(x) + \sum_{i=1}^{M} w_i f_i(x)$$

with *Lagrange multipliers* w_i. The *dual problem* is given by $\max_{w\geq 0} \min_{x\in\mathcal{X}} \mathcal{L}(x, w)$. The solution of the dual problem provides a lower bound of the solution of the primal problem and is always a concave problem, even if the primal is not convex.

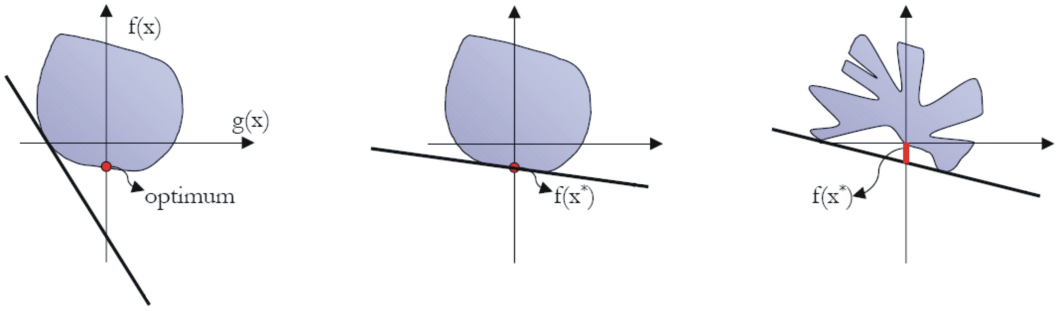


Figure 5.2: The set of possible pairs of $g(x)$ and $f(x)$ are shown as the blue region. Left: Any hyperplane which has normal $(w, 1)$ intersects the y-axis at the point $f(x^*) + w^\top > g(x^*)$ where x^* minimizes $\mathcal{L}(x, w)$ with respect to x. Middle: A hyperplane whose y intercept is equal to the minimum of $f(x)$ on the feasible set. The dual optimal value is equal to that of the primal. Right: No hyperplane can achieve the primal optimal value. The discrepancy between the primal and dual optima is called a *duality gap*. The dual optimum value is always a lower bound for the primal.

Figure 5.2 illustrates the graphical interpretation of duality: The optimal value is equal to the minimum crossing point on the y-axis. The dual problem seeks to find the half-space which contains the image of the problem and which has the greatest intercept with the $f(x)$ axis.

Finally, if the primal problem is not convex or not strictly feasible, it is often possible to bound the *duality gap* between the primal and the dual optimal values from above such that one can produce sub-optimal solutions to the primal problems whose cost is only a constant fraction away from optimality. This is the topic of convex relaxations.

Non-convex Quadratically Constrained Quadratic Programming: The central point is, that the Lagrangian dual of the general non-convex quadratically constrained quadratic problem is a semidefinite problem. Note that the "recipe" for relaxation heavily depends on the structure of the optimization problem [207; 179; 2]. Here we consider a

non-convex quadratically constrained quadratic program:

$$\begin{aligned} &\min_x \min x^\top A_0 x + 2b^\top x + c \\ \text{s.t.} \quad &x_i^\top A_i x + 2b_i^\top + c_i \leq 0 \qquad i = 1, \ldots, M. \end{aligned}$$

Due to its generality this problem is NP-hard. However the problem is efficiently solvable with a unique solution if $A_i \succeq 0$ for $i = 0, \ldots, M$.

In order to inquire the Lagrangian dual problem, we make a variable substitution to get the equivalent optimization:

$$\begin{aligned} &\min_y \min y^\top Q_0 y \\ \text{s.t.} \quad &y_i^\top Q_i y_i \leq 0 \qquad i = 1, \ldots, M \\ &y_0^2 = 1 \end{aligned}$$

where $y = [1\ x]^\top$ and

$$Q_i = \begin{bmatrix} c_i & b_i^\top \\ b_i & A_i \end{bmatrix}$$

Obviously the optimal values of both problems are equal. However the Lagrangian dual objective of the latter is

$$\begin{aligned} \mathcal{L}(y, w, t) &= y^\top Q(w, t) + t \quad \text{where} \\ Q(w, t) &= Q_0 + \textstyle\sum_{i=1}^M w_i Q_i - t \end{aligned}$$

Minimizing with respect to y, we obtain negative infinity if $Q(w, t)$ has at least one negative eigenvalue. The dual function $q(w, t)$ is given by

$$q(w, t) = \begin{cases} t & Q(w, t) \succeq 0 \\ -\infty & \text{otherwise} \end{cases} \tag{5.6}$$

and hence for the dual problem we get

$$\begin{aligned} &\max t \\ \text{s.t.} \quad &Q(w, t) = Q_0 + \textstyle\sum_{i=1}^M w_i Q_i - t \succeq 0 \\ &w \geq 0 \end{aligned}$$

This optimization is called a semidefinite program as the search is over the cone $\mathcal{S}_d$ of positive semidefinte matrices. The dual problem can be solved efficiently using IPMs [227; 228]. However for large d this is prohibitive time consuming.

Note that we can split the dual program as

$$\max_{w,t} \min_y \mathcal{L}(y, w, t) = \max_w \max_t \min_{y_0} \min_{y_1, \ldots, y_d} \mathcal{L}(y, w, t)$$

Then we ignore the maximization with respect to w and restrict ourselves to

$$\max_t \min_{y_0} \min_x \begin{bmatrix} y_0 \\ x \end{bmatrix} \begin{bmatrix} c & b^\top \\ b & Q(t) \end{bmatrix} \begin{bmatrix} y_0 \\ x \end{bmatrix} + t(1 - y_0^2)$$

Performing the minimization with respect to x, we either get negative infinity or, if the matrix is positive semidefinite, we get the Schur complement of the quadratic form

$$\max_t \min_{y_0} y_0^2(-b^\top Q(t)^{-1}b + c - t) + t$$

Obviously the saddle point of this problem is given when

$$t = -b^\top Q(t)^{-1}b + c$$
$$y_0^2 = 1.$$

Since that means

$$\max_t \min_{y_0} \min_x \begin{bmatrix} y_0 \\ x \end{bmatrix} \begin{bmatrix} c & b^\top \\ b & Q(t) \end{bmatrix} \begin{bmatrix} y_0 \\ x \end{bmatrix} + t(1 - y_0^2) = \min_x \begin{bmatrix} 1 \\ x \end{bmatrix} \begin{bmatrix} c & b^\top \\ b & Q(t) \end{bmatrix} \begin{bmatrix} 1 \\ x \end{bmatrix}$$

we arrive at the conclusion that the dual values with or without the additional ancillary variable y_0 are the same.

Furthermore let us consider the dual of the dual problem. Using $Z = yy^\top$ and duality again yields the semidefinite problem

$$\begin{aligned} &\min \mathrm{Tr}[Q_0 Z] && \text{s.t.} \\ \mathrm{Tr}[Q_i Z] &\leq 0 \qquad i = 1, \ldots, M \\ Z_{00} &= 1 \quad Z \succeq 0 \end{aligned}$$

Recall that the duality gap for semidefinite problems is zero whenever the primal is feasible and bounded.

Now we can demonstrate that this relaxation can be derived by dropping the non-convex constraints from the original primal program using the following lemma:

Lemma 1. *Let $y \in \mathbb{R}^d$. Then $Z = yy^\top$ if and only if Z is positive semidefinite and has rank 1.*

Proof. Let $Z = VDV^\top$ be the SVD of Z, i.e. $VV^\top = I$ with columns v_i and D is diagonal. Suppose that Z is positive semidefinite with $rankZ = 1$. Then without loss of generality we can assume that $d_{11} = 0$ and zero elsewhere. This implies

$$Z = d_{11}v_1v_1^\top = (\sqrt{d_{11}}v_1)(\sqrt{d_{11}}v_1)^\top$$

and we can set $y = \sqrt{d_{11}}v_1$. The converse is immediate. □

Then by means of the identity $y^\top Qy = \mathrm{Tr}[Qyy^\top]$ we can rewrite the original non-convex quadratic program as

$$\begin{aligned} &\min \mathrm{Tr}[QZ] && \text{s.t.} \\ \mathrm{Tr}[Q_i Z] &\leq 0 \qquad i = 1, \ldots, M \\ Z_{00} &= 1 \quad Z \succeq 0 \quad rankZ = 1 \end{aligned}$$

The rank constraint is non-convex, so a convex relaxation would be simply to drop it. This in fact would be the recipe for the semidefinite relaxation in our structural data analysis.

Convexification: Recall that with Π the matrix $I-\Pi$ is also idempotent and orthogonal. Thus we get the identity:

$$\|(I-\Pi)\widehat{U}c\|_2^2 = c^T\widehat{U}(I-\Pi)^2\widehat{U}c = c^T\widehat{U}(I-\Pi)\widehat{U}c = \mathrm{Tr}\left[\widehat{U}(I-\Pi)\widehat{U}X\right]. \tag{5.7}$$

Hence using (5.7), we consider the positive semidefinite matrix $X = cc^T \in \mathcal{X}^-$ with $rank X = 1$ as "new variable" of a relaxed and *linearized* version of the objective in (5.4). Here $\mathcal{X}^-$ denotes the feasible set of the variable X. Moreover we set $|X|_1 \stackrel{\text{def}}{=} \sum_{i,j=1}^{L} |X_{ij}|$. Then rewriting the constraints of the original objective in our new terms, we can derive the semidefinite relaxation in two steps: First due to the introduction of the new variable, we substitute the ℓ_1-constraint $\|c\|_1 \leq 1$ by $|X|_1 \leq 1$ and transform $\|\widehat{G}c\|_2 \leq \delta$ into $\mathrm{Tr}[\widehat{G}X\widehat{G}] \leq \delta^2$.

However the constraint $rank X = 1$ is non-convex and leads to a computationally hard problem [2]. In order to get an efficiently solvable problem we simply drop this constraint such that $X \in \mathcal{X}$ and $\mathcal{X}^- \subset \mathcal{X}$. The consequence is that for a minimization problem f we get

$$\min_{X\in\mathcal{X}^-} f(X)) \leq \min_{X\in\mathcal{X}} f(X))$$

Hence we come to the linear constrained problem

$$\widehat{\Pi} = \min_{\Pi}\max_{X}\left\{ \mathrm{Tr}\left[\widehat{U}(I-\Pi)\widehat{U}X\right] \;\middle|\; \begin{array}{c} 0 \preceq \Pi \preceq I,\ \mathrm{Tr}[\Pi] = m,\ rank\Pi = m; \\ X \succeq 0,\ |X|_1 \leq 1,\ \mathrm{Tr}[\widehat{G}X\widehat{G}] \leq \delta^2 \end{array} \right\}. \tag{5.8}$$

Yet the problem (5.8) is still not convex in Π. Therefore we remove the constraint $rank\Pi = m$ and finally arrive at

$$\widehat{P} = \min_{P}\max_{X}\left\{ \mathrm{Tr}\left[\widehat{U}(I-P)\widehat{U}X\right] \;\middle|\; \begin{array}{c} 0 \preceq P \preceq I,\ \mathrm{Tr}[P] = m, \\ X \succeq 0,\ |X|_1 \leq 1,\ \mathrm{Tr}[\widehat{G}X\widehat{G}] \leq \delta^2 \end{array} \right\}. \tag{5.9}$$

Note that $\widehat{P}$ of (5.9) is not a projector matrix. To provide an estimation of the projector Π^*, one can use the projector $\widehat{\Pi}$ onto the subspace spanned by m principal eigenvectors of $\widehat{P}$.

Finally we have to bound the error of the estimations $\widehat{P}$ and $\widehat{\Pi}$ of Π^* that stems from the semidefinite relaxation. To this end we need an identifiability assumption on the system $\{h_l\}$ of test functions as follows:

Assumption 9. *Suppose that there are vectors $c_1, ..., c_{\overline{m}}$, $m \leq \overline{m} \leq L$ such that $\|c_k\|_1 \leq 1$ and $Gc_k = 0$, $k = 1, ..., \overline{m}$, and non-negative constants $\mu^1, \ldots, \mu^{\overline{m}}$ such that*

$$\Pi^* \preceq \sum_{k=1}^{\overline{m}} \mu^k U c_k c_k^T U^T. \tag{5.10}$$

We denote $\mu^* = \mu^1 + \ldots + \mu^{\overline{m}}$.

In other words, if Assumption 9 holds, then the projector Π^* is μ^* times a convex combination of rank-one matrices Ucc^TU^T where c satisfies the constraints $Gc = 0$ and $\|c\|_1 \leq 1$. The assumption holds if U spans the whole data space [46].

Theorem 10. *Let Assumption 9 hold. Then an optimal solution $\widehat{P}$ of (5.9) satisfies*

$$\mathrm{Tr}\left[(I-\widehat{P})\Pi^*\right] \leq 4\mu^*\delta^2(\lambda_{\min}^{-1}(\Sigma)+1)^2. \tag{5.11}$$

Further, if $\widehat{\Pi}$ *is the projector onto the subspace spanned by* m *principal eigenvectors of* $\widehat{P}$*, then*

$$\|\widehat{\Pi} - \Pi^*\|_2^2 \leq \frac{8\mu^*\delta^2(\lambda_{\min}^{-1}(\Sigma)+1)^2}{1-4\mu^*\delta^2(\lambda_{\min}^{-1}(\Sigma)+1)^2} \tag{5.12}$$

(here $\|A\|_2 = \left(\sum_{i,j} A_{ij}^2\right)^{1/2} = \left(\mathrm{Tr}[A^T A]\right)^{1/2}$ *is the Frobenius norm of A).*

The proof of this theorem can be found in the appendix A. Note that Σ is the covariance matrix of the original distribution, not the data covariance matrix that is typically bad conditioned in high dimensions. In the next section we will give an intuitive introduction to the dual extrapolation method.

5.2 Objectives with Convex Structure

Since we have applied a semidefinite relaxation to the original problem (5.1) used to obtain "good" coefficients c_l for the estimation of vectors $\beta \in \mathcal{I}$, we arrive at a semidefinite and constrained convex large scale problem. In order to solve this type of problem we aim to apply the dual extrapolation method [167], that is designed to solve variational inequality problems (VIP) with monotone operators.

5.2.1 Variational Inequalities

The deterministic gradient algorithm in the dual space from above aims at solving convex-concave nonlinear optimization problems $f: E \to \mathbb{R}$ of the so called saddle point form

$$\min_{X\in\mathcal{X}} \max_{Y\in\mathcal{Y}} f(X,Y) \tag{5.13}$$

where $\mathcal{X}, \mathcal{Y} \subseteq dom f$ are convex and compact sets from a finite dimensional vector space E. With $dom f \subseteq \mathbb{R}^d$ we denote the domain of the objective f where f is continuously differentiable with Lipschitz continuous gradient, i.e. $f \in \mathcal{C}^{1,1}_{L_{\|\cdot\|}(f)}(\mathbb{R}^d \times \mathbb{R}^d, \mathbb{R})$ where $L_{\|\cdot\|}(f)$ is the Lipschitz constant. It is well known [18] that if f is a Lipschitz-continuous, convex function, then f has saddle points, i.e. points (x^*, y^*) such that

$$f(x^*, y) \leq f(x^*, y^*) \leq f(x, y^*)$$

for any $x \in \mathcal{X}$ and $y \in \mathcal{Y}$. Hence minimizing a convex function over a convex feasible set is to find the saddle points of a convex-concave Lagrange function. The existence of saddle points gives rise to a corresponding pair of convex problems such that the minimax inequality

$$\max_{y\in\mathcal{Y}} \min_{x\in\mathcal{X}} f(x,y) \leq \min_{x\in\mathcal{X}} \max_{y\in\mathcal{Y}} f(x,y)$$

holds [18]. The left hand side is called the dual problem (D) and the right hand side the primal problem (P). Moreover we can define a duality gap by

$$0 < DG(f) \stackrel{\text{def}}{=} \max_{y\in\mathcal{Y}} f(x,y) - \min_{x\in\mathcal{X}} f(x,y) \tag{5.14}$$

that in terms of the objectives of the primal and dual problem is the sum of the residuals and thus allows to determine the quality of numerical approximate solutions of (P) and (D). Obviously problem (5.9) is of type (5.13).

In order to explain the ideas behind the dual extrapolation method let us rewrite (5.13) as

$$F(x,y) = \begin{bmatrix} \frac{\partial}{\partial x} f(x,y) \\ -\frac{\partial}{\partial y} f(x,y) \end{bmatrix} \tag{5.15}$$

where F is a continuous nonlinear operator $F: \mathcal{X} \times \mathcal{Y} \to E^*$. Here E^* denotes the dual space of E with norm $\|s\|_* = \max_{x \in E}\{\langle s, x\rangle : \|x\| \leq 1\}$. $\mathcal{X}$ and $\mathcal{Y}$ are compact and convex feasible sets. In the sequel we use $z = [x,y]^T$ and $\mathcal{K} = \mathcal{X} \times \mathcal{Y}$. Recall that the problem of a (deterministic) variational inequality is to find a so called variational point $z \in \mathcal{K}$ such that

$$\text{VIP}(F,\mathcal{K}): \quad F(z)^T(z' - z) \leq 0, \quad \forall z \in \mathcal{K} \tag{5.16}$$

Since f is convex if and only if F is monotone on $\mathcal{K}$, i.e.

$$(z - z')^T(F(z) - F(z')) \geq 0 \quad \forall z, z' \in \mathcal{K} \tag{5.17}$$

we conclude that a unique solution of (5.13) can be obtained by solving (5.16) if F is monotone [68]. In turn f in (5.13) is a $\mathcal{C}^{1,1}$-function and convex-concave and we conclude that F is monotone.

5.2.2 Extragradient Methods

One strategy to solve a VIP(F,$\mathcal{K}$) with at most linear numerical convergence in d, is to apply a fixpoint algorithm motivated by a close link between the variational inequality and the projection: Consider the projection $\Pi_{\mathcal{K}}: E \to \mathcal{K}$ defined by

$$\Pi_{\mathcal{K}}(z) \stackrel{\text{def}}{=} \arg\min_{z' \in \mathcal{K}} \|z' - z\| \tag{5.18}$$

It is well known [68] that for all $z, z' \in \mathcal{K}$ the projection inequality holds

$$(z - \Pi_{\mathcal{K}}(z))(z' - \Pi_{\mathcal{K}}(z)) \leq 0 \tag{5.19}$$

Using the identity $F \equiv -[(I - F) - I]$ we obtain from (5.16) the inequality

$$[(I - F)(z') - z']^T(z - z') \leq 0 \tag{5.20}$$

Comparing (5.20) with (5.19) leads to the fixpoint equation

$$z = \Pi_{\mathcal{K}}((I - F)(z)) \tag{5.21}$$

The existence of a fixpoint follows by the fixpoint theorem of Brouwer for continuous operators F [187]. Obviously (5.21) is equivalent to $z \in \ker\,[I - \Pi_{\mathcal{K}} \circ (1 - F)]$. Analogously we conclude for a point $u \in \mathbb{R}^d$ such that z is its projection on $\mathcal{K}$, u must be an element of the kernel space of the adjoint operator, i.e.

$$u = (1 - F) \circ \Pi_{\mathcal{K}}(u) = (1 - F)(z) \tag{5.22}$$

Since (5.16) is scale invariant a factor $\gamma > 0$ can be introduced. Then substituting (5.22) into (5.21) gives

$$z \;\; = \;\; \Pi_{\mathcal{K}}(z - \alpha F \circ \Pi_{\mathcal{K}}((1 - \alpha F)(z))) \tag{5.23}$$

suggesting a numerical prox-type [93] iteration scheme where α can be considered as stepsize. Using the 2-norm in (5.18) adapts (5.23) to the Euclidean space. The efficiency of this so called primal extragradient algorithm hinges on the computational complexity of the Euclidean projection onto the feasible sets [90].

The Non-Euclidean Case: Let E^* be the dual space of E. Suppose we substitute in (5.23) the projector by the so called *prox-transform* $T_\beta(x,s)$. Intuitively, $T_\beta(x,s)$ tries to make a step from x in the direction of s penalized by $\beta V(x,y)$ with $\beta > 0$. To be more precise, let $d : \mathcal{K} \to \mathbb{R}$ be a given distance generating function. Motivated by the low computational complexity and geometrical considerations to be described in the spectahedron setup below, a typical choice for $d(\cdot)$ is a Bregman function [28]. We impose $d(\cdot)$ to be continuous and strongly convex with modulus $\alpha > 0$ on $\mathcal{K}$ with respect to the norm $\|\cdot\|$, i.e.

$$\langle \nabla d(z) - \nabla d(z'), z - z' \rangle \geq \alpha \|z - z'\|^2 \quad \forall z, z' \in \mathcal{K}.$$

As usual let us denote by

$$d^*_{\mathcal{K}} \stackrel{\text{def}}{=} \max_{x \in \mathcal{K}} \{\langle s, x \rangle - d(x) : \ s \in E^*\} \tag{5.24}$$

the conjugate or Fenchel-Legendre-Transformation (FLT) illustrated in figure 5.3 of $d(\cdot)$. Since $d(\cdot)$ is strongly convex, $d^*_{\mathcal{K}}(\cdot)$ is well defined, convex and differentiable at any $s \in E^*$ [187].

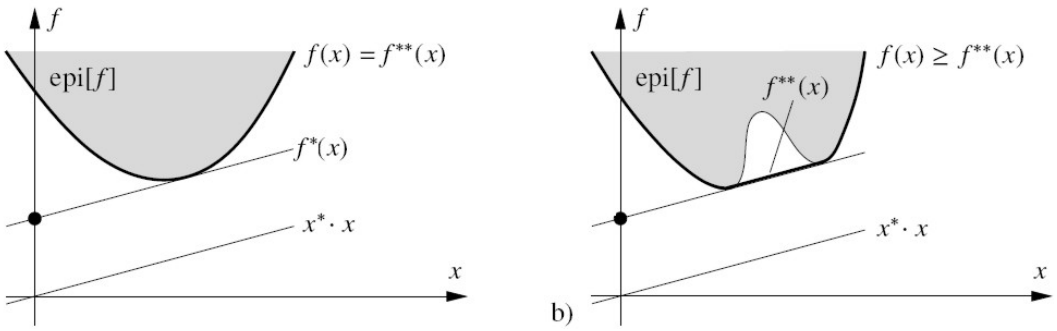


Figure 5.3: FLT for a convex (left) and a nonconvex function (right). The hypeplanes $\langle s, x \rangle - d(x)$ are always below $epi(f)$. By f^{**} we denote the biconjugate. Note that from $f(x) \geq f^{**}(x)$ it follows that $f^{**}(x)$ is the convex hull of f and $f^*(x)$ is a supporting hyperplane in x.

Note that for a strict convex function it holds

$$s = \partial_x f(x) \quad x = \partial_s f(s) \quad f(x) + f^*(s) = \langle s, x \rangle$$

Next we define

$$\widetilde{\mathcal{K}} \stackrel{\text{def}}{=} \{x \in \mathcal{K} \mid x = \nabla d^*_{\mathcal{K}}(s), s \in E^*\} \ \subseteq \ \mathcal{K}$$

and impose $d(x)$ to be differentiable at any $x \in \widetilde{\mathcal{K}}$. Here $\nabla d(z)$ denotes the gradient of $d(z)$. Now the scaled *prox-transform* $T_\beta(x,s) : \ E \times E^* \times \mathbb{R}^+ \to \widetilde{\mathcal{K}}$ can be introduced in terms of the distance generating function by

$$T_\beta(x,s) \stackrel{\text{def}}{=} \arg\min_{y \in \widetilde{\mathcal{K}}} \big\{ \langle s, y - x \rangle - \beta V(x,y) \big\} \tag{5.25}$$

where $V(x,y)$ is a local distance called *prox-function* associated with $d(\cdot)$ via

$$V(x,y) \stackrel{\text{def}}{=} d(y) - d(x) - \langle \nabla d(x), y - x \rangle > 0.$$

$\beta > 0$ is the scaling parameter. If $d(z)$ is a Bregman function, $V(x,y)$ is called a Bregman divergence. Consequently the prox-transform $T_\beta(x,s)$ as the Fenchel-Legendre transform of $\beta V(x,y)$ is well-defined and a $\mathcal{C}^{1,1}_{L_{\|\cdot\|}(f)}$-function, since $V(x,y)$ is strongly convex also. Geometrically, the prox-transform answers the question about the minimal shift down of the hyperplane $s = e^T x \in \mathbb{R}^d$ which places it below the graph of a given function to be linear approximated. Since $T_\beta(x,s)$ is a contraction [162], replacing $\Pi_{\mathcal{K}}(z)$ in (5.23) by $T_\beta(x,s)$ suggests a fixpoint iteration scheme also. Let us fix this more formally.

Dual Extrapolation Step: In sum the dual extrapolation algorithm adjusts the update step of the extragradient-type method to the geometry of the objective induced by the norm using Bregman functions. The update step $\mathcal{E}_{\beta,\alpha_k}(s)$ transforms an arbitrary point $s \in E^*$ by means of $T_\beta(x,s)$ into a new point s^+:

$$(x,y,s^+) = \mathcal{E}_{\beta,\alpha_k}(s) \Leftrightarrow \begin{cases} x = & T_\beta(\overline{x}, s) \\ y = & T_\beta(x, -\alpha_k F(x)) \\ s^+ = & s - \alpha_k F(y) \end{cases}$$

where it is assumed that an arbitrary point $\overline{x} \in \mathcal{K}$ is the center of $\mathcal{K}$. Obviously $\mathcal{E}_{\beta,\alpha_k}(s)$ is an update of an affine function, which can be considered as a local model of the objective. In order to describe its convergence properties, let us introduce the convex, restricted merit function [18] by

$$\Phi_D(x) \stackrel{\text{def}}{=} \max_{y \in \mathcal{K}} \{ \langle F(y), y - x \rangle \ : \ V(\overline{x}, y) < D \} \tag{5.26}$$

where $D > 0$ is a fixed parameter. (5.26) works as a measure of the quality of any point $x \in \mathcal{K}$. x is an approximate solution of (5.16) on $\mathcal{K}_D \stackrel{\text{def}}{=} \{ y \in \mathcal{K} : \ V(x,y) \leq D \}$ since on the one hand it holds $\Phi_D(x^*) = 0$ if and only if x^* solves (5.16). On the other hand if we define arbitrary search points [166] by

$$\widetilde{y}_n \stackrel{\text{def}}{=} \frac{1}{\sum_{k=0}^n \alpha_k} \sum_{k=0}^n \alpha_k y_k \tag{5.27}$$

with stepsizes $\alpha_k > 0$ we get for smooth variational inequalities the complexity estimate

$$\Phi_D(\widetilde{y}_n) \leq \frac{L_{\|\cdot\|}(F) D}{\alpha(k+1)} \tag{5.28}$$

that is unimprovable due to results from IBCT [160]. Here we have the Lipschitz constant

$$L_{\|\cdot\|}(F) = \max_{z,z' \in \mathcal{K}} \frac{\|F(z - z')\|_*}{\|z - z'\|}$$

on $\mathcal{K}$ and k is the current number of iteration. Moreover $\Phi_D(x)$ is well defined and convex on E [167]. The complexity of each step heavily depends on the costs of the objective evaluation and the computation of $T_\beta(x,s)$. In cases of special geometry of the feasible sets of the problem the latter is cheap.

The Spectahedron Setup: The configuration of the setup of the optimization problem has to attend to some conditions. First if $\mathcal{K}$ is bounded, the parameter D can be chosen as

$$D = \max_{y \in \mathcal{K}} V(\overline{x}, y)$$

such that the choice of $d(\cdot)$ influences the performance of the algorithm. Hence given the feasible set and the norm of the problem the choice of $d(\cdot)$ should minimize D. Second if f is Lipschitz continuous with $L_{\|\cdot\|_1}(f)$ such that the feasible set is the "full" simplex

$$\Delta_d^+ \stackrel{\text{def}}{=} \{x \in \mathbb{R}^d \mid 0 \leq x, \sum_{i=1}^{d} x_i \leq 1\} \tag{5.29}$$

then we distance generating function should represent the distribution of the components of the solution $x \in \mathcal{R}^d$ on Δ_d^+. Hence we choose as Bregman function the entropy distance function

$$d(x) = \sum_{i=1}^{d} x_i \ln(x_i) \tag{5.30}$$

If $\forall i \; 0 \leq x_i$, this choice covers the always positive and convex Kullback-Leibler divergence between the uniform distribution and the distribution given by the values of the coefficients x_i. However other choices for $d(\cdot)$ are possible [16]. Consequently [165; 164] it holds

$$V(x, y) = \sum_{i=1}^{d} y_i \ln\left(\frac{y_i}{x_i}\right) \tag{5.31}$$

and the i^{th} component of the prox-transform is given [165] by

$$T_\beta(x, s)_i = \frac{x_i \exp(s_i/\beta)}{\sum_{j=1}^{d} x_j \exp(s_j/\beta)} \tag{5.32}$$

This gives rise to the *spectahedron setup* in the matrix case, where A, B belong to the space S_d of $d \times d$ blockdiagonal matrices equipped with the Frobenius inner product $\langle A, B \rangle_F = \mathrm{Tr}[AB]$ and the trace norm $||A||_1 \stackrel{\text{def}}{=} \sum_{i=1}^{n} \lambda_i(A)$. Here $\lambda_i(A)$ denotes the i^{th} singular value of A. The dual matrix norm $||A||_*$ is given by the usual spectral norm. In this case the setup for the extrapolation method on $\mathcal{S}_d \stackrel{\text{def}}{=} \{A \in S_d \mid \mathrm{Tr}[A] \leq 1\}$ is completed [164; 119] by $D = \ln d$, $\alpha = 0.5$ and

$$d(A) = \sum_{i=1}^{d} \lambda_i(A) \ln(\lambda_i(A)) \tag{5.33}$$

is the matrix entropy. The prox-transform in the spectahedron case is obtained from (5.32) analogously.

We will now describe how to apply the dual extrapolation method in the spectahedron setup to the task of semi-parametric structural data analysis posed by SNGCA.

5.2.3 Application to Structural Data Analysis

Observe that due to the eigenvalue decomposition $\widehat{G}\widehat{G}^\top = \Gamma\Lambda\Gamma^\top$ the constraint $\mathrm{Tr}[\widehat{G}\widehat{G}^\top X] = 0$ claims that X vanishes outside the kernel of $\widehat{G}\widehat{G}^\top$, a subspace that is spanned by the columns of Γ corresponding to the non-zero eigenvalues of $\widehat{G}\widehat{G}^\top$. Hence we can define a matrix $Q \in \mathbb{R}^{L\times(L-d)}$ as submatrix of that columns of Γ that corresponds to the vanishing eigenvalues of $\widehat{G}\widehat{G}^\top$. Moreover we set $\widehat{V} = \widehat{U}Q$ and $X = QZQ^\top$. Using this definitions and $\mathcal{Z} \stackrel{\mathrm{def}}{=} \{Z \in \mathcal{S}_{L-d} \mid Z \succeq 0, \mathrm{Tr}[Z] \leq 1\}$ we get the optimization problem

$$\arg\min_{P\in\mathcal{P}} \arg\max_{Z\in\mathcal{Z}} \mathrm{Tr}[\widehat{V}^\top(I-P)\widehat{V}Z] \tag{5.34}$$

However in order to stabilize the computation against stochastic perturbations we use the convex [48] reformulation

$$\arg\min_{(P,W)\in\mathcal{P}\times\mathcal{W}} \arg\max_{(Z,Y)\in\mathcal{Z}\times\mathcal{Y}} \Big\{ \mathrm{Tr}[\widehat{V}^\top(I-P)\widehat{V}Z] + r\mathrm{Tr}[W(QZQ^\top - Y)] \Big\} \tag{5.35}$$

of the primal problem instead of (5.34) where $r \geq L\|U\|^2$ is the parameter of the quadratic Moreau-Yosida regularization [138]. Hence (5.35) is always smooth and convex in Z. The regularization of the dual problem is often done by the augmented Lagrangian technique [93]. The feasible sets $\mathcal{W}$ and $\mathcal{Y}$ in (5.35) are given by

$$\begin{aligned} \mathcal{W} &\stackrel{\mathrm{def}}{=} \{W \in \mathcal{S}_L \mid W \succeq 0, \mathrm{Tr}[W^2] \leq 1\} \\ \mathcal{Y} &\stackrel{\mathrm{def}}{=} \{Y \in \mathcal{S}_L \mid |Y_{ij}|_1 \leq 1\} \end{aligned}$$

In the sequel we will denote the objective in (5.35) as $f(P, Z, W, Y)$ also. It can be shown that (5.34) can be reduced to (5.35) in the sense of the following lemma.

Lemma 2. *Let $(\widehat{P}, \widehat{W}, \widehat{Z}, \widehat{Y})$ be a feasible δ-solution to (5.35), i.e. $\overline{f}(\widehat{P}, \widehat{W}) - \underline{f}(\widehat{Z}, \widehat{Y}) \leq \delta$ where*

$$\overline{f}(\widehat{P}, \widehat{W}) \stackrel{\mathrm{def}}{=} \max_{(Z,Y)\in\mathcal{Z}\times\mathcal{Y}} f(P, Z, W, Y) \tag{5.36}$$

$$\underline{f}(\widehat{Z}, \widehat{Y}) \stackrel{\mathrm{def}}{=} \min_{(P,W)\in\mathcal{P}\times\mathcal{W}} f(P, Z, W, Y) \tag{5.37}$$

Then the pair $(\widehat{P}, \widetilde{Z})$ is a feasible δ-solution to the problem (5.34), i.e. it holds that $\overline{g}(\widehat{P}) - \underline{g}(\widehat{Z}) \leq \delta$ where

$$\overline{g}(P) \stackrel{\mathrm{def}}{=} \max_{X\in\mathcal{X}, \mathrm{Tr}[\widehat{G}^\top\widehat{G}X]\leq\epsilon^2} \mathrm{Tr}[\widehat{U}^\top(I-P)\widehat{U}X] \tag{5.38}$$

$$\underline{g}(X) \stackrel{\mathrm{def}}{=} \min_{P\in\mathcal{P}, \mathrm{Tr}[\widehat{G}^\top\widehat{G}X]\leq\epsilon^2} \mathrm{Tr}[\widehat{U}^\top(I-P)\widehat{U}X] \tag{5.39}$$

The proof of this lemma follows directly from (5.35) and the definitions (5.36), (5.37), (5.38) and (5.39). Writing (5.9) as (5.35) is motivated by the fact that all the feasible sets are convex and admit evident distance generating functions $d(\cdot)$ such that (5.35) conforms to the required spectahedron setup for the dual extrapolation method.

5.3 Algorithmic Procedures

We will see in this section, that the total numerical effort of the new approach is given by $\mathcal{O}(LN^2 + L\log L)$ where every computation of the prox-transform costs $\mathcal{O}(L^3)$.

Recall that the feasible sets $\mathcal{P}$, $\mathcal{Z}$ and $\mathcal{Y}$ induce the spectahedron setup whereas in the case of computing the prox-transform (5.25) for the weight matrix $W \in \mathcal{W}$ the Euclidean unit ball setup is appropriate. In the sequel we will denote them as standard setups.

The prox-transform of Π: Using the matrix entropy as distance generating function, we have to solve

$$\begin{aligned} T_{\beta_\Pi}(\Pi, S) &= \arg\max_{Y\in\mathcal{P}}\left\{\mathrm{Tr}[S(Y-\Pi)] - \beta_\Pi \mathrm{Tr}\Big[\frac{Y}{m}\Big(\log\frac{Y}{m} - \log\frac{\Pi}{m}\Big)\Big]\right\} \\ &= \arg\max_{Y\in\mathcal{P}}\left\{\mathrm{Tr}\Big[Y\Big(S + \frac{\beta_\Pi}{m}\log\frac{\Pi}{m}\Big)\Big] - \beta_\Pi \mathrm{Tr}\Big[\frac{Y}{m}\log\frac{Y}{m}\Big]\right\} \end{aligned}$$

Since S is a symmetric matrix, we can compute the eigenvalue decomposition

$$S + \frac{\beta_\Pi}{m}\log\frac{\Pi}{m} = \Gamma\Lambda\Gamma^\top \tag{5.40}$$

with $\Lambda \stackrel{\mathrm{def}}{=} \mathrm{diag}(\lambda_1, \ldots, \lambda_L)$. Substituting (5.40) into the prox-transform leads to the equivalent problem

$$y^* = \arg\max_{0\le y\le 1, \sum_l y_l \le m} \lambda^\top y - \frac{\beta_\Pi}{m}\sum_{l=1}^{L} y_l \log\frac{y_l}{m} \tag{5.41}$$

such that $T_{\beta_\Pi}(\Pi, S) = \Gamma\mathrm{diag}(y^*)\Gamma^\top$. Using the Lagrangian dual of (5.41) we can obtain its solution componentwise from

$$y_l^* = \exp\Big(\beta_\Pi^{-1} s_l - w\Big) \ \wedge\ 1 \tag{5.42}$$

where the Lagrange multiplier w is set to get $\sum_l y_l^* = m$. This problem can be solved by a bisection method [18] in w.

The prox-transform of Z: Using the analogous argument as above we have to consider the optimization problem

$$y^* = \arg\max_{0\le y\le 1, \sum_l y_l \le 1} \lambda^\top y - \beta_Z \sum_{l=1}^{L} y_l \log y_l$$

In this case the prox-transform is analytically given [16] for each component by

$$y_l^* = \frac{\exp(\frac{s_l}{\beta_Z})}{\sum_l \exp(\frac{s_l}{\beta_Z})} \tag{5.43}$$

such that $T_{\beta_Z}(Z, S) = \Gamma\mathrm{diag}(y^*)\Gamma^\top$. Here Γ is obtained from the solution of the eigenvalue problem $S + \beta_Z \log Z = \Gamma\Lambda\Gamma^\top$.

The prox-transform of W: In the Euclidean case the distance-generating function is given by $d(X) = 0.5\mathrm{Tr}[X^\top X]$. Consequently we have to solve

$$T_{\beta_W}(W, S) = \arg\max_{\mathrm{Tr}[Y]\le 1}\left\{\mathrm{Tr}[S(Y-W)] - \frac{\beta_W}{2}\mathrm{Tr}[(Y-W)^\top(Y-W)]\right\}$$

Using the derivative of the trace the solution is given by

$$Y^* = \begin{cases} W + \beta_W^{-1} S & \text{if } \|W + \beta_W^{-1} S\|_2 \leq 1 \\ \frac{W+\beta_W^{-1} S}{\|W+\beta_W^{-1} S\|_2} & \text{otherwise.} \end{cases} \tag{5.44}$$

The prox-transform of Y: Again the distance generating function is the entropy. Since the constraint $\|Y\|_1 \leq 1$ is not smooth, we set $Y = U - Y$ with $0 \leq V_{ij}, U_{ij}$ for all components. The consequence of this representation is, that we have to solve the problem

$$T_{\beta_Y}(Y, S) = \arg\max_{0 \leq U_{ij}, V_{ij} \leq 1} \left\{ \mathrm{Tr}[S(U - V)] - \beta_Y \sum_{ij} \left[U_{ij} \log \frac{U_{ij}}{U_{ij}^0} + V_{ij} \log \frac{V_{ij}}{V_{ij}^0} \right] \right\}$$

To this end we set

$$a_{ij} = U_{ij} \exp(\beta_Y^{-1} S_{ij}) \qquad b_{ij} = V_{ij} \exp(\beta_Y^{-1} S_{ij})$$

Then in some sense the matrices U and V can be propagated inside the spectahedron by

$$U_{ij}^* = \frac{a_{ij}}{\sum_{ij} a_{ij} + b_{ij}} \qquad V_{ij}^* = \frac{b_{ij}}{\sum_{ij} a_{ij} + b_{ij}} \tag{5.45}$$

Finally we set $Y^* = U^* - V^*$. Now we are to describe all the details of the numerical implementation.

Initialization: Let ϵ be numerical accuracy. As initial values of the variables we choose $P_0 = \frac{m}{d} I$, $Z_0 = \frac{1}{L} I$, $U_{ij} = V_{ij} = (2L)^{-1}$ and $W = \mathbf{0}$, where $\mathbf{0} \in \mathbb{R}^{\mathbf{L} \times \mathbf{L}}$ and $\overline{L} = L - d$. For the step size we choose $\gamma_0 = \gamma_1 = 1$. $\overline{P} = 0$, $\overline{W} = 0$, $\overline{Z} = 0$, $\overline{Y} = 0$ are set as centers of the feasible sets. Set $\kappa_{up} = 1.4$ and $\kappa_{down} = 0.5$. The value m^* can be estimated by looking how many eigenvalues of $\widehat{P}$ in algorithm 7 are significant. Finally we set $d_P = \log d$, $d_W = 1$, $d_Z = \log(L - d)$ and $d_Y = \log(2L^2)$.

Stepsize and stopping rule: In comparison to subgradient schemes here it is not necessary for convergence that the stepsizes $\alpha_k > 0$ build up a divergent step size series, such that $\alpha_k \to 0$ for $k \to \infty$ and $\sum_{k=0}^{\infty} \alpha_k = \infty$ [168]. However the accuracy heavily depends on the choice of $\alpha_k > 0$. Let $\overline{x} \in \mathcal{X}$ be the prox center, $s_k \in E^*$ the corresponding gradient in x_k in the k^{th} iteration and s_k^{++} the extrapolated gradient. Moreover let

$$\begin{aligned} \Psi_\beta(z, s) &\stackrel{\text{def}}{=} \max_{x \in \mathcal{K}_D} \Big\{ \langle s, x - z \rangle - \beta V(z, x) \Big\} \\ \tau_k(D) &\stackrel{\text{def}}{=} \max_{x \in \mathcal{K}_D} \Big\{ \sum_{i=1}^{k} \alpha_i \langle F(y_i), y_i - x \rangle \Big\} \end{aligned}$$

Then it is shown in [167] that for smooth variational problems, it holds that

$$\begin{aligned} \tau_k(D) = \sum_{i=1}^{k} \alpha_i \langle F(y_i), y_i - \overline{x} \rangle + \max_{x \in \mathcal{K}_D} \Big\{ - \sum_{i=1}^{k} \alpha_i \langle F(y_i), x - \overline{x} \rangle \Big\} \\ \max_{x \in \mathcal{K}_D} \Big\{ \langle s, x - \overline{x} \rangle \Big\} \leq \beta D + \Psi_\beta(\overline{x}, s) \end{aligned} \tag{5.46}$$

For the restricted merit function $\Phi_D(x)$ defined in (5.26) this means

$$\Phi_D(\widetilde{y}_k) \leq \tau_k(D)(\sum_{i=1}^{k} \alpha_i)^{-1} \tag{5.47}$$

(5.46) and (5.47) can be used to justify a stepsize strategy as well as a termination criterion as follows: Suppose we want to test the current stepsize α_k. Depending on the setup of the variable of the objective we can compute in the k^{th} iteration

$$\zeta_k(x) \stackrel{\text{def}}{=} \alpha_k \langle s_k^{++}, x_k - x_0 \rangle - \beta V(s_k^{++}, x_0)$$

for every primal variable. Hence if $\overline{\zeta}_k \leq \overline{\zeta}_0 - \tau_k$ with $\overline{\zeta}_k = \zeta_k(P) + \zeta_k(Z) + \zeta_k(W) + \zeta_k(Y)$ the stepsize α_k is acceptable small. Moreover due to (5.48) and (5.47) a reasonable termination criterion is given by

$$\frac{\overline{\zeta}_k + \overline{d} + \sum_{i=1}^{k} \tau_i(D)}{\sum_{i=1}^{k} \alpha_i} \leq \epsilon$$

where $\overline{d} = \beta_P d_P + \beta_W d_W + \beta_Z d_Z + \beta_Y d_Y$.

Choice of scaling parameters: According to (5.46) and (5.47) we are interested in the lowest upper bound for the merit function. Hence in order to compute the scaling parameters of the prox-transform β_P, β_Z, β_Y and β_W we have to solve to following problem:

$$\begin{aligned} &\min \{\beta_P d_P + \beta_W d_W + \beta_Z d_Z + \beta_Y d_Y\} \quad \text{s.t.} \\ &\beta_P \beta_Z \geq K_{12}, \quad \beta_W \beta_Y \geq K_{24}, \quad \beta_W \beta_Z \geq K_{23} \end{aligned} \tag{5.48}$$

Depending to the standard setups [119] of the extrapolation method one finds [167] for the constants K_{12}, K_{24} and K_{23}:

$$\begin{aligned} K_{13} = 4\frac{L^2_{\|\cdot\|}(f(P,Z))}{\alpha_P \alpha_Z}, \quad L_{\|\cdot\|}(f(P,Z)) = \lambda_{\max}(V^\top V) \\ K_{23} = 4\frac{L^2_{\|\cdot\|}(f(W,Z))}{\alpha_W \alpha_Z}, \quad L_{\|\cdot\|}(f(W,Z)) = r \\ K_{24} = 4\frac{L^2_{\|\cdot\|}(f(W,Y))}{\alpha_W \alpha_Y}, \quad L_{\|\cdot\|}(f(W,Y)) = r \\ \alpha_P = \tfrac{1}{2m^2}, \quad \alpha_P = 0.5, \quad \alpha_Y = 0.5 \end{aligned}$$

The first two constraints are always active, since the scaling parameter have to be positive. Consequently we distinguish two cases: When all constraints are active and the solution of (5.48) is given by

$$\begin{aligned} \beta_W^2 = \frac{K_{23}(K_{23}d_Z + K_{24}d_Y)}{K_{13}d_P}, \quad \beta_W^2 = \frac{K_{23}(K_{23}d_Z + K_{24}d_Y)}{K_{13}d_P} \\ \beta_P = \frac{K_{13}}{d_Z}, \quad \beta_Y = \frac{K_{24}}{d_W} \end{aligned}$$

But if $\beta_W \beta_Z \geq K_{23}$ it holds:

$$\beta_P^2 = \frac{K_{13}d_Z}{d_P}, \quad \beta_W^2 = \frac{K_{24}d_Y}{W}, \quad \beta_Z^2 = \frac{K_{13}d_P}{d_Z}, \quad \beta_Y^2 = \frac{K_{24}d_W}{d_Y}$$

Using these parameter values and initial guesses for the primal variables we come to the following algorithm:

Algorithm 7: Adaptive Dual Extrapolation Method in Spectahedron Setup

Data: m, ϵ, κ_{up}, κ_{down}
Result: $\widehat{P}, \widehat{W}, \widehat{Y}, \widehat{Z}$
Initialization:
Choose the primal approximate solution variables $P_{k=0} \in \mathcal{P}$, $W_{k=0} \in \mathcal{W}$, $Y_{k=0} \in \mathcal{Y}$, $Z_{k=0} \in \mathcal{Z}$ as the prox centers of the corresponding feasible sets of the objective. Let $k = 0$ be the iteration index.
Compute the scaling parameter $\beta = (\beta_P, \beta_w, \beta_Y, \beta_Z)$ according to (5.48).
Chose a stepsize α_0 and a regularization parameter r according to (5.35).
Consider the objective in (5.35) and denote with

$$s_{P_0} \stackrel{\text{def}}{=} \alpha_0 \nabla_P f(P_0, W_0, Y_0, Z_0) \quad s_{W_0} \stackrel{\text{def}}{=} \alpha_0 \nabla_W f(P_0, W_0, Y_0, Z_0)$$
$$s_{Y_0} \stackrel{\text{def}}{=} -\alpha_0 \nabla_Y f(P_0, W_0, Y_0, Z_0) \quad s_{Z_0} \stackrel{\text{def}}{=} -\alpha_0 \nabla_Z f(P_0, W_0, Y_0, Z_0)$$

Compute the gradient $S(P_0, W_0, Y_0, Z_0) = (s_{P_0}, s_{W_0}, s_{Y_0}, s_{Z_0})^T$.
According to (5.41), (5.43), (5.44) and (5.45) compute the prox-transform

$$(P_0^+, W_0^+, Y_0^+, Z_0^+) = T_\beta(S(P_0, W_0, Y_0, Z_0), (P_0, W_0, Y_0, Z_0)).$$

and set $P = P_0^+$, $W = W_0^+$, $Y = Y_0^+$, $Z = Z_0^+$. Moreover let s denote $s = -\alpha_1 S(P_0, W_0, Y_0, Z_0)$. Finally, for later use we set

$$\zeta = (P_0^+, W_0^+, Y_0^+, Z_0^+)[s - \beta \log(P_0^+, W_0^+, Y_0^+, Z_0^+)^T] - (P_0, W_0, Y_0, Z_0)s.$$

while *1* **do**
  Evaluate $s_k = S(P, W, Y, Z)$ and set $P_k = P$, $W_k = W$, $Y_k = Y$, $Z_k = Z$.
  while *1* **do**
    Compute the corresponding mirror points

$$(P^+, W^+, Y^+, Z^+) = T_\beta(-\alpha_k s_k, (P_k, W_k, Y_k, Z_k)).$$

    Evaluate $s_k^+ = S(P^+, W^+, Y^+, Z^+)$ and make the extrapolation step

$$s_k^{++} = s - \alpha_k s_k^+.$$

    Renew the mirror points $(P, W, Y, Z) = T_\beta(s_k^{++}, (P_0, W_0, Y_0, Z_0))$.
    To control the size of α_k, compute

$$\begin{aligned} \zeta_k &= (P, W, Y, Z)[s - \beta \log(P, W, Y, Z)^T] - (P_0, W_0, Y_0, Z_0)s_0 \\ \tau_k &= \alpha_k (P^+ - P_0^*, W^+ - W_0^*, Y_0^+ - Y^+, Z_0^* - Z^+) s_k^+. \end{aligned}$$

    if $\zeta_k \le \zeta - \tau_k$ **then**
      Update $s = s_k^{++}$, $\alpha_{k+1} = \kappa_{up}\alpha_k$, $\zeta = \zeta_k$ and terminate the inner while-loop.
    else
      $\alpha_{k+1} = \kappa_{down}\alpha_k$
    end
  end
  Set $c_k = (\sum_{i=1}^k \alpha_i)^{-1}$ and $t_k = \sum_{i=1}^k \tau_i$.
  if $c_k(\zeta_k + (d_P, d_W, d_U, d_Y)\beta^T + t_k) \le \epsilon$ **then**
    terminate the outer while-loop.
  end
  Set $k = k + 1$.
end
Averaging: Set $(\widehat{P}, \widehat{W}, \widehat{Y}, \widehat{Z}) = c_k \sum_{i=1}^k \alpha_i (P_i, W_i, Y_i, Z_i)$

The number of iterations required [167] in algorithm 7 to archive (5.28) is $1+\rfloor \frac{LD}{\alpha\delta} \lfloor$. In sum the complete "semidefinite programming"-approach has the form:

Algorithm 8: full procedure of unified approach to SNGCA

Data: $\{X_i\}_{i=1}^N$,L,m
Result: $\widehat{\mathcal{I}}$

Normalization: The data $(X_i)_{i=1}^N$ are re-centered. Let $\sigma = (\sigma_1, \dots \sigma_d)$ be the standard deviations of the components of X_i. Then $Y_i = \mathrm{diag}(\sigma^{-1})X_i$ denotes the componentwise empirically normalized data.

Main Procedure:

Directional Sampling: The components of of $\omega_l^{(k)}$ are randomly chosen from $\mathcal{U}_{[-1,1]}$. Then $\omega_l^{(k)}$ are normalized to unit length.

Linear Estimation Procedure:
for *l=1* **to** L **do**

$\widehat{\eta}_l^{(k)} = \frac{1}{N}\sum_{i=1}^N \nabla h_{\omega_l^{(k)}}(Y_i)$

$\widehat{\alpha}_l^{(k)} = \frac{1}{N}\sum_{i=1}^N Y_i h_{\omega_l^{(k)}}(Y_i)$

end

Solve the relaxed and regularized semi-definite optimization problem:

$$\arg\min_{(P,W)\in\mathcal{P}\times\mathcal{W}} \arg\max_{(Z,Y)\in\mathcal{Z}\times\mathcal{Y}} \left[\mathrm{Tr}[\widehat{V}^\top (I-P)\widehat{V} Z] + r\,\mathrm{Tr}[W(QZQ^\top - Y)] \right]$$

stated in (5.35) by the dual extrapolation method.

Dimension Reduction:
Choose the relevant columns of $\widehat{P}$ according to some criterion for NonGaussianity.

In this form the total numerical complexity of the unified approach to SNGCA is $\mathcal{O}(N^2 L)$ for the linear estimation procedure and $\mathcal{O}(L \log L)$ for optimization step. The numerical convergence of the algorithm is $\mathcal{O}(\delta^{-2} L_{\|\cdot\|}(f)^2 \log L)$.

5.4 Numerical Simulations

The aim of this section is to compare the different approaches to SNGCA with other statistical methods of dimension reduction. To this end we will discuss the well known test densities from section 4.5 in order to demonstrate the progress that is made with the "semidefinite programming"-approach to SNGCA. To this end by SNGCA(2) we refer to the latter and by SNGCA(1) to the "convex projection"-approach .

Let us compare SNGCA(2) with PP and SNGCA(1) using the test data sets from above with respect to the estimation error

$$\mathcal{E}(\widehat{\mathcal{I}}, \mathcal{I}) \stackrel{\mathrm{def}}{=} \|\Pi_{\widehat{\mathcal{I}}} - \Pi_{\mathcal{I}}\|_F^2. \tag{5.49}$$

where $\|\cdot\|_F$ is the Frobenius norm. Each simulation is repeated 100 times. All simulations are done with the index 'tanh'. In the experiments the error measure $\mathcal{E}(\widehat{\mathcal{I}}, \mathcal{I})$ is used only to determine the final estimation error. All simulations other than whose with respect to model (C) are computed with a componentwise pre-whitening.

Since the optimizer used in PP tends to trap in local a minimum in each of the 100 simulations, PP is 10 times restarted with random starting points. The best result with respect to (5.49) is reported as the result of each PP-simulation. In all simulations the number of non-Gaussian dimensions is apriori given. In the next figure 5.4 we present boxplots of the error (5.49) of the methods PP, SNGCA(1) and SNGCA(2).

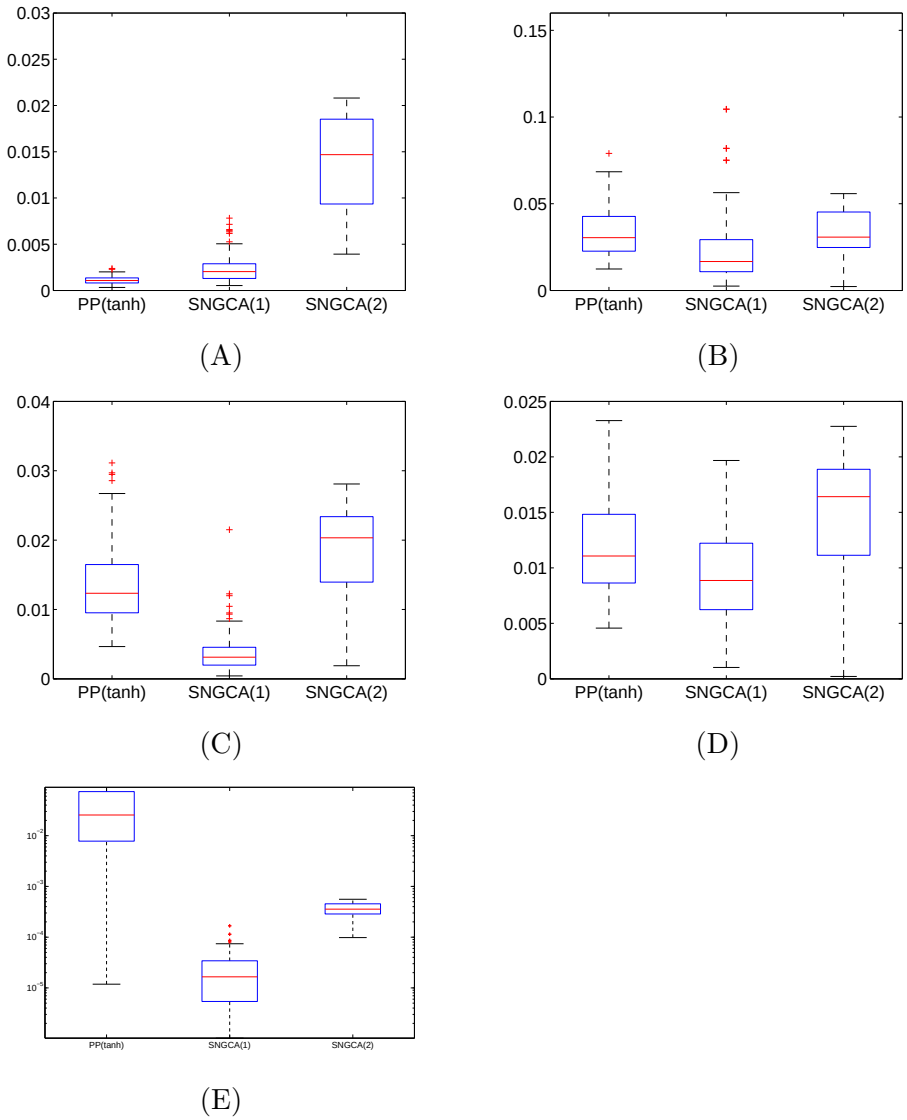


Figure 5.4: Performance comparison in 10 dimensions of PP and SNGCA(1) versus SNGCA(2) (with respect to the error criterion $\mathcal{E}(\widehat{\mathcal{I}},\mathcal{I})$) using the index 'tanh'. The doted line denotes the mean, the solid lines the variance of (5.49).

Concerning the results of SNGCA(2) we observe a slightly inferior performance compared to SNGCA(1). This is not surprisingly since gradient-type methods are well known to archive fast progress during the first iterations before they start jamming [93]. Since the size of the directional sampling is constant the lower variance in the case of model (B) and (E) is due to the fact, that the corresponding non-Gaussian components show a bigger degree of deviation from normality and hence are much easier to detect for SNGCA(2).

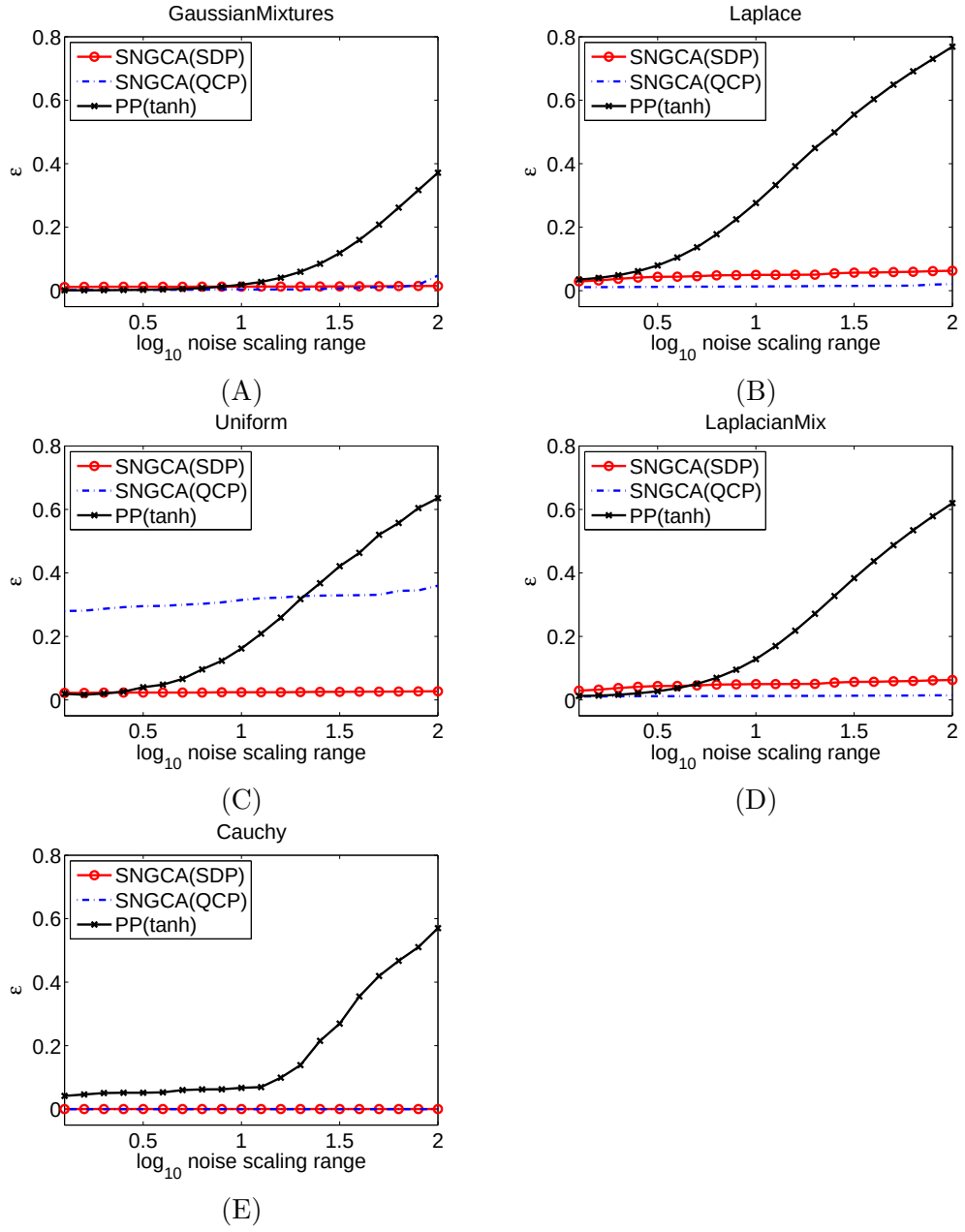


Figure 5.5: Results with respect to the test densities from section 4.5 in terms of $\mathcal{E}(\widehat{\mathcal{I}}, \mathcal{I})$ with deviations of Gaussian components following a geometrical progression on $[10^{-r}, 10^{r}]$ where r is the parameter on the abscissa) .

Now let us switch to the question of robustness of the estimation procedure with respect to a bad conditioning of the covariance matrix Σ of the data. In figure 5.5 we consider the same test data sets as above. The non-Gaussian coordinates always have variance unity, but the standard deviation of the 8 Gaussian dimensions now follows the geometrical progression $10^{-r}, 10^{-r+2r/7}, \ldots, 10^{r}$ where $r = 1, \ldots, 8$. Again we apply a componentwise whitening procedure to the data from the models (A), (B), (D), (E). We observe that the condition of the covariance matrix heavily influences the estimation error for the methods PP(tanh) but not for SNGCA(1) and SNGCA(2). The variants of SNGCA are independent of differences in the noise variance along different direction in most cases. Moreover SNGCA(2) gives a better result in case of the uniform components to the data density

than SNGCA(1).

The next figure 5.6 compares the behavior of SNGCA compared with PP and SNGCA as the number of standard and homogeneous Gaussian dimensions increases.

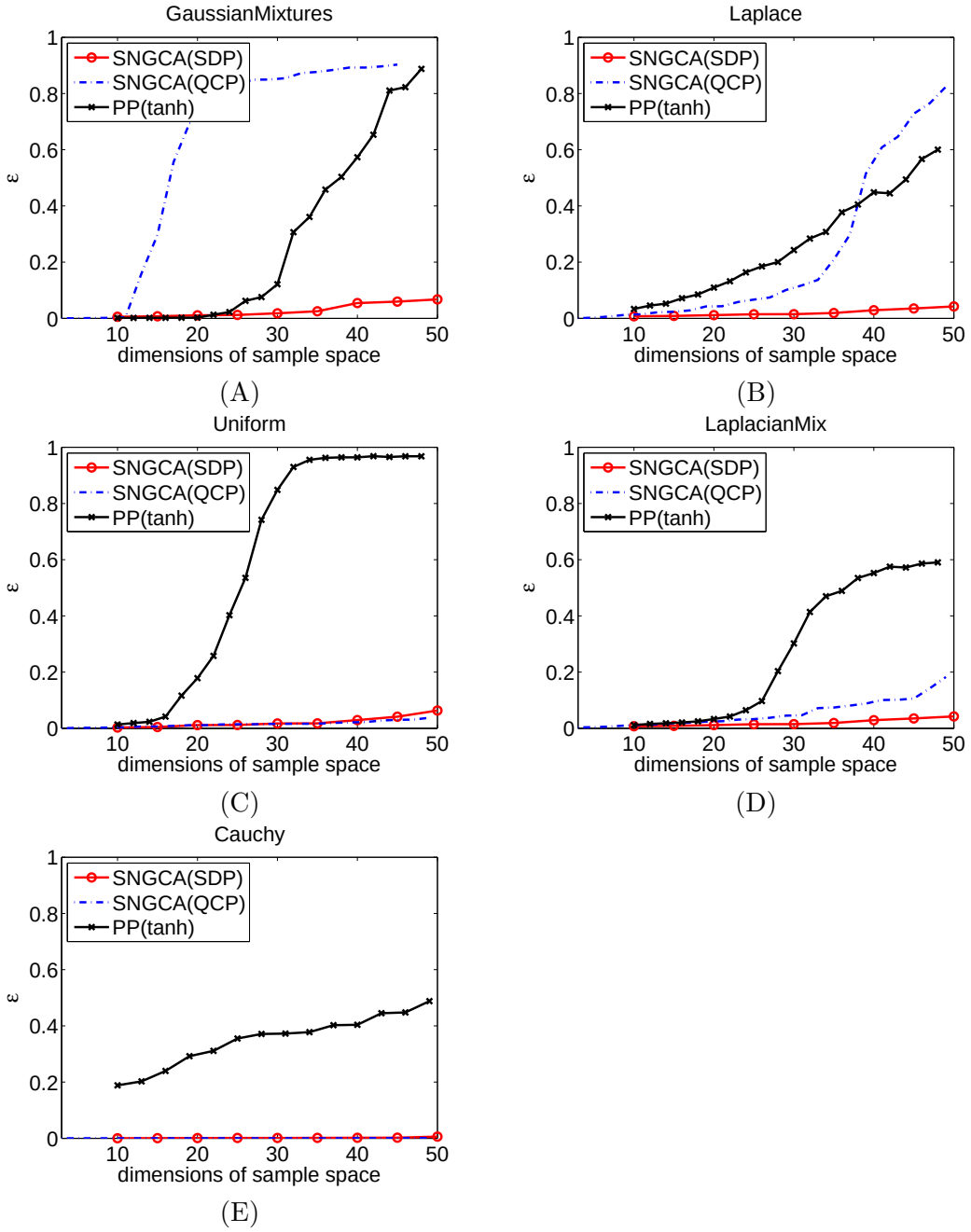


Figure 5.6: Results with respect to the test densities from section 4.5 in terms of $\mathcal{E}(\widehat{\mathcal{I}}, \mathcal{I})$ with increasing number of gaussian components.

As described above we use the test models with 2-dimensional non-Gaussian components with variance unity. We plot the mean of errors $\mathcal{E}(\widehat{\mathcal{I}}, \mathcal{I})$ over 100 simulations with respect to the test models (A) to (E). Again concerning the mean of errors $\mathcal{E}(\widehat{\mathcal{I}}, \mathcal{I})$ of PP we find a transition in the error criterion to a failure mode for the test models (A), (C) and (D) between $d = 30$ and $d = 40$ and between $d = 20$ and $d = 30$ respectively. For the test models (B), and (E) we found a relative continuous increase in $\mathcal{E}(\widehat{\mathcal{I}}, \mathcal{I})$. For the

methods SNGCA(1) the methods fails in case of model (A) after 10 dimensions and in case of model (B) after 35 dimensions. However SNGCA(2) is successful for all kinds of deviations from normality up to 50 dimensions. Moreover there is only a slight increase of the error function towards higher dimensions.

The results of the numerical simulation indicate that the new "semidefinite programming"-approach to SNGCA archives to exploit the information obtained from the test densities much better that other approaches to NonGaussian Component Analysis. Moreover the new algorithm gives promising results in comparison to other linear projective feature extraction methods.

In the final chapter we will apply the "semidefinite programming"-approach to SNGCA as a preprocessing step to the metastability analysis of large biomolecules.

Chapter 6

A Geometric Approach to Metastability Analysis

Introduction: In many cases one can observe that biological active molecules $\mathfrak{M}$ exhibit different large geometric structures $\mathfrak{G}$ on a length scale much larger than the diameter of the atoms. However the existence of $\mathfrak{G}$ do not imply a set of fixed positions of the atoms of $\mathfrak{M}$. Rather one can observe a local flexibility in the bonds between the atoms such that they exhibit local random vibrations around a stable geometric mean position. If there are more than only one large scale structures $\mathfrak{G}$ with life times much larger that the time scales of the local vibrations of the atoms, then we will call $\mathfrak{G}$ of $\mathfrak{M}$ metastable [197]. Therefore the term *conformation* has dynamical provenance since we refer to a set of geometrical structures as variations of the same global configuration of $\mathfrak{M}$. This configuration can be identified as connected subsets of state space. With $\mathfrak{G}$ we do not refer to a local minimum of some energy functional of $\mathfrak{M}$ or the corresponding thermodynamical ensemble used in molecular dynamics (MD) [5] to describe $\mathfrak{M}$. Hence conformations represent all molecules belonging the same large scale geometric structure. To find metastable conformations is thus an aggregation problem with respect to molecular states into conformational states that can be considered as clustering problem. We illustrate conformational changes in figure 6.1.

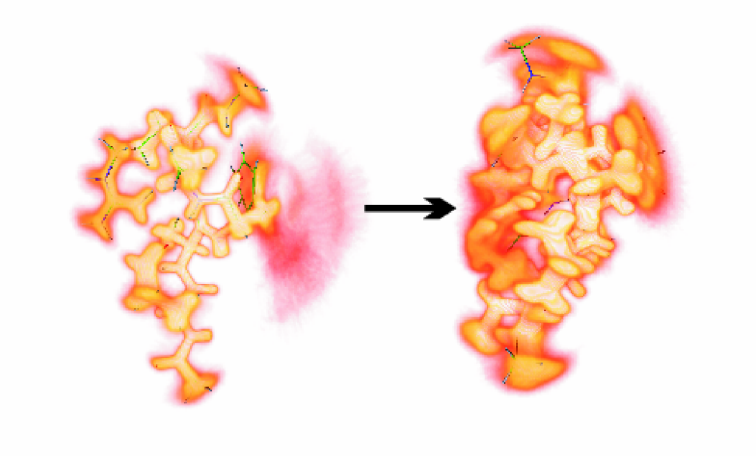


Figure 6.1: Changes of geometric large scale configurations of a biological active molecule with life times much longer than the time scale of the internal interactions between the atoms and the random perturbations of the molecules with the solvent visualized by AMIRA [214].

The observation from above motivates the separation of different time scales in the description of the dynamics of $\mathfrak{M}$ as well as the decomposition of the corresponding phase space into domains, that can not be defined in terms of some potential energy. Concerning the multiscales the shorter time scales typically range from femto- to picoseconds, while the larger time scale can be found on nanoseconds to seconds [64]. Hence transitions between different conformations of a molecule are rare statistical events on the macroscopic length scale compared to the fluctuations within each conformation on the microscopic scale. With respect to the phase space decomposition a paradigmatic concept to visualize this is the so called Ramachandran plot [182]. In the next exposure we mainly follow [155].

Dieder-Angles: The general motive to explain the folding process of $\mathfrak{M}$ goes back to the fact that the geometrical large scale structure of $\mathfrak{M}$ is essential for its biological function, i.e. the interaction of $\mathfrak{M}$ with the physiological environment [71]. As examples for biomolecules we primarily consider in this thesis peptides built up from so-called proteinogenic amino acids.

Recall that peptides are built up from an almost periodic sequence of proteinogenic amino acids, that itself consist in an amino group linked to a carboxyl group via an α-carbon C_α. C_α has a broad variety of side groups ranging from a H-atom over a CH_3-group to other more complex residuals determining the type of the amino acid. Frequently one amino- and one carboxyl-group bound electronically such that a so-called *peptide bond* emerge. Remarkably a peptide bond is planar an hence stable against rotations. If we call the repeated $C_\alpha - C - N$-chain connecting the amino acids in a peptide *the backbone* of a peptide, we can conclude that the only degrees of freedom for each adjacent pair of amino acids along the backbone of $\mathfrak{M}$ are the Φ-angle, is identified by the dihedral angle $C - N - C_\alpha - C$, and the Ψ-angle located along $N - C_\alpha - C - N$. For an analysis the macroscopic flipping process between different conformations this results in a description of conformational changes in terms of a sequence of (Φ, Ψ)-pairs along the backbone of $\mathfrak{M}$ [71]. Figure 6.2 illustrates the dieder-angles (Φ, Ψ) of the backbone.

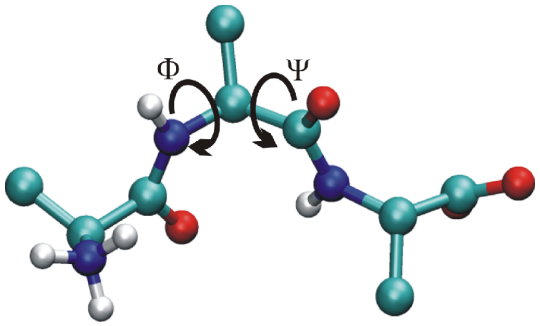


Figure 6.2: Since some of the bonds in $\mathfrak{M}$ are a planar and thus stiff peptid bond, the rotational degrees of freedom (Φ, Ψ) allow to describe the macroscopic folding process of a biomolecule as a change of the geometric configuration of the backbone of $\mathfrak{M}$.

However due to steric interactions and the emergence of H-bond bridging between different peptide bond units, (Φ, Ψ) are not completely free for a given $\mathfrak{M}$. Consequently there are "allowed" and "forbidden" domains for $\mathfrak{M}$ in the Ramachandran plane [183], which are similar for most of the amino acids. In other words the Ramachandran plane contain the most favorable energetic domains corresponding to the secondary structures of $\mathfrak{M}$.

In the next section we summarize the state-of-the-art explanation of the origination of conformational changes and of the dynamics on the smaller length scales. The latter one involves the introduction of concepts central to MD-simulations.

6.1 Conformational Dynamics of Biomolecular Systems

Molecular Dynamics: Classical MD-models describe $\mathfrak{M}$ by means of coupled ordinary differential equations for the N atoms [5]. The well known equations of motion have the following general form:

$$\begin{aligned} \frac{d}{dt}q(t) &= M^{-1}p(t), \\ \frac{d}{dt}p(t) &= -\nabla V(q(t)), \end{aligned} \tag{6.1}$$

where $(q,p) \in \mathbb{R}^{3N} \times \mathbb{R}^{3N}$ and (q,p) are the atomic positions and momenta, respectively. $M \in \mathbb{R}^{3N\times 3N}$ denotes the (assumed diagonal) mass matrix and $V : \mathbb{R}^d \to \mathbb{R}$, $q \mapsto V(q)$ is a differentiable potential energy function describing all the interactions between the atoms. The state space of the system $\mathfrak{M}$ is $\Gamma \subset \mathbb{R}^{6N}$. The Hamiltonian function

$$H(q,p) = \frac{1}{2}\, p^T M^{-1} p + V(q), \tag{6.2}$$

denotes the internal energy of the system in state (q,p) that is the total energy of the system and preserved by the dynamics in energetically closed systems.

However if the perspective on conformations in the above section is correct, a symplectic [136; 137] time integrator may take infinite long time to propagate a single molecule to every state in Γ that is allowed with respect to the physical constraints as constant volume, constant temperature and constant number of particles. Fortunately the hypothesis of ergodicity [150], that states an equivalence of time and particle averaging, allows to consider sets of copies of a molecules instead of single molecules, where to each copy a unique physical microstate is assigned. The macroscopic observable, representing the macroscopic properties, are obtained by averaging over the complete ensemble of systems. Although due to energy conservation, a single solution of the Hamiltonian system (6.1) can never be ergodic (wrt. the canonical measure), the equivalence in the thermodynamical limit is due to the fact that using Hamiltonian systems with randomized momenta allows to sample the state space Γ appropriately with respect to some prescribed statistical distribution. Consequently we can restrict ourselves to the consideration of the time evolution of ensembles, represented by a probability distribution of initial states.

To this end let Φ^t denote the flow in Γ associated with the Hamiltonian, such that the solution $x_t = (q_t, p_t)$ for the initial value $x_0 = (q_0, p_0)$ is given by $x_t = \Phi^t x_0$. Now consider an ensemble at constant temperature T with a constant number of molecules N and constant volume. Then the time evolution of the initial canonical ensemble is governed by the dynamics of the single molecules [196]. Then due to Liouville's Theorem implying conservation of probability, we get in terms of the ensemble:

$$\rho(x,t) = (\rho_0 \circ \Phi^{-t})(x)\,, \quad \text{with} \quad \rho_0 = \rho(\cdot, 0)\,. \tag{6.3}$$

The prominent case of a canonical density is the Boltzmann density

$$\rho_{\text{can}}(x) \stackrel{\text{def}}{=} \frac{1}{Z}\exp(-\beta H(x)) \tag{6.4}$$

$$Z \stackrel{\text{def}}{=} \int_\Gamma \exp(-\beta H(x))\mathrm{d}x\,, \tag{6.5}$$

where $\beta \stackrel{\text{def}}{=} 1/k_B T$ is the inverse temperature and k_B the Boltzmann constant. Recall that the particle dynamics and the ensemble dynamics are equivalent [49]. The time evolution of a probability density is governed by the Fokker-Planck equation.

The Langevin Equation: Unfortunately there are no general conditions to ensure the quality of the sampling of Γ, while the ergodicity of deterministic Hamiltonian systems is preserved [109; 94] in the simulations. Hence Langevin dynamics, arising from a stochastic modelling of $\mathfrak{M}$, is a promising alternative to describe the dynamics of $\mathfrak{M}$ under given physical constraints from above, since it also realizes a Boltzmann density [39].

A Langevin system [186] includes stochastic interaction with the environment in Γ:

$$\frac{d}{dt}q = M^{-1}p \qquad \frac{d}{dt}p = -\lambda_q V(q) - \gamma M^{-1}p + \sigma\dot{W}_t \tag{6.6}$$

It can be regarded as an open mechanical system where the stochastic term mimics the stimulation from the environment. In (6.6) $\gamma > 0$ denotes a friction constant and $F_{ext} = \sigma\dot{W}_t$ is the external forcing given by a $3N$-dimensional Brownian motion W_t. The noise σ models the influence of a surrounding heat bath and the friction $\dot{W}_t$ is chosen such as to counterbalance the energy fluctuations due to the noise. Recall that the Langevin dynamics is ergodic with respect to the invariant equilibrium (probability) measure

$$\mathrm{d}\nu(q,p) = \frac{1}{Z}\exp(-\beta H(q,p))\mathrm{d}q\mathrm{d}p,$$

if the fluctuation-dissipation relation $\beta = 2\gamma\sigma^{-2}$ holds [14].

Since the Langevin system is given by a linear stochastic equation (SDE), (6.6) has a unique and continuous solution, if the Brownian motion has finite second moments [9]. Furthermore suppose that a linear SDE

$$\mathrm{d}z(t) = A(t)z(t)\mathrm{d}t + a(t)\mathrm{d}t + \Sigma(t)\mathrm{d}W(t), \tag{6.7}$$

with drift coefficient $A:\ \mathbb{R}^d\times[t_0,T]\to\mathbb{R}^d$ and diffusion coefficient $\Sigma:\ \mathbb{R}^d\times[t_0,T]\to\mathbb{R}^{d\times m}$ has, for every initial value $z(t_0)=c$ independent of the increments $W(t)-W(t_0)$, $t_0\le t$, a unique solution in $[t_0,T]$ provided that $A(t), a(t), \Sigma(t)$ are measurable and bounded on $[t_0,T]$. Then the solution is a Markov process

$$z(t) = \Psi(t)\left(c + \int_{t_0}^t \Psi(s)^{-1}a(s)\mathrm{d}s + \int_{t_0}^t \Psi(s)^{-1}\Sigma(s)\mathrm{d}W(s)\right),$$

where $\Psi(t)$ is the solution of $\frac{d}{dt}\Psi(t) = A(t)\Psi(t)$ with $\Psi(t_0) = I$ [174]. Recall that a (time-continuous) stochastic process $\{X(t), 0\le t\}$ with state space Γ is called a Markov process [29] if for any $0 < t_0 < \cdots < t_< \le t_{k+1}$ and any $j, i_1, \cdots, i_k \in \Gamma$ it holds

$$I\!P(X(t_{k+1}) = j | X(t_k) = i_k, \cdots, X(t_1) = i_1) = I\!P(X(t_{k+1}) = j | X(t_k) = i_k). \tag{6.8}$$

A time-continuous Markov process is called homogeneous if the right hand side of (6.8) only depends on the increments $t_{k+1} - t_k$. A Markov process with discrete state space is called a Markov jump process. More technical details about the generation of the time series and the analysis of Langevin systems can be found in [201; 108; 198].

Consequently we finally arrive at the following picture of the dynamics of the conformational changes of $\mathfrak{M}$.

Conformational Changes: Concerning the analysis of large molecular systems modelled as a Langevin- or Smoluchovski-system, the free energy landscape of such thermodynamical systems decompose into particulary deep wells each containing many local minima [177; 72]. These wells are typically separated by relatively large barriers from each other measured on the scale of the thermal energy k_bT. Deep wells represent different almost invariant geometrical large scale structures [197]. In the configuration space the conformations are represented by clusters of stationary distributed points generated by a homogenous and time-reversible Markovian process. Since the shape of the free energy landscape and hence the canonical density are unknown, the essential degrees of freedom, in which the rare conformational changes occur, are sought. Due to the existence of "forbidden" domains for $\mathfrak{M}$ in the Ramachandran plane, we expect the interactions internal to $\mathfrak{M}$ to be confined to a tiny fraction of the full high-dimensional configuration space [6]. In particular the macroscopic dynamics is assumed to be a Markov jump process, hopping between the metastable sets of the state space while the microscopic dynamics within these sets mixes on much shorter time scales [201].

6.2 Hidden Markov Models with Gaussian Densities

The most important outcomes of the last sections are the existence of different time scales in the dynamics of observed stochastic process $\{X(t), 0 \leq t\}$ and that $\{X(t), 0 \leq t\}$ is a discrete and reversible Markovian process associated with $\mathfrak{M}$. In order to develop an almost geometrical approach to the metastability analyis of highdimensional biomolecules both results can be represented in the well known technique of parametric Hidden-Markov Models (HMM) [181].

Hidden-Markov Models: In comparison to other approaches to metastability (c.f. [234; 52]) HMMs abstain from a decomposition of the observation space. Instead it favors to fit a hidden and discrete Markovian stochastic process $\{Y_t, 0 \leq t\}$ switching between different conformations to the observable data by means of optimizing a high dimensional likelihood function. In particular the discrete observations $\{X_t, 0 \leq t\}$ depend on the current conformation, that correspond to a special parameterization of an assumed local model. The fitting itself by means of maximum likelihood (ML) estimation via expectation-maximization (EM) algorithm [51] and Viterbi-algorithm [231] is a little bit expensive and takes $\mathcal{O}(N(M^2 + Md^2) + Md^3)$ operations, where M is the number of model parameter [181].

Formally a HMM is given by $\mathcal{H} = \{\Gamma, \mathbb{R}^d, \mathbf{P}, \rho, I\!P_0\}$ where

$$I\!P_0 = \big(I\!P(Y_{t=1} = 1), \ldots, I\!P(Y_{t=1} = n)\big)$$

is the initial state distribution and

$$\mathbf{P} = \Big(I\!P(Y_{t+1} = j | Y_t = i)\Big)_{ij} \in \mathbb{R}^n \times \mathbb{R}^n \tag{6.9}$$

a transition matrix of the hidden Markov process, $\Gamma = \{1, 2, \ldots, j, \ldots, n\}$ is the finite state space of the hidden model with output densities $\rho = (\rho_1, \rho_2, \ldots, \rho_n)$. Algorithmically the computation of a HMM means parameter estimation of the output densities in the ML-sense in high dimensional parameter spaces. In this thesis we focus on HMMs with Gaussian output [143].

Parameter Estimation: In order to identify algorithmically the parameters of a HMM a nonlinear global optimization problem in small dimensions must be solved [70]: In the k^{th} step of the optimization process an EM-algorithm based on dynamical programming techniques [21] can determine the optimal parameters θ^* via maximizing the expectation

$$I\!E_{\theta,\theta_k} = I\!E[\log p(X, j|\theta)|X, \theta_k]$$

where $p(X, j|\theta)$ is the density of $I\!P(X, j|\theta)$ wrt. the hidden states. This can be rewritten as a sum over all hidden sequences [143]:

$$I\!E_{\theta,\theta_k} = \sum_{j=1}^{n} p(X, j|\theta_k) \log p(X, j|\theta)$$

In general the expectation-step of the EM-algorithm evaluates the expectation based on the given parameter estimate θ_k, while the maximization-step determines the refined model parameter set by the solution of

$$\theta_{k+1} = \arg\max_\theta I\!E_{\theta,\theta_k}.$$

Moreover if the complex system is ergodic, the estimators corresponding to a global maximum converge weakly to their estimated functions [99]. However the EM-algorithm is if only a local optimization method such that one has to provide an appropriate initial guess for the model parameters.

Next we have to determine the optimal sequence j^* of hidden metastable states for optimal parameters by means of the well-known Viterbi algorithm, which exploits again dynamic programming techniques to resolve in a recursive manner the optimization problem

$$j^* = \arg\max_j p(X, j|\theta^*)$$

The obtained optimal sequence j^* is called the *Viterbi path*. The algorithmic complexity of the Viterbi algorithm is $\mathcal{O}(n^2 N)$ [181].

Initialization of EM-Algorithm: In numerical simulations it turns out that the convergence of the EM-algorithm is dominated by the initial Viterbi path. Since we only aim to find a HMM model for the reduced data, we can use a non-parametric clustering algorithm to generate an appropriate guess for the initial Viterbi path. In the thesis we use an adaptive mean-shift clustering algorithm [35; 206]. The mean shift algorithm does not require prior knowledge of the number of clusters and does not constrain appropriate geometric shapes of the clusters. The basic idea here is to use radially symmetric kernels $k(x)$ for estimating the multivariate kernel density with window radius h according to

$$\rho(x) = \frac{\mathcal{O}(1)}{Nh^d} \sum_{i=1}^{N} k(\frac{X - X_i}{h})$$

The modes of the density function are located by means of $\nabla\rho(x) = 0$ with

$$\nabla\rho(x)) = \frac{2\mathcal{O}(1)}{Nh^{d+2}}\left[\sum_{i=1}^{N} -k'\Big(\|\frac{X - X_i}{h}\|^2\Big)\right]\left[\frac{\sum_{i=1}^{N} -x_i k'\Big(\|\frac{X-X_i}{h}\|^2\Big)}{\sum_{i=1}^{N} -k'\Big(\|\frac{X-X_i}{h}\|^2\Big)} - x\right].$$

The second brackets contain the mean shift. The mean shift vector always points toward the direction of the maximum increase in the density. The mean shift procedure is obtained by successive computation of the mean shift vector and then translation of the window by

$$x^{(k+1)} = x^{(k)} + m_h(x^{(k)})$$

The mean shift mode finding process is illustrated in figure 6.3:

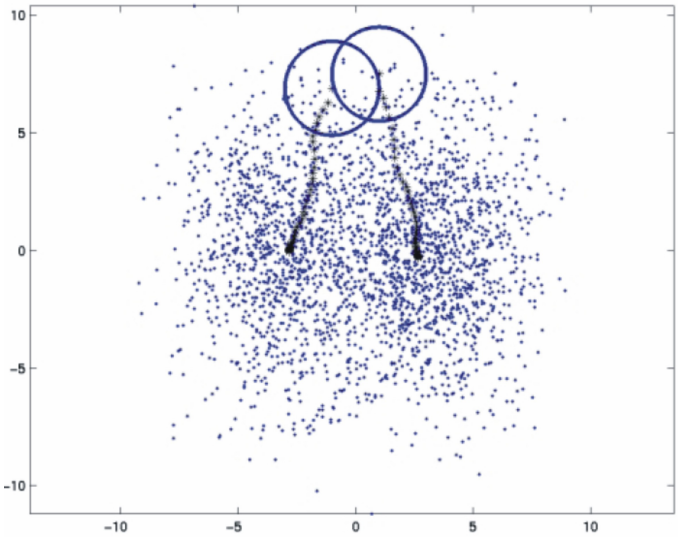


Figure 6.3: Illustration of non-parametric mean-shift mode finding process for componentwise normalized data in $\mathbb{R}^2$. The blue circles are the windows of the algorithm. The black stars are the centers od the windows.

In numerical simulations it turns out, that for componentwise normalized data the bandwidth h can be set to $h = 0.6$ for all data sets. A wrong choice of h would change the convergence if the clustering. In particular it leads to a wrong number of clusters.

Viterbi clustering: However the parameter fitting step requires the specification of the number M of hidden states, which, whenever the hidden states should be metastable states, is in general not known apriori. One practical way to overcome this problem is to assume a large number of hidden states in the HMM-step of the analysis. Then one performs the parameter fitting as described and conduct a further aggregation of the estimated transition matrix based on the resulting Viterbi path. This can be done by means of the set-oriented Perron Cluster Cluster Analysis (PCCA) (c.f. [107] and the references therein): A normalization of the transition matrix $\mathbf{P}$ introduced in (6.9) yields a stochastic (symmetric) matrix whose spectral properties can be used to identify the conformations as metastable sets of states of the obtained (reversible, homogenous and aperiodic) Markov chain. Note that $\mathbf{P}$ has to be estimated from the (reduced) data as proposed in [98]. Then the number of metastable states is given by the number of eigenvalues close to 1 that are separated from the rest of the (perturbed) spectrum by a small gap. This method is due to the fact that aggregating the hidden states into

conformation states also allows to aggregate the clusters into conformations. However the numerical complexity of this PCCA-step in one dimension is $\mathcal{O}(n^3)$ where n is the number of hidden states of the dynamical system.

6.3 Reduced Conformational Dynamics

In this section we explain the geometrical approach to metastability analysis and conduct some numerical simulations. Up to now we have described all the algorithmic tools additional to the SDP-approach of SNGCA that will be used to analyze the essential dynamics of high dimensional biomolecules. It is obvious that all tools are geometric except PCCA required to determine the number of (hidden) conformations realized in a given data set. Moreover the only apriori knowledge about the data used in our tools is the intrinsic dimension m of the SNGCA target space. It remains to try out the tools for extracting a multimodal component from a high dimensional data density and to compare them with other methods.

6.3.1 Clustering of Highdimensional Data

The popularity of the geometric clustering methods like e.g. K-means [148] is due to its low computational complexity of $\mathcal{O}(NkKd)$, where K is the number of clusters and k the number of K-means-iterations. While e.g. K-means often produces reasonable results in small dimensions, in high dimensions it has difficulties with outliers: Points which do not belong to any of the K clusters can move the estimated means away from the densest regions. Moreover K-means fails when high dimensional data deviates from properties as equal size, equal density, globular shape and very low noise. Other methods like agglomerative hierarchical clustering schemes, which are often thought to be superior to K-means for low-dimensional data, have similar problems. For example, the single linkage approaches are very vulnerable to noise and differences in density and group average. Complete linkage has trouble with differing densities and, unlike single linkage, cannot handle clusters of different shapes and sizes [147]. As expected from section 2.1 one source of these problems is the use of the Euclidean distance measure. There are many proposals to cope with the fact that the Euclidean distance does not work well in high dimensions. Some clustering algorithms use distance or similarity measures that work better for high dimensional data e.g., the cosine measure. Other approaches use iterative mode estimation techniques in the underlying feature space e.g. mean shift based clustering [77]. However to our current knowledge the question for a reliable clustering method for high dimensional data sets is still open.

Detecting Multimodality: Different linear projection methods for feature extraction indicate different approaches to solve the task of clustering high dimensional data. In the case of Principle Component Analysis there is a joint solution of the problem of dimension reduction and the cluster detection problem: A gap in the series of eigenvalues of $\mathbb{E}_N[XX^\top]$ indicates that the eigenvectors corresponding to the vanishing eigenvalues represent a basis for the nullspace of $\mathbb{E}_N[XX^\top]$. Hence projecting the data on the set of m eigenvectors of $\mathbb{E}_N[XX^\top]$ corresponding to the non-vanishing eigenvalues implies a solution to the problem of detecting multimodality only if the variance in the clustered part of the data is much higher than in the rest of the data. Obviously this is a limiting assumption using PCA as a pre-processing step for high dimensional clustering.

Compared to PCA using ICA also implies a joint solution of the problem of dimension reduction and the cluster detection problem. However due to the successive extraction

of non-Gaussian components it is supposed that the part of the given data distribution which shows most deviance from normality coincides with the clustered part of the data. Hence the ICA-solution to the problem of detecting multimodality does not depend on a special difference in the magnitude of the second moments of Gaussian noise and informative data components, but equates interestingness with NonGaussianity. However, in cluster analysis we are more interested in a departure from unimodality. Consequently the ICA-assumption may be too restrictive for some applications, since e.g. the entropy of the mixture of Gaussians is much closer to the entropy of a standard normal distribution than the entropy of the unimodal Pareto distribution [4].

Obviously for one-dimensional data any departure from an unimodal distribution implies departure from normality. But the converse is not true, since identifying the non-Gaussian components of $\rho(x)$ is a solution to a more general task and thus may be too rough to identify a cluster structure in real world examples with sufficient accuracy. But the dynamics of a polypeptide taken as an example for a biomelcular system revealing metastablity is typically modelled as an Markovian process [154], that implies a spatial separation of the metastable states in the high dimensional state space. Consequently a statistical measure sensitive to multimodality with appropriate properties as scale invariance and robustness to noise or outliers is sought. However it is well known that the cumulant indices are highly sensitive to outliers and favor heavy tails of the distributions [73]. Consequently higher moments of a distribution as skewness or kurtosis or other tests on non-normality will not be very helpful to determine a multimodal structure in the reduced data contained in the estimated target space $\mathcal{I}$.

The dip index: The are many non-parametric statistical tests for multimodality e.g. the dip test [85], the excess mass test [158], and Silvermans bootstrap test [209]. The dip test and the excess mass tests are equivalent in the sense that the excess mass statistic is exactly twice the dip statistic [33] in one dimension. Figure 6.4 illustrates the geometrical meaning of the dip statistics.

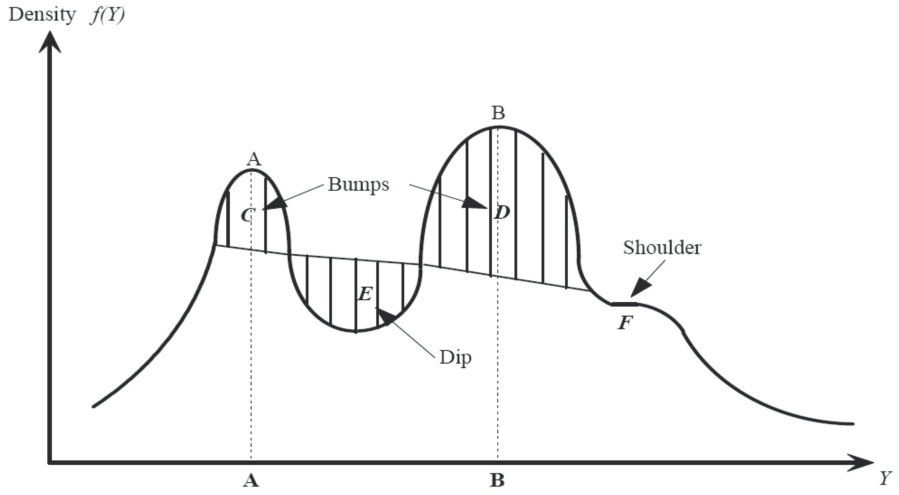


Figure 6.4: Illustration of the important points of a density: modes, bumps, dip and shoulder. Points A and B are modes, shaded areas C and D are bumps, area E is a dip and F is a shoulder point.

The dip test is extended to the case of multimodalily in [84]. However, compared to the Silverman test, the dip test is more conservative, since the dip test requires a greater mass

to signify a mode. Hence, it will be more likely to accept the unimodality using the dip test. For the numerical experiments to be described here we use the so called non-parametric dip-test [83; 85] as an index in order to identify successively a linear subspace $\mathcal{I}_M \subseteq \mathcal{I}$. This choice is due to the most important assumption in the analysis of metastability, the Markov assumption. Hence if SNGCA successfully detects the multimodal components we are interested in a target space $\mathcal{I}_M$ of "best cluster separation".

The dip test is the maximal difference between the empirical kernel distribution function and the unimodal distribution function that minimizes this difference. That is, the dip is the distance between the tightest fitting unimodal distribution function and the empirical distribution

$$F_N(x) \stackrel{\text{def}}{=} \frac{1}{N} \sum_{i=1}^{N} \mathbf{1}_{(X_i \leq x)} \tag{6.10}$$

with respect to the supremum norm. Moreover the dip-statistic is scale-invariant and robust against outliers [34]. In order to see that the dip index is a measure of departure from unimodality define the *Kolmogorov distance*

$$dist_{\mathcal{K}}(F, G) \stackrel{\text{def}}{=} \sup_{x \in \mathbb{R}} |F(x) - G(x)| \tag{6.11}$$

where F and G are bounded functions. Then we can define the one dimensional dip-statistics as follows:

Definition 1. *(dip-statistic)*
Let $\mathcal{U} \subset \mathfrak{D}(\mathbb{R})$ be the subset of unimodal distribution functions with respect to the class $\mathfrak{D}(\mathbb{R})$ of all distribution functions from $\mathbb{R}$. Then the dip of a distribution function F is given by $D(F) \stackrel{\text{def}}{=} \inf_{G \in \mathcal{U}} dist_{\mathcal{K}}(F, G)$.

Note that

$$D(F_1) \leq D(F_2) + \inf_{G \in \mathcal{U}} dist_{\mathcal{K}}(F, G)$$

and

$$D(F_1) \begin{cases} = 0 & \text{if } F \in \mathcal{U} \\ \geq 0 & \text{if } F \notin \mathcal{U} \,. \end{cases}.$$

Hence $D(F)$ is a measure of departure from unimodality.

In order to explain how the dip statistic works on $\mathbb{R}$, observe that for calculating algorithmically the tightest fitting unimodal distribution and the dip statistic, one can exploit the fact that an unimodal distribution on the real line is concave over the interval $]-\infty, a]$ and convex over $[a, \infty[$, where a is the position of the mode. Then denote by $G(x)$ the *greatest convex minorant* (GCM) of the distribution F on $]-\infty, a]$ the supremum over all real-valued convex functions $G(x)$ from $]-\infty, a]$, satisfying

$$G(x) \leq F(x), \ \forall \, x \in]-\infty, a]$$

The *lowest concave majorant* $L(x)$ (LCM) of the distribution F in $[a, \infty[$ is the infimum over all concave functions $L(x)$ satisfying

$$L(x) \geq F(x), \ \forall \, x \in [a, \infty[$$

Now the basic steps of the algorithm are instructive: Suppose the sample from the given distribution is listed in ascending order. Then:

- Set $x_L = X_1$, $x_U = X_N$ and $D(F_N) = 0$.
- Compute the GCM G and the LCM L with respect to F_N in the interval $[x_L, x_U]$. Let $\{g_i\}$ and $\{l_j\}$ denote the contact points of G and L with F_N respectively.
- If $d = dist_{\mathcal{K}}(G(g_i), L(g_i)) \geq dist_{\mathcal{K}}(G(l_i), L(l_i))$ for a given index i and the supremum occurs in g_j satisfying $l_j \leq g_i \leq l_{j+1}$, the set $x_i = g_i$ and $x_L = l_{j+1}$.
- If $d = dist_{\mathcal{K}}(G(l_i), L(l_i)) \geq dist_{\mathcal{K}}(G(g_i), L(g_i))$ for a given index i and the supremum occurs in g_j satisfying $g_j \leq l_i \leq g_{j+1}$, the set $x_i = g_i$ and $x_L = l_j$.
- If $d \leq D(F_N)$, then stop.
- Otherwise set

$$D(F_N) = \sup\Big\{D(F_N), \sup_{x_L \leq x \leq x_L^{old}} |G(x) - F_N(x)|, \sup_{x_U^{old} \leq x \leq x_U} |L(x) - F_N(x)|\Big\}$$

The computation of GCM and LCM is done by means of (6.10) and rather technical. More details about the algorithmic realization can be found in [83].

Furthermore it is shown in [83; 85] that the dip is asymptotically larger for the uniform distribution than for any distribution in a wide class of unimodal distributions: With growing sample size, the dip statistic for a unimodal distribution approaches zero while the dip of a any multimodal distribution approaches a positive constant. This leads to an algorithm with numerical complexity $\mathcal{O}(N)$ where N is the sample size. Figure 6.5 illustrates the computation of the dip index for a mixture of Gaussians with increasing distance of their mean values.

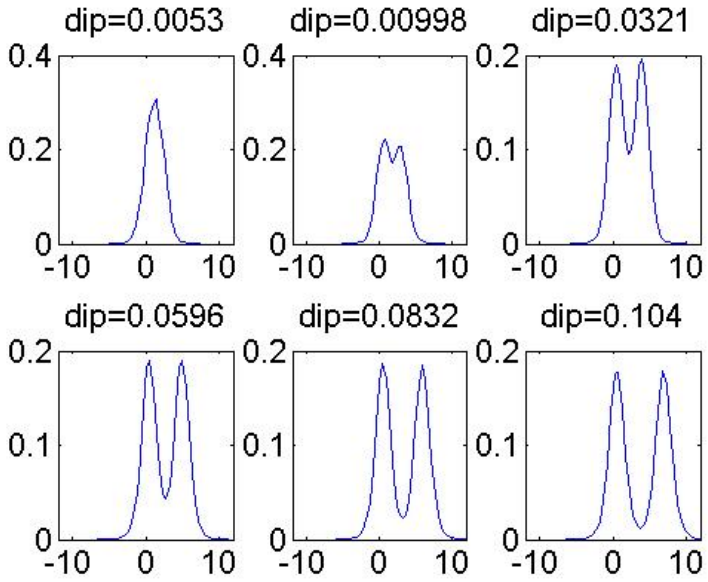


Figure 6.5: Illustration of the estimated values of the dip index significant for multimodality computed for a mixture of Gaussians with increasing distance of their mean values.

The dip statistic from above can be used to identify the cluster structure contained in the data projected to $\mathcal{I}$. To this end we project the data to every element of the basis of $\widehat{\mathcal{I}}$ given by the columns of $\widehat{\Pi}$. Computing the dip statistic with respect to the projected data allows to select a basis for $\mathcal{I}_M$ that not only contains non-Gaussian components of ρ but also the multimodal structure in the data. This step is inevitable since the solution of the semidefinite reformulation of (5.1) implies noting about any kind of ordering of the columns of $\widehat{\Pi}$. This means that the dimension reduction is essentially done using the dip-index. Finally we end up with the following algorithmic scheme:

Algorithm 9: NonGaussian Cluster Analysis

Data: $\{X_i\}_{i=1}^N$,ϵ,L,m
Result: m-dimensional basis of target space $\mathcal{I}$
Directional Sampling: As usual choose randomly a set of directions $\{\omega_l\}$, $l = 1 \dots L$ with $L >> d$ from $\mathcal{B}_d$.

Estimation: Compute $\widehat{\gamma}_l$ and $\widehat{\eta}_l$ according to (4.10) and (4.11) for $l = 1, \dots, L$.

Convex Optimization: Solve the corresponding linear constraint semi-definite problem to get an estimator $\widehat{\Pi}$ for the projector on the SNGCA target space $\mathcal{I}$.

Dimension Reduction: Project the data on the basis vectors of $\mathcal{I}$ provided by the column space of $\widehat{\Pi}$. Choose $p \leq m$ vectors with significant high dip index as ONB for the so called multimodal subspace $\mathcal{I}_M$.

For simplicity we call this algorithm the NonGaussian Clustering Analysis (NCA). Due the use of the dipp index NCA is highly selective to any cluster structure in the data.

The next task is to illustrate geometrically the differences between the projection methods described above using an artificial toy example of 3 non-spherical Gaussian clusters in $\mathbb{R}^3$ with $N = 1000$ points in every cluster. This is shown in figure 6.6.

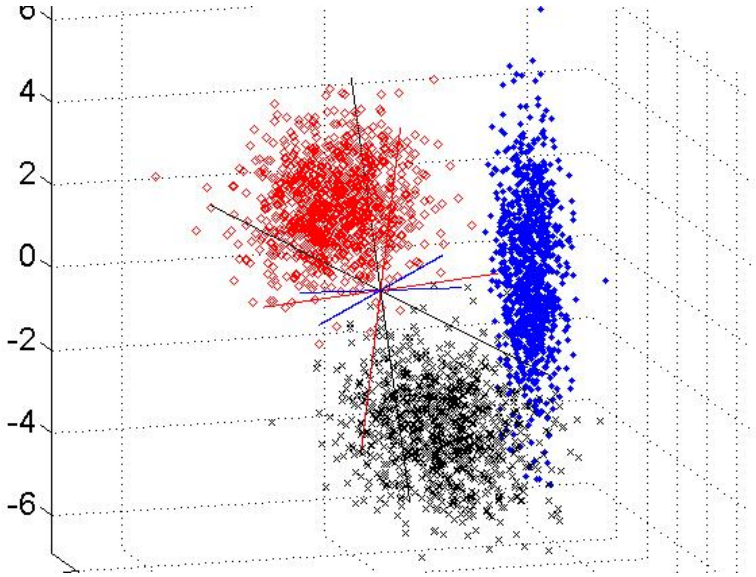


Figure 6.6: Original data consisting of 3 non-overlapping clusters. The colored axis are the basis provided by the concurrent methods: Red color indicates PCA, blue and black color ICA and NCA respectively.

In comparison it is demonstrated in figure 6.7 that a clustering of the same data as in figure 6.6 even in a low dimensional space according to the features of maximal data variance or NonGaussianty in principle differs from the clustering results obtained from NCA.

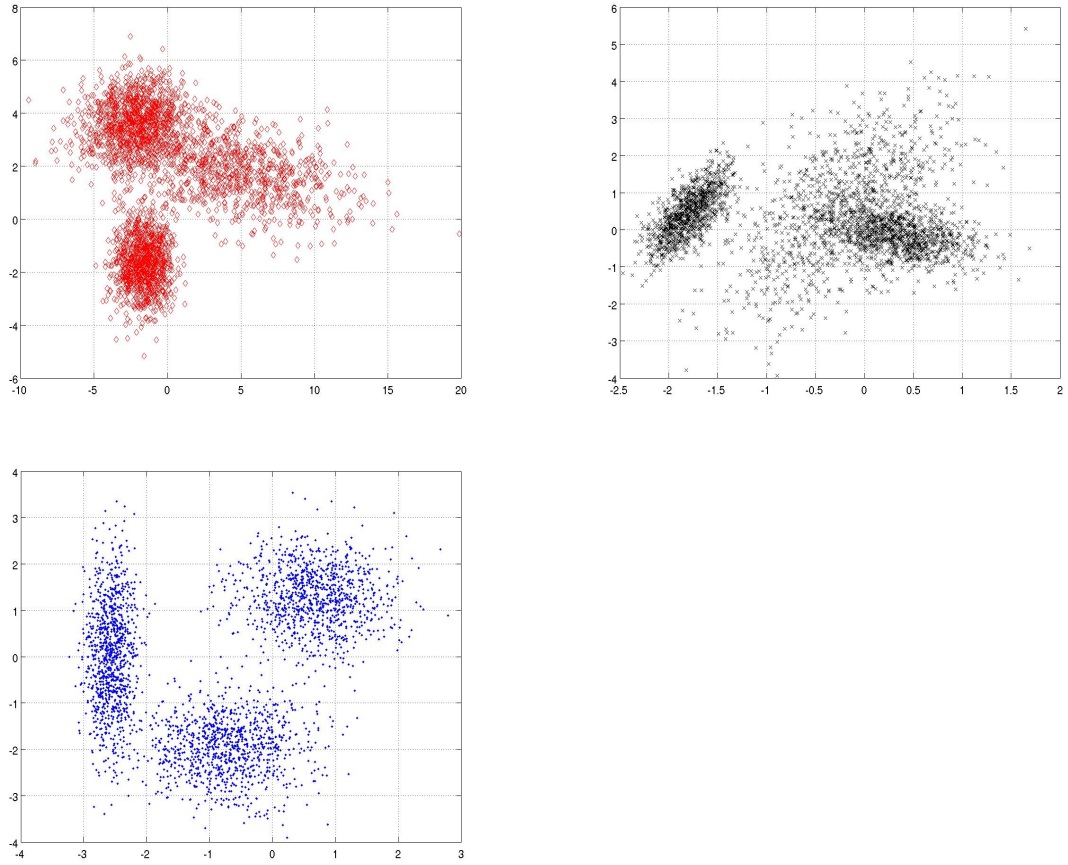


Figure 6.7: Illustration of general differences of comparable projective feature extraction methods from $\mathbb{R}^3$ to $\mathbb{R}^2$. On the upper left the data projected on the first two eigenvectors obtained from PCA are shown, the upper right and the lower left figure show the analogous result for ICA and NCA respectively. Obviously only NCA gives a sufficient separation of the cluster.

In the next section we will apply PCA, ICA and NCA described in algorithm 9 as a preprocessing step to a state-of-the art Markovian approach to the analysis of the metastable dynamical behavior of polypetides using Hidden Markov Models with Gaussian output combined with the Viterbi-clustering, that we have already described in section 6.2.

6.3.2 Metastability of Polypeptides

The problem of dimension reduction becomes crucial when dealing with e.g. data from molecular dynamics trajectories. Although chemical observations reveal that sometimes even for larger biomolecules the curse of dimensionality sometimes can be circumvented by exploiting the hierarchical structure of the dynamical properties of biomolecular systems e.g. [154], for many biomolecules a conformational analysis is possible only in low dimensions. In particular there are frequently relatively few independent essential degrees of freedom, needed to describe the conformational transitions and the rich spatial multiscale

structure induced by a structured potential energy landscape. However the essential degrees of freedom are typically unknown and have to be recovered from a non-hierarchical analysis of the large scale dynamical behavior of e.g. polypeptides.

6.3.3 Penta-Alanine

It is well known [182] that the effective dynamics of the real penta-alanine molecule showed in figure 6.8 occurs mainly in pairs of backbone torsion angles (Φ, Ψ) belonging to the same amino acid residue. Recall that electromagnetic repulsion between adjacent ligands and H-bond bridging between different peptide bonds serve as constraints for the rotation in these angles. The different global (secondary) structures of polypeptides (folded and unfolded) can hence be easily determined by the peptide angles, since other internal coordinates such as bond lengths, bond angles usually do not undergo changes of large amplitudes [157].

For each alanine amino acid we consider two of these backbone torsion angles shown in figure 6.8. These peptide angles pairs do not take arbitrary values, but adopt values in definite regions of the Ramachandran plane. They belong to various secondary structures of the molecule and shows the energetically preferred regions of a so called *dihedral pairs* of backbone angles.

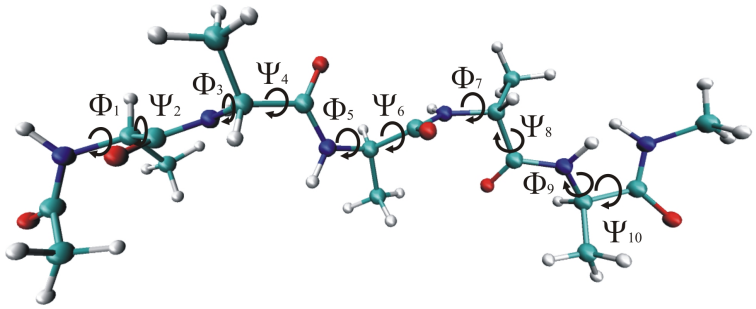


Figure 6.8: The figure shows the ten peptide angles of 5-alanine determining the secondary structure of 5-alanine, marked by $\Phi_1, \Psi_2, \ldots \Phi_9, \Psi_{10}$.

The time series to be analyzed consists in a 10 dimensional cyclic data set from the interval $]-180°, 180°]$ of all backbone torsion angles. The numerical simulation that corresponds to this time series, has been elaborately discussed in [157]. For convenience we only report the basic facts here: The simulation using a leap frog numerical integration scheme was done with explicit water using a thermostat of $300K$. The integration step size is $0.1ps$ whereas the complete integration time is $100ns$. Since we have a representation of the time series of 5-alanine in cyclic coordinates, the conformations of the molecule are not at best separated in the data. In order to minimize the periodicity in the data, the data were linear shifted such that the smallest local minimum of the data density is chosen as the point $-180°$. To this end we compute a adaptive kernel density estimation with Gaussian kernel [203] in each dimension of the time series.

Dimension Reduction: For the analysis with ICA we use the tanh-index. For PCA the reduced dimension m is obtained from a gap in the eigenvalue spectrum of the data covariance matrix. In the case of NCA the reduced dimension m is experimentally found using the gap size in the decreasing series of the dip index values. The same value for m is used in the case of ICA. For the detection of the non-Gaussian target space with NCA we found that the choice of the nonlinear test function is $h_\omega(x)$ is essential. Throughout

this section we use

$$h_\omega(x) = \frac{1}{1 + \exp(-\omega^\top x)} \tag{6.12}$$

The results obtained from PCA, ICA and NCA are controlled using the dip index in the following way: If the linear target space of each dimension reduction method contains the structure in the data set, we expect the dip index of the data projected on each basis vectors of the target space significantly higher than the index values of the data projected on each basis vector of its complement. In this sense the step of dimension reduction is essentially based on the idea of feature extraction.

The next figure 6.9 shows the values of the dip index computed from the projected data corresponding to the directions of decreasing maximum data variance and NonGaussianity respectively. Moreover in order to get some insight in the success of the dimension reduction method, we mark the target space of the reduction methods on focus using the eigenvalues and the entropy of the projected data respectively.

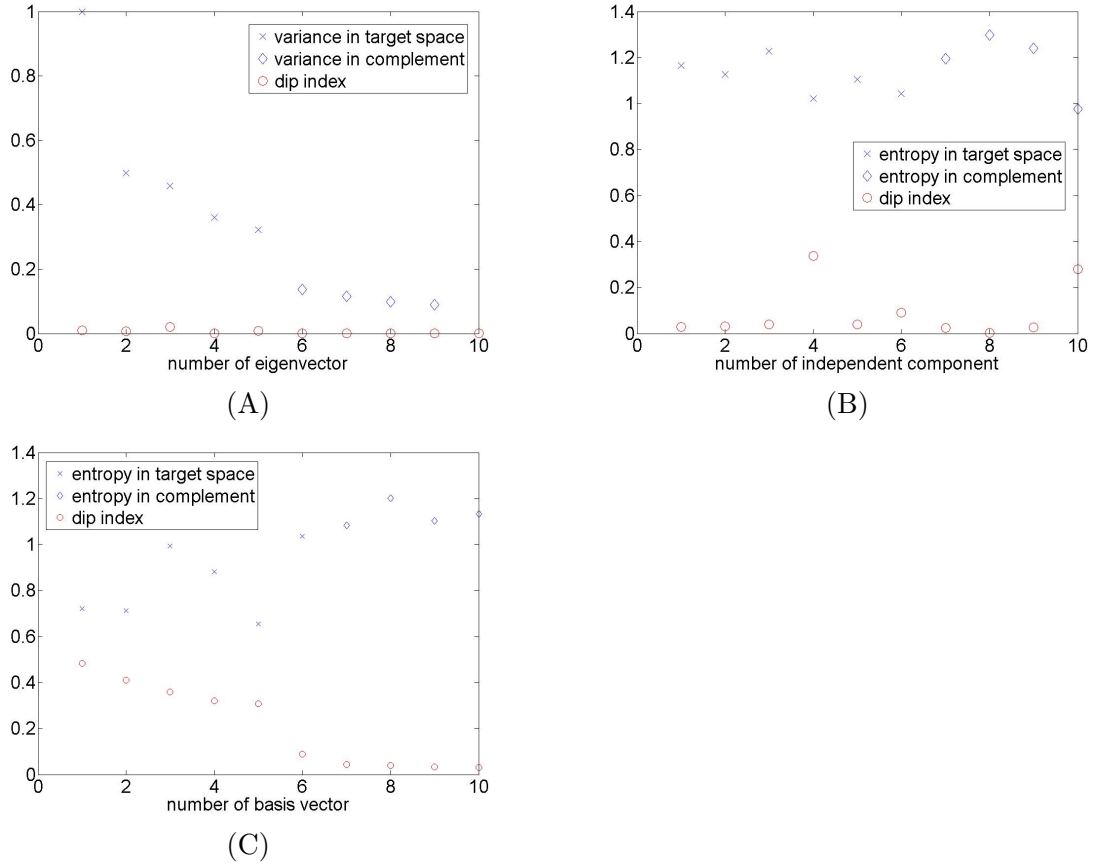


Figure 6.9: Comparison of feature extraction methods by means of the dip index and the estimated entropy of data, projected on the basis of $\mathcal{I}$: (A) shows the normed eigenvalues from PCA against the dip index, (B) the results from ICA and (C) the results from NCA.

The essential role of the dip index for the task of extracting the multimodal contribution to the data density ρ is illustrated in the next figures 6.10 and 6.11.

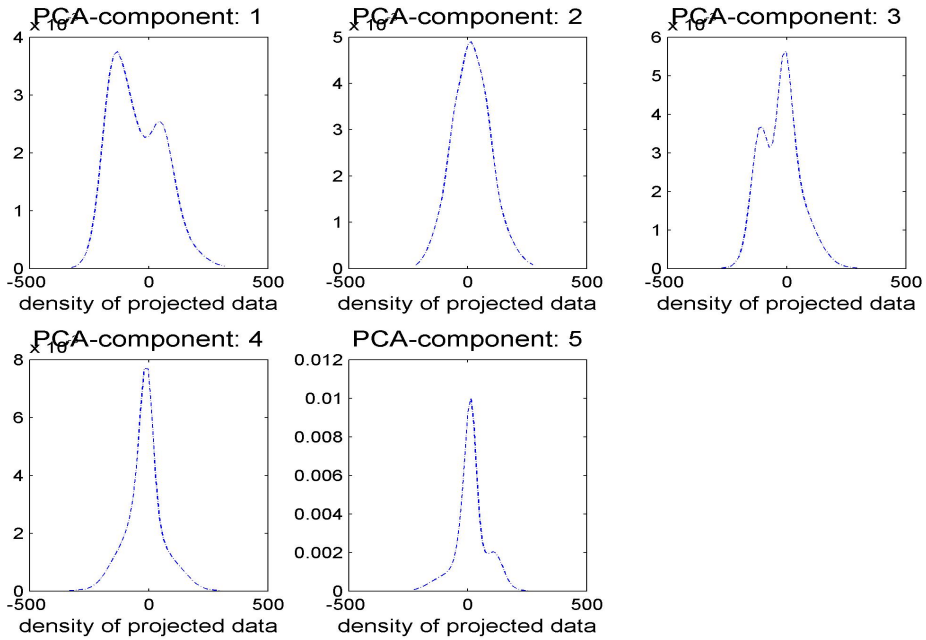


Figure 6.10: Densities, estimated by adaptive kernel methods [210] of the data from simulations of 5-alanine projected on the basis vectors of the PCA target space.

In the figures 6.10 and 6.11 we show the estimated densities of the data in the target space of the corresponding dimension reduction method projected on every of its basis vectors. The estimation is done with an adaptive kernel density estimation method using a Gaussian kernel [210].

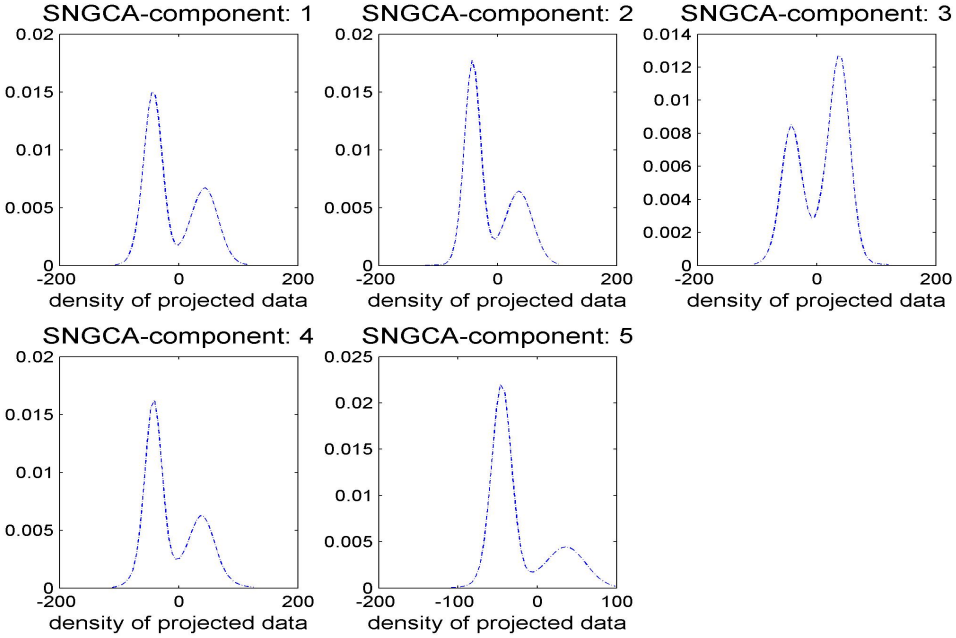


Figure 6.11: Estimated densities of the data from simulations of 5-alanine projected on the basis vectors of the NCA target space.

We observe for PCA in figure 6.9 and figure 6.10 that for the simulated data from 5-alanine described above the cluster structure in the data is not contained in the subspace of maximum data variance. Moreover with respect to ICA based on this observation in figure 6.9 we are forced to conclude that the subspace build up of directions with successively decreasing NonGaussianity contains only a small fraction of the multimodal distributed data. Furthermore multimodality is found in the complement of its target space. This means that NonGaussianity can be a misleading feature for the task of detecting multimodality. However in case of the cluster analysis with NCA figure 6.9 shows a gap in the decreasing series of dip-index values that corresponds to a 5 dimensional subspace $\mathcal{I}_m = \mathcal{I}$ of multimodal distributed data. We show the estimated density of the data in the target space $\mathcal{I}$ of NCA projected on every of its basis vectors in figure 6.11.

Obviously $\mathcal{I}$ is much better adapted to the position of the multimodal distribution part of the data than e.g. the subspace of maximal data variance provided by PCA. In order to affirm this result we compute the angle between the subspace U and V [79] defined by

$$\measuredangle(U,V) \stackrel{\text{def}}{=} \arccos(U^\top V) \quad [rad] \tag{6.13}$$

for all pairs of target spaces. For ICA get $\measuredangle(ICA, NCA) = 0.9320$ and for PCA we found $\measuredangle(PCA, NCA) = 1.3795$. We conclude that for 5-alanine PCA and ICA are not appropriate geometric pre-clustering methods, such that the attempt to understand the free energy surface of 5-alanine in terms of their low-dimensional projections will be deceptive. Consequently the following results on 5-alanine are obtained using NCA.

Analysis of Metastability: In the analysis of metastability the first step is to determine the number of clusters in the reduced data set. Using the the spectrum of the estimated symmetric transition matrix, figure 6.12 shows the Viterbi clustering result from PCCA of the Viterbi path obtained from the HMM-analysis discussed in the last section performed with 50 clusters. Here PCCA is used with a discretization of $[0, 2\pi]$ with 100 boxes. Alternative methods for determining the number of clusters are available [99].

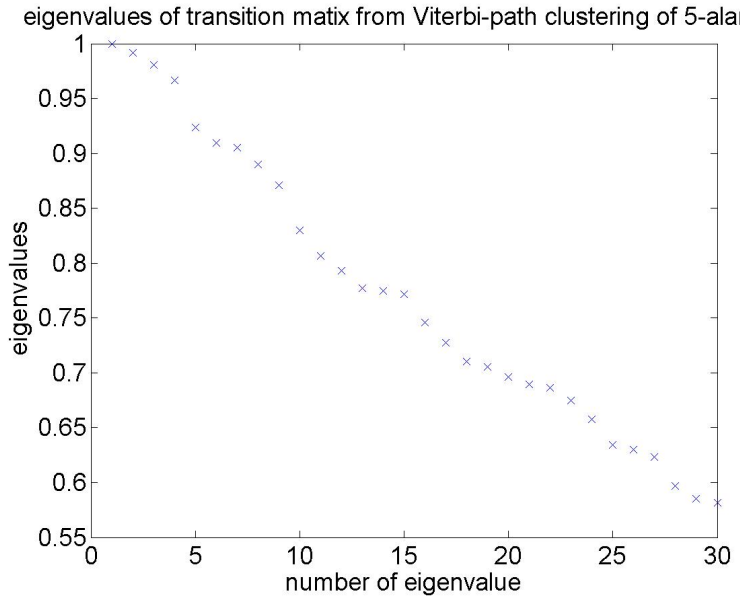


Figure 6.12: Plot of first 30 PCCA-eigenvalues from Viterbi-Path-clustering of 5-alanine after dimension reduction with SNGCA.

Using the gap criterion of the eigenvalues, reasonable values for the number of metastable states with dominating mean life time are $4, 9, 15$ or 16 respectively. This result coincides with a former hierarchical analysis of 5-alanine proposed in [154]. In order to get an insight of higher resolution in the dynamical behavior of 5-alanine we choose $n = 9$ as the number of metastable states. Using this tuning parameter we conduct a geometric analysis of the NCA-reduced data by means of the HMM-approach from above in a 5 dimensional subspace $\mathcal{I}$.

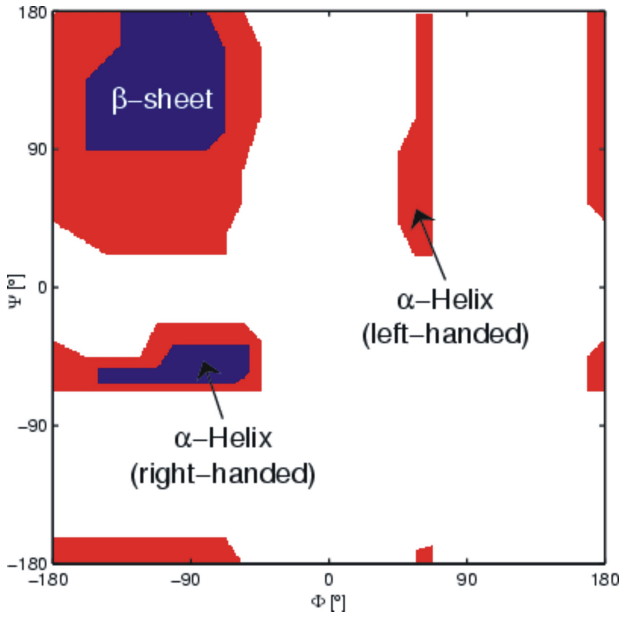


Figure 6.13: Schematic Ramachandran plot of penta-alanine.

In order to find a reasonable interpretation of our results that will enable a certain control of the metastability analysis, we use the Ramachandran plot of 5-alanine. The plot is obtained from physical considerations [182] and can be found in [154]. Since the free energy surface is pointwise given by

$$\Delta G(x) = -k_B T[\ln(I\!P(x)) - \ln(I\!P_{\max})] \tag{6.14}$$

where $I\!P_{\max}$ denotes the maximum of the whole distribution, the interesting information about the almost metastable states is given by the estimates of the density in the (Φ/Ψ)-plane with respect to the time series points belonging to its conformation. The above figure 6.13 shows the schematic Ramachandran plot of penta-alanine.

In order to get a first interpretation of the reduced dynamics we show in the figures 6.14 and 6.15 the empirical Ramachandran-plots of the first 4 conformations with dominating life time in descending order characterizing the effective dynamics of 5-alanine.

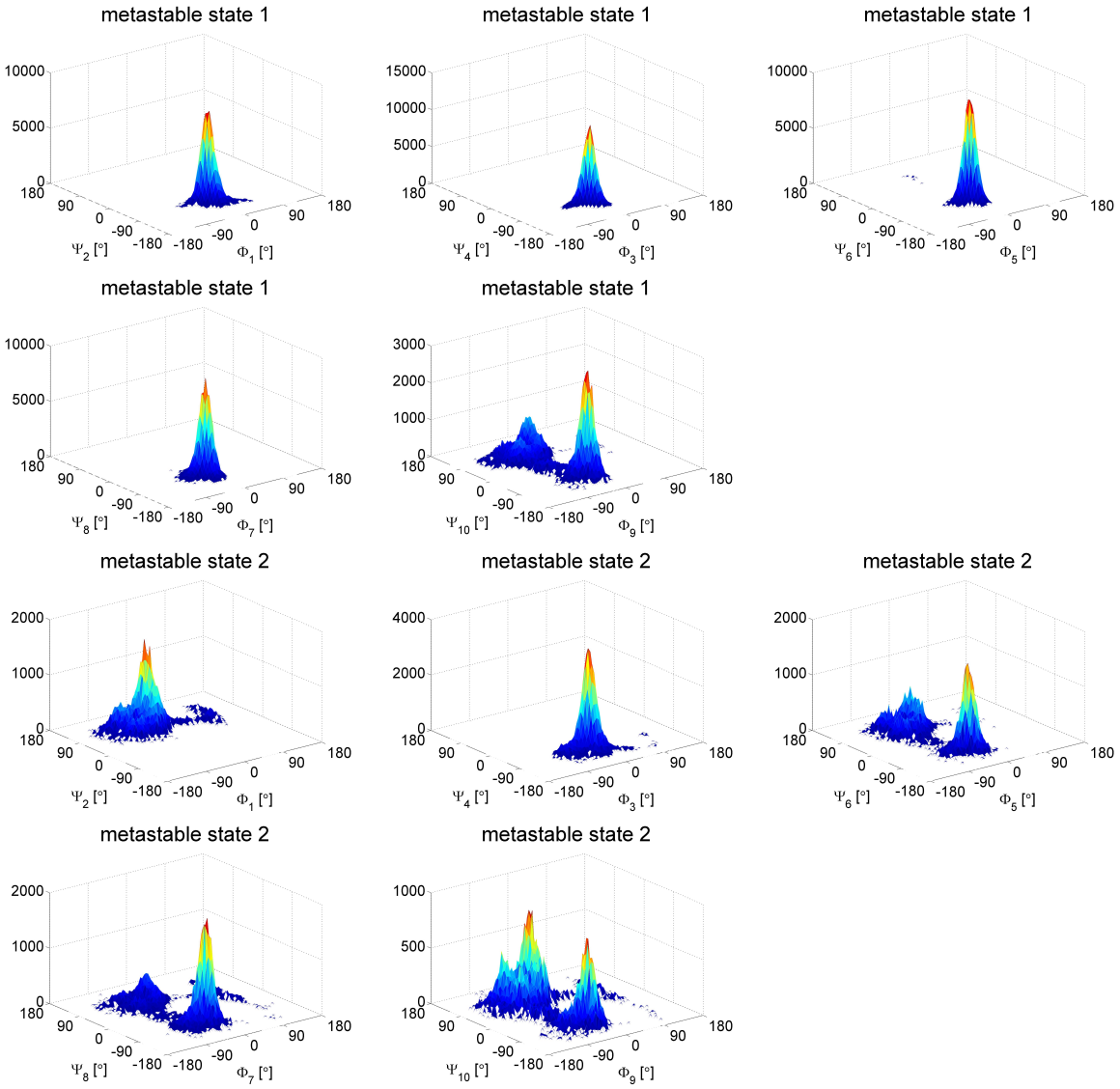


Figure 6.14: Empirical Ramachandran-plots of the first 2 conformations with dominating life time in descending order characterizing the effective dynamics of 5-alanine.

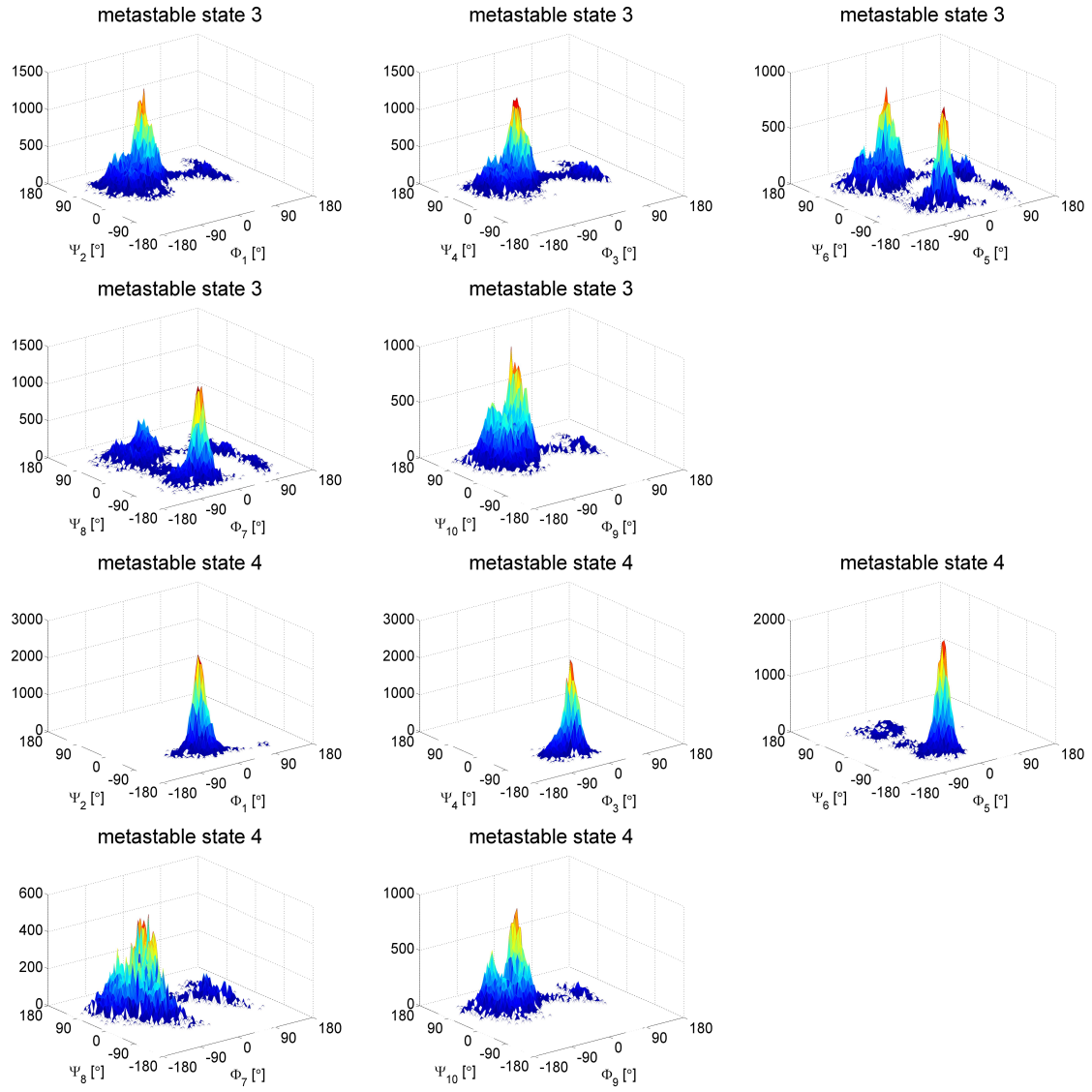


Figure 6.15: Empirical Ramachandran-plots of the conformations 3 and 4 with dominating life time in descending order characterizing the effective dynamics of 5-alanine.

Using the Ramachandran plot in figure 6.13 we can now compare the positions of the density peaks in the first conformation from figure 6.14 and 6.15 with the theoretical expectations: Obviously the first conformation corresponds to an α-helix structure.

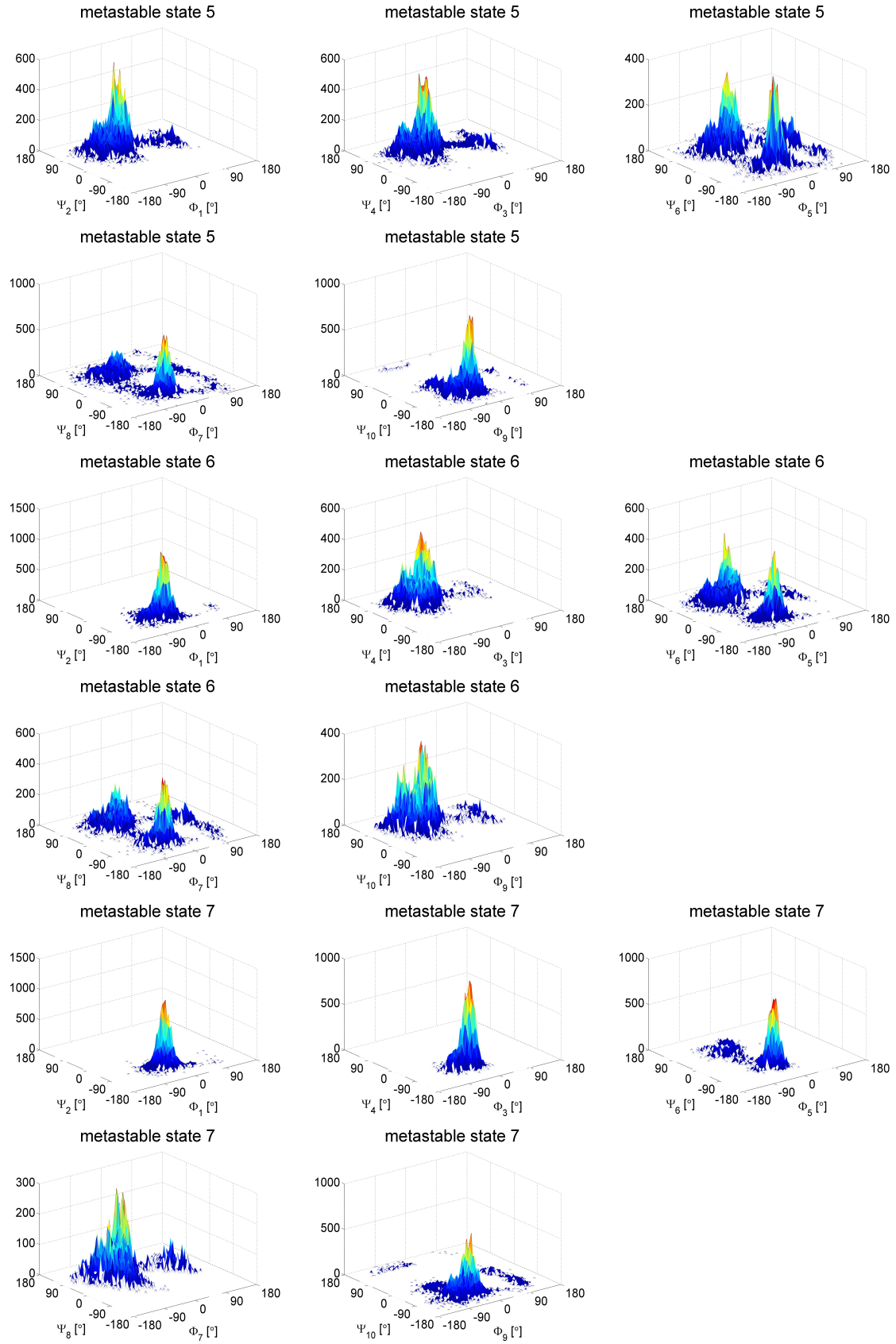


Figure 6.16: Empirical Ramachandran-plots of the conformations 5, 6 and 7 with dominating life time in descending order characterizing the effective dynamics of 5-alanine.

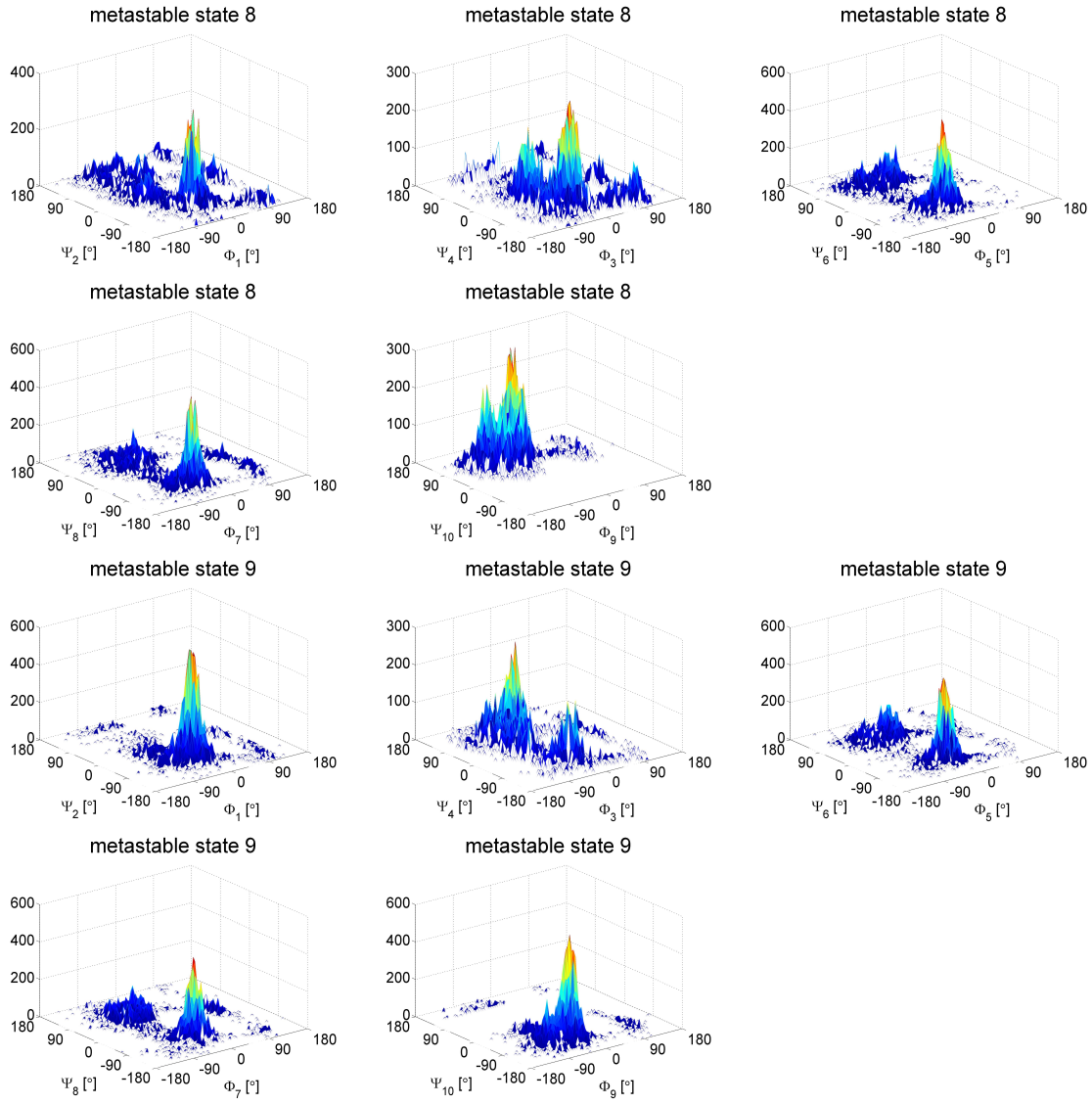


Figure 6.17: Empirical Ramachandran-plots of the conformations 8 and 9 with dominating life time in descending order characterizing the effective dynamics of 5-alanine.

Figures 6.16 and 6.17 show the empirical Ramachadran plots, i.e. the estimates of the densities corresponding to the conformations $5-9$ with dominating life time in descending order. We observe again separated favourable energetic regions in the Ramachandran plane, indicating a set of global structures, that are stable over long periods of time but not a single state corresponding to a local minimum of some energy function. A stable β-sheet conformation is not expected for 5-alanine, since it has a too short alanine amino acid chain. Other conformations identified in the reduced data set allow no unique assignment to a specific secondary structure. Unfortunately only the dominating metastable state has a unique geometrical interpretation since the other states exhibits a high flexibility with respect to dieder angles.

6.3.4 Octa-Alanine

Our next example is a times series from 8-alanine, presented in figure 6.18 and generated by a equilibrium molecular dynamics simulation with zwitterionic termini at temperature $T = 300K$ (courtesy to F. Noe). With the same motivation as in the example of 5-alanine, we consider the backbone dihedral angles in order to determine different conformations.

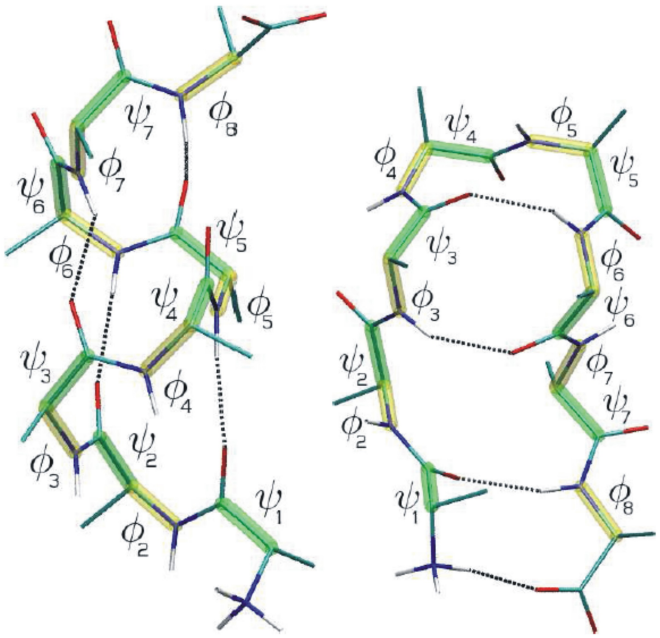
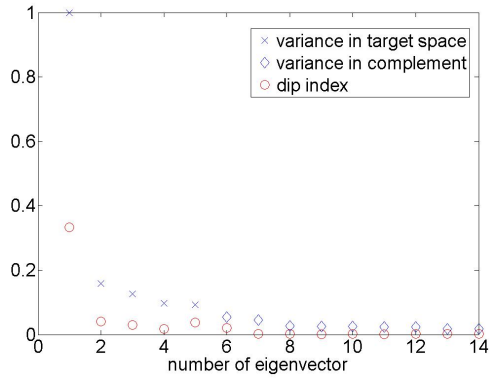


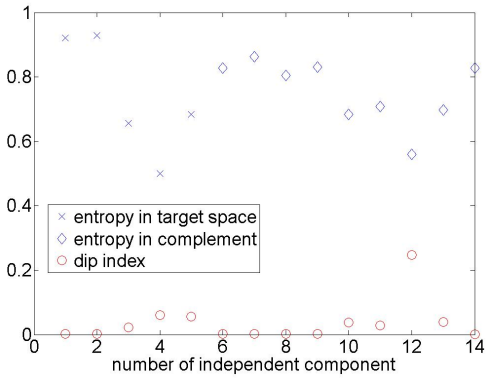
Figure 6.18: 8-alanine in $\alpha-helix$ (left) and $hair-pin$ configuration (right) representation with dihedral angles.

The time series consists in 14 dimensional cyclic data set of the all backbone torsion angles of 8-alanine. Therefore we conduct a linear transformation as described in the last section. The generation of the time series using CHARMM [30] is already described in [172]. Thus we will only summarize the main facts briefly. The simulation was done at $300K$ with implicit water by means of the solvent model ACE2 [194]. The time series comes with an integration step of $1fs$ using a symplectic Verlet integrator. The total trajectory length was $4\mu s$ and every $\tau = 50fs$ a set of coordinates was recorded.

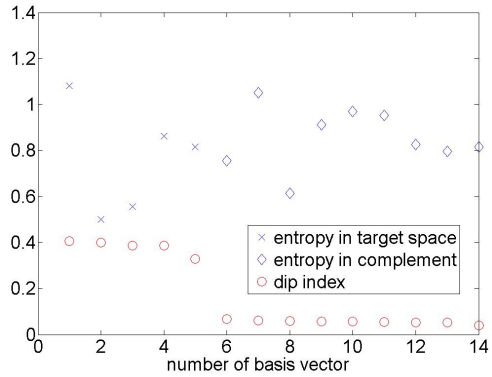
Dimension Reduction: Analog to the last numerical example we compare the dimension reduction methods PCA, ICA and NCA on 8-alanine using the dip index and the estimated entropy [135] described above. Again due to its essential role we use the dip index for the task of extracting the multimodal contribution. Figure 6.19 shows the values of the dip index corresponding to the directions of decreasing maximum data variance and NonGaussianity respectively. Again for the case of NCA the reduced dimension m is experimentally found using the gap size in the decreasing series of the dip index values. The same value for m is used in the case of ICA.



(A)



(B)



(C)

Figure 6.19: Comparison of feature extraction methods by means of the dip index and the estimated entropy of data, projected on the basis of $\mathcal{I}$: (A) shows the normed eigenvalues from PCA against the dip index, (B) the results from ICA and (C) the results from NCA.

Concerning the results of the dimension reduction we observe that the PCA-subspace determined by the number of eigenvalues close to 1 and well separated from the rest of the empirical spectrum only contains a small fraction of the multimodal contribution to the data density compared to the case of NCA. Here we found for the subspace angle $\measuredangle(PCA, NCA) = 1.4932$. In the case of ICA we found that the multimodal components are partially contained in the complement of the ICA-subspace. However we found $\measuredangle(ICA, NCA) = 1.5128$. In comparison to the 5-alanine example it seems, that also in the case of 8-alanine the dip index is a more reliable criterion to determine the interesting column space of Π^*, that identifies target space of SNGCA than the estimated entropy. Consequently PCA and ICA are not acceptable dimension reduction methods for 8-alanine. Using NCA we detect a 5-dimensional subspace $\mathcal{I}_m$ that contains the multimodal components of the data density ρ.

Analysis of Metastability: Again we determine the number of metastable states of the dimension reduced 8-alanine data set using a Viterbi clustering result from PCCA with a discretization of $[0, 2\pi]$ using 100 boxes of the Viterbi path obtained from the HMM-analysis performed with 50 clusters. The result is shown in the figure 6.20.

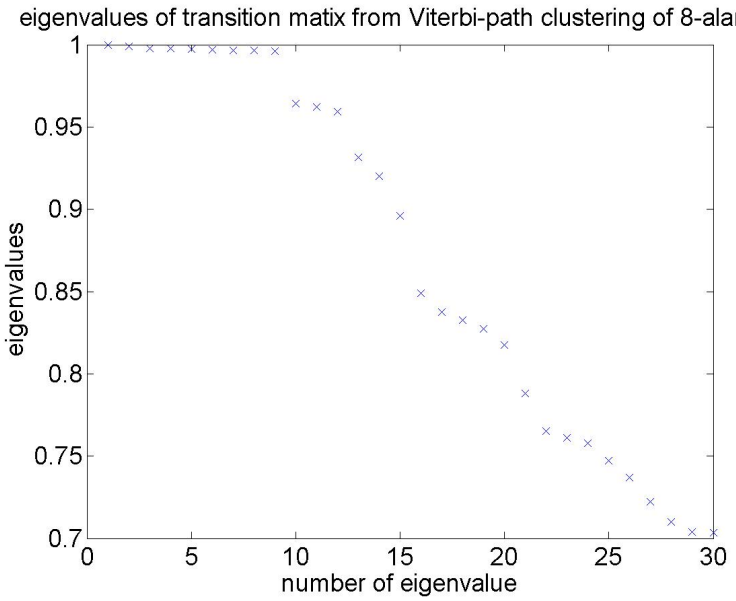


Figure 6.20: Plot of first 30 PCCA-eigenvalues from Viterbi-Path-clustering of 8-alanine after dimension reduction with SNGCA.

Using again the PCCA-eigenvalue gap criterion from above, we can identify of 9,12 or 15 metastable states with dominating mean life time respectively. The more conformations are accepted, the richer becomes the dynamical analysis. Again (6.14) motivates the use of the empirical Ramachandran-plots of 8-alanine in the (Φ, Ψ)-plane. In the following we present a dynamical analysis with 9 metastable clusters. Due to their unimodal densities the metastable states $1, 2, 5, 6, 8$ and 9 of 8-alanine with dominant life time in descending order have a unique geometrical interpretation.

In complete analogy to the simulation with respect to penta-alanine we present the empirical Ramachandran-plots of the different conformations with dominating life time in descending order characterizing the effective dynamics.

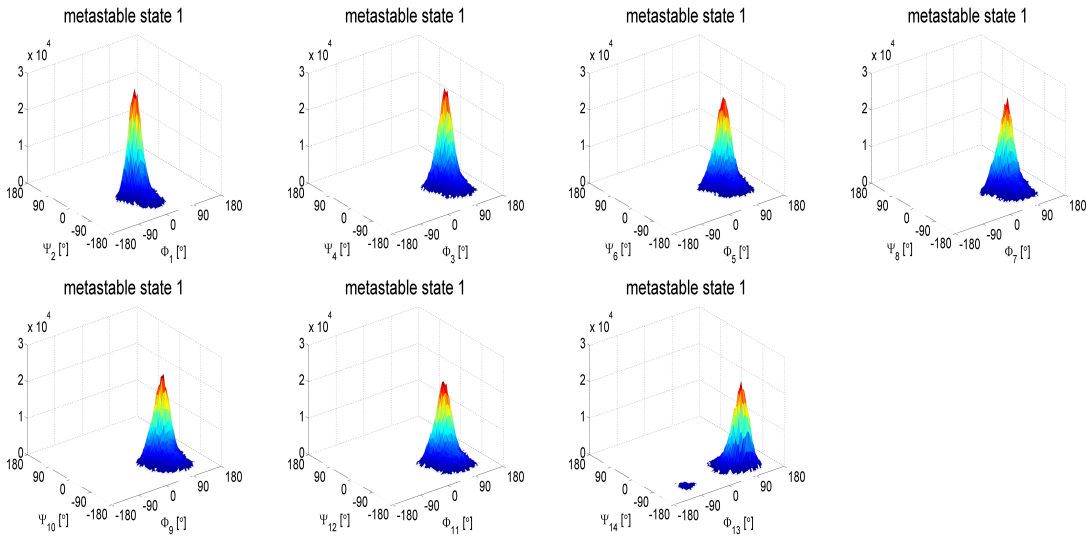


Figure 6.21: Empirical Ramachandran-plots of the first conformation with dominating life time in descending order characterizing the effective dynamics of 8-alanine.

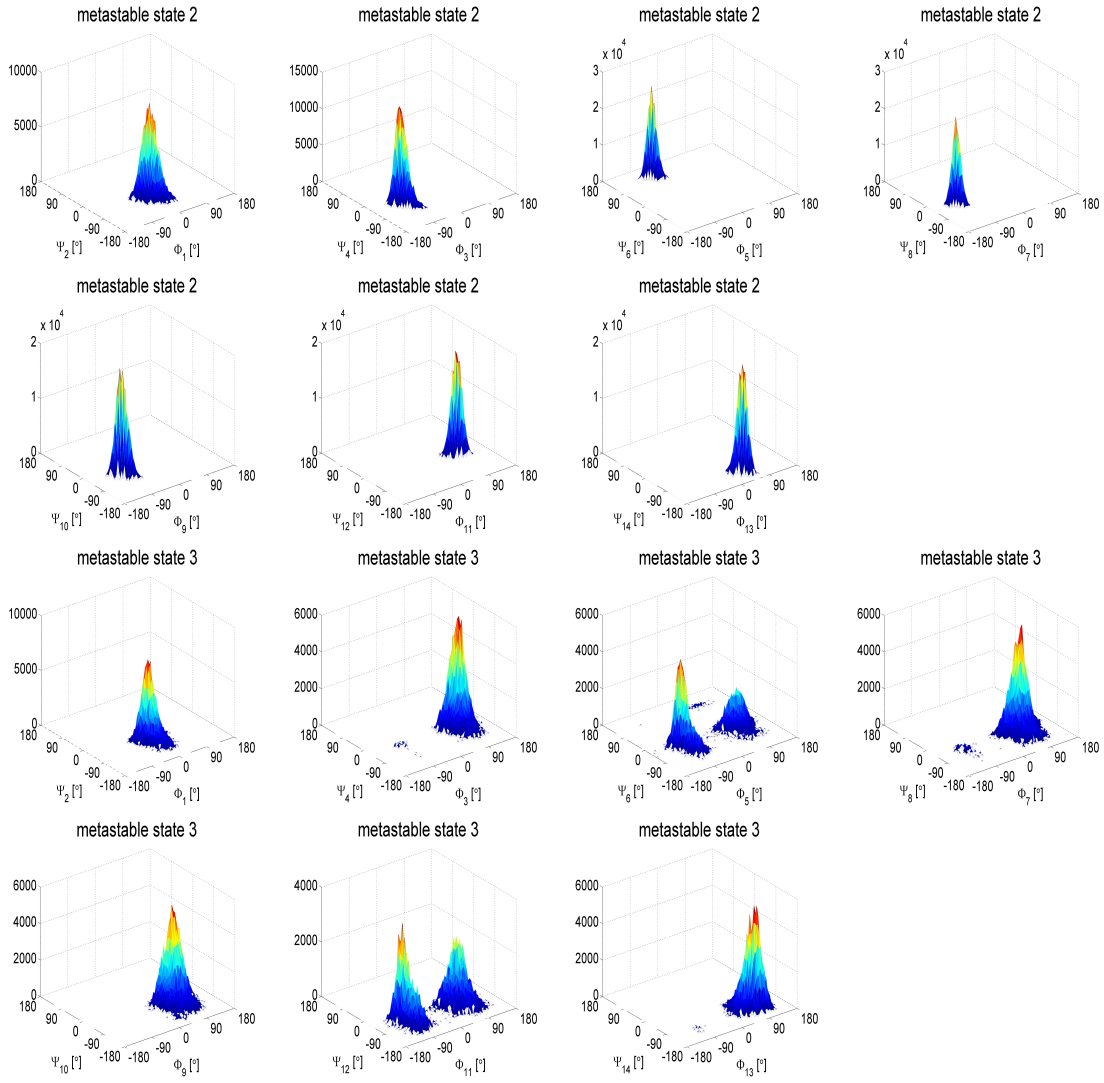


Figure 6.22: Empirical Ramachandran-plots of the conformations 2 and 3 with dominating life time in descending order characterizing the effective dynamics of 8-alanine.

Since the Ramachandran plot of the most favorable energetic regions is similar for many peptides [40], we use figure 6.13 again in order in find a reasonable interpretation of the results in the figures 6.21, 6.22, 6.23 and 6.24. In spite of the fact that the densities of the first two metastable states with dominant life time in descending order are unimodal only the second dominating state seems to be an α-helix configuration. The next figures show the states 4,5,6,7,8 and 9 metastable states again with dominant life time in descending order.

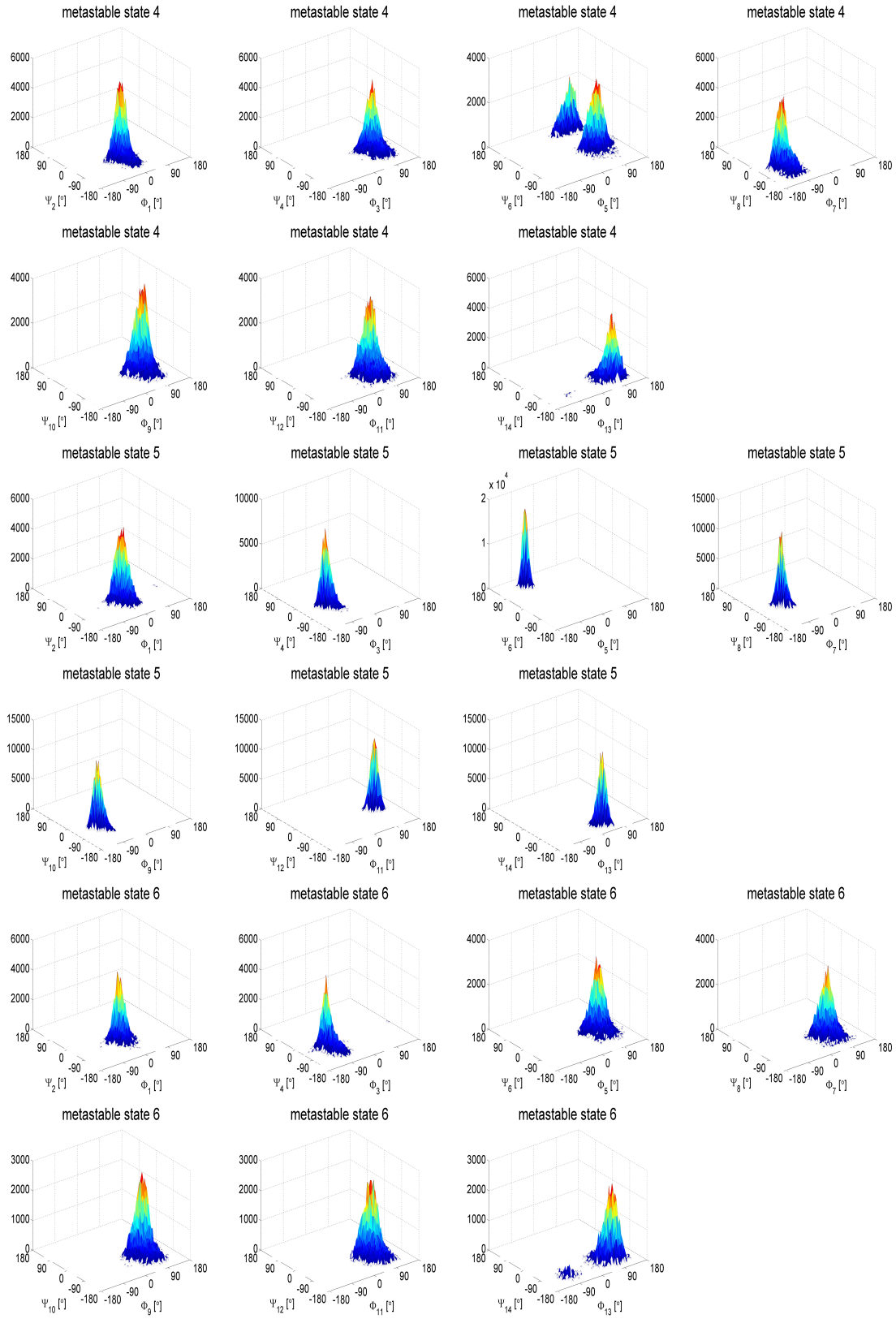


Figure 6.23: Empirical Ramachandran-plots of the conformations 4, 5 and 6 with dominating life time in descending order characterizing the effective dynamics of 8-alanine.

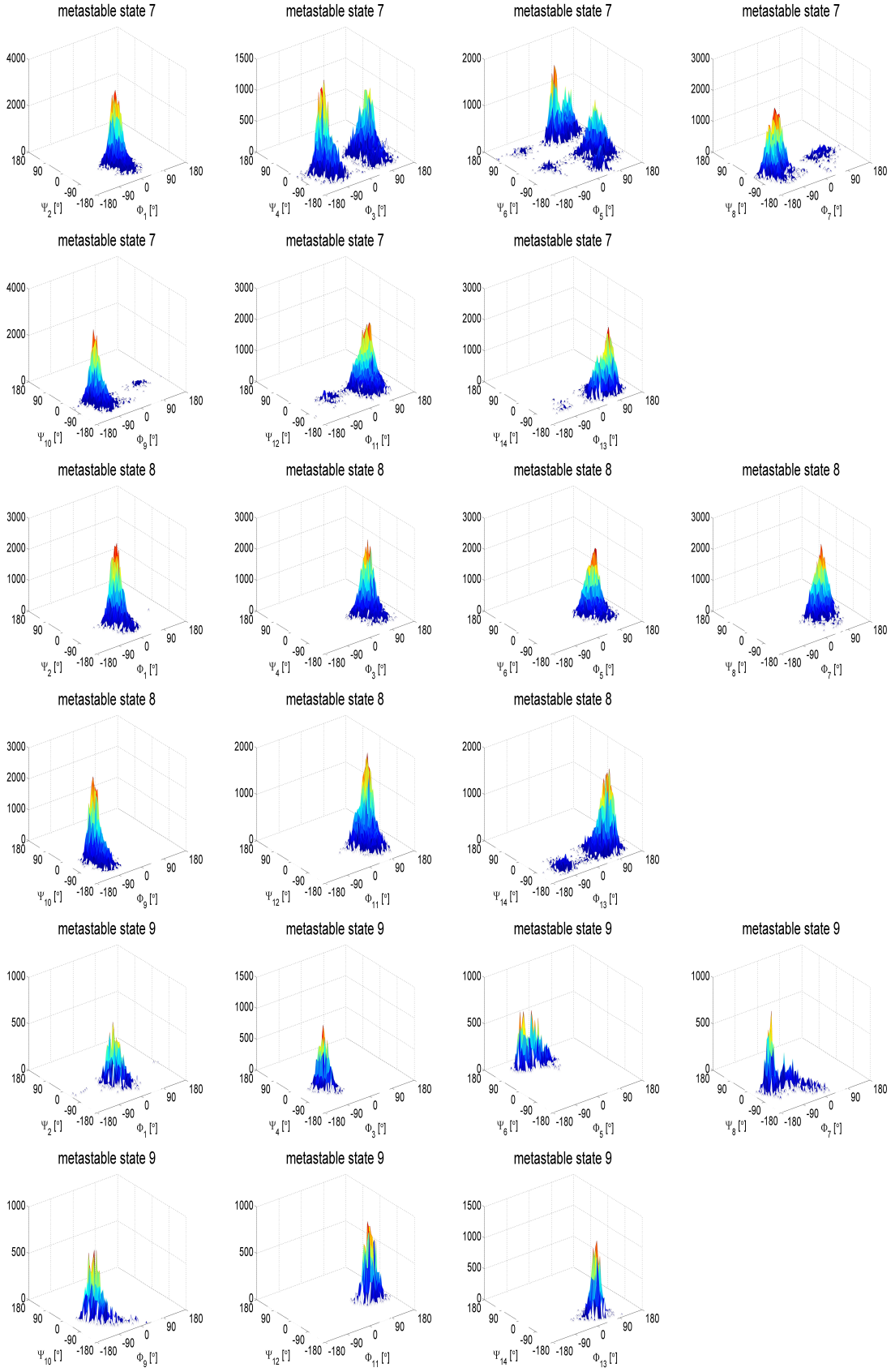


Figure 6.24: Empirical Ramachandran-plots of the conformations 7, 8 and 9 with dominating life time in descending order characterizing the effective dynamics of 8-alanine.

Again we conclude that only NCA is an acceptable pre-processing method for an analysis of the conformations of 8-alanine. However these good news for NCA can not be generalized: Which of the available dimension reduction methods will provide acceptable results only depends on the properties of every single data set. Unfortunately an obvious physical interpretation of the last conformation of 8-alanine is not available.

Chapter 7

Summary and Conclusion

In this thesis two formerly different projects have been merged and realized:

The first project is to develop an unsupervised, linear, projective feature extraction method that uses a completely different semi-parametric framework for dimension reduction than the common and well known Continuous Latent Variable Model. For this development it is required that almost none information represented by the original data is lost during the dimension reduction process and that the method is statistically sensitive and computationally cheap. It is shown that structural assumptions can be used in different and efficient ways to extract non-Gaussian components representing the information contained in the given high dimensional data distributed according to a stationary density. The best approach is realized using a combination of common relaxation and regularization methods with state-of-the-art dual gradient-type methods for semidefinite programming. By means of empirical process theory it is demonstrated that the statistical estimation error has rate of convergence $\mathcal{O}(\|\Sigma^{-1}\|_2, d)\frac{1}{\sqrt{N}}$. The whole SNGCA procedure has analytical complexity $\mathcal{O}(L \log L)$. However the numerical bottle neck is the arithmetical complexity of $\mathcal{O}(N^2 L + L^3)$ required for the data space sampling and the computation of the prox-transform. Using a broad variety of deviations from normality it was demonstrated, that SNGCA is superior to currently comparable feature extraction methods indicating the success of the semi-parametric framework.

The second project is to inquire the scope of an approach to the analysis of metastability that is almost geometric in the sense that only the metric relations between the points in the data space are used to detect a cluster structure in high dimensional data confined to a low-dimensional subspace, that represents the essential macroscopic dynamics of a biological active molecule. Due to the geometric origins of the curse of dimensionality described at the beginning of this thesis, the use of any metric in a common clustering algorithm will produce misleading clustering results even if a low dimensional set of independent vectors can be found to fully describe the cluster structure. Hence SNGCA is combined with a special index in the sense of projection pursuit that is sensitive only to a multimodal structure. Thus we come up to the so called NonGaussian Clustering Analysis, that works as a preprocessing step for a HMM-anaylsis combined with a Viterbi clustering. We have demonstrated that NCA is more efficient to extract a cluster structure from the data than other current popular methods. Finally we have applied the resulting almost geometrical approach to metastability to several biomolecular systems. Since in the reduced data space the extracted clusters are typically well separated, we found some evidence that the resulting data at least approximately fulfil the central assumption that the macroscopic dynamics is still Markovian.

Chapter 8

Zusammenfassung

Im ersten Teil dieser Arbeit wird eine vollständig datengesteuerte, lineare und projektive Methode der Merkmalsextraktion entwickelt. Sie beruht auf einer semiparametrischen Hypothese in Bezug auf die Datendichte und unterscheidet sich grundlegend von dem im linearen Fall typischerweise benutzten Continuous Latent Variable Model. Als Adäquatheitsbedingung wurde verlangt, daß so wenig wie möglich von der durch die Daten repräsentierten Information bei der Dimensionsreduktion verloren gehen darf. Weiter sollte die Methode auch in hohen Dimensionen sensitiv und mit wenig Zeitaufwand zu berechnen sein. Es wurde gezeigt, daß die semi-parametrischen Hypothese in verschieden effizienter Weise benutzt werden kann, Merkmale aus einer hochdimensionalen Dichte zu extrahieren. Als bester Zugang hat sich eine Methode erwiesen, die neuste Techniken der semidefiniten Programmierung benutzt. Mit den Mitteln der empirischen Prozeßtheorie wurde gezeigt, daß die Konvergenzrate des Schätzfehlers $\mathcal{O}(\|\Sigma^{-1}\|_2, d)\frac{1}{\sqrt{N}}$ ist. Der Aufwand des kompletten SNGCA-Algorithmus hat eine analytische Komplexität von $\mathcal{O}(L \log L)$. Der numerische Flaschenhals besteht jedoch in der arithmetischen Komplexität von $\mathcal{O}(N^2L+L^3)$, die beim Abtasten des Datenraums und der Berechnung der prox-Transformation anfällt. Ein Vergleich mit anderen, gegenwärtig populären, projektiven Methoden zeigt für eine Vielzahl verschiedener Abweichungen von der Normalverteilung, daß SNGCA im Moment die überlegene Methode ist. Das zweite Unterprojekt untersucht die Reichweite eines Zugangs zur Analyse von Metastabilität bei Biomolekülen, der soweit wie möglich geometrisch ist in dem Sinne, als nur die metrischen Relationen zwischen den Datenpunkten benutzt werden, um eine Clusterstruktur in einer stationären Verteilung von Punkten zu identifizieren, welche, auf einen niedrig dimensionalen Unterraum beschränkt, die essentielle, makroskopische Dynamik z.B. eines biologisch aktiven Moleküls repräsentiert. Aufgrund des geometrischen Ursprungs des sogenannten Fluchs der Dimension, liefern herkömmliche Clusteralgorithmen, die auf der Berechnung einer Metrik in hohen Dimensionen beruhen, jedoch typischerweise irreführende Ergebnisse. Dies gilt selbst dann, wenn die betreffenden Punkte faktisch auf einer niedrigdimensionalen Mannigfaltigkeit liegen. Aus diesem Grund wurde SNGCA mit einem Index im Sinne des projection-pursuit-Ansatzes kombiniert, der ausschließlich sensitiv ist gegenüber multimodalen Komponenten der vorgegebenen Dichte. Die entstandene Methode der NonGaussian Clustering Analysis wurde als Dimensionsreduktion vor einer Metastabilitätsanalyse auf der Basis von Hidden-Markov Modellen verwendet, was einen nahezu vollständig geometrischen Zugang zur Metastabilitätsanalyse bedeutet. Ein Vergleich verschiedener, und gegenwärtig populärer Methoden mit NCA zeigt, daß letztere besser als jene geeignet ist, Clusterstrukturen in hochdimensionalen Datensätzen zu detektieren. Insbesondere weist die gut ausgeprägte Separation der reduzierten Daten in Cluster bei verschiednen Simulationen von Biomolekülen darauf hin, daß die Dimensionsreduktion die der Metastabilitätsanalyze zugrunde liegende Markovannahme approximativ erhält.

Appendix A

Proofs

In this appendix we will give the proofs of the theorems used in this article.

A.1 Proof of Theorem 1

Theorem 1. *The density* $\rho(x)$ *for the model*

$$X = Y + Z \tag{A.1}$$

with the m*-dimensional signal* Y *and an independent Gaussian noise* Z *can be represented as*

$$\rho(x) = \phi_{\mu,\Sigma}(x) q(Tx). \tag{A.2}$$

where T *is a linear operator from* $T : \mathbb{R}^d \to \mathbb{R}^m$, $q(\cdot)$ *is some function on* $\mathbb{R}^m$ *and* $\phi_{\mu,\Sigma}$ *is the density of the Gaussian component.*

Proof. Consider the model (A.1) as well as the projectors $\Pi_{\mathcal{I}} : \mathbb{R}^d \to \mathbb{R}^m$ and $\Pi_{\mathcal{I}^\perp} : \mathbb{R}^d \to (\mathbb{R}^m)^\perp$. Let Z be decomposed in independent noise components $Z = Z_1 + Z_2$ with $Z_1 \stackrel{\text{def}}{=} \Pi_{\mathcal{I}} Z$ and $Z_2 \stackrel{\text{def}}{=} \Pi_{\mathcal{I}^\perp} Z$. Then due to (A.2) the model (A.1) can be written as $X = (\Pi_{\mathcal{I}} Y + Z_1) + Z_2$. According to our premises in (A.1) the noise is independent from the signal Y. Hence the density of $\Pi_{\mathcal{I}} Y + Z_1$ can be represented as the product $q(x_1)\phi(x_1)$ for some function q and the normal density $\phi(x_1)$, $x_1 \stackrel{\text{def}}{=} \Pi_{\mathcal{I}} x \in \mathbb{R}^m$. Furthermore due to their construction, we have independence of Z_1 and Z_2. Consequently the density ρ can be written as

$$\rho(x) = g(x_1)\phi(x_1)\phi(x_2) = g(x_1)\phi(x)$$

where $x_2 \stackrel{\text{def}}{=} \Pi_{\mathcal{I}^\perp} x$. Setting $T \stackrel{\text{def}}{=} \Pi_{\mathcal{I}}$ leads to $\ker(T) = \Pi_{\mathcal{I}^\perp}$.

Now suppose that Z is standard normal. Then we are done. Next suppose that Z is not standard normal and the covariance matrix Σ_Z is nondegenerated. Then we get the model

$$\Sigma_Z^{-\frac{1}{2}} X = \Sigma_Z^{-\frac{1}{2}} Y + \widetilde{Z} \tag{A.3}$$

where $\widetilde{Z} \stackrel{\text{def}}{=} \Sigma_Z^{-\frac{1}{2}} Z$ is standard normal. Using $\widetilde{X} \stackrel{\text{def}}{=} \Sigma_Z^{-\frac{1}{2}} X$ allows to repeat the argument from above with $\widetilde{\mathcal{I}} \stackrel{\text{def}}{=} \Sigma_Z^{-\frac{1}{2}} \mathcal{I}$ and $T = \Pi_{\mathcal{I}^\perp} \Sigma_Z^{-\frac{1}{2}}$. □

A.2 Proof of Theorem 3

Theorem 3. *Let X follow the distribution with the density $\rho(x)$ according to (4.1) and let $I\!EX = \mu = 0$. Suppose that $\psi \in \mathcal{C}^1(\mathbb{R}^d, \mathbb{R})$ is a function fulfilling the condition*

$$\gamma(\psi) \stackrel{\text{def}}{=} I\!E[X\psi(X)] = 0, \tag{A.4}$$

Define

$$\beta(\psi) \stackrel{\text{def}}{=} I\!E[\nabla\psi(X)] = \int \nabla\psi(x)\,\rho(x)dx, \tag{A.5}$$

where $\nabla_x\psi$ means the gradient of ψ. Then $\beta(\psi)$ belongs to $\mathcal{I}$. Moreover if (A.4) is not fulfilled, then there is a $\beta \in \mathcal{I}$ such that

$$\|\beta(\psi) - \beta\|_2 \leq \epsilon$$

where ϵ is the uniform error bound:

$$\epsilon = \left\|\Sigma^{-1}\int x\psi(x)\rho(x)\;dx\right\|_2. \tag{A.6}$$

Hence the distance between $\beta(\psi)$ and the non-Gaussian subspace $\mathcal{I}$ is uniformly bounded as given by (A.6).

Proof. The structural assumption (A.2) and the identity

$$\nabla_x \log\left[\phi_{\mu,\Sigma}(x)\right] = -\Sigma^{-1}(x-\mu)$$

imply

$$\begin{aligned}
- &\int \psi(x)\big[\nabla\log(\rho(x))\big]\rho(x)\,\mathrm{d}x = \\
- &\int \psi(x)\big[\nabla\log(q(Tx))\big]\rho(x)\,\mathrm{d}x - \int \psi(x)\big[\nabla\log(\phi_{\mu,\Sigma}(x))\big]\rho(x)\,\mathrm{d}x = \\
- &\int \psi(x)T^\top q'(Tx)\phi_{\mu,\Sigma}x\,\mathrm{d}x + \int \psi(x)\Sigma^{-1}(x-\mu)\rho(x)\,\mathrm{d}x
\end{aligned}$$

where $q'(x)$ denotes the gradient of $q(x)$. The vector $\beta(\psi)$ with

$$\beta(\psi) \stackrel{\text{def}}{=} -T^\top\int \psi(x)q'(Tx)\phi_{\mu,\Sigma}(x)\,\mathrm{d}x$$

obviously belongs to $\mathcal{I}$. Suppose now the condition (A.4) is fulfilled. Then it holds that

$$\Sigma^{-1}\left[\int x\psi(x)\rho(x)\,\mathrm{d}x - \mu\int\psi(x)\rho(x)\,\mathrm{d}x\right] = 0$$

Thus we know that $\beta(\psi) \in \mathcal{I}$. Otherwise if the condition (A.4) is not fulfilled, it follows from from (A.7) that there is a $\beta \in \mathcal{I}$ such that

$$\|\beta(\psi) - \beta\|_2 \;=\; \left\|\Sigma^{-1}\int(x-\mu)\psi(x)\rho(x)\,\mathrm{d}x\right\|_2. \tag{A.7}$$

Let $u \in \mathbb{R}^d$. Then it holds that

$$\int \psi(x+u)\rho(x)\,\mathrm{d}x = \int \psi(x)\rho(x-u)\,\mathrm{d}x \tag{A.8}$$

Using the regularity conditions on ψ and ρ, we differentiate $(A.8)$ with respect to u and use the identity

$$\nabla \log(\rho(x)) = \bigl[\nabla\rho(x)\bigr]\rho(x)$$

This yields:

$$\int \rho(x)\nabla\psi(x)\,\mathrm{d}x \;\; = \;\; -\int \psi(x)\bigl[\nabla \log(\rho(x))\bigr]\rho(x)\,\mathrm{d}x$$

From this consideration we get an equivalent expression for $\beta(\psi)$:

$$\beta(\psi) = \int [\nabla\psi(x)]\rho(x)\,\mathrm{d}x$$

□

A.3 Proof of Theorem 4

Theorem 4. *Suppose that f is continuously differentiable in w and for some fixed constant f_1^* and any $\omega \in \mathcal{B}_d$, $x \in \mathbb{R}^d$*

$$\begin{aligned} \operatorname{Var}\bigl[X_j\, f(X,\omega)\bigr] \le f_1^*, \quad \operatorname{Cov}\bigl[X_j\, \nabla_\omega f(X,\omega)\bigr] \le f_1^* I, \\ \operatorname{Var}\biggl[\frac{\partial}{\partial x_j} f(X,\omega)\biggr] \le f_1^*, \quad \operatorname{Cov}\biggl[\nabla_\omega \frac{\partial}{\partial x_j} f(X,\omega)\biggr] \le f_1^* I, \end{aligned}$$

Consider the (random) set

$$\mathscr{C} = \bigl\{c \in \mathbb{R}^L : \|c\|_1 \le 1,\, \widehat{\gamma}(c) = 0\bigr\}. \tag{A.9}$$

Then for any $\epsilon > 0$ there is a set $A \subset \Omega$ of probability at least $1-\epsilon$ such that on A for all $c \in \mathscr{C}$,

$$\bigl\|(I - \Pi^*)\widehat{\beta}(c)\bigr\|_2 \le \sqrt{d}\,\delta_N\bigl(1 + \|\Sigma^{-1}\|_2\bigr),$$

where

$$\delta_N = N^{-1/2} \inf_{\lambda \le \lambda_1^* N^{1/2}} \bigl\{5\mathfrak{n}_0 f_1^* \lambda + 2\lambda^{-1}\bigl[\mathfrak{e}_d + \log(2d/\epsilon)\bigr]\bigr\}$$

and $\mathfrak{e}_d = 4d\log 2$.

We use the following result from the empirical process theory (similar statements under slightly different assumptions can be found e.g. in [226]). Let $\mathcal{B}$ stand for the unit Euclidean ball, centered at the origin. Similarly, $B(\mu,\omega^\circ) = \{\omega : \|\omega - \omega^\circ\|_2 \le \mu\}$ is a ball of radius μ centered at ω°. For a function $q(\omega,x)$, denote $I\!\!E_N[q(\omega,X)] = N^{-1}\sum_{i=1}^N q(\omega,X_i)$.

Lemma 3. *Let $q(\omega,x)$ be a continuously differentiable function of $\omega \in \mathcal{B}_d$ and $x \in \mathbb{R}^d$ such that for every $\omega \in \mathcal{B}_d$*

$$\operatorname{Var}\bigl[q(\omega,X)\bigr] \le q^*, \quad \operatorname{Cov}\bigl[\nabla_\omega q(\omega,X)\bigr] \le q^* I, \tag{A.10}$$

with some $q^*, q^* > 0$. *Define*

$$\zeta(\omega) = N^{1/2}\big\{I\!E_N[q(\omega, X)] - I\!E[q(\omega, X)]\big\}$$

and $\zeta(\omega, \omega') = \zeta(\omega) - \zeta(\omega')$. *Then for any* $\mathfrak{n}_0 > 1$, *there is* $\lambda_1^* = \lambda_1^*(\mathfrak{n}_0) > 0$ *such that for any* $\omega^\circ \in \mathcal{B}_d$, $\mu \le 1$, *and* $\lambda \le \lambda_1^* N^{1/2}$

$$\log I\!E \exp\big[\lambda\zeta(\omega^\circ)\big] \le \mathfrak{n}_0 q^* \lambda^2/2, \tag{A.11}$$

$$\log I\!E \exp\Big[\frac{\lambda}{\mu} \sup_{\omega \in B(\mu, \omega^\circ)} \zeta(\omega, \omega^\circ)\Big] \le 2\mathfrak{n}_0 q^* \lambda^2 + \mathfrak{e}_d\,, \tag{A.12}$$

where $\mathfrak{e}_d = \sum_{k=1}^{\infty} 2^{-k} \log(2^{kd}) = 4d \log 2$. *Moreover, define*

$$\mathfrak{z}(\lambda) = \mathfrak{n}_0\big(q^*/2 + 2q^*\big)\lambda^2 + \mathfrak{e}_d.$$

Then for any $\epsilon > 0$

$$I\!P\bigg(\sup_{\omega \in \mathcal{B}_d} \zeta(\omega) \ge 2\lambda^{-1}\big[\mathfrak{z}(\lambda) + \log \epsilon^{-1}\big]\bigg) \le \epsilon.$$

Proof. Define for $\omega \in \mathcal{B}_d$

$$g_0(\lambda; \omega) = \log I\!E \exp\Big[\frac{\lambda}{\sqrt{\mathfrak{n}_0 q^*}}\big\{q(\omega, X_1) - I\!E[q(\omega, X_1)]\big\}\Big].$$

Then $g_0(\lambda; \omega)$ is analytic in λ and satisfies $g_0(0; \omega) = g_0'(0; \omega) = 0$. Moreover, the condition (A.10) implies $g_0''(0; \omega) < 1$. Therefore, there is some $\lambda_1^* > 0$ such that for any $\lambda_1 \le \lambda_1^*$ and any unit vector ω, it holds $g_0(\lambda_1; \omega) \le \lambda_1^2/2$. Independence of the X_i's implies (A.11) for $\lambda \le \lambda_1^* N^{1/2}(\mathfrak{n}_0 q^*)^{-1/2}$. In the same way, for $\omega, u \in \mathcal{B}_d$ define $\zeta(\omega, X) = \nabla_\omega q(\omega, X_1) - I\!E[\nabla_\omega q(\omega, X_1)]$ and

$$g(\lambda; \omega, u) = \log I\!E \exp\big[\frac{2\lambda u^\top}{\sqrt{\mathfrak{n}_0 q^*}} \zeta(\omega, X_1)\big].$$

Then similarly to the above, the function $g(\lambda; \omega, u)$ is analytic in λ and satisfies with some $\lambda_1^* > 0$, any $\lambda_1 \le \lambda_1^*$ and any unit vectors u and ω

$$g(\lambda_1; \omega, u) \le 2\lambda_1^2.$$

The bound (A.12) is derived from [211], Lemma 5.1. Independence of the X_i's yields for $\lambda \le \lambda_1^* N^{1/2}(\mathfrak{n}_0 q^*)^{-1/2}$

$$\log I\!E \exp\Big\{\frac{2\lambda}{\sqrt{\mathfrak{n}_0 q^*}} u^\top \nabla\zeta(\omega)\Big\} \le 2\lambda^2.$$

This means that the condition $(\mathscr{E}D)$ of [211] is verified and the result (A.12) follows from [211], Lemma 5.1. Introduce a random set $A = \{(\lambda/2)\sup_\omega \zeta(\omega) > \mathfrak{z}(\lambda) + \log \epsilon^{-1}\}$. and A^c is its complement. By the Cauchy-Schwartz inequality

$$\begin{aligned} I\!P(A^c) &\le I\!E \exp\Big\{\frac{\lambda}{2}\sup_\omega \zeta(\omega) - \mathfrak{z}(\lambda) - \log \epsilon^{-1}\Big\} \\ &\le \epsilon I\!E^{1/2} \exp\big\{\lambda\zeta(\omega^\circ) - \mathfrak{n}_0 q^* \lambda^2/2\big\} \\ &\quad \times I\!E^{1/2} \exp\big\{\lambda \sup_\omega \zeta(\omega, \omega^\circ) - 2\mathfrak{n}_0 q^* \lambda^2 - \mathfrak{e}_d\big\} \le \epsilon \end{aligned}$$

and the last result follows. □

The result of Lemma 3 can be easily extended to the case of a vector function $q(\omega, x) \in \mathbb{R}^d$:

$$I\!P\Big(\sup_{\omega \in \mathcal{B}_d} \|\zeta(\omega)\|_\infty \geq 2\lambda^{-1}\big[\mathfrak{z}(\lambda) + \log(d/\epsilon)\big] \Big) \leq \epsilon.$$

This fact can be obtained by applying Lemma 3 to each component of the vector $\zeta(\omega)$. The term $\log(d/\epsilon)$ is responsible for the overall deviation probability.

Let now $f(x, \omega)$ be a twice continuously differentiable function of $\omega \in \mathcal{B}_d$ and $x \in \mathbb{R}^d$ such that for every $j \leq d$, $\omega \in \mathcal{B}_d$, and $x \in \mathbb{R}^d$, it holds

$$\begin{aligned} \operatorname{Var}\big[X_j\, f(X,\omega)\big] &\leq f_1^*, \quad \operatorname{Cov}\big[X_j\, \nabla_\omega f(X,\omega)\big] \leq f_1^* I, \\ \operatorname{Var}\left[\frac{\partial}{\partial x_j} f(X,\omega)\right] &\leq f_1^*, \quad \operatorname{Cov}\left[\nabla_\omega \frac{\partial}{\partial x_j} f(X,\omega)\right] \leq f_1^* I, \end{aligned}$$

Then for any $\mathfrak{n}_0 > 1$, there is $\lambda_1^* = \lambda_1^*(\mathfrak{n}_0) > 0$ and for any $\epsilon > 0$, a random set A with $I\!P(A) \geq 1 - \epsilon$ such that on A it holds by Lemma 3

$$\begin{aligned} \sup_{\omega \in \mathcal{B}_d} \big\| I\!E_N[X f(X,\omega)] - I\!E[X f(X,\omega)] \big\|_\infty &\leq \delta_N, \\ \sup_{\omega \in \mathcal{B}_d} \big\| I\!E_N[\nabla_x f(X,\omega)] - I\!E[\nabla_x f(X,\omega)] \big\|_\infty &\leq \delta_N, \end{aligned}$$

where

$$\delta_N = N^{-1/2} \inf_{\lambda \leq \lambda_1^* N^{1/2}} \big\{ 5\mathfrak{n}_0 f_1^* \lambda + 2\lambda^{-1}\big[\mathfrak{e}_d + \log(2d/\epsilon)\big] \big\}.$$

By construction of vectors $\widehat{\gamma}_l$ and $\widehat{\eta}_l$, it holds on A

$$\max_{1 \leq l \leq L} \|\widehat{\gamma}_l - \gamma_l\|_\infty \leq \delta_N, \quad \max_{1 \leq l \leq L} \|\widehat{\eta}_l - \eta_l\|_\infty \leq \delta_N\,.$$

This implies for any $\|c\|_1 \leq 1$

$$\|\widehat{\gamma}(c) - \gamma(c)\|_\infty \leq \delta_N, \quad \|\widehat{\eta}(c) - \eta(c)\|_\infty \leq \delta_N.$$

The constraint $\widehat{\gamma}(\widehat{c}) = 0$ implies $\|\gamma(\widehat{c})\|_\infty \leq \delta_N$, thus

$$\|\gamma(\widehat{c})\|_2 \leq \sqrt{d}\, \delta_N,$$

and by (A.6)

$$\begin{aligned} & \big\| (I - \Pi^*) \widehat{\eta}(\widehat{c}) \big\|_2 \\ & \quad \leq \big\| (I - \Pi^*) \{ \widehat{\eta}(\widehat{c}) - \eta(\widehat{c}) \} \big\|_2 + \big\| (I - \Pi^*) \eta(\widehat{c}) \big\|_2 \\ & \quad \leq \big\| \widehat{\eta}(\widehat{c}) - \eta(\widehat{c}) \big\|_2 + \big\| \Sigma^{-1} \gamma(\widehat{c}) \big\|_2 \\ & \quad \leq \sqrt{d} \big(\delta_N + \big\| \Sigma^{-1} \big\|_2 \delta_N \big). \end{aligned}$$

A.4 Proof of Theorem 6

Theorem 6. *1. Let $\mathscr{S}$ be the convex envelope of the set $\{\pm\widehat{\beta}_j\}$, $j = 1, ..., J$, and let $\mathscr{E}_1(B)$ be an ellipsoid inscribed into $\mathscr{S}$, such that $\mathscr{E}_{\sqrt{d}}(B)$ is $\sqrt{d}$-rounding ellipsoid for*

$\mathscr{S}$. Then for any unit vector $v \perp \mathcal{I}$,

$$v^\top B^{-1} v \le \varrho^2.$$

2. If there is $\mu \in \mathbb{R}^J$ with $\mu_j \ge 0$ and $\sum_j \mu_j = 1$ such that

$$\lambda_m\Big(\sum_j \mu_j \beta_j \beta_j^\top\Big) \ge \lambda^* > 2\varrho^2,$$

where $\lambda_m(A)$ stands for the m-th principal eigenvalue of A, then

$$\lambda_m(B^{-1}) \ge \frac{\lambda^* - 2\varrho^2}{2\sqrt{d}}. \tag{A.13}$$

3. Moreover, let $\widehat{\Pi} \stackrel{\mathrm{def}}{=} \widehat{\Gamma}_m \widehat{\Gamma}_m^\top$ where Γ_m is the matrix of m principal eigenvectors of B^{-1}. Then

$$\|\widehat{\Pi} - \Pi^*\|_2^2 \le \frac{4\varrho^2 d\sqrt{d}}{\lambda^* - 2\varrho^2}.$$

Proof. Let $\mathscr{S}$ stand for the convex envelope of $\{\pm\widehat{\beta}_j\}_{j=1}^J$. As $\mathscr{E}_1(B)$ is inscribed in $\mathscr{S}$, its support function $\xi_{\mathscr{E}_1(B)}(x) = \max_{s \in \mathscr{E}_1(B)} s^\top x$ is majorated by that of $\mathscr{S}$:

$$\xi_{\mathscr{E}_1(B)}(v) \le \xi_{\mathscr{S}}(v) = \max_{j=1,\dots,J} |v^\top \widehat{\beta}_j|, \quad \text{for any } v \in \mathbb{R}^d.$$

Next, the support function of the ellipsoid $\mathscr{E}_1(B)$ is

$$\xi_{\mathscr{E}_1(B)}(v) = (v^\top B^{-1} v)^{1/2},$$

so that the condition $\|\widehat{\beta}_j - \beta_j\|_2 \le \varrho$ implies

$$v^\top B^{-1} v \le \max_{j=1,\dots,J} |v^\top \widehat{\beta}_j|^2 \le \varrho^2,$$

for any $v \perp \mathcal{I}$.

Let us prove the second claim of the proposition. Let Π^* be a projector onto $\mathcal{I}$. By the assumption of the proposition there exist coefficients μ_j with $\sum_j \mu_j \le 1$ such that

$$S \stackrel{\mathrm{def}}{=} \frac{1}{2}\Big[\sum_j \mu_j \beta_j \beta_j^\top - 2\varrho^2 \Pi^*\Big] \succeq 0.$$

This implies (A.13). Now, for any such S and its pseudo-inverse S^+, the ellipsoid, $\mathscr{E}_1^f(S^+)$ with

$$\mathscr{E}_1^f(S^+) = \{x \in \mathcal{I} \mid x^\top S^+ x \le 1\}$$

is inscribed into $\mathscr{S}$. Indeed, the support function $\xi_{\mathscr{E}_1^f(S^+)}(x) = (x^\top Sx)^{1/2}$ of this ellipsoid fulfills for $x \in \mathcal{B}_d$

$$\begin{aligned}\xi_{\mathscr{E}_1^f(S^+)}(x) &\le \left(\sum_j \mu_j\Big[\frac{1}{2}(x^\top\beta_j)^2 - \varrho^2\Big]\right)^{1/2}\\ &\le \left(\sum_j \mu_j\big|x^\top\widehat{\beta}_j\big|^2\right)^{1/2}\\ &\le \max_{1\le j\le J}|x^\top\widehat{\beta}_j| = \xi_{\mathscr{S}}(x),\end{aligned}$$

Now we are done: as the ellipsoid $\mathscr{E}_1^f(S^+)$ is inscribed into $\mathscr{S}$, it is contained in the concentric to $\mathscr{E}_1(B)$ ellipsoid $\mathscr{E}_{\sqrt{d}}(B)$ which covers $\mathscr{S}$.

To show the last statement of the theorem, observe that

$$\mathrm{Tr}\big[(\widehat{\Pi} - \Pi^*)^2\big] = 2(m - \mathrm{Tr}[\Pi^*\widehat{\Pi}]) = 2\mathrm{Tr}\big[(I - \Pi^*)\widehat{\Pi}\big].$$

On the other hand, using the second claim one gets

$$\begin{aligned}\mathrm{Tr}\big[(I - \Pi^*)\widehat{\Pi}\big] &\le (d-m)\sup_{v\perp\mathcal{I}} v^\top\widehat{\Pi}v\\ &\le (d-m)\sup_{v\perp\mathcal{I}}\frac{v^\top B^{-1}v}{\lambda_m(B^{-1})}\\ &\le \frac{2d^{3/2}\varrho^2}{\lambda^* - 2\varrho^2}.\end{aligned}$$

□

A.5 Proof of Theorem 7

Theorem 7. *Let $\mathscr{A}_\epsilon$ be a random set on which*

$$\max_l\|\gamma_l - \widehat{\gamma}_l\|_2 \le \epsilon, \qquad \max_l\|\eta_l - \widehat{\eta}_l\|_2 \le \epsilon.$$

and let β^* denote the "ideal aggregation" $\beta^* = \sum_l c_l^*\,\eta_l$. Then it holds:

$$\begin{aligned}\|\xi - \widehat{\beta}\|_2 &\le \|\xi - \beta^*\|_2 + \epsilon,\\ \|\Pi_{\mathcal{I}}(\xi - \widehat{\beta})\|_2 &\le \|\Pi_{\mathcal{I}}(\xi - \beta^*)\|_2 + (1 + C_1)\epsilon.\end{aligned}$$

Proof. Observe that on $\mathscr{A}_\epsilon$ the solution $c^* = \{c_l^*\}$ of the "ideal" optimization problem fulfills the constraint of the empirical one. Indeed,

$$\Big\|\sum_l c_l^*\,\widehat{\gamma}_l\Big\|_2 = \Big\|\sum_l c_l^*(\widehat{\gamma}_l - \gamma_l)\Big\|_2 \le \epsilon.$$

Therefore,

$$\Big\|\xi - \sum_l \widehat{c}_l\,\widehat{\eta}_l\Big\|_2 \le \Big\|\xi - \sum_l c_l^*\,\widehat{\eta}_l\Big\|_2.$$

because $\widehat{c}$ is the minimizer of such norm. It remains to mention that on $\mathscr{A}_\epsilon$

$$\Big\|\xi - \sum_l c_l^*\,\widehat{\eta}_l\Big\|_2 - \Big\|\xi - \sum_l c_l^*\,\eta_l\Big\|_2 \le \Big\|\sum_l c_l^*(\widehat{\eta}_l - \eta_l)\Big\|_2 \le \epsilon$$

and hence, for $\widehat{\beta} = \sum_l \widehat{c}_l\,\widehat{\eta}_l$

$$\big\|\xi - \widehat{\beta}\big\|_2 \le \big\|\xi - \beta^*\big\|_2 + \epsilon$$

and the first assertion follows. For second one use additionally that $(I - \Pi_{\mathcal{I}})\beta^* = 0$ and $\|(I - \Pi_{\mathcal{I}})\widehat{\beta}\|_2 \le C_1\epsilon$ on $\mathscr{A}_\epsilon$, see the proof of Theorem 4. □

A.6 Proof of Theorem 9

Assumption 8. *Suppose that there are vectors $c_1, ..., c_{\overline{m}}$, $m \le \overline{m} \le L$ such that $\|c_k\|_1 \le 1$ and $Gc_k = 0$, $k = 1, ..., \overline{m}$, and non-negative constants $\mu^1, \ldots, \mu^{\overline{m}}$ such that*

$$\Pi^* \preceq \sum_{k=1}^{\overline{m}} \mu^k U c_k c_k^T U^T. \tag{A.14}$$

We denote $\mu^* = \mu^1 + \ldots + \mu^{\overline{m}}$.

Theorem 9. *Let Assumption 8 hold. Then an optimal solution $\widehat{P}$ of (5.9) satisfies*

$$\mathrm{Tr}\Big[(I - \widehat{P})\Pi^*\Big] \le 4\mu^*\delta^2(\lambda_{\min}^{-1}(\Sigma) + 1)^2. \tag{A.15}$$

Further, if $\widehat{\Pi}$ is the projector onto the subspace spanned by m principal eigenvectors of $\widehat{P}$, then

$$\|\widehat{\Pi} - \Pi^*\|_2^2 \le \frac{8\mu^*\delta^2(\lambda_{\min}^{-1}(\Sigma) + 1)^2}{1 - 4\mu^*\delta^2(\lambda_{\min}^{-1}(\Sigma) + 1)^2} \tag{A.16}$$

(here $\|A\|_2 = \Big(\sum_{i,j} A_{ij}^2\Big)^{1/2} = \big(\mathrm{Tr}[A^TA]\big)^{1/2}$ is the Frobenius norm of A).

Proof. Let $X \in \mathbb{R}^{L\times L}$ be positive semidefinite with $|X|_1 \le 1$ and let Y be a symmetric square root of X, so that $X = Y^2$. If we denote y_i, $i = 1, .., L$ the columns of Y, the fact that $|X|_1 \le 1$ implies that

$$\sum_{1\le i,j\le L} |y_i^T y_j| \le 1.$$

We make here one trivial though useful observation: for any matrix $A \in \mathbb{R}^{d\times L}$, when denoting as above a_i the columns of A, we have

$$\begin{aligned}\mathrm{Tr}[A^TAX] &= \|AY\|_2^2 = \sum_{j=1}^L \Big|\sum_{i=1}^L a_i Y_{ij}\Big|^2 = \sum_{j=1}^L \Big[\sum_{i=1}^L\sum_{k=1}^L a_i^T a_k Y_{ij}Y_{jk}\Big] \\ &= \sum_{i=1}^L\sum_{k=1}^L a_i^T a_k \Big[\sum_{j=1}^L Y_{ij}Y_{jk}\Big] \le |A|_2^2 \sum_{i=1}^L\sum_{k=1}^L |y_i^T y_k| \le |A|_2^2. \end{aligned}\tag{A.17}$$

where for a matrix $A \in \mathbb{R}^{d\times L}$ with columns a_i, $i = 1, ..., L$, $|A|_2$ stands for the maximal column norm:

$$|A|_2 = \max_{1\le i\le L} \|a_i\|_2.$$

We can rewrite the problem (5.9) using $Y = X^{1/2}$, so that the objective function $\widehat{f}(X,P)$ of (5.9) becomes

$$\widehat{g}(Y,P) = \|(I-P)^{1/2}\widehat{U}Y\|_2^2.$$

Let now $(\widehat{X},\widehat{P})$ be the saddle point of (5.9). If $\widehat{f}(X,P)$ is the objective of (5.9) then

$$\widehat{f}(X,\widehat{P}) \le [\widehat{f}_* \equiv \widehat{f}(\widehat{X},\widehat{P})] \le \widehat{f}(\widehat{X},P),$$

for any feasible P and X. We denote $\widehat{Y} = \widehat{X}^{1/2}$.

Lemma 4. *Let $\widehat{P}$ be an optimal solution to (5.9), then*

$$\max_c \left\{ \|(I-\widehat{P})^{1/2}Uc\|_2 \mid \|c\|_1 \le 1,\ Gc = 0 \right\} \le 2(\lambda_{\min}^{-1}(\Sigma)+1)\delta. \tag{A.18}$$

Proof. To show (A.18) we write

$$\begin{aligned}
&\max_c \left\{ \|(I-\widehat{P})^{1/2}Uc\|_2 \mid \|c\|_1 \le 1,\ Gc = 0 \right\} \\
&\le \max_Y \left\{ \|(I-\widehat{P})^{1/2}UY\|_2 \,\middle|\, |Y^2|_1 \le 1,\ GY = 0 \right\} \\
&\le \max_Y \left\{ \|(I-\widehat{P})^{1/2}\widehat{U}Y\|_2 \,\middle|\, |Y^2|_1 \le 1,\ GY = 0 \right\} \\
&\quad + \max_Y \left\{ \|(I-\widehat{P})^{1/2}(\widehat{U}-U)Y\|_2 \,\middle|\, |Y^2|_1 \le 1,\ GY = 0 \right\} \\
&\le \max_Y \left\{ \|(I-\widehat{P})^{1/2}\widehat{U}Y\|_2 \,\middle|\, |Y^2|_1 \le 1,\ \|\widehat{G}Y\|_2 \le \delta \right\} + \delta \\
&= \|(I-\widehat{P})^{1/2}\widehat{U}\widehat{Y}\|_2 + \delta \le \|(I-\Pi^*)^{1/2}\widehat{U}\widehat{Y}\|_2 + \delta \\
&\le \|(I-\Pi^*)^{1/2}U\widehat{Y}\|_2 + 2\delta.
\end{aligned}$$

On the other hand, as $\|\widehat{G}\widehat{Y}\|_2 \le \delta$, we get

$$\|G\widehat{Y}\|_2 \le \|\widehat{G}\widehat{Y}\|_2 + \|(\widehat{G}-G)\widehat{Y}\|_2 \le \delta + |\widehat{G}-G|_2 \le 2\delta,$$

and by A.6,

$$\|(I-\Pi^*)U\widehat{Y}\|_2 \le 2\lambda_{\min}^{-1}(\Sigma)\delta.$$

This implies (A.18). □

Proof of theorem 9. We have due to (A.14) and (A.18):

$$\begin{aligned}
\mathrm{Tr}\left[(I-\widehat{P})\Pi^*\right] &= \mathrm{Tr}\left[(I-\widehat{P})^{1/2}\Pi^*(I-\widehat{P})^{1/2}\right] \\
&\le \sum_{k=1}^{\overline{m}} \mu^k \mathrm{Tr}\left[(I-\widehat{P})^{1/2}Uc_kc_k^TU^T(I-\widehat{P})^{1/2}\right] = \sum_{k=1}^{\overline{m}} \mu^k \|(I-\widehat{P})^{1/2}Uc_k\|_2^2 \\
&\le \sum_{k=1}^{\overline{m}} \mu_k \max_c \left\{ \|(I-\widehat{P})^{1/2}Uc\|_2^2 \mid \|c\|_1 \le 1,\ Gc = 0 \right\} \\
&= 4\mu^*\delta^2(\lambda_{\min}^{-1}(\Sigma)+1)^2,
\end{aligned}$$

what is (A.15).

Now let $\widehat{\lambda}_j$ and $\widehat{\theta}_j$, $j = 1,\dots,d$ be respectively the eigenvalues and the eigenvectors of $\widehat{P}$. Assume that $\widehat{\lambda}_1 \ge \widehat{\lambda}_2 \ge \ldots \ge \widehat{\lambda}_d$. Then $\widehat{P} = \sum_{j=1}^d \widehat{\lambda}_j\widehat{\theta}_j\widehat{\theta}_j^\top$ and $\widehat{\Pi} = \sum_{j=1}^m \widehat{\theta}_j\widehat{\theta}_j^\top$. Therefore,

on the one hand,

$$\begin{aligned}\mathrm{Tr}[\widehat{P}\Pi^*] &\le \sum_{j\le m}\widehat{\lambda}_j\mathrm{Tr}[\widehat{\theta}_j\widehat{\theta}_j^\top\Pi^*] + \widehat{\lambda}_m\sum_{j\le m}\mathrm{Tr}[\widehat{\theta}_j\widehat{\theta}_j^\top\Pi^*]\\ &= \sum_{j\le m}\widehat{\lambda}_j\mathrm{Tr}[\widehat{\theta}_j\widehat{\theta}_j^\top\Pi^*] + \widehat{\lambda}_m\mathrm{Tr}[(I-\widehat{\Pi})\Pi^*]\\ &= \sum_{j\le m}(\widehat{\lambda}_j-\widehat{\lambda}_m)\mathrm{Tr}[\widehat{\theta}_j\widehat{\theta}_j^\top\Pi^*] + m\widehat{\lambda}_m.\end{aligned}$$

Since $\mathrm{Tr}[\widehat{\theta}_j\widehat{\theta}_j^\top\Pi^*] = |\Pi^*\widehat{\theta}_j|^2 \le 1$, we get $\mathrm{Tr}[\widehat{P}\Pi^*] \le \sum_{j\le m}\widehat{\lambda}_j$. Taking into account the relations $\sum_{j\le d}\widehat{\lambda}_j \le m$, $\mathrm{Tr}[\Pi^*] = m$ and $(1-\widehat{\lambda}_{m+1})(I-\widehat{\Pi}) \preceq I-\widehat{P}$, we get

$$\lambda_{m+1} \le m - \sum_{j\le m}\widehat{\lambda}_j \le \mathrm{Tr}[(I-\widehat{P})\Pi^*] \le 4\mu^*\delta^2(\lambda_{\min}^{-1}(\Sigma)+1)^2,$$

and, therefore,

$$\mathrm{Tr}[(I-\widehat{\Pi})\Pi^*] \le \frac{4\mu^*\delta^2(\lambda_{\min}^{-1}(\Sigma)+1)^2}{1-4\mu^*\delta^2(\lambda_{\min}^{-1}(\Sigma)+1)^2}.$$

Now we are done, because due to $\mathrm{Tr}[\widehat{\Pi}] = \mathrm{Tr}[\Pi^*] = m$,

$$\|\widehat{\Pi}-\Pi^*\|_2^2 = \mathrm{Tr}[\widehat{\Pi}^2 - 2\widehat{\Pi}\Pi^* + (\Pi^*)^2] = 2m - 2\mathrm{Tr}[\widehat{\Pi}\Pi^*] = 2\mathrm{Tr}[(I-\widehat{\Pi})\Pi^*],$$

and we arrive at (A.16).

□

Appendix B

Statistical Tests

In this section we shortly report the statistical tests on normality used the dimension reduction step of the convex-projection approach to SNGCA.

In order to detect a significant asymmetry in the distribution of the original data projected on the semi-axis of the numerical approximation of the rounding ellipsoid $\mathcal{E}_{\sqrt{d}}$ we use the K^2-test according to D'Agostino-Pearson [244]. The D'Agostino-Pearson test computes how far the empirical skewness and kurtosis of the given data distribution differs from the value expected with a Gaussian distribution. The test statistic is approximately distributed according to the χ^2_2-distribution and its empirical data counterpart is given by

$$\begin{aligned}\widehat{K}^2 &= \mathcal{Z}^2(\sqrt{b_1}) + \mathcal{Z}^2(b_2)\\ \sqrt{b_1} &= \tfrac{1}{N}\sum_{i=1}^{N}\left(\tfrac{X_i-\mu}{\sigma}\right)^3\\ b_2 &= \tfrac{1}{N}\sum_{i=1}^{N}\left(\tfrac{X_i-\mu}{\sigma}\right)^4\end{aligned}$$

Here μ denotes the empirical mean, σ the empirical standard deviation of the data and $\mathcal{Z}(\cdot)$ denotes a normalizing transformations of skewness and kurtosis. The test is more powerful with respect to an asymmetry of a distribution.

Furthermore we use the EDF-test according to Anderson-Darling [8] with the modification of Stephens [215]: Let F_N be the empirical cumulative distribution function and F the assumed theoretical cumulative distribution function. The test statistics $\mathcal{T}$ measures the quadratic deviations between F_N and F:

$$\mathcal{T} = \int_{\mathbb{R}} [F_N(x) - F(x)]^2 \nu(x)\, dF$$

where $\nu(x)$ is the weighting function $\nu(x) = [F_N(x)(1 - F_N(x))]^{-1}$. In sum the data counterpart of $\mathcal{T}$ is given by

$$\widehat{\mathcal{T}} = c\left(-N - \sum_{i=1}^{N} \tfrac{[2i-1]}{N}\left[\log(F\left(\tfrac{X_i-\mu}{\sigma}\right) + \log(1 - F\left(\tfrac{X_{N-i+1}-\mu}{\sigma}\right)\right]\right)$$

$$c = \left(1 + \tfrac{0.75}{N} + \tfrac{2.25}{N^2}\right)$$

Again μ is the empirical mean and σ the empirical standard deviation of the data. We compute $\widehat{\mathcal{T}}$ to detect deviations from normality in the tails of the projected distributions. The test is rejected if $\widehat{\mathcal{T}}$ exceeds a critical value cv specific for a given level of significance:

α :	0.10	0.05	0.025	0.01	0.005
cv :	0.631	0.752	0.873	1.035	1.159

The last test, applied to the projected data is the Shapiro-Wilks test [204] based on a regression strategy in the version given by Royston [189; 190]:

$$\begin{aligned} W &= \frac{\left(\left[1-\frac{b^2}{\sigma^2(N-1)}\right]^{\lambda}-\mu\right)}{\sigma} \quad \sim \mathcal{N}(0,1) \\ b &= \sum_{i=1}^{N/2} a_{N-i+1}(X_{N-i+1}-x_i) \\ (a_1,\ldots,a_N) &= \frac{m^{\top}\Sigma^{-1}}{(m^{\top}\Sigma^{-1^{\top}}\Sigma^{-1}m)^{1/2}} \end{aligned}$$

In this test $m = (m_1, \ldots, m_n)$ denote the expected values of standard normal order statistics for a sample of size N and Σ is the corresponding covariance matrix.

Bibliography

[1] D. Achlioptas. *Symposion on Principles of Database Systems.*, chapter Database-friendly random projections., pages 274–281. 2001.

[2] Q. Zhao ad S.E. Karisch, F. Rendl, and H. Wolkowicz. Semidefinite programming relaxations for the quadratic assignment problem. *J. Comb. Optim.*, 2(71-109), 1998.

[3] H. Akaike. *Proceedings of the Second International Symposium on Information Theory Budapest*, chapter Information theory and an extension of the maximum likelihood principle., pages 267–281. B. N. Petrov (ed.), Akademiai Kiado, 1973.

[4] A.Krause and V. Liebscher. Multimodal projection pursuit using the dip statistic. *submitted*.

[5] M.P. Allen. *Computational Soft Matter: From Synthetic Polymers to Proteins*, chapter Introduction to Molecular Dynamics Simulation, pages 1–28. John von Neumann Institute for Computing, 2004.

[6] A. Amadei, A.B.M. Linssen, and H.J.C Berendsen. Essential dynamics on proteins. *Proteins*, 17:412–425, 1993.

[7] S. Amari. Natural gradient works efficiently in learning. *Neural Computation*, 10(2):251–276, 1998.

[8] F.J. Anscombe and W.J. Glynn. Distribution of kurtosis statistic for normal statistics. *Biometrika*, 70(1):227–234, 1983.

[9] L. Arnold. *Stochastic differential equations : theory and applications.* Wiley, New York, 1974.

[10] M. Attias. Independent factor analysis. *Neural Computation*, 11(2):803–851, 99.

[11] K. Ball. *An Elementary Introduction to Convex Geometry.* Mathematical Sciences Research Institute, 1977.

[12] D. J. Bartholomew. *Latent Variable Models and Factor Analysis.* Charles Griffin and Company Ltd., London, 1987.

[13] A. T. Basilevsky. *Statistical Factor Analysis and Related Methods: Theory and Applications.* Wiley and Sons, New York, 1994.

[14] C. Beck and G. Roepstorff. From dynamical systems to the langevin equation. *Physica A*, 145:1–14, 1987.

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

[16] A. Ben-Tal and A. Nemirovski. Non-euclidean restricted memory level method for large-scale convex optimization. *Mathematical Programming: Series A and B*, 102(3):407–456, 2005.

[17] Yoshua Bengio, Olivier Delalleau, and Nicolas Le Roux. The curse of highly variable functions for local kernel machines. In *In Advances in Neural Information Processing Systems 18*, page 2006. MIT Press, 2006.

[18] Bertsekas. *Nonlinear Programming.* Athena Scientific, 1999.

[19] K. Beyer, J. Goldstein, R. Ramakrishnan, and U. Shaft. *When is nearest neighbor meaningful?* In ICDT Conference, 1999.

[20] Patrick Billingsley. *Probability and Measure.* Wiley-Interscience, 3rd edition, 1995.

[21] Jeff A. Bilmes. A gentle tutorial of the EM algorithm and its application to parameter estimation for Gaussian Mixture and Hidden Markov Models. Technical report, International Computer Science Institute, Berkeley, 1998.

[22] C. M. Bishop. *Advances in Neural Information Processing Systems*, volume 11, chapter Bayesian PCA., pages 382–388. M. S. Kearns and S. A. Solla and D. A. Cohn (eds.) Cambridge MA, 1999.

[23] C.M. Bishop. *Neural Networks for Pattern Recognition.*, volume section 1.4. Oxford University Press, 1994.

[24] C.M. Bishop. *Pattern Recognition and Maschine Learning.* Information Science and Statistics. Springer, 2006.

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

[26] I. Borg and P. Groenen. *Modern multidimensional scaling, Theory and applications.* Springer-Verlag, New York, 1997.

[27] Stephen Boyd and Lieven Vandenberghe. *Convex Optimization.* Cambridge University Press, 2004.

[28] L. M. Bregman. A relaxation method of finding a common points of convex sets and its application to the solulation of problems in convex programming. *U.S.S.R. Comput. Math. Math. Phys.*, 7:620?631, 1967.

[29] Pierre Brémaud. *Markov Chains.* Springer, 2nd edition edition, 2008.

[30] B.R. Brooks, R.E. Bruccoleri, B.D. Olafson, D.J. States, S. Swaminathan, and M. Karplus. Charmm: A programm for macromelacular energy minimazation and dynamic calculations. *J. Comp. Chem.*, 4:187–217, 1983.

[31] Ioan Buciu and Ioan Nafornita. Linear and nonlinear dimensionality reduction techniques. *Journal of Studies in Informatics and Control*, 2008.

[32] E. Bura and R. D. Cook. Estimating the structural dimension of regressions via parametric inverse regression. *J. Roy. Statist. Soc. Ser. B*, 63(393-410), 2001.

[33] M-Y. Cheng and P. Hall. Calibrating the excess mass and dip tests of modality. *Journal of the Royal Statistical Society B*, 60(3):579–589, 1998.

[34] M-Y. Cheng and P. Hall. Mode testing in difficult cases. *Annals of Statistics*, 27:1294–1315, 1999.

[35] Y. Cheng. Mean shift, mode seeking, and clustering. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 17(790-799), 1995.

[36] John D Chodera, Nina Singhal, Vijay S Pande, Ken A Dill, and William C Swope. Automatic discovery of metastable states for the construction of Markov models of macromolecular conformational dynamics. *Journal of Computational Chemistry*, 126(15):155101, 2007.

[37] Moody Chu, , Moody Chu, and Robert Plemmons. Nonnegative matrix factorization and applications. *Bulletin of the International Linear Algebra Society*, 34:2–7, 2005.

[38] A. Cichocki and A. Amari. *Adaptive Blind Signal and Image Processing.* John Wiley and Sons, 1999.

[39] William T. Coffey, Yu. P. Kalmykov, and J. T. Waldron Waldron. *The Langevin Equation: With Applications in Physics, Chemistry and Electrical Engineering.* World Scientific, 1996.

[40] F. E. Cohen. Protein misfolding and prion diseases. *Journal of Molecular Biology*, 293(2):313–320, Oct 1999.

[41] P. Comon. Independent component analysis, a new concept? *Signal Processing*, (36):287–314, 1994.

[42] T. Conrad, A. Leichtle, A. Hagehülsmann, and E. Diederichs. Beating the noise: New statistical methods for detecting signals in maldi-tof spectra below noise level. *Lecture Notes in Computer Science*, 4216:119–128, 2006.

[43] R.D. Cook. Principal hessian directions revisited. *J. Am. Statist. Ass.*, 93:85–100, 1998.

[44] T.M. Cover and J.A. Thomas. *Elements of Information Theory.* Wiley Series in Telecommunications. Wiley and Sons, New York, 1991.

[45] T. F. Cox and M. A. A. Cox. *Multidimensional Scaling.* Chapman & Hall, London, 1994.

[46] A.S. Dalalyan, A. Juditsky, and V. Spokoiny. A new algorithm for estimating the effective dimension-reduction subspace. *The Journal of Machine Learning Research*, 9:1647–1678, 2008.

[47] S. Dasgupta and A. Gupta. An elementary proff of the johnson-lindenstrauss lemma. Technical Report 99-006, International Computer Science Institute Berkeley, Ca., 1999.

[48] J. Dattorro. *Convey Optimization and Euclidean Distance Geometry.* 2005.

[49] E.B. Davies. *One-parameter semigroups.* Academic Press, London, 1980.

[50] P. Dayan, G. E. Hinton, R. M. Neal, and R. S. Zemel. The helmholtz machine. *Neural Computation*, 7(5):889–904, 1995.

[51] A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society*, 39(1-38), 1977.

[52] P. Deuflhard, M. Dellnitz, O. Junge, and Ch. Schütte. Computation of essential molecular dynamics by subdivision techniques. In P. Deuflhard, J. Hermans, B. Leimkuhler, A. Mark, S. Reich, and R.D. Skeel, editors, *Computational Molecular Dynamics: Challenges, Methods, Ideas*, volume 4 of *Lecture Notes in Computational Science and Engineering*, pages 94–111. Springer, Berlin, 1999.

[53] P. Deuflhard and M. Weber. Robust Perron cluster analysis in conformation dynamics. *Linear Algebra and its Applications*, 398:161–184, 2005.

[54] Peter Deuflhard, Wilhelm Huisinga, Alexander Fischer, and Christof Schütte. Identification of almost invariant aggregates in reversible nearly uncoupled Markov chains. *Linear Algebra and its Applications*, 315:39–59, 2000.

[55] L. Devroye, L. Györfi, and G. Lugosi. *A Problabilistic Theory of Pattern Recognition.* Springer, 1996.

[56] P. Diaconis and D. Friedman. Asymptotics of graphical projection pursuit. *Annals of Statistics*, 12(3):793–815, 1984.

[57] K.I. Diamantaras and S.Y. Kung. *Principal Component Neural Networks.* John Wiley, 1996.

[58] E. Diederichs, A. Juditski, V. Spokoiny, and C. Schuette. Sparse nongaussian component analysis. *IEEE Transact. Inform. Theory*, (7):5249–5262, 2009.

[59] C. Domeniconi and D. Gunopulos. An efficient approach for approximating multidimensional range queries and nearest neighbor classiffication in large datasets. *In Proc. 18th International Conf. on Machine Learning*, pages 98–105, 2001.

[60] D.L. Donoho. Unconditional bases are optimal bases for data compression and for statistical estimation. *Appl. Computational Harmonic analysis*, 1:100–115, 1993.

[61] D.L. Donoho. High-dimensional data analysis: the curses and blessings of dimensionality. *Aide-Memoire of the lecture in AMS conference "Math challenges of 21st Centrury"*, 2000.

[62] D.L. Donoho. Sparse components of images and optimal atomic decomposition. *Constructive Approximation*, 17:353–382, 2001.

[63] R. Durbin, R. Szeliski, and A. Yuille. An analysis of the elastic net approach to the traveling salesman problem. *Neural Computation*, 326:689–691, 1989.

[64] R. Elber and M. Karplus. Multiple conformational states of proteins: a molecular dynamics analysis of Myoglobin. *Science*, 235(4786):318–321, 1987.

[65] E. Erwin, K. Obermayer, and K.J. Schulten. Self-organizing maps: ordering, convergence properties and energy functions. *Biological Cybernetics*, 67(1):47–55, 1992.

[66] P.F. Evangelista, M.J. Embrechts, and B.K. Szymanski. Taming the curse of dimensionality in kernels and novelty detection. In A. Abraham, B.d. Baets, M. Köppen, and B. Nickolay, editors, *Applied soft CcomputingG Technologies: The Challenge of Complexity*, volume 14 of *Advances in Soft Computing*. Springer Verlag, Berlin, 2006.

[67] B. S. Everitt. *An Introduction to Latent Variable Models.* Monographs on Statistics and Applied Probability. Chapman and Hall, London,, 1984.

[68] F. Faccinei and J.-S. Pang. *Finite-Dimensional Variational Inequalities And Complementarity Problems.*, volume I and II of *Springer Series in Operations Research.* Springer, 2003.

[69] Jianqing Fan and I. Gijbels. *Local Polynomial Modelling and Its Applications.* Chapman and Hall, 1996.

[70] Alexander Fischer, Sonja Waldhausen, Illia Horenko, Eike Meerbach, and Christof Schütte. Identification of biomolecular conformations from incomplete torsion angle observations by hidden Markov models. *Journal of Computational Chemistry*, 28(15):2453–2464, 2007.

[71] H. Frauenfelder and B. McMahon. Dynamics and function of proteins: the search for general concepts. *Proc Natl Acad Sci U S A*, 95(9):4795–4797, Apr 1998.

[72] H. Frauenfelder and B.H. McMahon. Energy landscape and fluctations in proteins. *Ann. Phys. (Leipzig)*, 9:655–667, 2000.

[73] J. H. Friedman. Exploratory projection pursuit. *J. Amer. Stat. Assoc.*, 82:249–266, 1987.

[74] J. H. Friedman and W. Stuetzle. Projection pursuit regression. *J. Amer. Stat. Assoc.*, 76:817–823, 1981.

[75] K. Fukunaga. *Statistical pattern recognition.* Academic Press, 2nd edition, 1990.

[76] J.E. Gentle. *Elements of Computational Statistics.* Statistics and Computing. Springer, New York, 2002.

[77] B. Georgescu, I.Shimshoni, and P. Meer. *Mean Shift Based Clustering in High Dimensions: A Texture Classification Example*, pages 456–463. 2003.

[78] N.E. Goljandina, V.V. Nekrutkin, and A.A. Zhigljavsky. *Analysis of Time Series Structure: SSA and related technique.* Chapman and Hall (CRS), Boca Raton, 2001.

[79] G. H. Golub and C. F. Van Loan. *Matrix Computations.* Johns Hopkins Studies in Mathematical Sciences. The Johns Hopkins University Press, 3rd edition, 1996.

[80] Yu. Golubev. Asymptotic minimax estimation of regression function in additive model. *Problems Inform. Transmission*, 28(2):3–15, 1992.

[81] I. Guyon and A. Elisseeff. An introduction to variable and feature selection. *Journal of Machine Learning Research*, 3:1157–1182, 2003.

[82] P. Hall. Projection pursiut methods. *Ann. Statist.*, 17:589–605, 1989.

[83] J. A. Hartigan. The dip test of unimodality. *Applied Statistics*, 34(3):320–325, 1985.

[84] J. A. Hartigan. *Data Analysis: Scientific Modeling and Practical Application.*, chapter Testing for antimodes., pages 169–181. W. Gaul and O. Optiz and M. Schader, 2000.

[85] J. A. Hartigan and P. M. Hartigan. The dip test of unimodality. *Annals of Statistics Volume*, 13(1):70–84, 1985.

[86] T. Hastie, J. Taylor, R. Tibshirani, and G. Walther. Forward stagewise regression and the monotone lasso. *preprint*, 2006.

[87] T. J. Hastie and W. Stuetzle. Principal curves. *J. Amer. Stat. Assoc.*, 84:502–516, 1989.

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

[89] T. J. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning.* Springer Series in Statistcs. Springer, 2001.

[90] Bingsheng He and Li-Zhi Liao. Improvements of some projection methods for monotone nonlinear variational inequalities. *Journal of Optimization Theory and Applications*, 112:111–128, 2002.

[91] R. Hecht-Nielsen. *Neurocomputing.* Addison-Wesley Reading MA, 1991.

[92] R. Hecht-Nielsen. *Computational Intelligence: Imitating Life.*, chapter Context vectors: general purpose approximate meaning representations self-organized from raw data., pages 43–56. J.M. Zurada and R.J. Marks II, and C.J. Robinson, IEEE Press, 1994.

[93] J.B. Hiriart-Urruty. *Convex Analysis and Minimuzation Algorithms*, volume I and II of *A Series of Comprehensive Studies in Mathematics.* Springer, 1993.

[94] B.L. Holian and W.G. Hoover. Numerical test of the Liouville equation. *Phys. Rev. A*, 34(5):4229–4239, 1986.

[95] I. Horenko, E. Dittmer, F. Lankas, J. Maddocks, P. Metzner, and C. Schütte. Macroscopic dynamics of complex metastable systems: Theory, algorithms, and application to B-DNA. *SIAM Journal on Applied Dynamical Systems*, 7:532–560, 2008.

[96] I. Horenko, E. Dittmer, and A. Fischer Ch. Schütte. Automated model reduction for complex systems exhibiting metastability. *SIAM Multiscale Modeling and Simulation*, 5(3):802–827, 2006.

[97] I. Horenko, E. Dittmer, and C. Schütte. Reduced stochastic models for complex molecular systems. *Comp. Vis. Sci.*, 9(2):89–102.

[98] I. Horenko, E. Dittmer, C. Schütte, and A. Fischer. Automated model reduction for complex systems exhibiting metastability. *SIAM Multiscale Modeling and Simulation*, 2005.

[99] I. Horenko and Ch. Schütte. Likelihood-based estimation of multidimensional langevin models and its application to biomolecular dynamics. *Submitted to Mult. Mod. Sim.*, 2006.

[100] Illia Horenko, Evelyn Dittmer, A. Fischer, and Ch. Schütte. Automated model reduction for complex systems exhibiting metastability. *Multiscale Modeling and Simulation*, 5(3):802–827, 2006.

[101] J.L. Horowitz and V.G. Spokoiny. An adaptive, rate-optimal test of a parametric model against a nonparametric alternative. *Econometrica*, 69:599–631, 2001.

[102] M. Hristache, A. Juditsky, J. Polzehl, and V. Spokoiny. Structure adaptive approach for dimension reduction. *Ann. Statist.*, 29(6):1537–1566, 2001.

[103] M. Hristache, A. Juditsky, and V. Spokoiny. Direct estimation of the index coefficient in a single-index model. *Ann. Statist.*, 29(3):595–623, 2001.

[104] S. Dasgupta D.J. Hsu and N. Verma. A concentration theorem for projections. Twenty-Second Conference on Uncertainty in Artificial Intelligence (UAI).

[105] P. J. Huber. *Robust Statistics.* Wiley Series in Probability and Mathematical Statistics. John Wiley and Sons, 1981.

[106] P. J. Huber. Projection pursuit. *The Annals of Statistics*, 13(2):435–475, 1985.

[107] W. Huisinga. *Metastability of Markovian systems: A transfer operator based approach in applications to molecular dynamics.* PhD Thesis, Fachbereich Mathematik und Informatik, Freie Universität Berlin, 2001.

[108] W. Huisinga, S. Meyn, and C. Schütte. Phase transitions and metastability in Markovian and molecular systems. *ANNAP*, 14(1):419–458, 2002.

[109] P.H. Hünenberger. Thermostat algorithms for molecular dynamics simulations. In C. Holm and K. Kremer, editors, *Advanced Computer Simulation: Approaches for Soft Matter Sciences I*, volume 173 of *Advances in Polymer Science*, pages 105–149. Springer, Berlin, 2005.

[110] A. Hyvärinen. Survey on independent component analysis. *Neural Computing Surveys*, 2:94–128, 1999.

[111] A. Hyvärinen, J. Karhunen, and E. Oja. *Independent Component Analysis.* Wiley Series in Probability and Mathematical Statistics. John Wiley and Sons, 2001.

[112] A. Hyvärinen and E. Oja. Independent component analysis: Algorithms and applications. *Neural Networks*, 13:411–420, 1999.

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

[114] F. John. *Extremum problems with inequalities as subsidiary conditions.*, volume Reprinted in: Fritz John, Collected Papers Volume 2 of *Birkhäuser, Boston*, pages 543–560. J. Moser, 1985.

[115] G. H. John, R. Kohavi, and K. P Pfleger. *Proceedings of the 11th International Conference on Machine Learning.*, chapter Irrelevant features and the subset selection problem., pages 121–129. Morgan Kaufmann, 1994.

[116] W.B. Johnson and J. Lindenstrauss. *Conference in modern analysis and probability.*, volume 26 of *Contemporary Mathematics*, chapter Extensions of Lipshitz mapping into Hilbert space., pages 189–206. Amer. Math. Soc., 1984.

[117] I.T. Jolliffe. *Principal Component Analysis.* Springer Series in Statistics. Springer, Berlin and New York, 2nd edition, 2002.

[118] H.C. Thode Jr. *Testing for Normality.* Marcel Dekker, New York., 2002.

[119] A. Juditsky, A. Nemirovski, and C. Tauvel. Solving variational inequalites with stochastic mirror-prox-algorithm. *preprint*, 2008.

[120] S. Kaski. Dimensionality reduction by random mapping: fast similarity computation for clustering. *Proc. IEEE International Joint Conference on Neural Networks*, 1:413–418, 1998.

[121] M. Kendall. *A Course in the Geometry of n Dimensions.* Charles Griffin and Company Ltd London, 1961.

[122] L. G. Khachiyan and M. J. Todd. On the complexity of approximating the maximal inscribed ellipsoid for a polytope. *Mathematical Programming*, 61:137–159, 1993.

[123] L.G. Khachiyan. Rounding of polytopes in the real number model of computation. *Mathematics of Operations Research*, 21(2):307–320, 1996.

[124] K.C. Kiwiel. An aggregate subgradient method for non-smooth convex minimization. *Mathematical Programming*, 27:320–341, 1983.

[125] K.C. Kiwiel, T. Larson, and P.O. Lindberg. The efficiency of ballstep subgradient level methods for convex optimization. *Mathematics of Operations Research*, 24(237-254), 1999.

[126] R. Kohavi and G. John. *Feature Extraction, Construction and Selection: A Data Mining Perspective.*, chapter The wrapper approach. H. Liu and H. Motoda, Springer Verlag, 1998.

[127] T. Kohonen. *Self-Organizing Maps.* Springer-Verlag, 1995.

[128] T. Kohonen. Self organization of massive document collection. *IEEE Transactions on Neural Networks*, 11(3):574–585, May 2000.

[129] T. Kohonen. The self organizing map. *Proc. IEEE*, 78(9):1464–1480, 90.

[130] I. Kotsia, S. Zafeiriou, and I. Pitas. *Biometrics and Identity Management*, volume 5372 of *Lecture Notes in Computer Science*, chapter Discriminant Non-negative Matrix Factorization and Projected Gradients for Frontal Face Verification., pages 82–90. Springer Berlin, 2008.

[131] S. B. Kotsiantis. Supervised machine learning: A review of classification techniques. *Informatica*, 31:249–268, 2007.

[132] Edwin K.P.Chong and Stanislaw H.Zak. *An Introduction to Optimization.* John Wiley and Sons, 2nd edition, 2001.

[133] J. B. Kruskal. Non metric multidimensional scaling : a numerical method. *Psychometrika*, 19:115–129, 1964.

[134] W. J. Krzanowski. *Principles of Multivariate Analysis: A User's Perspective.*, volume 3 of *Oxford Statistical Science Series.* Oxford University Press, New York, Oxford, 1988.

[135] Erik Learned-Miller and III. John W. Fisher. Ica using spacings estimates of entropy. *Journal of Machine Learning Research*, 4:1271–1295, 2003.

[136] B. Leimkuhler and S. Reich. Symplectic integration in constrained Hamiltonian systems. *Math. Comp.*, 63:589–605, 1994.

[137] B. Leimkuhler and R.D. Skeel. Symplectic numerical integrators in constrained Hamiltonian systems. *J. Comput. Phys.*, 112:117–125, 1994.

[138] C. Lemarechal and C. Sagastizabal. Practical aspects of the moreau yosida regularization: Theoretical prelimiaries. *SIAM Journal on Optimization*, 7(2):867–895, 1997.

[139] B. Li, T. Bengtsson, and P. Bickel. Curse of dimensionality revisited: Collapse of importance sampling in very large scale systems. technical report 696, Department of Statistics, University of California-Berkeley, http://www.stat.berkeley.edu/tech-reports/696.pdf, 2005.

[140] K.C. Li. Sliced inverse regression for dimension reduction. *J. Am. Statist. Ass.*, 86:316–342, 1991.

[141] K.C. Li. On principal hessian directions for data visualisation and dimension reduction: another application of stein's lemma. *Ann. Statist.*, 87:1025–1039, 1992.

[142] K.C. Li. High dimensional data analysis via the sir/phd approach. Lecture Notes, 2000.

[143] L. A. Liporace. Maximum likelihood estimation for multivariate observations of markov sources. *IEEE Transact. Inform. Theory*, 28(5):729–734, 1982.

[144] H. Liu and H. Motoda. Feature transformation and subset selection. *IEEE Intelligent Systems*, 13(2):26–34, 1998.

[145] M.S. Lobo, L. Vandenberghe, and S. Boyd H. Lebret. Applications of second-order cone programming. *Linear Algebra and its Applications*, 284:198–228, 1998.

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

[147] L. Ertoz M. Steinbach and V. Kumar. *New Vistas in Statistical Physics*, chapter Challenges of Clustering High Dimensional Data - Applications in Econophysics, Bioinformatics, and Pattern Recognition. Springer-Verlag, 2003.

[148] D. MacKay. *Information Theory, Inference and Learning Algorithms.* Cambridge University Press, 2003.

[149] C.L. Mallows. Some comments on cp. *Technometrics*, 15:661–675, 1973.

[150] J.C. Mattingly and A.M. Stuart. Geometric ergodicity of some hypo-elliptic diffusions for particle motions. *Markov Process. Related Fields.*, 8(2):199–214, 2001.

[151] P. McCullagh and J.A. Nelder. *Generalized linear models.* Chapman and Hall, 1989.

[152] A. D. R. McQuarrie and C.-L. Tsai. *Regression and Time Series Model Selection.* World Scientific, 1998.

[153] E. Meerbach. *Off- and Online Detection of Dynamical Phases in Time Series.* PhD thesis, Free University of Berlin, 2008.

[154] E. Meerbach, E. Dittmer, I. Horenko, and C. Schütte. *Computer Simulations in Condensed Matter Systems: From Materials to Chemical Biology.*, volume 703 of *Lecture Notes in Physics*, chapter Multiscale Modelling in Molecular Dynamics: Biomolecular Conformations as Metastable States. M. Ferrario and G. Ciccotti and K. Binder, 2006.

[155] E. Meerbach and C. Schütte. Sequential change point detection in molecular dynamics trajectories. *Submitted to Multiscale Modeling and Simulation*, 2008.

[156] M. Mizuta. *Dimension Reduction Methods*, chapter 6, pages 566–89. J.E. Gentle and W. Härdle, and Y. Mori (eds.): Handbook of Computational Statistics., 2004.

[157] Y. Mu, P. Nguyen, and G. Stock. Energy landscape of a small peptide revealed by dihedral angle principal component analysis. submitted.

[158] D.W. Müller and G. Sawitzki. Excess mass estimates and tests for multimodality. *Journal of the American Statistical Association*, 86:738–746, 1991.

[159] P. Narendra and K. Fukanaga. A branch and bound algorithm for feature subset selection. *IEEE Trans. Computers*, 29(9):917–922, 1977.

[160] A. Nemirovski. Informational-based complexity of linear operator equations. *Journal of Complexity*, 8:153–175, 1992.

[161] A. Nemirovski. *Lectures on Probability Theory and Statistics, Ecole d'ete de Probabilities de Saint-Flour XXVIII - 1998*, volume 1738 of *Lecture Notes in Mathematics*, chapter Topics in Non-Parametric Statistics. M. Emery and A. Nemirovski and D. Voiculescu, 2000.

[162] A. Nemirovski. Prox-method with rate of convergence o(1/t) for variational inequalities with lipschitz continuous monotone operators and smooth convex-concave saddle point problems. *SIAM Journal on Optimization*, 15:229–251, 2004.

[163] A. Nemirovski and D. Yudin. *Problem Complexity and Method Efficiency in Optimization.* J. Wiley and Sons, 1983.

[164] Yu. E. Nesterov. Excessive gap technique in non-smooth convex minimization. *SIAM Journal on Optimization*, 16(1):235 – 249, 2005.

[165] Yu. E. Nesterov. Smooth minimization of non-smooth functions. *Mathematical Programming: Series A and B*, 103(1):127–152, 2005.

[166] Yu. E. Nesterov. Solving strongly monotone variational and quasi-variational inequalities. *Discussion Paper 2006/107, CORE, Catholic University of Louvain, Louvain-la-Neuve, Belgium*, 2006.

[167] Yu. E. Nesterov. Dual extrapolation and its applications for solving variational inequalities and related problems. *Mathematical Programming: Series A and B*, 109(2):319 – 344, 2007.

[168] Yu. E. Nesterov. Primal-dual subgradient methods for convex problems. *Mathematical Programming*, 10.1007/s10107-007-0149-x, 2007.

[169] Yu. E. Nesterov. Rounding of convex sets and efficient gradient methods for linear programming problems. *Optimization Methods and Software*, 23(1):109–128, 2007.

[170] Yu. Nesterow. *Introductory Lectures on Convex Optimization*, volume 87 of *Applied Optimization*. Kluwer Academic, 2004.

[171] A. Ng, M. Jordan, and Y. Weiss. *in: Proc. Advances in Neural Information Processing*, chapter On spectral clustering: Analysis and an algorithm. 2001.

[172] F. Noe, I. Horenko, C. Schütte, and J.C. Smith. Hierarchical analysis of conformational dynamics in biomolecules: transition networks of metastable states. *J Chem Phys.*, 126(15), 2007.

[173] Frank Noé, Illia Horenko, Christof Schütte, and Jeremy C Smith. Hierarchical analysis of conformational dynamics in biomolecules: transition networks of metastable states. *J Chem Phys*, 126(15):155102, Apr 2007.

[174] Bernt Øksendal. *Stochastic Differential Equations: An Introduction with Applications*. Springer, December 2005.

[175] p. Demartines. *Analyse de donnée par réseaux de neurones auto-organisées.* PhD thesis, Institut National Polytechnique de Grenoble, 1994.

[176] J. Karhunen P. Pajunen and E. Oja. The nonlinear pca criterion in blind source separation: Relations with other approaches. *Neurocomputing*, 22:5–22, 98.

[177] J. Pillardy and L. Piela. Molecular dynamics on deformed energy hypersurfaces. *J.Phys.Chem.*, 99:11805–11812, 1995.

[178] B.T. Polyak. A general method for solving extremal problems. *Soviet Math.Doklady*, 174:33–36, 1967.

[179] S. Polyak, F. Rendl, and H. Wolkowicz. A recipe for semidefinite relaxation for (0,1)-quadratic programming. *J. Global Optim.*, 7(1):51–73,, 1995.

[180] P. Pudil, J. Novovicova, and J. Kittler. Floating search methods in feature selection. *Pattern Recognition Letters*, 15(1):1119–1125, 1994.

[181] L.R. Rabiner. A tutorial on hidden Markov models and selected applications in speech recognition. In *Proceedings of the IEEE*, volume 77, pages 257–286, 1989.

[182] G. N. Ramachandran and V. Sasiskharan. Conformations of polypeptides and proteins. *Advan. Prot. Chem.*, 23:283–427, 1968.

[183] G. N. Ramachandran and V. Sasiskharan. Conformations of polypeptides and proteins. *Advan. Prot. Chem.*, 23:283–427, 1968.

[184] J. Renegar. *A Mathematical View of Interior-Point Methods in Convex Optimization.*, volume 3 of *MPS-SIAM Series on Optimization*. SIAM, Philadelphia, PA, 2001.

[185] B. D. Ripley. *Pattern Recognition and Neural Networks*. Cambridge University Press, Cambridge, U.K.,, 1996.

[186] H. Risken. *The Fokker-Planck equation : methods of solution and applications*. Springer, Berlin, 1996.

[187] R.T. Rockafellar. *Variational Analysis*, volume 317 of *Grundlehren der mathematischen Wissenschaften*. Springer, Berlin, 1998.

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

[189] J.P. Royston. An extension of shapiro and wilks' w test for normality to large samples. *Applied Statistics*, 31:115–124, 1982.

[190] J.P. Royston. The w test for normality. *Applied Statistics*, 21:176–180, 1982.

[191] D. B. Rubin and D. T. Thayer. Em algorithms for ml factor analysis. *Psychometrika*, 47(1):69–76, 1982.

[192] D. B. Rubin and D. T. Thayer. More on em for ml factor analysis. *Psychometrika*, 48(2):253–257, 1983.

[193] J.W. Sammon. A nonlinear mapping for data analysis. *IEEE Transactions on Computers*, C-18:401–409, 1969.

[194] M. Schaefer and M. Karplus. A comprehensive analytical treatment of continuum electrostatics. *J. Chem. Phys.*, 100:1578–1599, 1996.

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

[196] C. Schütte. *Conformational Dynamics: Modelling, Theory, Algorithm, and Application to Biomolecules.* Habilitation Thesis, Fachbereich Mathematik und Informatik, Freie Universität Berlin, 1998.

[197] C. Schütte and W. Huisinga. *Biomolecular Conformations can be identified as metastable sets of melcular dynamics.*, volume Computational Chemistry of *Handbook of Numerical Analysis*, pages 699–744. P.G. Ciaret and J.-L. Lions, 2003.

[198] Ch. Schütte and W. Huisinga. On conformational dynamics induced by langevin processes. In B. Fiedler, editor, *International conference on Differential Equations*, volume 2, pages 1247–1262, 2000.

[199] Ch. Schütte, W. Huisinga, and P. Deuflhard. Transfer operator approach to conformational dynamics in biomolecular systems. In B. Fiedler, editor, *Ergodic Theory, Analysis, and Efficient Simulation of Dynamical Systems*, pages 191–223. Springer, 2001.

[200] Ch. Schütte, W. Huisinga, and P. Deuflhard. Transfer operator approach to conformational dynamics in biomolecular systems. In B. Fiedler, editor, *Ergodic Theory, Analysis, and Efficient Simulation of Dynamical Systems*, pages 191–223. Springer Verlag, 2001.

[201] Christof Schütte and Illia Horenko. Likelihood-based estimation of multidimensional Langevin models and its application to biomolecular dynamics. *Multiscale Modeling and Simulation*, 2008. Accepted.

[202] Christof Schütte and Wilhelm Huisinga. *Biomolecular Conformations can be Identified as Metastable Sets of Molecular Dynamics*, pages 669–744. Handbook of Numerical Analysis X. Elsevier, 2003. Special Volume: Computational Chemistry.

[203] D. W. Scott. *Multivariate Density Estimation. Theory, Practice, and Visualization.* Wiley Series in Probability and Mathematical Statistics. John Wiley and Sons, New York, London, Sydney, 1992.

[204] S.S. Shapiro and M.B. Wilk. An analysis of variance test for normality. *Biometrika*, 52:591–611, 1965.

[205] R. N. Shepard. The analysis of proximities: multidimensional scaling with an unknown distance function. *Psychometrika*, 27(2):125–140, 1962.

[206] I. Shimshoni, B. Georgescu, and P. Meer. *Nearest-Neighbor Methods in Learning and Vision: Theory and Practice*, chapter Adaptive Mean Shift Based Clustering in High Dimensions, pages 203–220. G. Shakhnarovich, T. Darrell and P. Indyk, 2006.

[207] N. Z. Shor. Quadratic optimization problems. *Soviet Journal of Circuits and Systems Sciences*, 25(6):1–11, 1987.

[208] N.Z. Shor. Generalized gradient descent with application to block programming. *Kibernetika*, 3, 1967.

[209] B. W. Silverman. Using kernel density estimates to investigate multimodality. *Journal of the Royal Statistical Society, Ser. B.*, 43:97–99, 1981.

[210] B. W. Silverman. *Density Estimation for Statistics and Data Analysis.*, volume 26 of *Monographs on Statistics and Applied Probability.* Chapman and Hall, London, New York, 1986.

[211] V. Spokoiny. A penalized exponential risk bound in parametric estimation. *http://arxiv.org/abs/0903.1721*, 2009. WIAS-preprint.

[212] V. Spokoiny, G. Blanchard, M. Sugiyama, M.Kawanabe, and Klaus-Robert Müller. In search of non-Gaussian components of a high-dimensional distribution. *Journal of Machine Learning Research*, 7:247–282, 2006.

[213] V.G. Spokoiny. Adaptive hypothesis testing using wavelets. *Ann. Statist.*, 24:2477–2498, 1996.

[214] D. Stalling, M. Westerhoff, and H.-C. Hege. *The Visualization Handbook*, chapter Amira: A Highly Interactive System for Visual Data Analysis (Ch. 38), pages 749–767. Elsevier Academic Press, 2004.

[215] M. A. Stephens. *Goodness of Fit Techniques.*, chapter Tests based on Goodness of Fit. D'Agostino, R. B. and Stephens, M. A., 1986.

[216] C.J. Stone. Optimal global rates of convergence for nonparametric regression. *Annals of Statistics*, 10(4):1040–1053, 1982.

[217] M. Sugiyama. *ICML*, chapter Local Fisher discriminant analysis for supervised dimensionality reduction., pages 905–912. William W. Cohen and Andrew Moore, 2006.

[218] A. Ben Tal and A. Nemirovski. *Lectures on Modern Convex Optimization.*, volume 1 of *MPS/ SIAM Series on Optimization*. SIAM, Philadelphia, 2001.

[219] F. Tao, S.Z. Li, Heung-Yeung S, and Z. HongJiang. Local non-negative matrix factorization as a visual representation. The 2nd International Conference on Development and Learning, 2002. Proceedings., vol. 2 2002. pp.178-183.

[220] F.J. Theis, P. Georgiev, and A. Cichocki. Robust sparse component analysis based on a generalized hough transform. *EURASIP Journal on Advances in Signal Processing*, 2007.

[221] R. Tibshirani. Regression shrinkage and selection via the lasso. *J. Royal. Statist. Soc B.*, 58(1):267–288, 1996.

[222] M.E. Tipping and C.M. Bishop. Mixtures of probabilistic principal component analyzers. *Neural Computation*, 11(2):442–482, 1999.

[223] M.E. Tipping and C.M. Bishop. Probabilistic principal component analysis. *Journal of the Royal Statistical Society*, 61(3):611–634, 99.

[224] M.J. Todd and E.A. Yildrim. On khachiyan's algorithm for the computation of minimum volume enclosing ellipsoids. *Technical Report, preprint*, 2005.

[225] J.F. Traub, G.W. Wasilkowski, and H. Wozniakowski. *Information-Based Complexity.* Computer Science and Scientific Computing. Academic Press, San Diego, 1988.

[226] A. van der Vaart and J.A. Wellner. *Weak Convergence and Empirical Proccesses.* Springer Series in Statistics. Springer - New York, 1996.

[227] L. Vandenberghe and S. Boyd. Semidefinite programming.. *SIAM Review*, 38(1):49–95, 1996.

[228] L. Vandenberghe and S. Boyd. Applications of semidefinite programming. *Applied Numerical Mathematics*, 29:283–299, 1999.

[229] V.N. Vapnik. *Statistical Learning Theory.* Wiley-Interscience, 1998.

[230] R. Vidal, Y. Ma, and S. Sastry. Generalized principal component analysis. *IEEE Transaction on Pattern Analysis and Maschine Intelligence*, 27(12):1–15, 2005.

[231] Andrew J. Viterbi. Error bound for convolutional codes and an asymptotically optimum decoding algorithm. *IEEE Transactions on Information Theory.*, 13(2):260–269, 1967.

[232] L. Wasserman. *All of Nonparametric Statistics.* Springer Texts in Statistcs. Springer, 2006.

[233] A.R. Webb. *Statistical Pattern Recognition.* Wiley, New York, 2nd edition, 2002.

[234] M. Weber. *Meshless Methods in Conformation Dynamics.* PhD thesis, Freie Universit/"at Berlin, 2005.

[235] E. J. Wegman. Hyperdimensional data analysis using parallel coordinates. *J. Amer. Stat. Assoc.*, 85(411):664–675, 1990.

[236] H. Wold. *Soft Modeling. The Basic Design and Some Extensions.*, volume 2 of *Systems Under Indirect Observation*, pages 1–53. K.-G. Jöreskog and H. Wold, North-Holland, Amsterdam, 1982.

[237] S. Wold, S. Hellberg, M. Sjostrom, and H. Wold. *PLS Model Building: Theory and applications. PLS modeling with latent variables in two or more dimensions.* 1987.

[238] H. Wolkowicz, R. Saigal, and L. Vandenberghe. *Handbook of Semidefinite Programming.* Kluwer Academics Mass., 2000.

[239] Y. Xia, H. Tong, W. K. Li, and L.-X. Zhu. An adaptive estimation of dimension reduction space. *Journal of Royal Statistical Society, Ser. B*, 64:363–410, 2001.

[240] J.-Ph. Vial Yu. Nesterov. Homogeneous analytic center cutting plane methods for convex problems and variational inequalities. *SIAM J.Optim.*, 9(3):707–728, 1999.

[241] D.B. Yudin and A.B. Nemirovsky. Computational complexity and efficiency of methods for solving convex extremum problems. *Ekonomika i matem. metody*, XII(2):357–369, 1976.

[242] D.B. Yudin and A.B. Nemirovsky. Estimating the computational complexity ofmathematical programming problems. *Ekonomika i matem. metody*, XII(1):128–142, 1976.

[243] D.B. Yudin and A.B. Nemirovsky. Efficiency of random search in control problems. *Izvestiya Akad. Nauk SSSR tekhnischeskaya kibernetika*, 3:3–17, 1977.

[244] J.H. Zar. *Biostatistical Analysis, (2nd ed.).* NJ: Prentice-Hall, Englewood Cliffs., 1999.

[245] C. Zong. *Strange Phenomena in Convex and Discrete Geometry.* Springer, New York, 1996.

[246] H. Zou, T. Hastie, and R. Tibshirani. Sparse principal component analysis. *Journal of Computational and Graphical Statistics*, 2004.