# Sparse NonGaussian Component Analysis*

Elmar Diederichs[1], Anatoli Juditski[3],
Vladimir Spokoiny[2], Christof Schütte[1]

[1]Institute for Mathematics and Informatics, Free University Berlin
Arnimallee 6, 14195 Berlin, Germany

[2]Weierstrass Institute and Humboldt University
Mohrenstr. 39, 10117 Berlin, Germany

[3]INRIA Grenoble, Rhône-Alpes
655 avenue de l'Europe, 38330 Montbonnot, France

April 26, 2007

**Abstract**

This article proposes a new approach to non-gaussian component analysis (NGCA). NGCA is already in use in high dimensional data analysis. It identifies non-gaussian components in the data as a preprocessing step for efficient information processing and is essentially based on Independent Component Analysis (ICA) and Principle Component Analysis (PCA). Instead we suggest an iterative and structure adaptive approach to non-gaussian component analysis (SNGCA) which is based on iterative convex projection. As an alternative to the use of principle components, SNGCA uses a statistical procedure which combines the computation of the Löwner-John ellipsoid and statistical tests for normality as a tool for reducing the dimension of the data space.

*keywords:* reduction of dimensionality, model reduction, sparsity, variable selection, principle component analysis, structural adaption, convex projection

*Mathematical Subject Classification:* 62G05, 60G10, 60G35, 62M10, 93E10

*Supported by DFG research center Matheon "Mathematics for key technologies" (FZT 86) in Berlin.

# 1 Introduction

Today many mathematical applications in econometrics or biology 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. On the one hand this means an exponential increase of the computational burden for many methods. On the other hand the sparsity of the data in high dimensions entails that data thin out in the local neighborhood of a given point $x$. Hence statistical methods are not reliable in high dimensions [32] if the sample size remains of the same order. This problem is usually referred to as "curse of dimensionality" [10]. The standard approach to deal with the high dimensional data is based of one or another *structural assumption* which allows to reduce the complexity or intrinsic dimension of the data without significant loss of statistical information [22], [20].

In order to illustrate this task 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 $\mathbb{R}^d$ with $m \leq d, d \in \mathbb{N}$. The additional degrees of freedom may stem from the influence of a variety of uncontrolled components for instance 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. Consequently a lower dimensional, compact representation that according to some criterion, captures the interesting information in the original data, is sought. If the perturbations do not mask the original model, dimension reduction techniques may be appropriate for understanding the underlying phenomena of interest [8]. This is illustrated in figure 1.1.

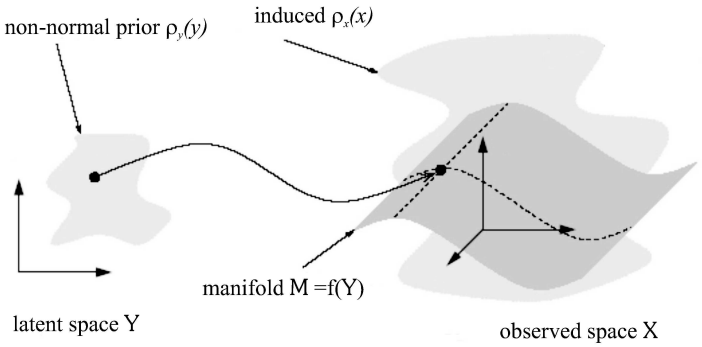


Figure 1.1: basic idea of dimension reduction

The problem of dimension reduction typically 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. Concerning both tasks the structure adaptive approach of SNGCA, presented in this article, is

an unsupervised, completely data driven, linear method. We will not assume any apriori knowledge about the density $\rho(x)$ of the original data or about the spatial distribution of the informative data lying on the manifold in focus. Moreover we will demonstrate in section (5.1) that SNGCA does not depend on a special difference in the magnitude of the second moments of Gaussian noise and informative data components. In the following sections we will see, that all these are advantages of SNGCA wrt. other dimension reduction methods e.g. Principal Component Analysis [14], Independent Component Analysis [12] or Singular Spectrum Analysis [9].

In this paper we assume that we have a sample of metric data actually lying approximately on a $m \leq d$ dimensional linear manifold $\mathcal{I} \subseteq \mathbb{R}^d$ of the finite dimensional Euclidean space. In order to reduce the dimensionality of the data it is sought to find a mapping from the original data space onto this manifold.

The article is organized as follows. Section 2 describes the considered set-up while Section 3 explains the main ideas and steps of the proposed approach. The formal description of the algorithm is given in Section 4. In Section 5 we will demonstrate the features of the algorithm and its behavior by means of some artificial test examples. Moreover we compare the results of SNGCA with that of FastICA, which is an implementation of the Independent Component Analysis (ICA) or Projection Pursuit (PP) method [12]. The performance of the procedure is demonstrated by the analysis of data sets, obtained from simulations of some realistic examples. In the appendix we present a short theoretical study of some features of the SNGCA.

# 2 Theoretical Framework

We will now give a formal representation of the idea of low-dimensional informative data embedded in a high-dimensional noise and a short sketch of the features of SNGCA. To this end, let us assume that we have $N$ metric data in $\mathbb{R}^d$, given by a set of independent and identically distributed (i.i.d) random variables $\{X_i\}_{i=1}^N \in \mathbb{R}^d$. Let $\mathcal{X}$ denote the space of the data. Furthermore we assume, that these random variables are distributed according to an unknown density $\rho(x)$.

## 2.1 General setting of the method

In this section we specify the considered model and the problem on focus.

**Semi-parametric framework:** According to the very popular parametric approach the density function $\rho(x)$ belongs to a parametric family $\mathcal{F} = \{\rho_\theta(x)\}$ where $\theta$ is an finite dimensional parameter which uniquely identifies $\rho(x)$. Then an algorithmic procedure to find a reliable value for an estimator $\widehat{\theta}$ of $\theta$ has to be applied to the data. However the execution time to evaluate a multivariate density typically has an exponential growth in the number

of dimensions [25]. 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. More precisely, 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.

**Non-informative Gaussian components:** As well as in NGCA [27] it is assumed here, that the structure of a data set is represented by non-Gaussian components of its density $\rho(x)$. The Gaussian components of $\rho(x)$ are considered as entropy maximizing and consequently as non-informative noise [4]. It is well known that for many high-dimensional clouds of points, most low-dimensional projections are approximately gaussian [5]. However, important structure in the data set is often contained in a linear subspace which is the affine hull of directions whose one-dimensional projected distribution is far from normality. Note that the suggested way of treating the Gaussian distribution as a noise exclude the use of the classical *Principle Component Analysis* (PCA) for searching the informative components because PCA heavily relies on the Gaussian distribution of the data and it looks at the directions with the largest variance.

**NGCA against ICA:** Another popular way of reducing the complexity of the model is the so called *Independent Component Analysis* (ICA). The basic assumption behind this method is that the whole distribution can be decomposed as a product of univariate ones. For identifiability reason all of them out of maybe one should be non-Gaussian. Every non-Gaussian direction can be identified by a local optimization over linear projection directions wrt. some characteristic which quantifies a certain deviation from normality [25]. Popular examples are curtosis or negentropy, [12]. The methodological problem with the ICA approach is the unrealistic product structure of the whole distribution and the requirement of non-Gaussianity for all the components.

In comparison, NGCA [27] and proposed here SNGCA allow for cross-dependence of the non-Gaussian components and for presence of a full dimensional Gaussian part (noise). 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 NGCA 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.

## 2.2 Semi-parametric framework

We will now introduce a semi-parametric framework, that allows to estimate the target space $\mathcal{I}$ as well as to determine the intrinsic dimension of $\mathcal{I}$.

The realization of this search for non-Gaussian components depends on a semi-parametric framework already presented in [27]. According to the semi-parametric framework, we assume the following stationary data model. Let $X_1, \ldots, X_N \in \mathbb{R}^d$ be i.i.d. random observable, distributed according to a structured density of the following form:

$$\rho(x) = \phi_{\mu,\Sigma}(x) q(Tx). \tag{2.1}$$

Here $\phi_{\mu,\Sigma}$ 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

$$\mathcal{I} = Ker(T)^{\perp} = range(T^{\top}). \tag{2.2}$$

(2.1) is not a unique representation. However the $m$-dimensional linear subspace $\mathcal{I} \subset \mathbb{R}^d$ is uniquely defined by (2.2). $\mathcal{I}$ contains the non-Gaussian distributed data and is called the *non-Gaussian subspace*. By analogy with the regression case [3, 19, 18], 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 (2.1) can be regarded as the distribution of the low dimensional *signal* $Z$ corrupted by a full dimensional Gaussian noise $\varepsilon_N$ :

$$X = Z + \varepsilon_N.$$

The next section motivates and describes the main steps of the proposed SNGCA method.

# 3 SNGCA procedure

For simplicity we assume below that the mean of the data is zero. This is easily achieved by removing the empirical mean from the data. In the sequel with $\mathbb{E}[X]$ we denote the expectation in $X$ and $\mathbb{E}_N[\cdot]$ denotes the empirical mean, i.e. for any function $f(x)$ on $\mathbb{R}^d$ we set

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

Next we will explain how elements $\beta \in \mathcal{I}$ can be estimated from the data without estimating the parameters $\mu$, $\Sigma$ and $q$ of $\rho$ in (2.1).

## 3.1 Estimation of the vectors from non-Gaussian subspace

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

**Theorem 1.** *Let $X$ follow the distribution with the density $\rho(x)$ according to (2.1) and let $\mathbb{E}[X] = 0$. Suppose that $\psi \in \mathcal{C}^1(\mathbb{R}^d, \mathbb{R})$ is a function[1] fulfilling the condition*

$$\gamma(\psi) := \mathbb{E}[X\psi(X)] = 0, \tag{3.3}$$

*Define*

$$\beta(\psi) := \mathbb{E}_x\big[\nabla_x \psi(X)\big] = \int \nabla_x \psi(x)\, \rho(x)\, dx, \tag{3.4}$$

*where $\nabla_x \psi$ means the gradient of $\psi$ in $x$. Then $\beta(\psi)$ belongs to $\mathcal{I}$. Moreover if (3.3) is not fulfilled, then there is a $\beta \in \mathcal{I}$ such that*

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

*where $\epsilon$ is the uniform error bound:*

$$\epsilon = \Big\| \Sigma^{-1} \int x\psi(x)\rho(x)\, dx \Big\|_2 . \tag{3.5}$$

*Hence the distance between $\beta(\psi)$ and the non-Gaussian subspace $\mathcal{I}$ is uniformly bounded as given by (3.5).*

Equivalently one can state the result (3.5) in the form

$$\|(\mathbf{1}_d - \Pi_{\mathcal{I}})\beta(\psi)\|_2 \le \epsilon$$

where $\mathbf{1}_d$ 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 idea of the estimation procedure is the algorithmic realization of (3.4) and (3.3) in Theorem 1. To fulfill (3.3), it was already suggested in [27] to start from some smooth function $h(x)$ and to build $\psi(x)$ in the form $\psi(x) = h(x) - \alpha^\top x$ from the data where the vector $\alpha$ is selected to provide the condition $\mathbb{E}_N[X\psi(X)] = 0$. A problem with this approach is that it requires to operate with the empirical covariance matrix which can be a hard numerical and analytical problem in the case of a big dimension $d$.

In this article we apply a slightly different approach: Theorem 1 relies on the vectors $\gamma(\psi)$ and $\beta(\psi)$ which in turn depend on the unknown density $\rho$. However, both vectors are integrals w.r.t. $\rho$. Therefore, they can be easily estimated from the data by using their empirical counterparts:

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

[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 continuous differentiable.

To this end we will construct $\psi(x)$ as a convex combination of smooth functions from the data. Let $\{\omega_l\}_{l=1}^L$ be a given set of unit vectors $\omega_l \in \mathbb{R}^d$. Then define

$$\psi_{h,c}(x) := \sum_{l=1}^{L} c_l \, h_{\omega_l}(x) \tag{3.6}$$

where the vector of coefficients $c = \{c_l\}_{l=1}^L$ fulfills $\|c\|_1 \leq 1$. Moreover we chose $h_\omega \in \mathcal{C}^{1,1}(\mathbb{R}^d \times \mathbb{R}^d, \mathbb{R})$ of the form

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

where $\omega \in \mathbb{R}^d$ is a unit vector and $\lambda > 0$. The function $h \in \mathcal{C}^{1,1}(\mathbb{R}, \mathbb{R})$ should be informative wrt. non-Gaussian components. For example the non-Gaussian-Gaussian distinction can be exploited using the higher moments of the data distribution. The multiplier $e^{-\lambda\|x\|^2/2}$ ensures that $h_\omega(x)$ is bounded and integrable w.r.t. the density $\rho$ over the whole space $\mathbb{R}^d$.

Now let $h'$ denote the derivative of $h$ and let us define

$$\eta_{\omega_l} := \mathbb{E}\big\{\nabla_x h_{\omega_l}(X)\big\} = \mathbb{E}\big\{\omega_l h'(\omega_l^\top X) e^{-\lambda\|x\|^2/2} - \lambda X h_{\omega_l}(X)\big\}$$

Then using definition (3.6) this yields

$$\beta(\psi_{h,c}) = \sum_{l=1}^{L} c_l \mathbb{E}\big\{\nabla h_{\omega_l}(X)\big\} = \sum_{l=1}^{L} c_l \, \eta_{\omega_l} \, . \tag{3.7}$$

Similarly with $\gamma_{\omega_l} := \mathbb{E}\big\{X h_{\omega_l}(X)\big\}$

$$\gamma(\psi_{h,c}) = \sum_{l=1}^{L} c_l \mathbb{E}\big\{X h_{\omega_l}(X)\big\} = \sum_{l=1}^{L} c_l \, \gamma_{\omega_l} \, . \tag{3.8}$$

The data counterpart of these expressions playing the central role in the algorithm of SNGCA is given by

$$\begin{aligned}
\widehat{\gamma}_{\omega_l} &= \mathbb{E}_N\big\{X h_{\omega_l}(X)\big\} = \frac{1}{N}\sum_{i=1}^{N} X_i h_{\omega_l}(X_i) \\
\widehat{\eta}_{\omega_l} &= \mathbb{E}_N\big\{\nabla h_{\omega_l}(X)\big\} = \omega_l \frac{1}{N}\sum_{i=1}^{N} h'(\omega_l^\top X_i) - \lambda \widehat{\gamma}_l \\
\widehat{\beta}(\psi_{h,c}) &= \sum_{l=1}^{L} c_l \, \widehat{\eta}_{\omega_l} = \frac{1}{N}\sum_{l=1}^{L} c_l \omega_l \sum_{i=1}^{N} h'(\omega_l^\top X_i) - \lambda \sum_{l=1}^{L} c_l \, \widehat{\gamma}_l .
\end{aligned} \tag{3.9}$$

We will describe in the next section how to determine the coefficients $\{c_l\}_{l=1}^L$ by means of a convex projection approach.

**Choice of $h$:** One may consider different parameter-dependent families of symmetric and non-symmetric, smooth functions $h$. For the numerical simulations later shown in this article, we use the families

$$\begin{aligned} h(t) &= \tanh(t) && \text{(hyperbolic tangent)} \\ h(t) &= (1+t^2)^{-1}\exp(t) && \text{(asymmetric Gauss).} \end{aligned}$$

**Choice of the directions $\omega$:** The choice of these directions is crucial to the algorithm because the estimated vectors $\widehat{\beta}$ are searched as aggregations of the vectors $\omega_l$. We use a Monte-Carlo or quasi Monte-Carlo sampling from the uniform distribution $\mathcal{U}(x)$ for $\omega_l$. Due to the sparsity of high dimensional data this step becomes more and more computationally expensive with increasing number of dimensions.

## 3.2 Convex projection

The approach described in the last section may lead to the following implementation of (3.4) and (3.3): with given vectors $\omega_1, \ldots, \omega_L$, compute $\widehat{\gamma}_l$ and $\widehat{\eta}_l$ for all $l$ and consider the convex combinations $\widehat{\beta} = \sum_l c_l \widehat{\eta}_l$ under the constraint $\sum_l c_l \widehat{\gamma}_l = 0$. The vector $\widehat{\beta}$ is informative if its length is larger in order than the accuracy of approximation $\varepsilon$. So, one may try to maximize the length of $\widehat{\beta}$ over the $\ell_1$-ball $\|c\|_1 \le 1$ under the constraint $\sum_l c_l \widehat{\gamma}_l = 0$. But this method is not appropriate by two reasons: first, such a problem of maximization a convex function over a convex set cannot be solved by a computation cheap interior point method. Second, such a way focus on the major non-Gaussian direction(s) and discards the less pronounced directions.

However a well known result from the empirical process theory claims that $\widehat{\gamma}_\omega$ approximates the unknown vector $\gamma_\omega$ with the accuracy of order $N^{-1/2}$. Moreover, this result can be stated uniformly over all $\omega$ from the unit ball $\mathcal{B}_d$ in $\mathbb{R}^d$, see Theorem 4 in the appendix. The same holds for the differences $\widehat{\eta}_\omega - \eta_\omega$. The use of convex combinations $\psi_{h,c}(x) = \sum_l h(\omega_l^\top x)$ allows to extend this accuracy of approximation $\varepsilon$ on the difference $\widehat{\beta}(\psi_{h,c}) - \beta(\psi_{h,c})$. This motives an alternative method to compute the empirical counterpart of $\beta(\psi_{h,c})$ in (betaestim), called *convex projection*: For a given direction $\xi$ consider an optimization problem

$$\min_{\|c\|_1 \le 1} \Big\| \xi - \sum_{l=1}^{L} c_l \widehat{\eta}_l \Big\|_2^2 \qquad \text{subject to} \qquad \sum_{l=1}^{L} c_l \widehat{\gamma}_l = 0. \tag{3.10}$$

Note that under the latter constraint, the sum $\sum_l c_l \widehat{\eta}_l$ is a convex combination of the vectors $\omega_l$, see (3.9). Moreover, it is well known [6, 7], [29, 34] that such a convex optimization realizes a numerical stable, continuous shrinkage technique and thus leads to a *sparse* solution in which only few coefficients $c_l$ are different from zero. Consequently (3.10) suppresses directions $\omega_l$ which are uninformative about $\mathcal{I}$. However it is a limitation of SNGCA that in the first iteration $k = 1$ of SNGCA, $\xi$ is randomly chosen from $\mathcal{U}(x)$.

In Theorem 2 from the appendix it is shown that the convexity condition $\sum_l c_l \leq 1$ leads to the claims that there is a value $\varepsilon = \sqrt{C_1/N}$ for a fixed constant $C_1$ and a random set $\mathcal{A}_{\varepsilon_r}$ of a dominating probability such that $\|(\mathbf{1}_d - \Pi_{\mathcal{I}})\widehat{\beta}\|_2 \leq \varepsilon_r$ for all such constructed vectors $\widehat{\beta}$. Consequently the idea of the procedure is to repeat this construction for different combinations of $\xi, \omega_1, \ldots, \omega_L$ leading to a family of estimated vectors $\widehat{\beta}$. Each of this vectors belong to the target space $\mathcal{I}$ up to an estimation error $\varepsilon_r$. Then the subspace $\mathcal{I}$ can be recover from the set of $\widehat{\beta}$'s.

Obviously (3.10) is a non-smooth optimization problem. Since we are only interested in the directions $\widehat{\beta}$ an equivalent but smooth and convex version of (3.10) is given by

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

The estimation procedure of elements from the target space described here solves only this smooth problem. In the next section we will describe how SNGCA aims to solve the remaining task of dimension reduction, the reconstruction of a low dimensional linear subspace from the set of $\{\widehat{\beta}_j\}_{j=1}^J$.

### 3.3 Reduction of dimensionality

The first step of the SNGCA procedure consists in estimating the set of vectors $\widehat{\beta}_j$, $j = 1, \ldots, J$. The basic property of these vectors is given by Theorem 2 of the appendix: with a dominating probability $\|(\mathbf{1}_d - \Pi_{\mathcal{I}})\widehat{\beta}\|_2 \leq \varepsilon$ for some small $\varepsilon$. The next important step of the SNGCA procedure is to recover the subspace $\mathcal{I}$ from the estimated vectors $\widehat{\beta}_j$. 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 $\sum_{j=1}^J \|(\mathbf{1}_d - \Pi_{\mathcal{I}})\widehat{\beta}_j\|_2^2$ over all projectors $\Pi_{\mathcal{I}}$ of a given rank $m$, i.e.

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

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 $\widehat{D} := \sum_{j=1}^{J} \widehat{\beta}_j \widehat{\beta}_j^\top$ :

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

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

$$\mathcal{E}(\widehat{\mathcal{I}}, \mathcal{I}) := \|\Pi_{\widehat{\mathcal{I}}} - \Pi_{\mathcal{I}}\|_{Frob}^2 \tag{3.11}$$

where $\|\cdot\|_{Frob}$ 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

$$\|\widehat{D} - D\|_2^2 = tr(\widehat{D} - D)^2 \leq \varepsilon^2.$$

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

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

This condition is easily verified if some significant fraction of the vectors $\beta_j$ are significant (informative) in the sense $\|\beta_j\|_2 \geq \kappa$ with some fixed $\kappa > 0$. However, if the most of vectors $\beta_j$ 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 $\mathcal{I}$.

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

$$\mathcal{S} := \{\widehat{\beta}_1, -\widehat{\beta}_1, \widehat{\beta}_2, -\widehat{\beta}_2, \ldots\}. \tag{3.12}$$

In the direction orthogonal to the linear subspace $\mathcal{I}$, this set expanded only with the distance not larger than $\varepsilon$ while the for the direction within $\mathcal{I}$ we expect some informative vectors. This leads to the idea of building an ellipsoid which contains $\mathcal{S}$ and hence its convex hull $conv(\mathcal{S})$ and 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 2.** *(Existence and Uniqueness) [13]*
*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 $\widehat{B}$:

$$\mathcal{E}_{\sqrt{d}} := \{x \in \mathbb{R}^d : \ \|\widehat{B}^{-1/2}x\|_2 \leq 1\} \tag{3.13}$$

Finding the matrix $\widehat{B}$ is a convex optimization problem. The numerical algorithm 4.2 to compute $\widehat{B}$ is presented in the next section.

However the numerical results 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 $\xi$ the length of the informative vectors is no longer correlated with small values of

$$\|(\mathbf{1}_d - \Pi_{\mathcal{I}})\widehat{\beta}(\psi)\|_2$$

In higher dimensions this leads typically to the situation when some of the longest semi-major axis of $\mathcal{E}_{\sqrt{d}}$ are also non-informative and nearly orthogonal to $\mathcal{I}$. Motivated by this observation we propose to identify the semi-axis of $\mathcal{E}$ close to $\mathcal{I}$ using statistical tests on normality.

**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 within the target space. Hence the basic idea is to apply statistical tests on normality wrt. the significance level $\alpha$ to the original data from $\mathbb{R}^d$ projected on every semi-axis of $\mathcal{E}_{\sqrt{d}}$. If the hypothesis of normality is rejected wrt. the projected data, the corresponding semi-axis is used as a basis vector for the reduced target space $\mathcal{I}$.

Since statistical tests specialized for a certain deviation 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 [33] to identify a significant asymmetry in the projected distribution and the EDF-test according to Anderson-Darling [1] with the modification of Stephens [28], 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 [26] based on a regression strategy in the version given by Royston [23, 24]. Once we have classified the semi-axis of $\mathcal{E}_{\sqrt{d}}$ as being close to the target space we can use the identified subset of axis in the structural adaption step to be described in the next section.

## 3.4 Structural adaptation

The quality of recovering the target non-Gaussian subspace $\mathcal{I}$ heavily depends on the quality of sampling the directions $\xi_j$ and $\omega_l$. At the beginning of algorithm, we have no prior information about $\mathcal{I}$ and therefore, sample them 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 the *structurally adaptation* iterative procedure [11]: the step of estimating the vectors $\{\widehat{\beta}_j\}_{j=1}^J$ and the step of estimating subspace $\mathcal{I}$ are iterated, the estimated structural information given by $\widehat{\mathcal{I}}$ is used to improve the quality of estimating the vectors $\widehat{\beta}_j$ in the next iteration of SNGCA which in turn allows to better estimate the target space $\mathcal{I}$.

Statistically this structural adaptation idea is justified by Theorem 3 from the appendix. If the sampling directions $\{\xi_j\}_{j=1}^J$ and $\{\omega_l\}_{l=1}^L$ are informative then the corresponding vectors $\eta_l = \mathbb{E}\nabla h_{\omega_l}(X)$ 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:

$$\{c_l^*\} = \arg\min_{\|c\|_1 \le 1} \Big\|\xi - \sum_{l=1}^L c_l \eta_l\Big\|_2$$
$$\text{subject to} \qquad \sum_{l=1}^L c_l \gamma_l = 0$$

is also informative. The message of Theorem 3 is that in this situation the estimated vector $\widehat{\beta}$ delivers as much information as $\beta^*$ up to a small error of estimation. So, a successful sampling of the data space $\mathcal{X}$ increases the fraction of informative vectors $\widehat{\beta}_j$ and hence, the final quality of estimating the subspace $\mathcal{I}$.

In our implementation, 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.

# 4 Algorithms

This section presents the formal description of the SNGCA algorithm.

## 4.1 Basic subprocedures of SNGCA

**Whitening:** As a preprocessing step the SNGCA procedure uses a componentwise whitening of the data. To this end let $\sigma = (\sigma_1, \dots \sigma_d)$ be the standard deviations of the data components of $x_1, \dots, x_d$. Then the componentwise whitening of the data is done by

$$Y_i = \mathrm{diag}(\sigma) X_i$$

for $i = 1, \dots, N$.

**Estimation of the vectors from non-Gaussian subspace:** Here we repeat in more detail the estimation procedure already presented in section (3.1).

**Algorithm 4.1:** ESTIMATION$(Y, L, J)$

**comment:** linear estimation of $\beta(\psi)$

**Sampling**: choice of the measurement directions

**Estimation**:

**for** $j = 1$ **to** $J$

**for** $l = 1$ **to** $L$

Compute:

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

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

**end loop on** $l$

Convex projection:

$$\text{Solve} \quad \widehat{c}_j = \arg\min_{\|c\|_1 \le 1} \ \|\xi_j - \textstyle\sum_{l=1}^{L} c_{jl}\, \widehat{\eta}_{jl}\|_2^2$$

$$\text{subject to} \qquad \textstyle\sum_{l=1}^{L} c_{jl}\, \widehat{\gamma}_{jl} = 0$$

Compute:

$$\widehat{\beta}_j = \textstyle\sum_{l=1}^{L} \widehat{c}_{jl}\, \widehat{\eta}_{jl}$$

**end loop on** $j$

In the sequel we will call the directions $\omega_l$ and $\xi_j$ the measurement directions.

**Reduction of dimensionality:** Let us consider again the symmetrized set $\mathcal{S} = \{\widehat{\beta}_1, -\widehat{\beta}_1, \widehat{\beta}_2, -\widehat{\beta}_2, \ldots\}$ already defined in (3.12). From theorem 2 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 [16]. Even though the latter problem can be solved using interior-point-methods in $\mathcal{O}(d^3 \log N)$ iterations, their use is advisable only in the case of a full-dimensional ellipsoid [30].

However we expect the computation of the MVEE to be numerically bad conditioned. Hence SNGCA uses a regularized version of an algorithm recently proposed in [21] to compute an approximation of the MVEE. For convenience we repeat that algorithm here:

**Algorithm 4.2:** $\sqrt{d}$-ROUNDING$(\{\widehat{\beta}_j\}_{j=1}^J)$

**comment:** computation of the $\sqrt{d}$-rounding of the MVEE

Let $\delta_i^{k^*} = \max_{1\leq j\leq J} \langle \widehat{\beta}_j, \widehat{B}_i\widehat{\beta}_j\rangle$ 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 index of the loop.

**while**

**if** $\delta_i^{k^*} \leq C \cdot d$ , where $1 \leq C$ is a tuning parameter

**then** stop the loop

**else**
compute the updates:
$x_i = \widehat{B}_i\widehat{\beta}_{k^*}$
$\widehat{B}_{i+1} = \frac{1}{1-t_i}\Big(\widehat{B}_i - \frac{t_i}{1+\nu_i} x_i x_i^\top\Big)$
$\delta_{i+1}^{k^*} = \frac{1}{1-t_i}\Big(\delta_i^{k^*} - \frac{t_i}{1+\nu_i}\langle\widehat{\beta}_{k^*}, x_i\rangle^2\Big)$

The next algorithm (4.3) reports the pseudocode for constructing a basis of the target space from the estimated elements:

**Algorithm 4.3:** DIMENSIONREDUCTION$(\widehat{B}, \alpha)$

**comment:** discard elements from the basis of eigenvectors of $\widehat{B}$

Let $\widehat{V}$ be the matrix of eigenvectors $\widehat{v}_i$ from $\widehat{B}$ computed according to algorithm 4.2.

**for** $i = 1$ **to** $d$

**do**
Project the data orthogonal on $\widehat{v}_i$ .
Compute tests on normality of the projected data.

Discard every eigenvector with associated normal distributed projected data.

The reduction of the dimensionality described above allows us to restart the algorithm wrt. to an already identified estimator $\widehat{\mathcal{I}}$ .

**Structural Adaption:** In algorithm 4.1 we start with a random initialization of the non-parametric estimator (3.9) by means of a Monte-Carlo sampling of the directions $\omega_{jl}$ and $\xi_j$ containing only rough and poor information about the image of the operator $T$ in (2.1). 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 as described in section 3.4. The procedure is described in detail in the pseudocode of algorithm 4.4.

---

**Algorithm 4.4:** StructualAdaption$(\widehat{B}, n_1, n_2, J, L)$

---

**comment:** structural adaption of the estimation of $\beta(\psi_h)$

Let $\widehat{V}$ be the matrix of eigenvectors $\widehat{v}_i$ from $\widehat{B}$ and suppose that there are already $k$ iterations completed.

For the initialization of 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

$$\begin{array}{ll} \xi_j := \sum_{s=1}^m z_{j,s} \widehat{v}_{i_s} & \text{for} \quad 1 \leq j \leq n_1 < J \\ \omega_l := \sum_{s=1}^m u_{l,s} \widehat{v}_{i_s} & \text{for} \quad 1 \leq l \leq n_2 < L \end{array}$$

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

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

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 $n_2 > L - d$. In this case we set $n_1 = J - d$ or $n_2 = L - d$ for every iteration $k = k + 1$.

---

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

In the next section we will describe, how SNGCA makes use of the algorithms 4.1, 4.3 and 4.4 in order to realize an iterative estimation procedure of $\mathcal{I}$. We will close the following subsection by demonstrating the improvement of the estimation error between subsequent iterations of SNGCA.

## 4.2 Full description

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.

*Input:* Data points $(X_i)_{i=1}^N \in \mathbb{R}^d$
*Parameters:* numbers $J$, $L$ of measurement directions; significance level $\alpha$

**Whitening:**
The data $(X_i)_{i=1}^N$ are recentered by subtracting the empirical mean. Let $\sigma = (\sigma_1, \dots \sigma_d)$ be the standard deviations of the components of $X_i$. Then $Y_i = \mathrm{diag}(\sigma) X_i$ denotes the componentwise empirically whitened data.

**Main Procedure:** (Loop on $k$)

**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 adaption approach described in algorithm 4.4. Then $\xi_j^{(k)}$ and $\omega_{jl}^{(k)}$ are normalized to unit length.

**Estimation:**
Loop on $j = 1, \dots, J$:

Loop on $l = 1, \dots, L$: Compute the estimators

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

End Loop on $l$.

Compute the coefficients $\{c_l\}_{l=1}^L$ by solving the second-order conic optimization problem (SOCP) (3.10):

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

End Loop on $j$.

Compute the estimator $\widehat{\beta}_j = \sum_{l=1}^{L} (\widehat{c_l^+} - \widehat{c_l^-})\widehat{\eta}_{jl}$.

**Dimension reduction:**
Compute the symmetric matrix $\widehat{B}$ defining the approximation of the Löwner-John ellipsoid $\mathcal{E}$ in (3.13) according to algorithm 4.2. Reduce the basis of $\mathcal{X}$ according to algorithm 4.3.

*Output:* $\widehat{\mathcal{I}}$

Figure 4.1: Pseudocode of the full SNGCA algorithm.

**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 $\omega_i$ 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:

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

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 $\chi$ in the dimension reduction step is set to $\chi = 3$.

**Stopping criterion:** Suppose that $\mathcal{I}$ is apriori given. Then the convergence of SNGCA can be measured according to the criterion (3.11). More precisely we assume convergence if the improvement of the error measured by (3.11) from one iteration to the next one is less than $\delta$ percent of the error in the former iteration.

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

**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 $\eta_{jl}$ and $\gamma_{jl}$ in algorithm 4.1 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 4.2 takes $\mathcal{O}(d^2 2J \log(2J))$ operations [21].

3. For the optimization step in 4.1 we use a commercial solver [2] 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.

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

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 [15] 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 [17].

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

**Illustration of one-step-improvement:** We will now illustrate the iterative gain of information about the EDR space. To this end we use the projection of $\widehat{\beta}_j$ to the EDR-space in order to demonstrate, how the algorithm works.

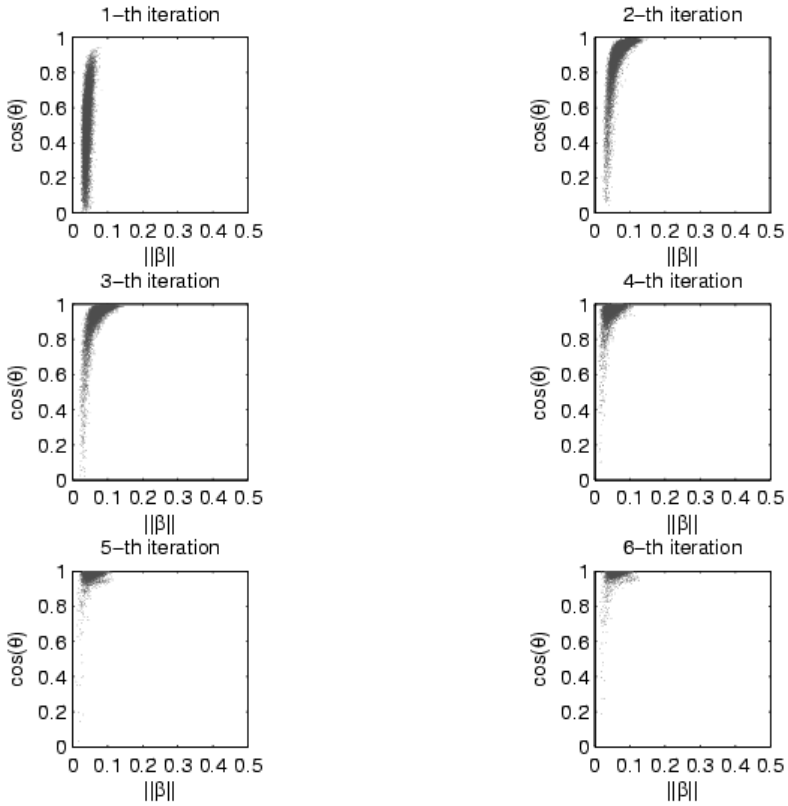


Figure 4.2: illustrative plots of SNGCA applied to toy 20 dimensional data of type **(C)** (see section 5): We show $\|\widehat{\beta}\|$ vs. $\cos(\theta(\widehat{\beta}, \mathcal{I}))$ for different iterations of the algorithm where $\mathcal{I}$ is the apriori known EDR-space.

Figure 4.2 shows that $dist(\widehat{\beta}, \widehat{\mathcal{I}})$ 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 and hence depends on the size of the MC-sampling of the measurement directions.

# 5 Numerical results

The aim of this section is to compare SNGCA with other statistical methods of dimension reduction. The reported results from Projection Pursuit (PP) and NGCA were already published in [27].

## 5.1 Synthetic Data

Each of the following test data sets includes 1000 samples in 10 dimension and each sample consists of 8-dimensional independent, standard and homogeneous Gaussian distributions. The other 2 components of each sample are non-Gaussian with variance unity. The densities of the non-Gaussian components are chosen as follows:

**(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.

**(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 $\lambda = 1$.

That means, that the non-normal distributed data are located in a linear subspace.

In the sequel we compare SNGCA with PP and NGCA using the test data sets from above and the estimation error defined in (3.11). Each simulation 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 $maxIter = 3log(d)$ 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 wrt. model (C) are computed with a componentwise pre-whitening.

Figure 5.1 illustrates the densities of the non-Gaussian components of the test data. For all numerical experiments reported in this article the dimension of the target space $\mathcal{I}$ is apriori given as a tuning parameter for the algorithm.

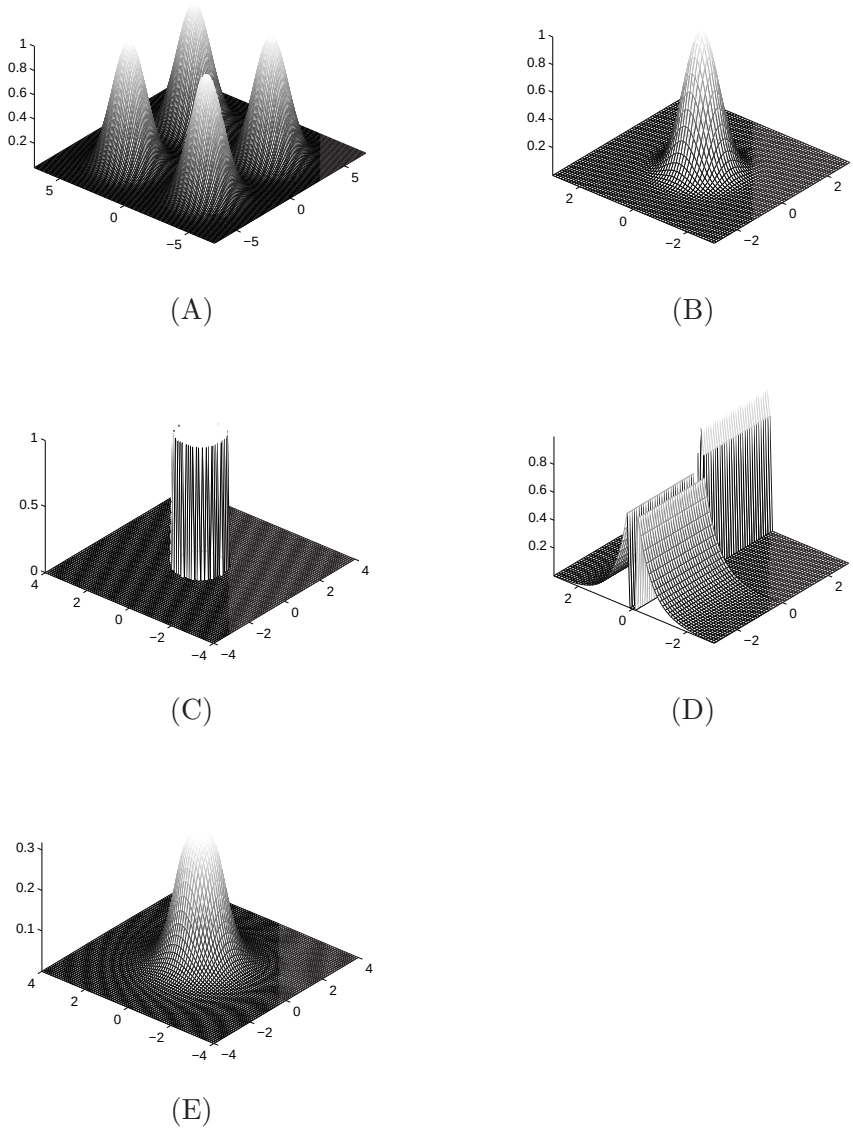


Figure 5.1: densities of the non-Gaussian components: (A) 2 D independent Gaussian mixtures, (B) 2 D isotropic super-Gaussian, (C) 2 d isotropic uniform and (D) dependent 1 d Laplacian with additive 1 D uniform, (E) 2 d isotropic sub-Gaussian

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 wrt. (3.11) 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 5.2 we present boxplots of the error (3.11) of the methods PP, NGCA and SNGCA.

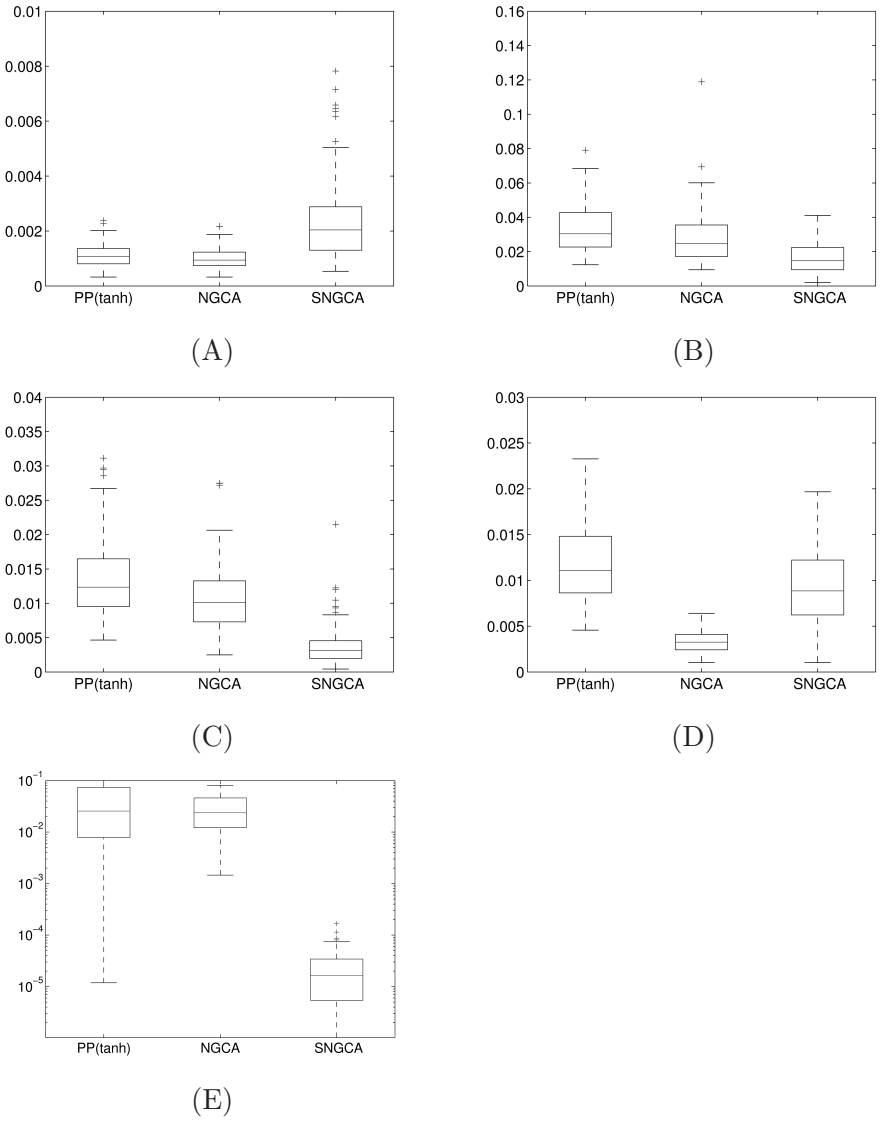


Figure 5.2: performance comparison in 10 dimensions of PP and NGCA versus SNGCA (wrt. the error criterion $\mathcal{E}(\widehat{\mathcal{I}},\mathcal{I})$ ) using the index $tanh(x)$. The doted line denotes the mean, the solid lines the variance of (3.11).

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. The smallest possible error of SNGCA heavily depends on the quality and the size of the initial MC-sampling as well as on the number of allowed iterations. In order to illustrate this interpretation we

report the progress of SNGCA wrt. estimation error $\mathcal{E}(\mathcal{I}, \widehat{\mathcal{I}})$ in each iteration for every test model.

| $j$ | $\mu_\epsilon$ | $\sigma_\epsilon^2$ |
|---|---|---|
| 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)

| $j$ | $\mu_\epsilon$ | $\sigma_\epsilon^2$ |
|---|---|---|
| 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)

| $j$ | $\mu_\epsilon$ | $\sigma_\epsilon^2$ |
|---|---|---|
| 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)

| $j$ | $\mu_\epsilon$ | $\sigma_\epsilon^2$ |
|---|---|---|
| 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)

| $j$ | $\mu_\epsilon$ | $\sigma_\epsilon^2$ |
|---|---|---|
| 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 5.1: Progress of SNGCA for test models in 10 dimensions with increasing number $j$ of iterations. The empirical mean of the error $\mathcal{E}(\widehat{\mathcal{I}}, \mathcal{I})$ defined in (3.11) is denoted by $\mu_\epsilon$ and $\sigma_\epsilon^2$ 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 $\Sigma$ of the data. In figure 5.3 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).

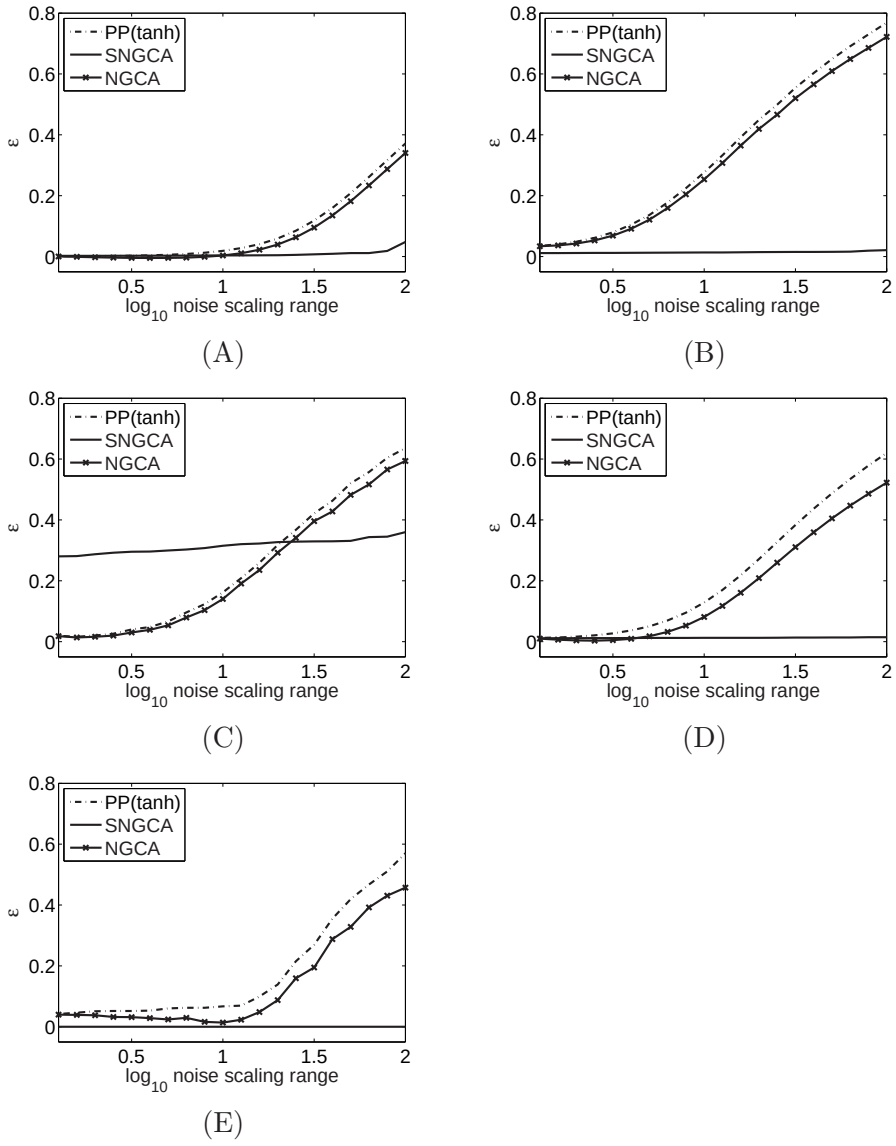


Figure 5.3: results wrt. $\mathcal{E}(\widehat{\mathcal{I}}, \mathcal{I})$ with deviations of Gaussian components following a geometrical progression on $[10^{-r}, 10^{r}]$ where $r$ is the parameter on the abscissa) .

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 direction in most cases. Only the detection of the uniform distribution by SNGCA is influenced by the condition of $\Sigma$. The next figure 5.4 compares the behavior of SNGCA compared with PP and SNGCA 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 wrt. the test models (A) to (E).

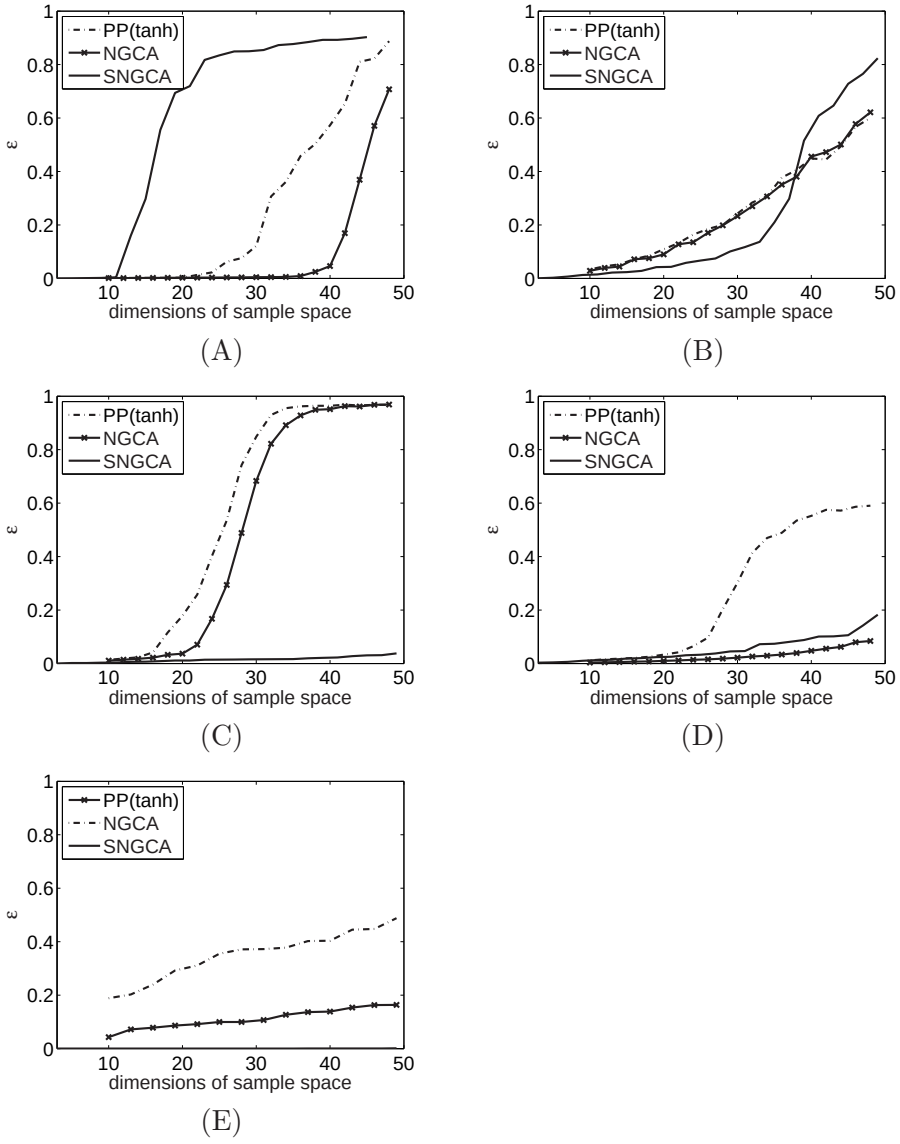


Figure 5.4: results wrt. $\mathcal{E}(\widehat{\mathcal{I}},\mathcal{I})$ with increasing number of gaussian components.

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 $d = 40$.

**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 wrt. the models (A) and (B).

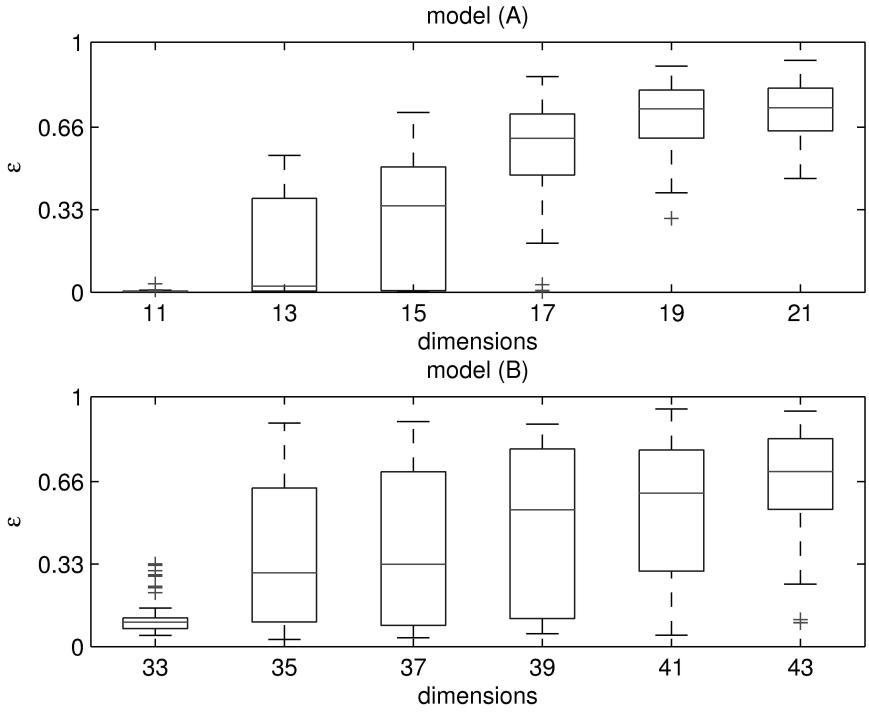


Figure 5.5: failure modes of SNGCA - upper figure: model (A) - lower figure: model(B)

Figure 5.5 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 $\sigma_{\mathcal{E}}^2$ 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 $\sigma_{\mathcal{E}}^2$ 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. Currently the performance of SNGCA is limited by the Monte-Carlo sampling of the measurement directions.

## 5.2 Application to real life examples

We consider a simulating of a mixture of oil and gas flowing under high pressure through a pipeline. Under these physical conditions different phases of the oil-gas-mixture may exist at the same time in the phase space $\Gamma$. Only some of these phase configurations in $\Gamma$ are stable over long periods of time. Consequently one expects some clusters of points in $\Gamma$ indicating the physical state of the mixture. The $12$-dimensional data set, obtained by numerical simulations of a stationary physical model, was already used before for testing techniques of dimension reduction [2]. The data set comes with a subset of training data and a subset of test data. The length of the time series is $1000$ in each dimension.

The task with this data is to find the clusters representing the stable configurations in the training data set. It is not known a priori if some dimensions are more relevant than others. However is known apriori that the data is divided into $3$ classes, indicated by different shapes of the data points.

The cluster information is not used in finding the EDR-space. Again we compare SNGCA with NGCA and PP using the tanh index. For PP and NGCA the results are shown in figure 5.6. They were already published in [27].

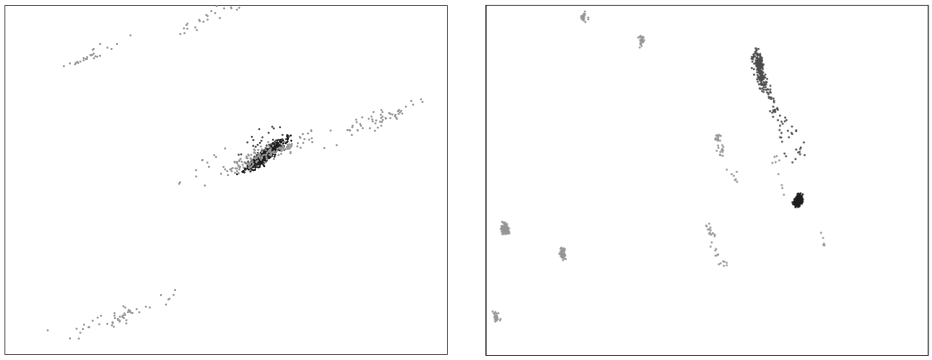

Figure 5.6: left: 2D projection of the "oil flow" data manually chosen from 3D projection obtained from by vanilla FastICA methods using the tanh index - right: projection obtained by NGCA using a combination of Fourier, tanh, Gauss-pow3 indices

Figure 5.6 shows a slice through $\Gamma$ such that the structure in the data set become visible: Using NGCA we can distinguish $10 - 11$ clusters versus at most $5$ for the PP method with index tanh.

For the SNGCA method the results are shown in the figure 5.7. SNGCA identifies 3 non-Gaussian dimensions. All figures are rotated by hand such that the separation of the cluster is illustrated at best. The next figure 5.7 shows the result of the oil-flow data obtained from SNGCA using a combination of the tanh and the asymmetric Gauss index. In this case we can distinguish $10 - 11$ clusters versus at most $5$ for the PP methods. Moreover we confirm the result of

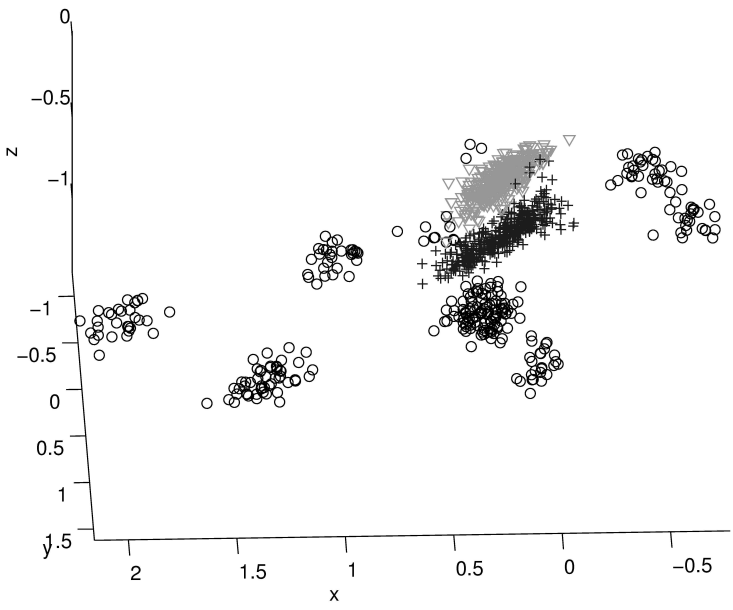


Figure 5.7: phase configurations of the "oil flow" data with apriori cluster mapping induced by crosses, circles and triangles obtained by SNGCA using a combination of asymmetric-Gauss and the tanh index

NGCA on the data set. The clusters are clearly separated from each other on the SNGCA projection. Only on the PP projection they are partially confounded in one single cluster. By applying the projection obtained from SNGCA to the test data, we found the cluster structure to be relevant. We conclude that SNGCA gives a more relevant estimation of $\mathcal{I}$ than PP. However it is found that the family of functions $h_\omega(x)$ is an important tuning parameter in SNGCA: If we use only the tanh-index, we found only 6-7 cluster are identified and they are partially confounded. Hence a combination should be used in order to cope with symmetric data distributions.

# 6 Conclusion

We proposed a slightly more sophisticated approach to non-Gaussian component analysis as already proposed in [27]. As well as NGCA the suggested method is based on a semi-parametric framework for separation an uninteresting multivariate Gaussian noise subspace from a linear subspace, where the data are non-Gaussian distributed. Both methods assume that the non-Gaussian contribution to the data density contains the structure in a given data set. The combined strategy of convex projection and structural adaption provides promising results of SNGCA. Moreover SNGCA provides an estimate for the dimension of the non-Gaussian subspace. However the method is limited by the quality and the computational effort of Monte-Carlo sampling of the measurement directions.

**Acknowledgement:** We are grateful to Gilles Blanchard from the FIRST.IDA Fraunhofer Institute Berlin for helpful discussions as well as the permission to republish the results of NGCA in this paper as well as to Jörg Polzehl from WIAS Institute Berlin. Finally the authors acknowledge financial support from the DFG research center MATHEON "Mathematics for key technologies" (FZT 86) in Berlin.

# A Statistical tests

In this section we shortly report the statistical tests on normality used the dimension reduction step of SNGCA (see section 3.3).

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 [33]. 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 $\chi^2_2$-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 $\mu$ denotes the empirical mean, $\sigma$ the empirical standard deviation of the data and $\mathcal{Z}(\cdot)$ denotes a normalizing transformations of skewness and kurtosis and. The test is more powerful wrt. an asymmetry of a distribution.

Furthermore we use the EDF-test according to Anderson-Darling [1] with the modification of Stephens [28]: 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 $\mu$ is the empirical mean and $s$ 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:

| $\alpha$ : | 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 [26] based on a regression strategy in the version given by Royston [23, 24]:

$$\begin{aligned} W &= \frac{\left(\left[1-\frac{b^2}{\sigma^2(N-1)}\right]^{\lambda}-\mu\right)}{\sigma} \quad \sim \mathcal{N}(0,1) \\ b &= \sum_{i=1}^{N/2} a_{N-i+1}(X_{N-i+1}-x_i) \\ (a_1,\ldots,a_N) &= \frac{m^\top \Sigma^{-1}}{(m^\top \Sigma^{-1^\top}\Sigma^{-1}m)^{1/2}} \end{aligned}$$

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 $\Sigma$ is the corresponding covariance matrix.

# B Theoretical study

In this appendix we will give the proofs of the theorems used in this article.

**Theorem 1.** *Let $X$ follow the distribution with the density $\rho(x)$ according to (2.1) and let $\mathbb{E}X=0$. Suppose that $\psi\in\mathcal{C}^1(\mathbb{R}^d,\mathbb{R})$ is a function fulfilling the condition*

$$\gamma(\psi):=\mathbb{E}[X\psi(X)]=0, \tag{B.14}$$

*Define*

$$\beta(\psi):=\mathbb{E}\nabla\psi(X)=\int\nabla\psi(x)\,\rho(x)\,dx, \tag{B.15}$$

*where $\nabla_x\psi$ means the gradient of $\psi$. Then $\beta(\psi)$ belongs to $\mathcal{I}$. Moreover if (B.14) is not fulfilled, then there is a $\beta\in\mathcal{I}$ such that*

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

*where $\epsilon$ is the uniform error bound:*

$$\epsilon=\left\|\Sigma^{-1}\int x\psi(x)\rho(x)\,dx\right\|_2. \tag{B.16}$$

*Hence the distance between $\beta(\psi)$ and the non-Gaussian subspace $\mathcal{I}$ is uniformly bounded as given by (B.16).*

*Proof.* (2.1) and the identity

$$\nabla_x \log\left[\phi_{\mu,\Sigma}(x)\right] = -\Sigma^{-1}(x-\mu)$$

imply

$$\begin{aligned} - & \int \psi(x)\big[\nabla \log(\rho(x))\big]\rho(x)\, dx = \\ - & \int \psi(x)\big[\nabla \log(q(Tx))\big]\rho(x)\, dx - \int \psi(x)\big[\nabla \log(\phi_{\mu,\Sigma}(x))\big]\rho(x)\, dx = \\ - & \int \psi(x) T^\top q'(Tx)\phi_{\mu,\Sigma} x\, dx + \int \psi(x)\Sigma^{-1}(x-\mu)\rho(x)\, dx \end{aligned}$$

where $q'(x)$ denotes the gradient of $q(x)$. The vector $\beta(\psi)$ with

$$\beta(\psi) \;:=\; -T^\top \int \psi(x) q'(Tx)\phi_{\mu,\Sigma}(x)\, dx$$

obviously belongs to $\mathcal{I}$. Suppose now the condition (B.14) is fulfilled. Then it holds that

$$\Sigma^{-1}\left[\int x\psi(x)\rho(x)\, dx - \mu \int \psi(x)\rho(x)\, dx\right] = 0$$

Thus we know that $\beta(\psi) \in \mathcal{I}$. Otherwise if the condition (B.14) is not fulfilled, we it follows from from (B.17) that there is a $\beta \in \mathcal{I}$ such that

$$\|\beta(\psi) - \beta\|_2 \;=\; \left\|\Sigma^{-1}\int (x-\mu)\psi(x)\rho(x)\, dx\right\|_2 . \tag{B.17}$$

Let $u \in \mathbb{R}^d$. Then it holds that

$$\int \psi(x+u)\rho(x)\, dx = \int \psi(x)\rho(x-u)\, dx \tag{B.18}$$

Using the regularity conditions on $\psi$ and $\rho$, we differentiate (B.18) with respect to $u$ and use the identity

$$\nabla \log(\rho(x)) = \big[\nabla\rho(x)\big]\rho(x)$$

This yields:

$$\int \rho(x)\nabla\psi(x)\, dx \;=\; -\int \psi(x)\big[\nabla \log(\rho(x))\big]\rho(x)\, dx$$

From this consideration we get an equivalent expression for $\beta(\psi)$:

$$\beta(\psi) = \int [\nabla\psi(x)]\rho(x)\, dx$$

□

## B.1 Estimation accuracy

This section presents some upper bounds on the accuracy of approximating the target non-Gaussian subspace $\mathcal{I}$ by the vectors $\widehat{\beta}$ which come out of our algorithm.

By $\mathcal{B}_d$ we denote the unit ball in $\mathbb{R}^d$. Let $h(\omega, x)$ be a continuously differentiable function of $\omega \in \mathcal{B}_d$ and $x \in \mathbb{R}^d$ such that $h(\omega, x)$, $\nabla_x h$, $\nabla_\omega h$ and $\nabla_\omega \nabla_x h$ are uniformly continuous and bounded on $\mathcal{B}_d \times \mathbb{R}^d$ and let some set of unit vectors $\omega_1, \ldots, \omega_L$ and also a unit vector $\xi$ be given. Consider the optimization problem

$$\{\widehat{c}_l\} = \arg\min_{\|c\|_1 \le 1} \Big\| \xi - \sum_l c_l \, \widehat{\eta}_l \Big\|_2 \qquad \text{subject to} \qquad \Big\| \sum_l c_l \, \widehat{\gamma}_l \Big\|_2 = 0 \tag{B.19}$$

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 $\|(\mathbf{1}_d - \Pi_{\mathcal{I}})\widehat{\beta}\|_2$. Again $\Pi_{\mathcal{I}}$ means the orthogonal projector on $\mathcal{I}$.

**Theorem 2.** *There is a constant $C_1$ depending on function $h$ and the dimension $d$ only such that*

$$\mathbb{E}\|(\mathbf{1}_d - \Pi_{\mathcal{I}})\widehat{\beta}\|_2^2 \le C_1/N \, .$$

*Proof.* For the proof we apply one result from the theory of empirical processes given in Theorem 4 below in the appendix. Corollary (B.21) of this theorem applied to every coordinate of the vectors $\widehat{\gamma}_l$ yields in the obvious way that

$$\mathbb{E} \max_l \|\widehat{\gamma}_l - \gamma_l\|_2^2 C_0/N.$$

This yields for any $c \in \mathbb{R}^L$ with $\|c\|_1 \le 1$

$$\mathbb{E}\Big\| \sum_l c_l \, \widehat{\gamma}_l - \sum_l c_l \, \gamma_l \Big\|_2^2 \le C_0/N.$$

Similarly

$$\mathbb{E}\Big\| \sum_l c_l \, \widehat{\eta}_l - \sum_l c_l \, \eta_l \Big\|_2^2 \le C_0/N. \tag{B.20}$$

For the solution $\widehat{c}$ of (B.19) this yields in view of the constraint $\sum_l c_l \widehat{\gamma}_l = 0$ that

$$\mathbb{E}\Big\| \sum_l \widehat{c}_l \, \gamma_l \Big\|_2^2 \le C_0/N.$$

The result of Theorem 1 ensures that

$$\mathbb{E}\Big\| (\mathbf{1}_d - \Pi_{\mathcal{I}}) \sum_l \widehat{c}_l \, \eta_l \Big\|_2^2 \le C_0 \|\Sigma^{-1}\|.$$

This yields the result in view of (B.20). □

## B.2 Significance bound

Unfortunately the result of Theorem 2 only explains the distance from the constructed vector $\widehat{\beta}$ from the target space and tells nothing about the projection of $\widehat{\beta}$ onto $\mathcal{I}$. In particular, if this projection is small then the vector $\widehat{\beta}$ is not informative.

Now we aim to present a result which gives some sufficient conditions for vector $\widehat{\beta}$ to be informative. Consider the couple of optimization problems

$$\begin{aligned} \{c_l^*\} &= \arg\min_{\|c\|_1 \le 1} \Big\|\xi - \sum_l c_l\,\eta_l\Big\|_2 \qquad \text{subject to} \qquad \Big\|\sum_l c_l\,\gamma_l\Big\|_2 = 0, \\ \{\widehat{c}_l\} &= \arg\min_{\|c\|_1 \le 1} \Big\|\xi - \sum_l c_l\,\widehat{\eta}_l\Big\|_2 \qquad \text{subject to} \qquad \Big\|\sum_l c_l\,\widehat{\gamma}_l\Big\|_2 \le \varepsilon_r \end{aligned}$$

where $\varepsilon_r \ge 0$ is a relaxation parameter. The second problem can be viewed as the empirical counterpart of the first one. Define also $\beta^* = \sum_l c_l^*\,\eta_l$. By Theorem 1 $\beta^*$ belongs to the non-Gaussian space $\mathcal{I}$. The vector $\beta^*$ is the *convex projection* of $\xi$ onto $\mathcal{I}$. It is natural to measure the significance of $\beta^*$ by the value $\|\xi - \beta^*\|_2$: $\beta^*$ is significant if $\|\xi - \beta^*\|_2 \le 1 - \delta$ for some $\delta$ which is larger in order than $\varepsilon_r$. Alternatively, $\|\Pi_{\mathcal{I}}(\xi - \beta^*)\|_2$ can be considered. Here significance means that $\|\Pi_{\mathcal{I}}\xi\|_2 \ge \delta$ and $\|\Pi_{\mathcal{I}}(\xi - \beta^*)\|_2 \le (1-\delta)\|\Pi_{\mathcal{I}}\xi\|_2$. The next result shows that if $\beta^*$ is significant, then $\widehat{\beta}$ is significant as well.

**Theorem 3.** *Let $\mathcal{A}_{\varepsilon_r}$ be a random set on which*

$$\max_l \|\gamma_l - \widehat{\gamma}_l\|_2 \le \varepsilon_r, \qquad \max_l \|\eta_l - \widehat{\eta}_l\|_2 \le \varepsilon_r.$$

*Then*

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

*Proof.* Observe that on $\mathcal{A}_{\varepsilon_r}$ the solution $c^* = \{c_l^*\}$ of the "ideal" optimization problem fulfills the constraint of the empirical one. Indeed,

$$\Big\|\sum_l c_l^*\,\widehat{\gamma}_l\Big\|_2 = \Big\|\sum_l c_l^*(\widehat{\gamma}_l - \gamma_l)\Big\|_2 \le \varepsilon_r.$$

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 $\mathcal{A}_{\varepsilon_r}$

$$\Big\|\xi - \sum_l c_l^*\,\widehat{\eta}_l\Big\|_2 - \Big\|\xi - \sum_l c_l^*\,\eta_l\Big\|_2 \le \Big\|\sum_l c_l^*(\widehat{\eta}_l - \eta_l)\Big\|_2 \le \varepsilon_r$$

and hence, for $\widehat{\beta} = \sum_l \widehat{c}_l\,\widehat{\eta}_l$

$$\big\|\xi - \widehat{\beta}\big\|_2 \le \big\|\xi - \beta^*\big\|_2 + \varepsilon_r$$

and the first assertion follows. For second one use additionally that $(\mathbf{1}_d - \Pi_{\mathcal{I}})\beta^* = 0$ and $\|(\mathbf{1}_d - \Pi_{\mathcal{I}})\widehat{\beta}\|_2 \le C_1\varepsilon_r$ on $\mathcal{A}_{\varepsilon_r}$, see the proof of Theorem 2. □

## B.3 Uniform error bound

Here we present one useful result from the empirical process theory and some its corollaries. Similar statements under a bit different assumptions can be found e.g. in [31].

**Theorem 4.** *Let $f(\omega, x)$ be a continuously differentiable function of $\omega \in \mathcal{B}_d$ and $x \in \mathbb{R}^d$ such that $f(\omega, x)$ and $\nabla_\omega f(\omega, x)$ are uniformly continuous and bounded on $\mathcal{B}_d \times \mathbb{R}^d$. Define*

$$\zeta(\omega) = N^{1/2}\bigl\{\mathbb{E}_N f(\omega, X) - \mathbb{E} f(\omega, X)\bigr\}$$

*and $\zeta(\omega, \omega') = \zeta(\omega) - \zeta(\omega')$. Then there are two constants $\mathfrak{n}_0 > 0$ and $\lambda^* > 0$ such that for any $\omega^\circ \in \mathcal{B}_d$ and any $\mu > 0$, $\lambda \le \lambda^*$*

$$\log \mathbb{E} \exp\Bigl[\frac{\lambda}{\mu} \sup_{\omega \in B(\mu, \omega^\circ)} \zeta(\omega, \omega^\circ)\Bigr] \le \mathfrak{n}_0 \lambda^2 + \mathfrak{e}_d$$

*where $B(\mu, \omega) = \{\omega' : \|\omega - \omega'\|_2 \le \mu\}$ and $\mathfrak{e}_d = \sum_{k=1}^{\infty} 2^{-k} \log(2^{kd})$.*

Before proving this result, we present some useful corollaries. As application with $\mu = 1$ and $\omega^\circ = 0$ yields the bounded exponential moments for the supremum of differences $\zeta(\omega, \omega^\circ)$ over all $\omega \in \mathcal{B}_d$. This immediately implies the result about the supremum of the $\zeta(\omega)$:

$$\mathbb{E}\bigl|\sup_{\omega \in \mathcal{B}_d} \zeta(\omega)\bigr|^2 \le C_0 \tag{B.21}$$

for some fixed constant $C_0$.

*Proof.* Define for $\omega, \omega' \in \mathcal{B}_d$

$$\xi(\omega, \omega') = \frac{\zeta(\omega, \omega')}{\|\omega - \omega'\|_2}.$$

With $u = (\omega - \omega')/\|\omega - \omega'\|$

$$\xi(\omega, \omega') = \int_0^1 u^\top \nabla \zeta(\omega + tu) dt.$$

The conditions of the theorem easily imply that there is some $\lambda_1 > 0$ such that for any $\lambda \le \lambda_1$ and any unit vectors $u$ and $\omega$

$$g(\lambda; \omega, u) := \log \mathbb{E} \exp\Bigl[\lambda u^\top \{\nabla_\omega f(\omega, X_1) - \mathbb{E} \nabla_\omega f(\omega, X_1\}\Bigr] \le \mathfrak{n}_0 \lambda^2$$

where $\mathfrak{n}_0$ depends only on $\lambda_1$ and on the upper bound of the gradient $\nabla_\omega f(\omega, x)$. Indeed, it suffices to note that $g(\lambda; \omega, u)$ is analytic in $\lambda$ and satisfies $g(0; \omega, u) = 0$ and $g'_\lambda(0; \omega, u) = 0$. Independence of the $X_i$'s yields for $\lambda \le \lambda_1 N^{1/2}$

$$\log \mathbb{E}\bigl\{\exp \lambda u^\top \nabla \zeta(\omega)\bigr\} = N g(\lambda N^{-1/2}; \omega, u) \le N \mathfrak{n}_0 \, (\lambda N^{-1/2})^2 = \mathfrak{n}_0 \lambda^2.$$

Therefore, for $\lambda \le \lambda_1 N^{1/2}$

$$\begin{aligned}\log \mathbb{E}\exp\{2\lambda\xi(\omega,\omega')\} &= \log \mathbb{E}\exp\{\lambda\int_0^1 u^\top \nabla\zeta(\omega+tu)dt\} \\ &\le \int_0^1 \log \mathbb{E}\exp\{\lambda u^\top \nabla\zeta(\omega+tu)dt\} \le \mathfrak{n}_0\lambda^2. \quad \text{(B.22)}\end{aligned}$$

The rest of the proof is based on the standard chaining argument (see e.g., [31]). For any integer $k \ge 0$, there exists a $2^{-k}\mu$-net $\mathcal{D}_k(\mu)$ in the ball $B(\mu,\omega^\circ)$ having the cardinality $\mathbb{N}(2^{-k}\mu,\mu) \le 2^{kd}$ in the sense that

$$B(\mu,\omega^\circ) \subset \bigcup_{\omega\in\mathcal{D}_k(\mu)} B(2^{-k}\mu,\omega).$$

Using the nets $\mathcal{D}_k(\mu)$ with $k = 1,\dots,K-1$, one can construct a chain connecting an arbitrary point $\omega$ in $\mathcal{D}_K(\mu)$ and $\omega^\circ$. It means that one can find points $\omega_k \in \mathcal{D}_k(\mu)$, $k = 1,\dots,K-1$, such that $\|\omega_k - \omega_{k-1}\|_2 \le 2^{-k+1}\mu$ for $k = 1,\dots,K$. Here we denoted for $\omega_K = \omega$, and $\omega_0 = \omega^\circ$. Notice that $\omega_k$ can be constructed recurrently $\omega_{k-1} = \tau_{k-1}(\omega_k)$, $k = K,\dots,1$, where

$$\tau_{k-1}(\omega) = \operatorname*{arg\,min}_{\omega'\in\mathcal{D}_{k-1}(\mu)} \|\omega-\omega'\|_2.$$

It obviously holds

$$\zeta(\omega,\omega^\circ) = \sum_{k=1}^K \zeta(\omega_k,\omega_{k-1}).$$

In view of the definition of $\xi(\cdot,\cdot)$

$$\zeta(\omega_k,\omega_{k-1}) = \|\omega_k-\omega_{k-1}\|_2 \times \xi(\omega_k,\omega_{k-1}) = 2\mu c_k \xi(\omega_k,\omega_{k-1})$$

with $c_k = c_k(\omega) = \|\omega_k-\omega_{k-1}\|_2/(2\mu) \le 2^{-k}$, thus resulting in

$$\begin{aligned}\sup_{\omega\in\mathcal{D}_K(\mu)} \zeta(\omega,\omega^\circ) &\le \sum_{k=1}^K \sup_{\omega'\in\mathcal{D}_k(\mu)} \zeta(\omega',\tau_{k-1}(\omega')) \\ &\le 2\mu\sum_{k=1}^K \sup_{\omega'\in\mathcal{D}_k(\mu)} c_k(\omega')\xi(\omega',\tau_{k-1}(\omega')). \quad \text{(B.23)}\end{aligned}$$

Next, we use the elementary inequality

$$\log \mathbb{E}\exp\Big(\sum_{k=1}^K \lambda_k\xi_k\Big) \le \sum_{k=1}^K \lambda_k \log \mathbb{E}\mathrm{e}^{\xi_k} \quad \text{(B.24)}$$

which holds for any r.v.'s $\xi_k$ and any nonnegative coefficients $\lambda_k$ with $\Lambda = \sum_{k=1}^K \lambda_k \le 1$. Its proof is based on the convexity of $\mathrm{e}^x$ and concavity of $x^\Lambda$

$$\begin{aligned}\mathbb{E}\exp\Big[\Lambda\frac{1}{\Lambda}\sum_{k=1}^K \lambda_k\big(\xi_k - \log\mathbb{E}\mathrm{e}^{\xi_k}\big)\Big] &\le \mathbb{E}^\Lambda \exp\Big[\frac{1}{\Lambda}\sum_{k=1}^K \lambda_k\big(\xi_k-\log\mathbb{E}\mathrm{e}^{\xi_k}\big)\Big] \\ &\le \Big[\frac{1}{\Lambda}\sum_{k=1}^K \lambda_k \mathbb{E}\exp\big(\xi_k-\log\mathbb{E}\mathrm{e}^{\xi_k}\big)\Big]^\Lambda = 1.\end{aligned}$$

Combining (B.23) and (B.24) and (B.22) yields in view of $\sum_{k=1}^{K} 2^{-k} \le 1$ and $2^k c_k(\omega) \le 1$

$$\begin{aligned}
&\log\mathbb{E}\exp\Big[\frac{\lambda}{\mu}\sup_{\omega\in\mathcal{D}_K(\mu)}\zeta(\omega,\omega^\circ)\Big] \\
&\quad\le \log\mathbb{E}\exp\Big[2\lambda\sum_{k=1}^{K}\sup_{\omega'\in\mathcal{D}_k(\mu)} c_k(\omega')\xi(\omega',\tau_{k-1}(\omega'))\Big] \\
&\quad\le \sum_{k=1}^{K} 2^{-k}\log\Big[\mathbb{E}\sup_{\omega'\in\mathcal{D}_k(\mu)}\exp\big\{2^k c_k(\omega')\times 2\lambda\xi(\omega',\tau_{k-1}(\omega'))\big\}\Big] \\
&\quad\le \sum_{k=1}^{K} 2^{-k}\log\Big[\sum_{\omega'\in\mathcal{D}_k(\mu)}\mathbb{E}\exp\big\{2^k c_k(\omega')\times 2\lambda\xi(\omega',\tau_{k-1}(\omega'))\big\}\Big] \\
&\quad\le \sum_{k=1}^{K} 2^{-k}\big\{\log\mathbb{N}(2^{-k}\mu,\mu)+\mathfrak{n}_0\lambda^2\big\}.
\end{aligned}$$

These inequalities yield

$$\begin{aligned}
\log\mathbb{E}\exp\Big[\frac{\lambda}{\mu}\sup_{\omega\in B(\mu,\omega^\circ)}\zeta(\omega,\omega^\circ)\Big] &= \lim_{K\to\infty}\log\mathbb{E}\exp\Big[\frac{\lambda}{\mu}\sup_{\omega\in\mathcal{D}_K(\mu)}\zeta(\omega,\omega^\circ)\Big] \\
&\le \sum_{k=1}^{\infty} 2^{-k}\big\{\mathfrak{n}_0(\lambda)+\log\mathbb{N}(2^{-k}\mu,\mu)\big\} \le \mathfrak{n}_0\lambda^2+\mathfrak{e}_d
\end{aligned}$$

where $\mathfrak{e}_d = \sum_{k=1}^{\infty} 2^{-k}\log(2^{kd})$ which completes the proof of the theorem. □

# References


[1] F.J. Anscombe and W.J. Glynn. Distribution of kurtosis statistic for normal statistics. *Biometrika*, 70(1):227–234, 1983.

[2] M. Svensen C.M. Bishop and C.K.I. Wiliams. Gtm: The generative topographic mapping. *Neural Computation*, 10(1):215–234, 1998.

[3] R.D. Cook. Principal hessian directions revisited. *J. Am. Statist. Ass.*, 93:85–100, 1998.

[4] T.M. Cover and J.A. Thomas. *Elements of Information Theory.* Wiley Series in Telecommunications. Wiley and Sons, New York, 1991.

[5] P. Diaconis and D. Friedman. Asymptotics of graphical projection pursuit. *Annals of Statistics*, 12(3):793–815, 1984.

[6] D.L. Donoho. Unconditional bases are optimal bases for data compression and for statistical estimation. *Appl. Computational Harmonic analysis*, 1(100-115), 1993.

[7] D.L. Donoho. Sparse components of images and optimal atomic decomposition. *Constructive Approximation*, 17:353–382, 2001.

[8] B. S. Everitt. *An Introduction to Latent Variable Models.* Monographs on Statistics and Applied Probability. Chapman and Hall, London,, 1984.

[9] 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.

[10] T. Hastie, R. Tibshirani, and J. Friedman. *The elements of statistical learning.* Springer Series in Statistcs. Springer, 2001.

[11] M. Hristache, A. Juditsky, J. Polzehl, and V. Spokoiny. Structure adaptive approach for dimension reduction. *Ann. Statist.*, 29(6):1537–1566, 2001.

[12] A. Hyvärinen. Survey on independent component analysis. *Neural Computing Surveys*, 2:94–128, 1999.

[13] 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.

[14] I.T. Jolliffe. *Principal Component Analysis.* Springer Series in Statistics. Springer, Berlin and New York, 2nd edition, 2002.

[15] H.C. Thode Jr. *Testing for Normality.* Marcel Dekker, New York., 2002.

[16] 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.

[17] Erik Learned-Miller and III. John W. Fisher. Ica using spacings estimates of entropy. *Journal of Machine Learning Research*, 4:1271–1295, 2003.

[18] K.C. Li. Sliced inverse regression for dimension reduction. *J. Am. Statist. Ass.*, 86:316–342, 1991.

[19] 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.

[20] 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.

[21] Yu. E. Nesterov. Rounding of convex sets and efficient gradient methods for linear programming problems. *Discussion Paper 2004-4, CORE, Catholic University of Louvain, Louvain-la-Neuve, Belgium*, 2004.

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

[23] J.P. Royston. An extension of shapiro and wilks' w test for normality to large samples. *Applied Statistics*, 31:115–124, 1982.

[24] J.P. Royston. The w test for normality. *Applied Statistics*, 21:176–180, 1982.

[25] 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.

[26] S.S. Shapiro and M.B. Wilk. An analysis of variance test for normality. *Biometrika*, 52:591–611, 1965.

[27] 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*, preprint TR05-003, 2005.

[28] M. A. Stephens. *Goodness of Fit Techniques*, chapter Tests based on Goodness of Fit. D'Agostino, R. B. and Stephens, M. A., 1986.

[29] R. Tibshirani. Regression shrinkage and selection via the lasso. *J. Royal. Statist. Soc B.*, 58(1):267–288, 1996.

[30] M.J. Todd and E.A. Yildrim. On khachiyan's algorithm for the computation of minimum volume enclosing ellipsoids. *Technical Report, preprint*, 2005.

[31] A. van der Vaart and J.A. Wellner. *Weak Convergence and Empirical Proccesses.* Springer Series in Statistics. Springer – New York, 1996.

[32] L. Wasserman. *All of Nonparametric Statistics.* Springer Texts in Statistcs. Springer, 2006.

[33] J.H. Zar. *Biostatistical Analysis, (2nd ed.).* NJ: Prentice-Hall, Englewood Cliffs., 1999.

[34] H. Zou, T. Hastie, and R. Tibshirani. Sparse principal component analysis. *Journal of Computational and Graphical Statistics*, 2004.