

Sparse non Gaussian component analysis by semidefinite programming

Elmar Diederichs · Anatoli Juditsky ·
Arkadi Nemirovski · Vladimir Spokoiny

Received: 9 March 2010 / Accepted: 28 January 2013 / Published online: 3 April 2013
© The Author(s) 2013

Abstract Sparse non-Gaussian component analysis is an unsupervised linear method of extracting any structure from high-dimensional distributed data based on estimating a low-dimensional non-Gaussian data component. In this paper we discuss a new approach with known apriori reduced dimension to direct estimation of the projector on the target space using semidefinite programming. The new approach avoids the estimation of the data covariance matrix and overcomes the traditional separation of element estimation of the target space and target space reconstruction. This allows to reduced the sampling size while improving the sensitivity to a broad variety of deviations from normality. Moreover the complexity of the new approach is limited to $O(d \log d)$. We also discuss the procedures which allows to recover the structure when its effective dimension is unknown.

Keywords Dimension reduction · Non-Gaussian components analysis · Feature extraction

Editor: Jennifer Dy.

E. Diederichs (✉) · V. Spokoiny
Weierstrass Institute, Mohrenstr. 39, 10117 Berlin, Germany
e-mail: diederich@wias-berlin.de

V. Spokoiny
e-mail: spokoiny@wias-berlin.de

A. Juditsky
LJK, Université J. Fourier, BP 53, 38041 Grenoble cedex 9, France
e-mail: juditsky@imag.fr

A. Nemirovski
ISyE, Georgia Institute of Technology, Atlanta, GA 30332, USA
e-mail: nemirovs@isye.gatech.edu

V. Spokoiny
Humboldt University, Berlin Germany

V. Spokoiny
Moscow Institute of Physics and Technology, Moscow, Russia

1 Introduction

Numerous statistical applications are confronted today with the so-called curse of dimensionality (cf. Hastie et al. 2001; Wasserman 2006). Using high-dimensional datasets implies an exponential increase of computational effort for many statistical routines, while the data thin out in the local neighborhood of any given point and classical statistical methods become unreliable. When a random phenomenon is observed in the high dimensional space \mathbb{R}^d the “intrinsic dimension” m covering degrees of freedom associated with same features may be much smaller than d . Then introducing *structural assumptions* allows to reduce the problem complexity without sacrificing any statistical information (Mizuta 2004; Roweis and Saul 2000). This study focuses on structural analysis of high dimensional unlabeled data. It is well known that the projection of high dimensional data on a randomly selected low-dimensional subspace is nearly normal. This fact can be intuitively explained by the central limit theorem. At the same time, interesting data features such as multimodality, limited support, or heavy tails are essentially non-Gaussian. These “deviations from normality” are easily observable in low-dimensional data (say, up to dimension 3), but they are hardly detectable when the data is high-dimensional (when data dimension exceeds 10). Our structural assumption is that the informative non-Gaussian part of the data is (approximately) located in a low-dimensional linear subspace \mathcal{I} while the non-informative noisy Gaussian part of the data distribution is full dimensional. If such non-Gaussian subspace were known, one could reduce data dimension and complexity by projecting the data onto it without losing any substantial information. This assumption naturally leads to the problem of recovering the low dimensional informative non-Gaussian subspace from the original data. In what follows we focus on the problem of estimating the projector Π onto an m -dimensional non-Gaussian subspace from d -dimensional data when $m \ll d$. This is a semiparametric statistical problem with a target parameter of dimension $O(dm)$ (the projection matrix) and a general m -dimensional nonparametric nuisance (the non-Gaussian data distribution). An important feature of the problem is that the target space—the set of m -dimensional projectors Π in \mathbb{R}^d —is not convex, so efficient methods of convex optimization cannot be applied to this problem directly, and the recovery of the parameter requires application of computationally expensive relaxation techniques. Another challenging task is the design of an estimation routine which is robust with respect to the unknown covariance matrix of the Gaussian component of the data. These characteristics of the problem make it hard even in the case of moderate dimension d .

Modern statistical literature offers two types of methods for solving a similar dimension reduction problem. The principal component analysis (PCA) looks for the subspace of the highest data variability but is limited to Gaussian data and is inefficient when recovering the non-Gaussian features. Another well known statistical technique, the independent component analysis (ICA) is based on the rather restrictive assumption that the full dimensional data distribution is a superposition of independent one-dimensional non-Gaussian sources; see e.g. Hyvärinen (1999) and references therein.

A novel approach to recovering the non-Gaussian subspace have been proposed recently in Blanchard et al. (2006), Kawanabe et al. (2007). The procedure called NGCA (for *Non-Gaussian Component Analysis*) decomposes the original problem of dimension reduction into two tasks: the first one is to extract from the data a number of candidate vectors $\hat{\beta}_j$ which are “close” to \mathcal{I} . The second is to recover an estimation $\hat{\Pi}$ of the projector Π^* on \mathcal{I} from the collection $\{\hat{\beta}_j\}$. An essential disadvantage of the original NGCA estimation procedure is that it uses the ICA algorithm to construct the “initial guess” for \mathcal{I} . As a result, it shares with the ICA its drawbacks. In particular, NGCA (same as projection pursuit of

Hyvärinen 1999) is extremely sensitive to the condition number of the covariance matrix of the Gaussian component. To overcome this problem, the NGCA approach has been then developed into SNGCA (for *Sparse NGCA*), in Diederichs et al. (2009). The SNGCA algorithm does not rely upon the ICA, and appears to be more stable with respect to the parameters of the distribution of the Gaussian component. However, that procedure relies upon solving a convex optimization problem per candidate vector $\hat{\beta}_j$, and its numerical cost becomes prohibitive for problem dimension dimensions $d \geq 50$. Furthermore, the iterative nature of the SNGCA makes the theoretical study of the resulting estimation procedure extremely involved. In this paper we discuss a new method of NGCA which represents the problem of recovering the non-Gaussian subspace as a minmax problem. Special relaxation arguments allow to reduce this problem to a semidefinite convex-concave saddle-point optimization which can be effectively solved by modern methods of convex optimization. When compared to previous implementations of the SNGCA in Blanchard et al. (2006), Kawanabe et al. (2007), Diederichs et al. (2009), the new approach “shortcuts” the intermediary stages and makes better use of available information for estimation of \mathcal{I} . Furthermore, it allows to treat in a transparent way the case of unknown dimension m of the target space \mathcal{I} . Finally, the representation of the estimator $\hat{\mathcal{I}}$ as a solution of a semidefinite optimization problem allows to provide a rigorous theoretical study of the accuracy of the proposed method. In particular, our theoretical results in Sect. 3 claim root- n consistency of $\hat{\mathcal{I}}$ (up to a “log-factor”) and hence its near optimality.

The paper is organized as follows: in Sect. 2 we present the setup of SNGCA and briefly review some existing techniques. Then in Sect. 3 we introduce the new approach to recovery of the non-Gaussian subspace and analyze its accuracy. Further in Sect. 4 we describe the use of the dual extrapolation scheme for linear matrix games for solving the semidefinite optimization problem. Finally we provide a simulation study in Sect. 5, where we compare the performance of the proposed algorithm SNGCA to that of some known projective methods of feature extraction.

2 Sparse non-Gaussian component analysis

2.1 The setup

The *Non-Gaussian Component Analysis* (NGCA) approach is based on the assumption that a high dimensional distribution tends to be normal in almost any randomly selected direction. This intuitive fact can be justified by the central limit theorem when the number of directions tends to infinity. It leads to the NGCA-assumption: the data distribution is a superposition of a full dimensional Gaussian distribution and a low dimensional non-Gaussian component. In many practical problems like clustering or classification, the Gaussian component is uninformative and it is treated as noise. The approach suggests to identify the non-Gaussian component and to use it for the further analysis.

The NGCA set-up can be formalized as follows; cf. Blanchard et al. (2006). Let X_1, \dots, X_N be i.i.d. from a distribution \mathbf{P} in \mathbb{R}^d describing the random phenomenon of interest. We suppose that \mathbf{P} possesses a density ρ w.r.t. the Lebesgue measure on \mathbb{R}^d , which can be decomposed as follows:

$$\rho(x) = \phi_{\mu, \Sigma}(x)q(Tx). \quad (1)$$

Here $\phi_{\mu, \Sigma}$ denotes the density of the multivariate normal distribution $\mathcal{N}(\mu, \Sigma)$ with parameters $\mu \in \mathbb{R}^d$ (expectation) and $\Sigma \in \mathbb{R}^{d \times d}$ positive definite (covariance matrix). The function

$q : \mathbb{R}^m \rightarrow \mathbb{R}$ with $m \leq d$ is positive and bounded. $T \in \mathbb{R}^{m \times d}$ is an unknown linear mapping. We refer to $\mathcal{I} = \text{range } T$ as *target* or *non-Gaussian subspace*. Note that though T is not uniquely defined, \mathcal{I} is well defined, same as the Euclidean projector Π^* on \mathcal{I} . In what follows, unless it is explicitly specified otherwise, we assume that the *effective dimension* m of \mathcal{I} is known *a priori*. For the sake of simplicity we assume that the expectation of X vanishes: $\mathbf{E}[X] = 0$.

The model (1) allows for the following interpretation (cf. Sect. 2 of Blanchard et al. 2006): suppose that the observation $X \in \mathbb{R}^d$ can be decomposed into $X = Z + \xi$, where Z is an “informative low-dimensional signal” such that $Z \in \mathcal{I}$, \mathcal{I} being an m -dimensional subspace of \mathbb{R}^d , and ξ is independent and Gaussian. One can easily show (see, e.g., Lemma 1 of Blanchard et al. 2006) that in this case the density of X can be represented as (1).

2.2 Basics of SNGCA estimation procedure

The estimation of \mathcal{I} relies upon the following result, proved in Blanchard et al. (2006): suppose that the function q is smooth, then for any smooth function $\psi : \mathbb{R}^d \rightarrow \mathbb{R}$ the assumptions of (1) and $\mathbf{E}[X] = 0$ ensure that for

$$\beta(\psi) \stackrel{\text{def}}{=} \mathbf{E}[\nabla \psi(X)] = \int \nabla \psi(x) \rho(x) dx, \tag{2}$$

there is a vector $\beta \in \mathcal{I}$ such that

$$\|\beta(\psi) - \beta\|_2 \leq \|\Sigma^{-1} \mathbf{E}[X \psi(X)]\|_2$$

where $\nabla \psi$ denotes the gradient of ψ and $\|\cdot\|_p$ is the standard ℓ_p -norm on \mathbb{R}^d . In particular, if ψ satisfies $\mathbf{E}[X \psi(X)] = 0$, then $\beta(\psi) \in \mathcal{I}$. Consequently

$$\|(I - \Pi^*)\beta(\psi)\|_2 \leq \left\| \Sigma^{-1} \int x \psi(x) \rho(x) dx \right\|_2, \tag{3}$$

where I is the d -dimensional identity matrix and Π^* is the Euclidean projector on \mathcal{I} .

The above result suggests the following two-stage estimation procedure: first compute a set of estimates $\{\widehat{\beta}_\ell\}$ of elements $\{\beta_j\}$ of \mathcal{I} , then recover an estimation of \mathcal{I} from $\{\widehat{\beta}_\ell\}$. This heuristic has been first used to estimate \mathcal{I} in Blanchard et al. (2006). To be more precise, the construction implemented in Blanchard et al. (2006) can be summarized as follows: let for a family $\{h_\ell\}$, $\ell = 1, \dots, L$ of smooth bounded (test) functions on \mathbb{R}^d

$$\gamma_\ell \stackrel{\text{def}}{=} \mathbf{E}[X h_\ell(X)], \quad \eta_\ell \stackrel{\text{def}}{=} \mathbf{E}[\nabla h_\ell(X)], \tag{4}$$

and let

$$\widehat{\gamma}_\ell \stackrel{\text{def}}{=} N^{-1} \sum_{i=1}^N X_i h_\ell(X_i), \quad \widehat{\eta}_\ell \stackrel{\text{def}}{=} N^{-1} \sum_{i=1}^N \nabla h_\ell(X_i) \tag{5}$$

be their “empirical counterparts”. The set of “approximating vectors” $\{\widehat{\beta}_\ell\}$ used in Blanchard et al. (2006) is as follows: $\widehat{\beta}_\ell = \widehat{\eta}_\ell - \widehat{\Sigma}^{-1} \widehat{\gamma}_\ell$, $\ell = 1, \dots, L$, where $\widehat{\Sigma}$ is an estimate of the covariance matrix Σ . The projector estimation at the second stage is $\widehat{\Pi} = \sum_{j=1}^m e_j e_j^T$, where e_j , $j = 1, \dots, m$, are m principal eigenvectors of the matrix $\sum_{\ell=1}^L \widehat{\beta}_\ell \widehat{\beta}_\ell^T$. A numerical study,

provided in Blanchard et al. (2006), has shown that the above procedure can be used successfully to recover \mathcal{I} . On the other hand, such implementation of the two-stage procedure possesses two important drawbacks: it relies upon the estimation of the covariance matrix Σ of the Gaussian component, which can be hard even for moderate dimensions d . Poor estimation of Σ then will result in badly estimated vectors $\widehat{\beta}_\ell$, and as a result, poorly estimated \mathcal{I} . Further, using the eigenvalue decomposition of the matrix $\sum_{\ell=1}^L \widehat{\beta}_\ell \widehat{\beta}_\ell^T$ entails that the variance of the estimation $\widehat{\Pi}$ of the projector Π^* on \mathcal{I} is proportional to the number L of test-functions. As a result, the estimation procedure is restricted to utilizing only relatively small families $\{h_\ell\}$, and is sensitive to the initial selection of “informative” test-functions.

To circumvent the above limitations of the approach of Blanchard et al. (2006) a different estimation procedure has been proposed in Diederichs et al. (2009). In that procedure the estimates $\widehat{\beta}$ of vectors from the target space are obtained by the method, which was referred to as *convex projection*. Let $c \in \mathbb{R}^L$ and let

$$\beta(c) = \sum_{\ell=1}^L c^\ell \eta_\ell, \quad \gamma(c) = \sum_{\ell=1}^L c^\ell \gamma_\ell.$$

Observe that $\beta(c) \in \mathcal{I}$ conditioned that $\gamma(c) = 0$. Indeed, if $\psi(x) = \sum_\ell c^\ell h_\ell(x)$, then $\sum_\ell c^\ell \mathbf{E}[Xh_\ell(X)] = 0$, and by (3),

$$\eta(c) = \sum_\ell c^\ell \mathbf{E}[\nabla h_\ell(X)] \in \mathcal{I}.$$

Therefore, the task of estimating $\beta \in \mathcal{I}$ reduces to that of finding a “good” corresponding coefficient vector. In Diederichs et al. (2009) vectors $\{\widehat{c}_j\}$ are computed as follows: let

$$\widehat{\eta}(c) = \sum_{\ell=1}^L c^\ell \widehat{\eta}_\ell \quad \text{and} \quad \widehat{\gamma}(c) = \sum_{\ell=1}^L c^\ell \widehat{\gamma}_\ell, \quad \ell = 1, \dots, L$$

and let $\xi_j \in \mathbb{R}^d, j = 1, \dots, J$ constitute a set of *probe* unit vectors. Then it holds

$$\widehat{c}_j = \arg \min_c \{ |\xi_j - \widehat{\eta}(c)|_2 \mid \widehat{\gamma}(c) = 0, |c|_1 \leq 1 \}, \tag{6}$$

and we set $\widehat{\beta}_j = \widehat{\beta}(\widehat{c}_j) = \sum_\ell \widehat{c}_j^\ell \widehat{\eta}_\ell$. Then \mathcal{I} is recovered by computing m principal axes of the minimal volume ellipsoid (Fritz-John ellipsoid) containing the estimated points $\{\pm \widehat{\beta}_j\}_{j=1}^J$.

The recovery of $\widehat{\mathcal{I}}$ through the Fritz-John ellipsoid (instead of eigenvalue decomposition of the matrix $\sum_\ell \widehat{\beta}_j \widehat{\beta}_j^T$) allows to bound the estimation error of \mathcal{I} by the maximal error of estimation $\widehat{\beta}$ of elements of the target space (cf. Theorem 3 of Diederichs et al. 2009), while the ℓ_1 -constraint on the coefficients \widehat{c}_j allows to control efficiently the maximal stochastic error of the estimations $\widehat{\beta}_j$ (cf. Theorem 1 of Diederichs et al. 2009; Spokoiny 2009). On the other hand, that construction heavily relies upon the choice of the probe vectors ξ_j . Indeed, in order to recover the projector on \mathcal{I} , the collection of $\widehat{\beta}_j$ should comprise at least m vectors with non-vanishing projection on the target space. To cope with this problem a multi-stage procedure has been used in Diederichs et al. (2009): given a set $\{\xi_j\}_{k=0}$ of probe vectors an estimation $\widehat{\mathcal{I}}_{k=0}$ is computed, which is used to draw new probe vectors $\{\xi_j\}_{k=1}$ from the vicinity of $\widehat{\mathcal{I}}_{k=0}$; these vectors are employed to compute a new estimation $\widehat{\mathcal{I}}_{k=1}$, and so on. The iterative procedure improves significantly the accuracy of the recovery of \mathcal{I} . Nevertheless, the choice of “informative” probe vectors at the first iteration $k = 0$ remains a challenging task and hitherto is a weak point of the procedure.

3 Structural analysis by semidefinite programming

In the present paper we discuss a new choice of vectors β which solves the initialization problem of probe vectors for the SNGCA procedure in quite a particular way. Namely, the estimation procedure we are to present below does not require any probe vectors at all.

3.1 Informative vectors in the target space

Further developments are based on the following simple observation. Let η_ℓ and γ_ℓ be defined as in (4), and let $U = [\eta_1, \dots, \eta_L] \in \mathbb{R}^{d \times L}$, $G = [\gamma_1, \dots, \gamma_L] \in \mathbb{R}^{d \times L}$. Using the observation in the previous section we conclude that if $c \in \mathbb{R}^L$ satisfies $Gc = \sum_{\ell=1}^L c^\ell \gamma_\ell = 0$ then $Uc = \sum_{\ell=1}^L c^\ell \eta_\ell$ belongs to \mathcal{I} . In other words, if Π^* is the Euclidean projector on \mathcal{I} , then

$$(I - \Pi^*)Uc = 0.$$

Suppose now that the set $\{h_\ell\}$ of test functions is rich enough in the sense that vectors Uc span \mathcal{I} when c spans the subspace $Gc = 0$. Recall that we assume the dimension m of the target space to be known. Then projector Π^* on \mathcal{I} is *fully identified* as the optimal solution to the problem

$$\Pi^* = \arg \min_{\Pi} \max_c \left\{ \left| (I - \Pi)Uc \right|_2^2 \left| \begin{array}{l} \Pi \text{ is a projector on an} \\ m\text{-dimensional subspace of } \mathbb{R}^d \\ c \in \mathbb{R}^L, Gc = 0 \end{array} \right. \right\}. \tag{7}$$

In practice vectors γ_ℓ and η_ℓ are not available, but we can suppose that their “empirical counterparts”—vectors $\widehat{\gamma}_\ell, \widehat{\eta}_\ell, \ell = 1, \dots, L$ can be computed, such that for a set A of probability at least $1 - \varepsilon$,

$$|\widehat{\eta}_\ell - \eta_\ell|_2 \leq \delta_N, \quad |\widehat{\gamma}_\ell - \gamma_\ell|_2 \leq \nu_N, \quad \ell = 1, \dots, L. \tag{8}$$

Indeed, it is well known (cf., e.g., Lemma 1 in Diederichs et al. 2009 or van der Vaart and Wellner 1996) that if functions $h_\ell(x) = f(x, \omega_\ell), \ell = 1, \dots, L$, are used, where f is continuously differentiable, $\omega_\ell \in \mathbb{R}^d$ are vectors on the unit sphere and f and $\nabla_x f$ are bounded, then (8) holds with

$$\begin{aligned} \delta_N &= C_1 \max_{x \in \mathbb{R}^d, |\omega|_2=1} |\nabla_x f(x, \omega)|_2 N^{-1/2} \sqrt{\min\{d, \ln L\} + \ln \varepsilon^{-1}}, \\ \nu_N &= C_2 \max_{x \in \mathbb{R}^d, |\omega|_2=1} |xf(x, \omega)|_2 N^{-1/2} \sqrt{\min\{d, \ln L\} + \ln \varepsilon^{-1}}, \end{aligned} \tag{9}$$

where C_1, C_2 are some absolute constants depending on the smoothness properties and the second moments of the underlying density.

Then for any $c \in \mathbb{R}^L$ such that $|c|_1 \leq 1$ we can control the error of approximation of $\sum_\ell c_\ell \gamma_\ell$ and $\sum_\ell c_\ell \eta_\ell$ with their empirical versions. Namely, we have on A :

$$\max_{|c|_1 \leq 1} \left| \sum_\ell c_\ell (\widehat{\eta}_\ell - \eta_\ell) \right|_2 \leq \delta_N \quad \text{and} \quad \max_{|c|_1 \leq 1} \left| \sum_\ell c_\ell (\widehat{\gamma}_\ell - \gamma_\ell) \right|_2 \leq \nu_N.$$

Let now $\widehat{U} = [\widehat{\eta}_1, \dots, \eta_L]$, $\widehat{G} = [\widehat{\gamma}_1, \dots, \widehat{\gamma}_L]$. When substituting \widehat{U} and \widehat{G} for U and G into (7) we come to the following minmax problem:

$$\min_{\Pi} \max_c \left\{ |(I - \Pi)\widehat{U}c|_2^2 \mid \begin{array}{l} \Pi \text{ is a projector on an } m\text{-dimensional} \\ \text{subspace of } \mathbb{R}^d \\ c \in \mathbb{R}^L, |c|_1 \leq 1, |\widehat{G}c|_2 \leq \varrho \end{array} \right\}. \tag{10}$$

Here we have substituted the constraint $Gc = 0$ with the inequality constraint $|\widehat{G}c|_2 \leq \varrho$ for some $\varrho > 0$ in order to keep the optimal solution c^* to (7) feasible for the modified problem (10) (this will be the case with probability at least $1 - \varepsilon$ if $\varrho \geq \nu_N$).

As we will see in a moment, when c runs the ν_N -neighborhood of intersection C_N of the standard hyperoctahedron $\{c \in \mathbb{R}^L, |c|_1 \leq 1\}$ with the subspace $\widehat{G}c = 0$, vectors $\widehat{U}c$ span a close vicinity of the target space \mathcal{I} .

3.2 Solution by semidefinite relaxation

Note that (10) is a hard optimization problem. Namely, the candidate maximizers c_i of (10) are the extreme points of the set $C_N = \{c \in \mathbb{R}^L, |c|_1 \leq 1, |\widehat{G}c|_2 \leq \nu_N\}$, and there are $O(L^d)$ of such points. In order to be efficiently solvable, the problem (10) is to be “reduced” to a convex-concave saddle-point problem, which is, to the best of our knowledge, the only class of minmax problems which can be solved efficiently (cf. Nemirovski and Yudin 1983).

Thus the next step is to transform the problem in (10) into a convex-concave minmax problem using the *Semidefinite Relaxation* (or SDP-relaxation) technique (see e.g., Ben Tal and Nemirovski 2001, Chap. 4). We obtain the relaxed version of (10) in two steps. First, let us rewrite the objective function (recall that $I - \Pi$ is also a projector, and thus an idempotent matrix):

$$|(I - \Pi)\widehat{U}c|_2^2 = c^T \widehat{U}^T (I - \Pi)^2 \widehat{U}c = c^T \widehat{U}^T (I - \Pi) \widehat{U}c = \text{trace}[\widehat{U}^T (I - \Pi) \widehat{U}X],$$

where the positive semidefinite matrix $X = cc^T$ is the “new variable”. The constraints on c can be easily rewritten for X :

1. the constraint $|c|_1 \leq 1$ is equivalent to $|X|_1 \leq 1$ (we use the notation $|X|_1 = \sum_{i,j=1}^L |X_{ij}|$);
2. because X is positive semidefinite, the constraint $|\widehat{G}c|_2 \leq \varrho$ is equivalent to $\text{trace}[\widehat{G}X\widehat{G}^T] \leq \varrho^2$.

The only “bad” constraint on X is the rank constraint: $\text{rank } X = 1$, and we simply remove it. Now we are done with the variable c and we arrive at

$$\min_{\Pi} \max_X \left\{ \text{trace}[\widehat{U}^T (I - \Pi) \widehat{U}X] \mid \begin{array}{l} \Pi \text{ is a projector on an } m\text{-dimensional} \\ \text{subspace of } \mathbb{R}^d \\ X \succeq 0, |X|_1 \leq 1, \text{trace}[\widehat{G}X\widehat{G}^T] \leq \varrho^2 \end{array} \right\}.$$

Let us recall that an m -dimensional projector Π is exactly a symmetric $d \times d$ matrix of rank $\Pi = m$ and $\text{trace } \Pi = m$, with the eigenvalues $0 \leq \lambda_i(\Pi) \leq 1, i = 1, \dots, d$. Once again we remove the “difficult” rank constraint $\text{rank } \Pi = m$ and finish with

$$\min_P \max_X \left\{ \text{trace}[\widehat{U}^T (I - P) \widehat{U}X] \mid \begin{array}{l} 0 \leq P \leq I, \text{trace } P = m, \\ X \succeq 0, |X|_1 \leq 1, \text{trace}[\widehat{G}X\widehat{G}^T] \leq \varrho^2 \end{array} \right\} \tag{11}$$

(we write $P \leq Q$ if the matrix $Q - P$ is positive semidefinite). There is no reason for an optimal solution \widehat{P} of (11) to be a projector matrix. If an estimation of Π^* which is itself

a projector is needed, one can use instead the projector $\widehat{\Pi}$ onto the subspace spanned by m principal eigenvectors of \widehat{P} .

Note that (11) is a linear matrix game with bounded convex domains of its arguments—positive semidefinite matrices $P \in \mathbb{R}^{d \times d}$ and $X \in \mathbb{R}^{L \times L}$.

We are about to describe the accuracy of the estimation $\widehat{\Pi}$ of Π^* . To this end we need an identifiability assumption on the system $\{h_\ell\}$ of test functions as follows:

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

$$\Pi^* \preceq \sum_{k=1}^{\overline{m}} \mu^k U c_k c_k^T U^T. \tag{12}$$

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

In other words, if Assumption 1 holds, then the true projector Π^* on \mathcal{I} is $\mu^* \times$ convex combination of rank-one matrices $U c c^T U^T$ where c satisfies the constraint $Gc = 0$ and $|c|_1 = 1$.

Theorem 1 Suppose that the true dimension m of the subspace \mathcal{I} is known and that $\varrho \geq \nu_N$ as in (8). Let \widehat{P} be an optimal solution to (11) and let $\widehat{\Pi}$ be the projector onto the subspace spanned by m principal eigenvectors of \widehat{P} . Then with probability $\geq 1 - \varepsilon$:

(i) for any c such that $|c|_1 \leq 1$ and $Gc = 0$,

$$|(I - \widehat{\Pi})Uc|_2 \leq \sqrt{m+1}((\varrho + \nu_N)\lambda_{\min}^{-1}(\Sigma) + 2\delta_N);$$

(ii) further, if Assumption 1 holds then

$$\text{trace}[(I - \widehat{P})\Pi^*] \leq \mu^*((\varrho + \nu_N)\lambda_{\min}^{-1}(\Sigma) + 2\delta_N)^2, \tag{13}$$

and

$$\begin{aligned} \|\widehat{\Pi} - \Pi^*\|_2^2 &\leq 2\mu^*(\lambda_{\min}^{-1}(\Sigma)(\varrho + \nu_N) + 2\delta_N)^2 \tau, \\ \tau &= (m+1) \wedge (1 - \mu^*(\lambda_{\min}^{-1}(\Sigma)(\varrho + \nu_N) + 2\delta_N)^2)^{-1} \end{aligned} \tag{14}$$

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

Note that if we were able to solve the minimax problem in (10), we could expect its solution, let us call it $\widetilde{\Pi}$, to satisfy with high probability

$$|(I - \widetilde{\Pi})Uc|_2 \leq (\varrho + \nu_N)\lambda_{\min}^{-1}(\Sigma) + 2\delta_N$$

(cf. the proof of Lemma 1 in the Appendix). If we compare this bound to that of the statement (i) of Theorem 1, we conclude that the loss of the accuracy resulting from the substitution of (10) by its treatable approximation (11) is bounded with $\sqrt{m+1}$. In other words, the ‘‘price’’ of the SDP-relaxation in our case is $\sqrt{m+1}$ and does not depend on problem dimensions d and L . Furthermore, when Assumption 1 holds true, we are able to provide the bound on the accuracy of recovery of projector Π^* which is seemingly as good as if we were using instead of $\widehat{\Pi}$ the solution $\widetilde{\Pi}$ of (10).

Suppose now that the test functions $h_\ell(x) = f(x, \omega_\ell)$ are used, with ω_ℓ on the unit sphere of \mathbb{R}^d , that $\varrho = \nu_N$ is chosen, and that Assumption 1 holds with “not too large” μ^* , e.g., $\mu^* \leq \frac{1}{2}(\varrho + \nu_N)\lambda_{\min}^{-1}(\Sigma) + \delta_N$. When substituting the bounds of (9) for δ_N and ν_N into (14) we obtain the bound for the accuracy of the estimation $\widehat{\Pi}$ (with probability $1 - \epsilon$):

$$\|\widehat{\Pi} - \Pi^*\|_2^2 \leq C(f)\mu^*N^{-1}(\min(d, \ln L) + \ln \epsilon^{-1})$$

where $C(f)$ depends only on f . This bound claims the root- N consistency in estimation of the non-Gaussian subspace with the log-price for relaxation and estimation error.

3.3 Case of unknown dimension m

The problem (11) may be modified to allow the treatment of the case when the dimension m of the target space is unknown *a priori*. Namely, consider for $\rho \geq 0$ the following problem

$$\min_{P,t} \left\{ t \mid \begin{array}{l} \text{trace } P \leq t, \max_X \text{trace}[\widehat{U}^T(I - P)\widehat{U}X] \leq \rho^2, 0 \leq P \leq I, \\ X \geq 0, \|X\|_1 \leq 1, \text{trace}[\widehat{G}X\widehat{G}^T] \leq \varrho^2. \end{array} \right\} \quad (15)$$

The problem (15) is closely related to the ℓ_1 -recovery estimator of sparse signals (see, e.g., the tutorial (Candès 2006) and the references therein) and the trace minimization heuristics widely used in the Sparse Principal Component Analysis (SPCA) (cf. d’Aspremont et al. 2007, 2008). As we will see in an instant, when the parameter ρ of the problem is “properly chosen”, the optimal solution \widehat{P} of (15) possesses essentially the same properties as that of the problem (11).

A result analogous to that in Theorem 1 holds:

Theorem 2 *Let \widehat{P} , \widehat{X} and $\widehat{t} = \text{trace } \widehat{P}$ be an optimal solution to (15) (note that (15) is clearly solvable), $\widehat{m} = \lfloor \widehat{t} \rfloor$,¹ and let $\widehat{\Pi}$ be the projector onto the subspace spanned by \widehat{m} principal eigenvectors of \widehat{P} . Suppose that $\varrho \geq \nu_N$ as in (8) and that*

$$\rho \geq \lambda_{\min}^{-1}(\Sigma)(\varrho + \nu_N) + \delta_N. \quad (16)$$

Then with probability at least $1 - \epsilon$:

(i)

$$\widehat{t} \leq m \quad \text{and} \quad \|(I - \widehat{\Pi})U_C\|_2 \leq \sqrt{m+1}(\rho + 2\delta_N);$$

(ii) *furthermore, if Assumption 1 hold then*

$$\text{trace}[(I - \widehat{P})\Pi^*] \leq \mu^*(\rho + \delta_N)^2,$$

and

$$\|\widehat{\Pi} - \Pi^*\|_2^2 \leq 2\mu^*(\rho + \delta_N)^2[(m+1) \wedge (1 - \mu^*(\rho + \delta_N)^2)^{-1}] \quad (17)$$

(here $\|A\|_2 = (\sum_{i,j} A_{ij}^2)^{1/2}$ is the Frobenius norm of A).

¹Here $\lfloor a \rfloor$ is the smallest integer $\geq a$.

The proof of the theorems is postponed until the [Appendix](#).

The estimation procedure based on solving (15) allows to infer the target subspace \mathcal{I} without *a priori* knowledge of its dimension m . When the constraint parameter ρ is close to the right-hand side of (16), the accuracy of the estimation will be close to that, obtained in the situation when dimension m is known. However, the accuracy of the estimation heavily depends on the precision of the available (lower) bound for $\lambda_{\min}(\Sigma)$. In the high-dimensional situation this information is hard to acquire, and the necessity to compute this quantity may be considered as a serious drawback of the proposed procedure.

4 Solving the saddle-point problem (11)

We start with the following simple observation: by using bisection or Newton search in ρ (note that the objective of (15) is obviously convex in ρ^2) we can reduce (15) to a small sequence to feasibility problems, closely related to (11): given t_0 report, if exists, P such that

$$\max_X \left\{ \begin{array}{l} \text{trace}[\widehat{U}^T(I - P)\widehat{U}X] \leq \rho^2, \quad 0 \leq P \leq I, \quad \text{trace} P \leq t_0, \\ X \geq 0, \quad |X|_1 \leq 1, \quad \text{trace}[\widehat{G}X\widehat{G}^T] \leq \varrho^2 \end{array} \right\}.$$

In other words, we can easily solve (15) if for a given m we are able to find an optimal solution to (11). Therefore, in the sequel we concentrate on the optimization technique for solving (11).

4.1 Dual extrapolation algorithm

In what follows we discuss the dual extrapolation algorithm of Nesterov (2007) for solving a version of (11) in which, with a certain abuse, we substitute the inequality constraint $\text{trace} \widehat{G}X\widehat{G}^T \leq \varrho^2$ with the equality constraint $\text{trace}[\widehat{G}X\widehat{G}^T] = 0$. This way we come to the problem:

$$\min_{P \in \mathcal{P}} \max_{X \in \mathcal{X}} \text{trace}[\widehat{U}^T(I - P)\widehat{U}X] \tag{18}$$

where

$$\mathcal{X} = \{X \in \mathbf{S}^L, X \geq 0, |X|_1 \leq 1, \text{trace}[\widehat{G}^T\widehat{G}X] = 0\}$$

(here \mathbf{S}^L stands for the space of $L \times L$ symmetric matrices) and

$$\mathcal{P} = \{P \in \mathbf{S}^d, 0 \leq P \leq I, \text{trace}[P] \leq m\}.$$

Observe first that (18) is a matrix game over two convex subsets (of the cone) of positive semidefinite matrices. If we use a large number of test functions, say $L^2 \sim 10^6$, the size of the variable X rules out the possibility of using the interior-point methods. The methodology which appears to be adequate in this case is that behind dual extrapolation methods, recently introduced in Nemirovski (2004), Lu et al. (2007), Nesterov (2005, 2007). The algorithm we use belongs to the family of subgradient descent-ascent methods for solving convex-concave games. Though the rate of convergence of such methods is slow—their precision is only $\mathcal{O}(1/k)$, where k is the iteration count, their iteration is relatively cheap, what makes the methods of this type appropriate in the case of high-dimensional problems when the high accuracy is not required.

We start with the general dual extrapolation scheme of Nesterov (2007) for linear matrix games. Let \mathbb{E}^n and \mathbb{E}^m be two Euclidean spaces of dimension n and m respectively, and let $\mathcal{A} \subset \mathbb{E}^n$ and $\mathcal{B} \subset \mathbb{E}^m$ be closed and convex sets. We consider the problem

$$\min_{x \in \mathcal{A}} \max_{y \in \mathcal{B}} \langle x, Ay \rangle + \langle a, x \rangle + \langle b, y \rangle. \tag{19}$$

Let $\|\cdot\|_x$ and $\|\cdot\|_y$ be some norms on \mathbb{E}^n and \mathbb{E}^m respectively. We say that d_x (resp., d_y) is a distance-generating function of \mathcal{A} (resp., of \mathcal{B}) if d_x (resp., d_y) is strongly convex modulus α_x (resp., α_y) and differentiable on \mathcal{A} (resp., on \mathcal{B}).² Let for $z = (x, y)$ $d(z) = d_x(x) + d_y(y)$ (note that d is differentiable and strongly convex on $\mathcal{A} \times \mathcal{B}$ with respect to the norm, defined on $\mathcal{A} \times \mathcal{B}$ according to, e.g. $\|z\| = \|x\|_x + \|y\|_y$). We define the *prox-function* V of $\mathcal{A} \times \mathcal{B}$ as follows: for $z_0 = (x_0, y_0)$ and $z = (x, y)$ in $\mathcal{A} \times \mathcal{B}$ we set

$$V(z_0, z) \stackrel{\text{def}}{=} d(z) - d(z_0) - \langle \nabla d(z_0), z - z_0 \rangle. \tag{20}$$

Next, for $s = (s_x, s_y)$ we define the *prox-transform* $T(z_0, s)$ of s :

$$T(z_0, s) \stackrel{\text{def}}{=} \arg \min_{z \in \mathcal{A} \times \mathcal{B}} [s, z - z_0] - V(z_0, z). \tag{21}$$

Let us denote $F(z) = (-A^T y - a, Ax + b)$ the vector field of descend-ascend directions of (19) at $z = (x, y)$ and let \bar{z} be the minimizer of d over $\mathcal{A} \times \mathcal{B}$. Given vectors $z_k, z_k^+ \in \mathcal{A} \times \mathcal{B}$ and $s_k \in E^*$ at the k -th iteration, we define the update z_{k+1}, z_{k+1}^+ and s_{k+1} according to

$$\begin{aligned} z_{k+1} &= T(\bar{z}, s_k), \\ z_{k+1}^+ &= T(z_{k+1}, \lambda_k F(z_{k+1})), \\ s_{k+1} &= s_k + \lambda_k F(z_{k+1}^+), \end{aligned}$$

where $\lambda_k > 0$ is the current stepsize. Finally, the current approximate solution \widehat{z}_{k+1} is defined with

$$\widehat{z}_{k+1} = \frac{1}{k+1} \sum_{i=1}^{k+1} z_i^+.$$

The key element of the above construction is the choice of the *distance-generating function* d in the definition of the prox-function. It should satisfy two requirements:

- let D be the variation of V over $\mathcal{A} \times \mathcal{B}$ and let α be the parameter of strong convexity of V with respect to $\|\cdot\|$. The complexity of the algorithm is proportional to D/α , so this ratio should be as small as possible;
- one should be able to compute efficiently the solution to the auxiliary problem (21) which is to be solved twice at each iteration of the algorithm.

Note that the prox-transform preserve the additive structure of the distance-generating function. Thus, in order to compute the prox-transform on the feasible domain $\mathcal{P} \times \mathcal{X}$ of (18) we need to compute its “ P and X components”—the corresponding prox-transforms on

²Recall that a (sub-)differentiable on \mathcal{F} function f is called strongly convex on \mathcal{F} with respect to the norm $\|\cdot\|$ of modulus α if $\langle f'(x) - f'(y), x - y \rangle \geq \alpha \|x - y\|^2$ for all $x, y \in \mathcal{F}$.

\mathcal{P} and cX . There are several evident choices of the prox-functions d_P and d_X of the domains \mathcal{P} and \mathcal{X} of (18) which satisfy the first requirement above and allow to attain the optimal value $O(\sqrt{m \ln d \ln L})$ of the ratio D/α for the prox-function V of (18). However, for such distance-generating functions there is no known way to compute efficiently the X -component of the prox-transform T in (21) for the set \mathcal{X} . This is why in order to admit an efficient solution the problem (18) is to be modified one more time.

4.2 Modified problem

We act as follows: first we eliminate the linear equality constraint which, taken along with $X \geq 0$, says that $X = Q^T Z Q$ with $Z \geq 0$ and certain Q ; assuming that the d rows of \widehat{G} are linearly independent, we can choose Q as an appropriate $(L - d) \times L$ matrix satisfying $Q Q^T = I$ (the orthogonal basis of the kernel of \widehat{G}). Note that from the constraints on X it follows that $\text{trace}[X] \leq 1$, whence

$$\text{trace}[Q^T Z Q] = \text{trace}[Z Q Q^T] = \text{trace}[Z] \leq 1.$$

Thus, although there are additional constraints on Z as well, Z belongs to the *standard spectahedron*

$$\mathcal{Z} = \{Z \in \mathbf{S}^{L-d}, Z \geq 0, \text{trace}[Z] \leq 1\}.$$

Now can rewrite our problem equivalently as follows:

$$\min_{P \in \mathcal{P}} \max_{Z \in \mathcal{Z}, |Q^T Z Q|_1 \leq 1} \text{trace}[\widehat{U}^T (I - P) \widehat{U} (Q^T Z Q)]. \tag{22}$$

Let, further,

$$\mathcal{W} = \{W \in \mathbf{S}^L, \|W\|_2 \leq 1\}, \quad \text{and} \quad \mathcal{Y} = \{Y \in \mathbf{S}^L, |Y|_1 \leq 1\}.$$

We claim that the problem (22) can be reduced to the saddle point problem

$$\min_{(P,W) \in \mathcal{P} \times \mathcal{W}} \max_{(Z,Y) \in \mathcal{Z} \times \mathcal{Y}} \underbrace{\{\text{trace}[\widehat{U}^T (I - P) \widehat{U} Y] + \lambda \text{trace}[W(Q^T Z Q - Y)]\}}_{F(P,W; Z,Y)} \tag{23}$$

provided that λ is not too small.

Now, “can be reduced to” means exactly the following:

Proposition 1 *Suppose that $\lambda > L|\widehat{U}|_2^2$, where $|U|_2$ is the maximal Euclidean norm of columns of U . Let $(\widehat{P}, \widehat{W}; \widehat{Z}, \widehat{Y})$ be a feasible solution ϵ -solution to (23), that is*

$$(\widehat{P}, \widehat{W}; \widehat{Z}, \widehat{Y}) \in (\mathcal{P}, \mathcal{W}; \mathcal{Z}, \mathcal{Y}), \quad \text{and} \quad \overline{F}(\widehat{P}, \widehat{W}) - \underline{F}(\widehat{Z}, \widehat{Y}) \leq \epsilon$$

where

$$\overline{F}(P, W) = \max_{(Z,Y) \in \mathcal{Z} \times \mathcal{Y}} F(P, W; Z, Y), \quad \underline{F}(Z, Y) = \min_{(P,W) \in \mathcal{P} \times \mathcal{W}} F(P, W; Z, Y).$$

Then setting

$$\tilde{Z} = \begin{cases} \widehat{Z}, & \text{if } |Q^T Z Q|_1 \leq 1, \\ |Q^T Z Q|_1^{-1} \widehat{Z} & \text{otherwise,} \end{cases}$$

the pair (\hat{P}, \tilde{Z}) is a feasible ϵ -solution to (22). Specifically, we have $(\hat{P}, \tilde{Z}) \in \mathcal{P} \times \mathcal{Z}$ with $|Q^T \tilde{Z} Q|_1 \leq 1$, and

$$\overline{G}(\hat{P}) - \underline{G}(\tilde{Z}) \leq \epsilon,$$

where

$$\begin{aligned} \overline{G}(P) &= \max_{Z \in \mathcal{Z}, |Q^T Z Q|_1 \leq 1} \text{trace}[\hat{U}^T (I - P) \hat{U} Q^T Z Q]; \\ \underline{G}(Z) &= \min_{P \in \mathcal{P}} \text{trace}[\hat{U}^T (I - P) \hat{U} Q^T Z Q]. \end{aligned}$$

The proof of the proposition is given in the Appendix A.3.

Note that feasible domains of (23) admit evident distance-generating functions. We provide the detailed computation of the corresponding prox-transforms in the Appendix A.4.

5 Numerical experiments

In this section we compare the numerical performance of the presented approach, which we refer to as SNGCA(SDP) with other statistical methods of dimension reduction on the simulated data.

5.1 Structural adaptation algorithm

We start with some implementation details of the estimation procedure. We use the choice of the test functions $h_\ell(x) = f(x, \omega_\ell)$ for the SNGCA algorithm as follows:

$$f(x, \omega) = \tanh(\omega^T x) e^{-\alpha \|x\|_2^2 / 2},$$

where $\omega_\ell, l = 1, \dots, L$ are unit vectors in \mathbb{R}^d .

We implement here a multi-stage variant of the SNGCA (cf. Diederichs et al. 2009). At the first stage of the SNGCA(SDP) algorithm (Algorithm 1) we assume that the directions ω_ℓ are drawn randomly from the unit sphere of \mathbb{R}^d . At each of the following stages we use the current estimation of the target subspace to “improve” the choice of directions ω_ℓ as follows: we draw a fixed fraction of ω ’s from the estimated subspace and draw randomly over the unit square the remaining ω ’s. The simulation results below are presented for the estimation procedure with three stages. The size of the set of test function is set to $L = 10d$, and the target accuracy of solving the problem (11) is set to $1e - 4$.

5.2 Experiment description

Each simulated data set $X^N = [X_1, \dots, X_N]$ of size $N = 1000$ represents N independent and identically distributed (i.i.d.) realizations of a random vectors X of dimension d . Each simulation is repeated 100 times and we report the average over 100 simulations Frobenius norm of the error of estimation of the projection on the target space. In the examples below only $m = 2$ components of X are non-Gaussian with unit variance, other $d - 2$ components of X are independent standard normal random variables. The densities of the non-Gaussian components are chosen as follows:

Algorithm 1: SNGCA (SDP)

```

% Initialization:
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  $X_i$ . We denote  $Y_i = \text{diag}(\sigma^{-1})X_i$  the standardized data.
Set the current estimator  $\widehat{\Pi}_0 = I_d$ .

% Main iteration loop:
for  $i = 1$  to  $I$  do
    Sample a fraction of  $\omega^{(i)}$ 's from the normal distribution  $N(0, \widehat{\Pi}_{i-1})$  (zero mean,
    with covariance matrix  $\widehat{\Pi}_{i-1}$ ), sample the remaining  $\omega^{(i)}$ 's from  $N(0, I_d)$ , then
    normalize to the unit length;

    % Compute estimations of  $\eta_\ell$  and  $\gamma_\ell$ 
    for  $\ell = 1$  to  $L$  do
         $\widehat{\eta}_\ell^{(i)} = \frac{1}{N} \sum_{j=1}^N \nabla h_{\omega_\ell^{(i)}}(Y_j)$ ;
         $\widehat{\gamma}_\ell^{(i)} = \frac{1}{N} \sum_{j=1}^N Y_j h_{\omega_\ell^{(i)}}(Y_j)$ ;
    end

    Solve the corresponding problem (11) and update the estimation  $\widehat{\Pi}_i$ ;
end

```

- (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: a component of X , say X_1 , follows the Laplace distribution $\mathcal{L}(1)$ and the other is a dependent uniform $\mathcal{U}(c, c + 1)$, where $c = 0$ for $|X_1| \leq \ln 2$ and $c = -1$ otherwise.
- (E) Dependent sub-Gaussian: 2-dimensional isotropic Cauchy distribution with density proportional to $\lambda(\lambda^2 - x^2)^{-1}$ where $\lambda = 1$.

We provide the 2-d density plots of the non-Gaussian components on Fig. 1. We start with comparing the presented algorithm with Projection Pursuit (PP) method (Hyvärinen 1999) and the NGCA for $d = 10$. The results are presented on Fig. 2 (the corresponding results for PP and NGCA has been already reported in Diederichs et al. 2009 and Blanchard et al. 2006). Since the minimization procedure of PP tends to be trapped in a local minimum, in each of the 100 simulations, the PP algorithm is restarted 10 times with random starting points. The best result is reported for each PP-simulation. We observe that SNGCA(SDP) outperforms NGCA and PP in all tests.

In the next simulation we study the dependence of the accuracy of the SNGCA(SDP) on the noise level and compare it to the corresponding data for PP and NGCA. We present on Fig. 3 the results of experiments when the non-Gaussian coordinates have unit variance, but the standard deviation of the components of the 8-dimensional Gaussian distribution follows the geometrical progression $10^{-r}, 10^{-r+2r/7}, \dots, 10^r$ where $r = 1, \dots, 8$. The conditioning of the covariance matrix heavily influences the estimation error of PP(tanh) and NGCA, but

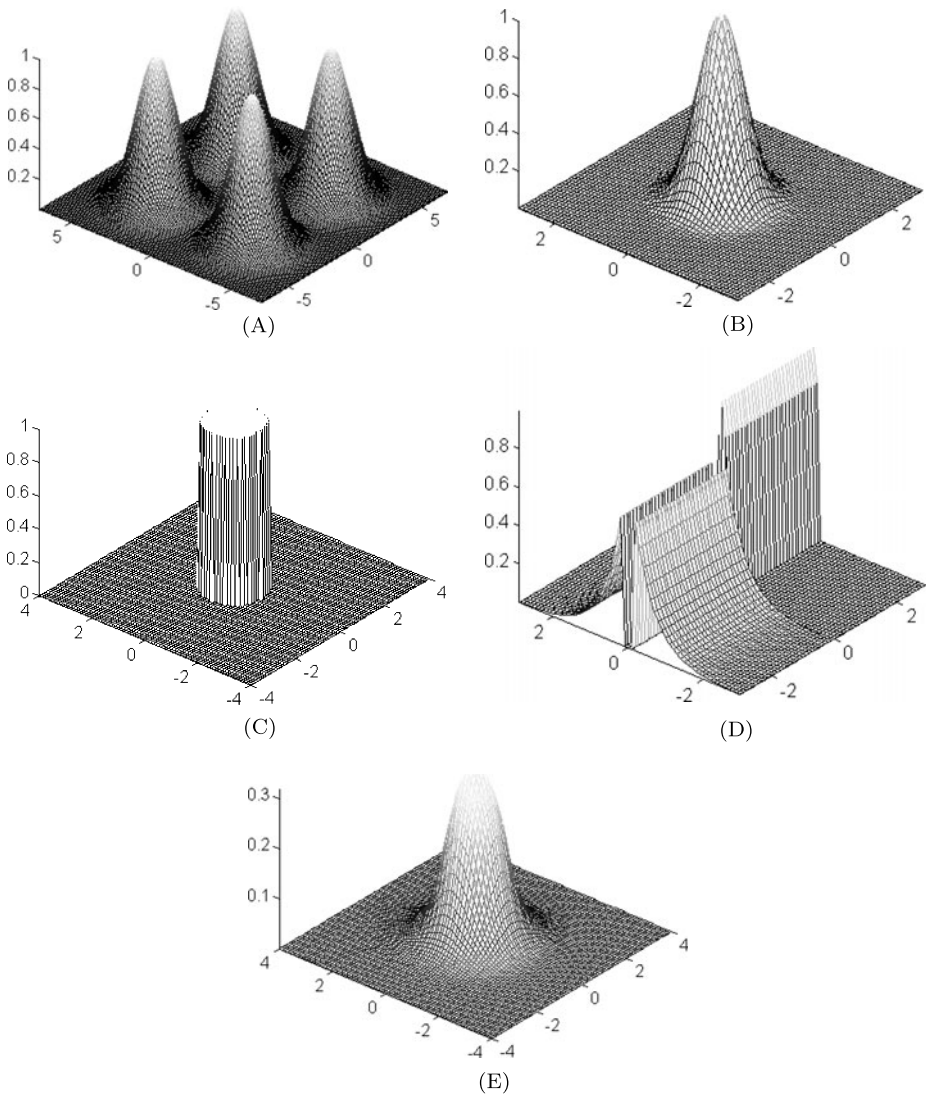


Fig. 1 (A) independent Gaussian mixtures, (B) isotropic super-Gaussian, (C) isotropic uniform and (D) dependent 1d Laplacian with additive 1d uniform, (E) isotropic sub-Gaussian

not that of SNGCA(SDP). The latter method appears to be insensitive to the differences in the noise variance along different directions in all test cases.

Next we compare the behavior of SNGCA(SDP), PP and NGCA as the dimension of the Gaussian component increases. On Fig. 4 we plot the mean error of estimation against the problem dimension d .

For PP and NGCA methods we observe that the estimation becomes meaningless (the estimation error explodes) already for $d = 30$ – 40 for the models (A), (C) and for $d = 20$ – 30 of the model (D). In the case of the models (B) and (E) we observe the progressive increase of

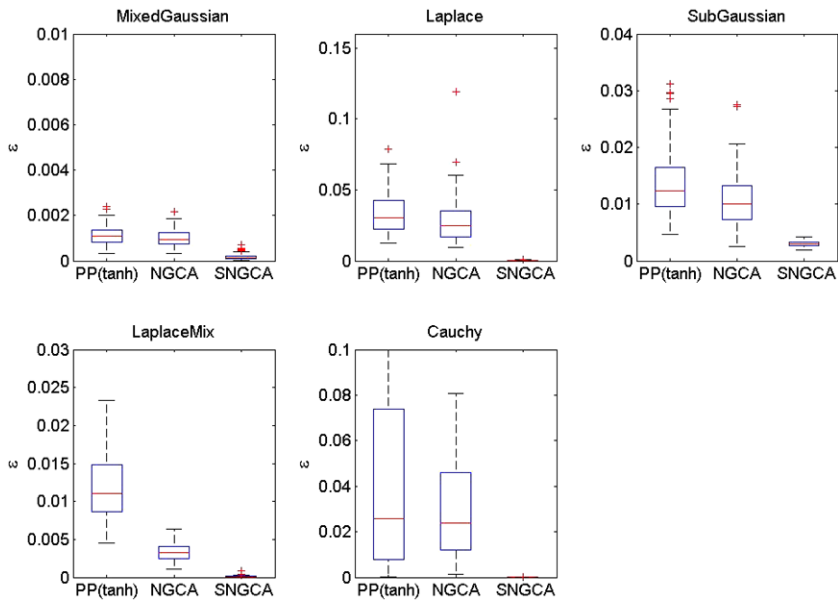


Fig. 2 Comparison of Projection Pursuit (PP), NGCA and the new approach SNGCA(SDP)

the error for methods PP and NGCA. The proposed method SNGCA(SDP) behaves robustly with respect to the increasing dimension of the Gaussian component for all test models.

5.3 Application to geometric analysis of metastability

Some biologically active molecules exhibit different large geometric structures at the scale much larger than the diameter of the atoms. If there are more than one such structures with the life span much larger than the time scale of the local atomic vibrations, the structure is called metastable conformation (Schütte and Huisinga 2003). In other words, metastable conformations of biomolecules can be seen as connected subsets of state-space. When compared to the fluctuations within each conformation, the transitions between different conformations of a molecule are rare statistical events. Such multi-scale dynamic behavior of biomolecules stem from a decomposition of the free energy landscape into particularly deep wells each containing many local minima (Pillardry and Piela 1995; Frauenfelder and McMahon 2000). Such wells represent different almost invariant geometrical large scale structures (Amadei et al. 1993). 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 (Horenko and Schütte 2008). Since the shape of the energy landscape and the invariant density of the Markov process are unknown, the “essential degrees of freedom”, in which the rare conformational changes occur, are of importance.

We will now illustrate that SNGCA(SDP) is able to detect a multimodal component of the data density as a special case of non-Gaussian subspace in high-dimensional data obtained from molecular dynamics simulation of oligopeptides.

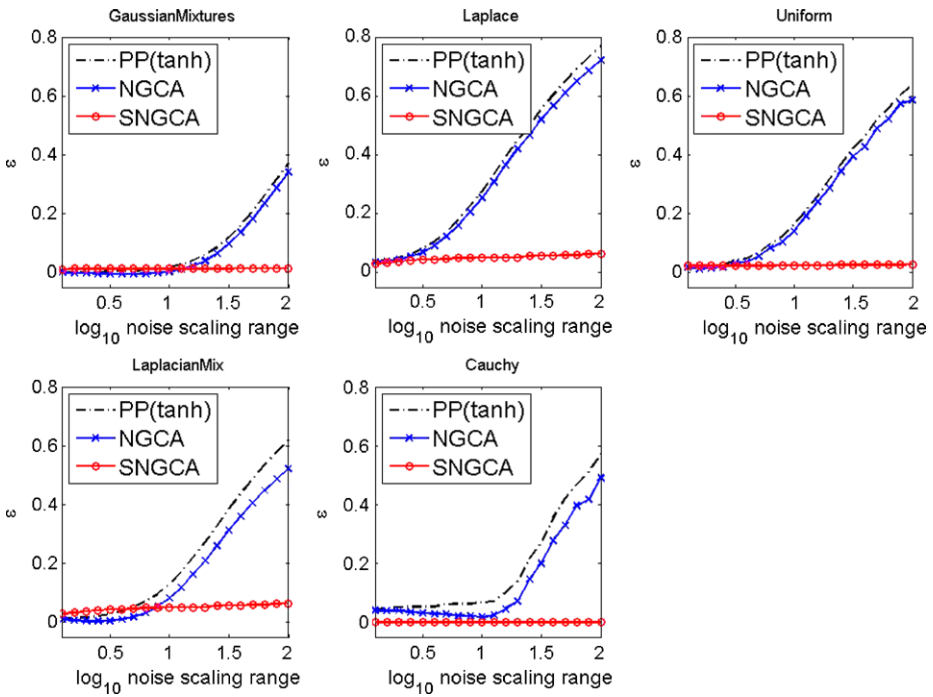


Fig. 3 Estimation error with respect to the standard deviation of Gaussian components following a geometrical progression on $[10^{-r}, 10^r]$ where r is the parameter on the abscissa

Clustering of 8-alanine The first example is a times series, generated by an equilibrium molecular dynamics simulation of 8-alanine. We only consider the backbone dihedral angles in order to determine different conformations.

The time series of dimension $d = 14$ consists of the cyclic data set of all backbone torsion angles. The simulation using CHARMM was done at $T = 300$ K with implicit water by means of the solvent model ACE2 (Schaefer and Karplus 1996). A symplectic Verlet integrator with integration step of 1 fs was used; the total trajectory length was 4 μ s and every $\tau = 50$ fs a set of coordinates was recorded.

The dimension reduction reported in the next Fig. 5 was obtained using SNGCA(SDP) for a given dimension $m = 5$ of the target space containing the multimodal component. A concentration of the clustered data in the target space of SNGCA may be clearly observed. In comparison, the complement of the target space is almost completely filled with Gaussian noise.

Clustering of a 3-peptide molecule In the next example we investigate Phenylalanyl-Glycyl-Glycine Tripeptide, which is assumed to realize all of the most important folding mechanisms of polypeptides (Reha et al. 2005). The simulation is done using GROMACS at $T = 300$ K with implicit water. An integration step of a symplectic Verlet integrator is set to 2 fs, and every $\tau = 50$ fs a set of $d = 31$ dihed angles was recorded. As in the previous experience, the dimension of the target space is set to $m = 5$.

Figure 7 shows that the clustered data can be primarily found in the target space of SNGCA(SDP).

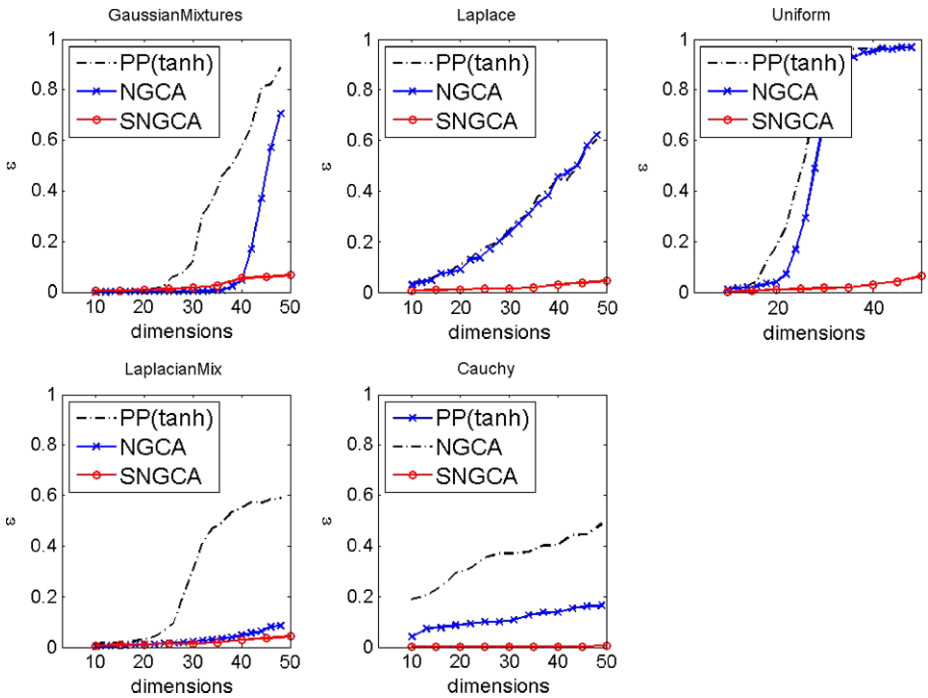


Fig. 4 Mean-square estimation error vs problem dimension d

We have studied a new approach to non-Gaussian component analysis. The suggested method, same as the techniques proposed in Blanchard et al. (2006), Diederichs et al. (2009) has two stages: on the first stage certain linear test functionals of unknown distribution are computed as a sampling of the data space and afterwards this information is used to recover the non-Gaussian subspace. The novelty of the proposed approach resides in the new method of non-Gaussian subspace identification, based upon semidefinite relaxation. The new procedure allows to overcome the main drawbacks of the previous implementations of the NGCA and seems to improve significantly the accuracy of estimation. Nevertheless the proposed algorithm is computationally demanding. While the first-order optimization algorithm we propose allows to treat efficiently the problems which are far beyond the reach of classical SDP-optimization techniques, the numerical difficulty seems to be one of the main practical limitation of the proposed approach. Furthermore the new approach lacks the ability to adapt itself to the underlying type of nonGaussian component. Both aspects including new rates of convergence for iterative adaptation of SNGCA to some former estimation result are currently in preparation.

Acknowledgements Elmar Diederichs acknowledges financial support from the DFG research center Matheon “Mathematics for key technologies” (FZT 86) in Berlin. Vladimir Spokoyny is supported by Predictive Modeling Laboratory, MIPT, RF government grant, ag. 11.G34.31.0073. Financial support by the German Research Foundation (DFG) through the Collaborative Research Center 649 “Economic Risk” is gratefully acknowledged. Critics and suggestions of three anonymous referees helped a lot in improving the paper.

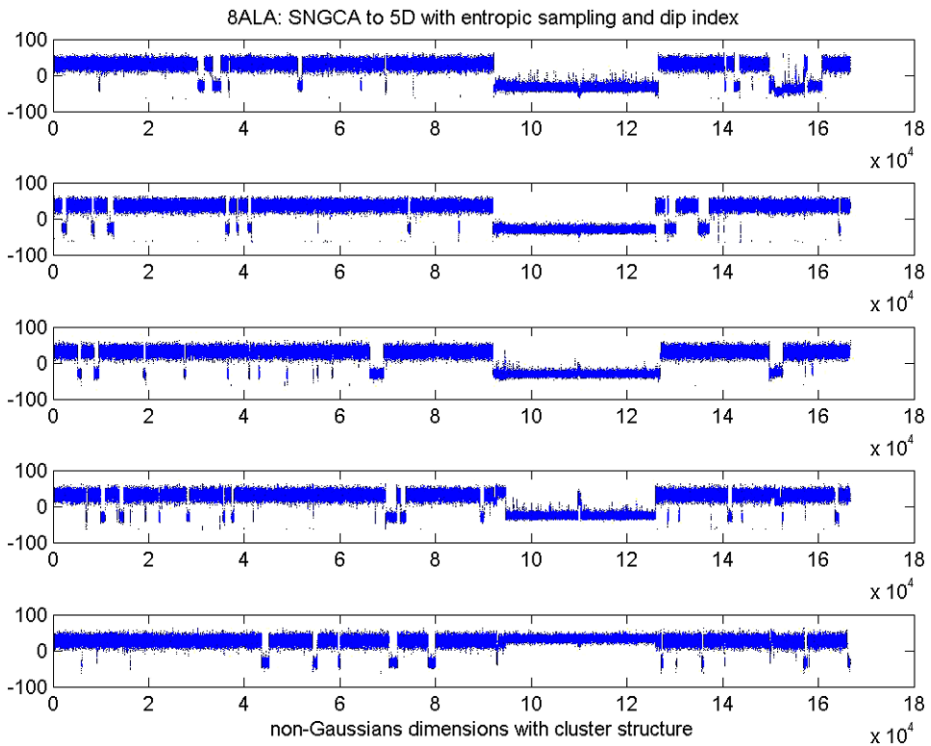


Fig. 5 Low dimensional multimodal component of 8-alanine

Appendix

Let $X = X^T \in \mathbb{R}^{L \times L}$ be positive semidefinite with $|X|_1 \leq 1$, and let $Y = X^{1/2}$ be the symmetric positive semidefinite square root of X . If we denote $y_i, i = 1, \dots, L$ the columns of Y , then $|X|_1 \leq 1$ implies that

$$\sum_{1 \leq i, j \leq L} |y_i^T y_j| \leq 1.$$

We make here one trivial though useful observation: for any matrix $A \in \mathbb{R}^{d \times L}$, when denoting $B = A^T A$, we have

$$\|AY\|_2^2 = \text{trace}(A^T A X) = \text{trace}[B X] = \sum_{j=1}^L \sum_{i=1}^L B_{ji} X_{ij} \leq \max_{ij} |B_{ij}| = |A|_2^2. \tag{24}$$

(Recall that for a matrix $A \in \mathbb{R}^{d \times L}$ with columns $a_i, i = 1, \dots, L$, $|A|_2$ stands for the maximal column norm: $|A|_2 = \max_{1 \leq i \leq L} |a_i|_2$.)

We can rewrite the problem (11) using $Y = X^{1/2}$, so that the objective

$$\widehat{f}(X, P) = \text{trace}[\widehat{U}^T (I - P) \widehat{U} X]$$

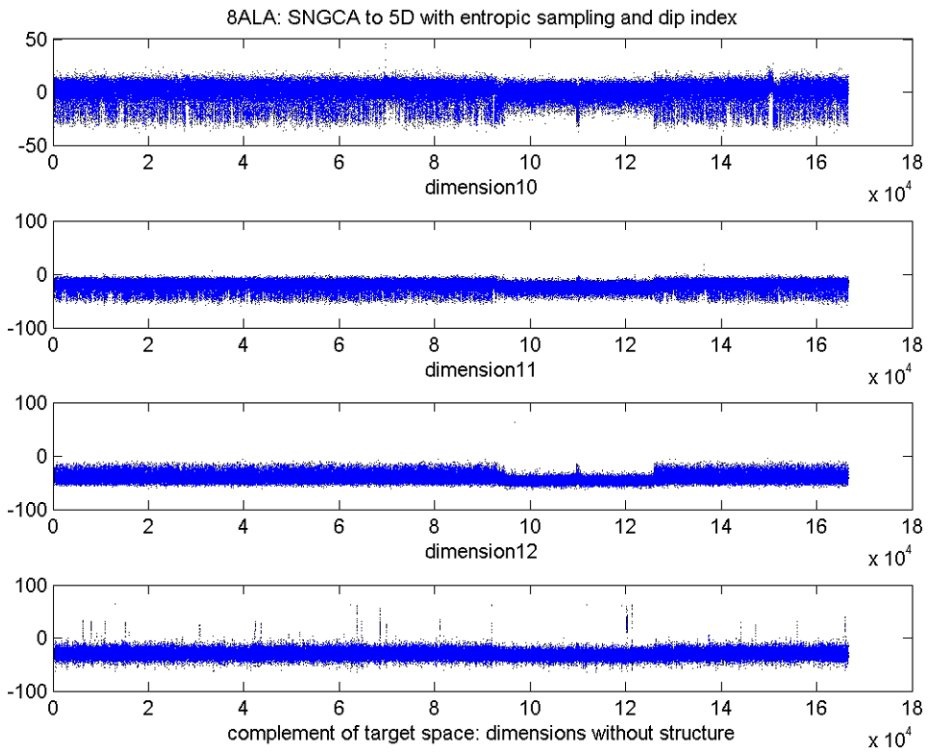


Fig. 6 Gaussian noise in the complement of the SNGCA target space

of (11) becomes

$$\widehat{g}(Y, P) = \|(I - P)^{1/2}\widehat{U}Y\|_2^2.$$

Let now $(\widehat{X}, \widehat{P})$ be a saddle point of (11). Namely, we have for any feasible P and X :

$$\widehat{f}(X, \widehat{P}) \leq [\widehat{f}_* \equiv \widehat{f}(\widehat{X}, \widehat{P})] \leq \widehat{f}(\widehat{X}, P).$$

We denote $\widehat{Y} = \widehat{X}^{1/2}$.

In what follows we suppose that vectors γ_ℓ and η_ℓ , $\ell = 1, \dots, L$ satisfy (8). In other words, it holds $|\widehat{U} - U|_2 \leq \delta_N$ and $|\widehat{G} - G|_2 \leq \gamma_N$.

A.1 Proof of Theorem 1

Lemma 1 *Let \widehat{P} be an optimal solution to (11). Then*

$$\max_c \{ |(I - \widehat{P})^{1/2}Uc|_2 \mid |c|_1 \leq 1, Gc = 0 \} \leq \lambda_{\min}^{-1}(\Sigma)(\rho + \nu_N) + 2\delta_N. \tag{25}$$

Proof We write:

$$\begin{aligned} & \max_c \{ |(I - \widehat{P})^{1/2}Uc|_2 \mid |c|_1 \leq 1, Gc = 0 \} \\ & \leq \max_Y \{ |(I - \widehat{P})^{1/2}UY\|_2 \mid |Y^2|_1 \leq 1, GY = 0 \} \end{aligned}$$

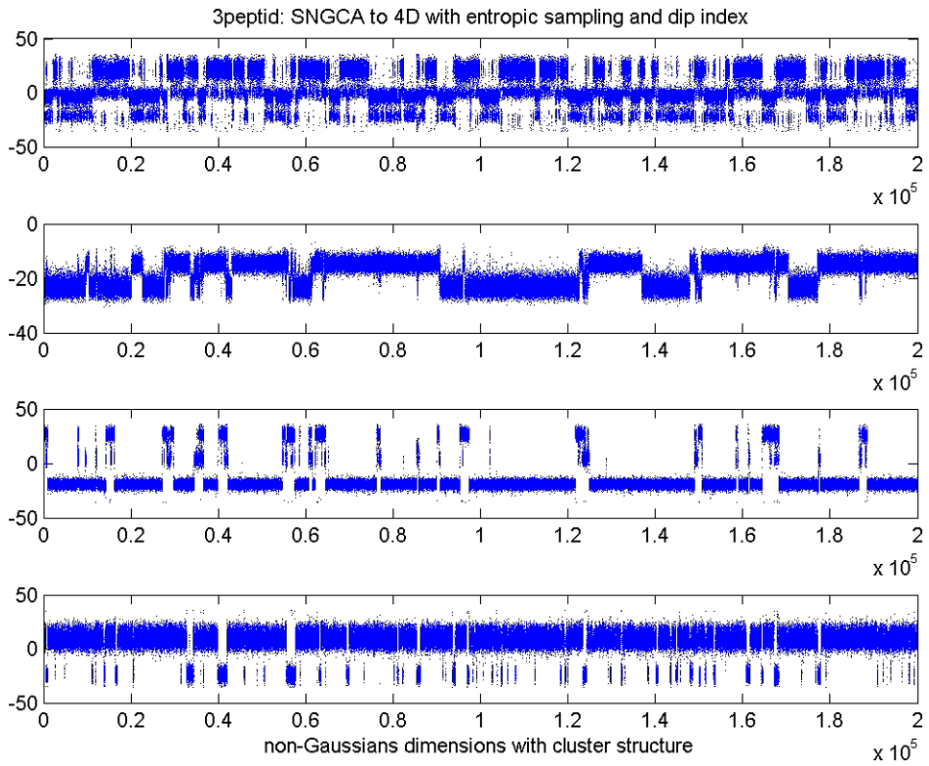


Fig. 7 Low dimensional multimodal component of 3-peptide

$$\begin{aligned}
 &\leq \max_Y \{ \|(I - \hat{P})^{1/2} \hat{U} Y\|_2 \mid |Y^2|_1 \leq 1, GY = 0 \} \\
 &\quad + \max_Y \{ \|(I - \hat{P})^{1/2} (\hat{U} - U) Y\|_2 \mid |Y^2|_1 \leq 1, GY = 0 \} \\
 &\leq \max_Y \{ \|(I - \hat{P})^{1/2} \hat{U} Y\|_2 \mid |Y^2|_1 \leq 1, \|\hat{G} Y\|_2 \leq \varrho \} \\
 &\quad + \|(I - \hat{P})^{1/2} (\hat{U} - U)\|_2 \quad (\text{by (24)}) \\
 &= \|(I - \hat{P})^{1/2} \hat{U} \hat{Y}\|_2 + \delta_N \leq \|(I - \Pi^*)^{1/2} \hat{U} \hat{Y}\|_2 + \delta_N \quad (\text{due to } 0 \leq I - \hat{P} \leq I) \\
 &\leq \|(I - \Pi^*)^{1/2} U \hat{Y}\|_2 + 2\delta_N \quad (\text{again by (24)}).
 \end{aligned}$$

On the other hand, as $\|\hat{G} \hat{Y}\|_2 \leq \nu_N$, we get

$$\|\hat{G} \hat{Y}\|_2 \leq \|\hat{G} \hat{Y}\|_2 + \|(\hat{G} - G) \hat{Y}\|_2 \leq \varrho + \|\hat{G} - G\|_2 \leq \varrho + \nu_N,$$

and by (3),

$$\|(I - \Pi^*) U \hat{Y}\|_2 \leq \lambda_{\min}^{-1}(\Sigma)(\varrho + \nu_N).$$

This implies (25). □

We now come back to the proof of the theorem. Let $\hat{\lambda}_j$ and $\hat{\theta}_j$, $j = 1, \dots, d$ be respectively the eigenvalues and the eigenvectors of \hat{P} . Assume that $\hat{\lambda}_1 \geq \hat{\lambda}_2 \geq \dots \geq \hat{\lambda}_d$. Then

$\widehat{P} = \sum_{j=1}^d \widehat{\lambda}_j \widehat{\theta}_j \widehat{\theta}_j^T$ and $\widehat{\Pi} = \sum_{j=1}^m \widehat{\theta}_j \widehat{\theta}_j^T$. Let $\beta = Uc$ for c such that $|c|_1 \leq 1$ and $Gc = 0$. We have

$$\begin{aligned} \beta^T (I - \widehat{P})\beta &= \sum_{j=1}^m (1 - \widehat{\lambda}_j) (\widehat{\theta}_j^T \beta)^2 + \sum_{j>m} (1 - \widehat{\lambda}_j) (\widehat{\theta}_j^T \beta)^2 \\ &\geq \sum_{j>m} (1 - \widehat{\lambda}_j) (\widehat{\theta}_j^T \beta)^2 \geq (1 - \widehat{\lambda}_{m+1}) (\widehat{\theta}_j^T \beta)^2 \\ &= (1 - \widehat{\lambda}_{m+1}) \beta^T (I - \widehat{\Pi})\beta = (1 - \widehat{\lambda}_{m+1}) |(I - \widehat{\Pi})\beta|_2^2. \end{aligned}$$

Since, for obvious reasons, $\widehat{\lambda}_{m+1} \leq \frac{m}{m+1}$, it applies (i) due to (25).

Let us show (ii). We have due to (12) and (25):

$$\begin{aligned} \text{trace}[(I - \widehat{P})\Pi^*] &= \text{trace}[(I - \widehat{P})^{1/2} \Pi^* (I - \widehat{P})^{1/2}] \\ &\leq \sum_{k=1}^m \mu_k \text{trace}[(I - \widehat{P})^{1/2} U c_k c_k^T U^T (I - \widehat{P})^{1/2}] \\ &= \sum_{k=1}^m \mu_k |(I - \widehat{P})^{1/2} U c_k|_2^2 \\ &\leq \sum_{k=1}^m \mu_k \max_c \{ |(I - \widehat{P})^{1/2} U c|_2^2 \mid |c|_1 \leq 1, Gc = 0 \} \\ &= \mu^*(\lambda_{\min}^{-1}(\Sigma)(\varrho + \nu_N) + 2\delta_N)^2, \end{aligned} \tag{26}$$

which is (13).

Note that $\text{trace}[\widehat{P}\Pi^*] \leq \sum_{j \leq m} \widehat{\lambda}_j$ (cf., e.g., Corollary 4.3.18 of Horn and Johnson 1985), thus by (26),

$$\widehat{\lambda}_{m+1} \leq m - \sum_{j \leq m} \widehat{\lambda}_j \leq \text{trace}[(I - \widehat{P})\Pi^*] \leq \mu^*(\lambda_{\min}^{-1}(\Sigma)(\varrho + \nu_N) + 2\delta_N)^2.$$

On the other hand,

$$\begin{aligned} \text{trace}[(I - \widehat{P})\Pi^*] &= \sum_{j=1}^m (1 - \widehat{\lambda}_j) \widehat{\theta}_j^T \Pi^* \widehat{\theta}_j + \sum_{j>m} (1 - \widehat{\lambda}_j) \widehat{\theta}_j^T \Pi^* \widehat{\theta}_j \\ &\geq \sum_{j>m} (1 - \widehat{\lambda}_j) \widehat{\theta}_j^T \Pi^* \widehat{\theta}_j \geq (1 - \widehat{\lambda}_{m+1}) \widehat{\theta}_j^T \Pi^* \widehat{\theta}_j \\ &= (1 - \widehat{\lambda}_{m+1}) \text{trace}[(I - \widehat{\Pi})\Pi^*], \end{aligned}$$

and we conclude that

$$\text{trace}[(I - \widehat{\Pi})\Pi^*] \leq \frac{\text{trace}[(I - \widehat{P})\Pi^*]}{1 - \widehat{\lambda}_{m+1}} \leq \frac{\mu^*(\lambda_{\min}^{-1}(\Sigma)(\varrho + \nu_N) + 2\delta_N)^2}{1 - \mu^*(\lambda_{\min}^{-1}(\Sigma)(\varrho + \nu_N) + 2\delta_N)^2}.$$

Now, using the relation $\text{trace} \widehat{\Pi} = \text{trace} \Pi^* = m$, we come to

$$\|\widehat{\Pi} - \Pi^*\|_2^2 = \text{trace}[\widehat{\Pi}^2 - 2\widehat{\Pi}\Pi^* + (\Pi^*)^2] = 2m - 2\text{trace}[\widehat{\Pi}\Pi^*] = 2\text{trace}[(I - \widehat{\Pi})\Pi^*],$$

and we arrive at (14).

A.2 Proof of Theorem 2

Let now \widehat{P} , \widehat{X} and $\widehat{t} = \text{trace } \widehat{P}$ be a triplet of optimal solution to (15).

Lemma 2 *Let \widehat{P} be an optimal solution to (15).*

- (i) *In the premises of the theorem Π^* is a feasible solution of (15) and $\text{trace } \widehat{P} \leq \text{trace } \Pi^* = m$.*
- (ii) *We have*

$$\max_c \{ |(I - \widehat{P})^{1/2} U c|_2 \mid |c|_1 \leq 1, Gc = 0 \} \leq \rho + \delta_N. \tag{27}$$

Proof We act as in the proof of Lemma 1: to verify (i) we observe that

$$\begin{aligned} & \max_X \{ \text{trace} [\widehat{U}^T (I - \Pi^*) \widehat{U} X] \mid X \succeq 0, |X|_1 \leq 1, \text{trace} [\widehat{G} X \widehat{G}^T] \leq \varrho^2 \} \\ &= \max_Y \{ \| (I - \Pi^*) \widehat{U} Y \|_2^2 \mid |Y^2|_1 \leq 1, \|\widehat{G} Y\|_2 \leq \varrho \} \\ &\leq \max_Y \{ (\| (I - \Pi^*) U Y \|_2 + \delta_N)^2 \mid |Y^2|_1 \leq 1, \|\widehat{G} Y\|_2 \leq \varrho \} \\ &\leq (\lambda_{\min}^{-1}(\Sigma)(\varrho + \nu_N) + \delta_N)^2. \end{aligned}$$

Thus, if $\rho \geq \lambda_{\min}^{-1}(\Sigma)(\varrho + \nu_N) + \delta_N$, Π^* is a feasible solution of (15) and, as a result, $\text{trace } \widehat{P} \leq \text{trace } \Pi^*$. To show (ii) it suffices to note that

$$\begin{aligned} & \max_c \{ |(I - \widehat{P})^{1/2} U c|_2 \mid |c|_1 \leq 1, Gc = 0 \} \\ &\leq \max_Y \{ \| (I - \widehat{P})^{1/2} U Y \|_2 \mid |Y^2|_1 \leq 1, GY = 0 \} \\ &\leq \max_Y \{ \| (I - \widehat{P})^{1/2} \widehat{U} Y \|_2 \mid |Y^2|_1 \leq 1, GY = 0 \} + |(I - \widehat{P})^{1/2} (\widehat{U} - U)|_2 \\ &\leq \max_Y \{ \| (I - \widehat{P})^{1/2} \widehat{U} Y \|_2 \mid |Y^2|_1 \leq 1, \|\widehat{G} Y\|_2 \leq \varrho \} + \delta_N \leq \rho + \delta_N \end{aligned}$$

because of the feasibility of \widehat{P} . □

Now using the bound $\widehat{m} \leq m$ we complete the proof following exactly the lines of the proof of Theorem 1.

A.3 Proof of Proposition 1

Observe that

$$\begin{aligned} \underline{E}(\widehat{Z}, \widehat{Y}) &= \min_{(P, W) \in \mathcal{P} \times \mathcal{W}} \{ \text{trace} [B^T (I - P) B Q^T \widehat{Z} Q] + \lambda \text{trace} [W (Q^T \widehat{Z} Q - \widehat{Y})] \} \\ &= \min_{P \in \mathcal{P}} \{ \text{trace} [B^T (I - P) B Q^T \widehat{Z} Q] - \lambda \| Q^T \widehat{Z} Q - \widehat{Y} \|_2 \} \\ &\leq \min_{P \in \mathcal{P}} \{ \text{trace} [B^T (I - P) B Q^T \widehat{Z} Q] \} = \underline{G}(\widehat{Z}); \end{aligned} \tag{28}$$

and

$$\begin{aligned} \overline{F}(\widehat{P}, \widehat{W}) &= \max_{(Z, Y) \in \mathcal{Z} \times \mathcal{Y}} \{ \text{trace}[B^T(I - \widehat{P})BQ^T ZQ] + \lambda \text{ trace}[\widehat{W}(Q^T ZQ - Y)] \} \\ &\geq \max_{Z \in \mathcal{Z}, |Q^T ZQ|_1 \leq 1, Y=Q^T ZQ} \{ \text{trace}[B^T(I - \widehat{P})BQ^T ZQ] \\ &\quad + \lambda \text{ trace}[\widehat{W}(Q^T ZQ - Y)] \} \\ &= \overline{G}(\widehat{P}). \end{aligned}$$

Assume first that $|Q^T \widehat{Z}Q|_1 \leq 1$. In this case $\widetilde{Z} = \widehat{Z}$ and

$$\epsilon \geq \overline{F}(\widehat{P}, \widehat{W}) - \underline{F}(\widehat{Z}, \widehat{Y}) = \overline{G}(\widehat{P}) - \underline{G}(\widehat{Z}) = \overline{G}(\widehat{P}) - \underline{G}(\widetilde{Z})$$

(the second, is given by (28)), as claimed. Now assume that $s = Q^T \widehat{Z}Q|_1 > 1$. We have already established the first equality of the following chain:

$$\begin{aligned} \underline{F}(\widehat{Z}, \widehat{Y}) &= \min_{P \in \mathcal{P}} \{ \text{trace}(B^T(I - P)BQ^T \widehat{Z}Q) - \lambda \|Q^T \widehat{Z}Q - \widehat{Y}\|_2 \} \\ &\leq \min_{P \in \mathcal{P}} \left\{ \text{trace}[B^T(I - P)BQ^T \widehat{Z}Q] - \frac{\lambda}{L} |Q^T \widehat{Z}Q - \widehat{Y}|_1 \right\} \\ &\leq \min_{P \in \mathcal{P}} \left\{ \text{trace}[B^T(I - P)BQ^T \widehat{Z}Q] - \frac{\lambda}{L} (s - 1) \right\} \\ &= \min_{P \in \mathcal{P}} \left\{ s \text{ trace}[B^T(I - P)BQ^T \widetilde{Z}Q] - \frac{\lambda}{L} (s - 1) \right\} \\ &\leq \min_{P \in \mathcal{P}} \left\{ \text{trace}[B^T(I - P)BQ^T \widetilde{Z}Q] + \underbrace{(s - 1) |B^T(I - P)B|_\infty |Q^T \widetilde{Z}Q|_1}_{\leq (s-1)|B^T B|_\infty = (s-1)|B|_2^2} \right. \\ &\quad \left. - \frac{\lambda}{L} (s - 1) \right\} \\ &\leq \min_{P \in \mathcal{P}} \{ \text{trace}[B^T(I - P)BQ^T \widetilde{Z}Q] \} = \underline{G}(\widetilde{Z}), \end{aligned}$$

where the concluding \leq is readily given by the definition of λ .³ Further, we have already seen that

$$\overline{F}(\widehat{P}, \widehat{W}) \geq \overline{G}(\widehat{P}).$$

Consequently,

$$\epsilon \geq \overline{F}(\widehat{P}, \widehat{W}) - \underline{F}(\widehat{Z}, \widehat{Y}) \geq \overline{G}(\widehat{P}) - \underline{G}(\widetilde{Z}),$$

as claimed. □

³We denote $|A|_\infty = \max_{ij} |A_{ij}|$.

A.4 Computing the prox-transform

Recall that because of the additivity of the distance-generating function d the computation of the prox-transform on the set $\mathcal{P} \times \mathcal{W} \times \mathcal{Z} \times \mathcal{Y}$ can be decomposed into independent computations on the four domains of (23).

Prox-transform on \mathcal{P} The proxy-function of \mathcal{P} is the matrix entropy:

$$d(P_0, P) = \beta_P \operatorname{trace} \left[\frac{P}{m} \left(\ln \left(\frac{P}{m} \right) - \ln \left(\frac{P_0}{m} \right) \right) \right] \quad \text{for } P, P_0 \in \mathcal{P}, \beta_P > 0.$$

To compute the corresponding component of T we need to find, given $S \in \mathbf{S}^d$,

$$\begin{aligned} T_{\beta}(P_0, S) &= \arg \max_{P \in \mathcal{P}} \left\{ \operatorname{trace}[S(P - P_0)] - \beta_P \operatorname{trace} \left[\frac{P}{m} \left(\ln \left(\frac{P}{m} \right) - \ln \left(\frac{P_0}{m} \right) \right) \right] \right\} \\ &= \arg \max_{P \in \mathcal{P}} \left\{ \operatorname{trace} \left[\left(S + \frac{\beta_P}{m} \ln \left(\frac{P_0}{m} \right) \right) P \right] - \beta_P \operatorname{trace} \left[\frac{P}{m} \ln \left(\frac{P}{m} \right) \right] \right\}. \end{aligned} \quad (29)$$

By the symmetry considerations we conclude that the optimal solution of this problem is diagonal in the basis of eigenvectors of $S + \frac{\beta_P}{m} \ln \left(\frac{P_0}{m} \right)$. Thus the solution of (29) can be obtained as follows: compute the eigenvalue decomposition

$$S + \frac{\beta_P}{m} \ln \left(\frac{P_0}{m} \right) = \Gamma \Lambda \Gamma^T$$

and let λ be the diagonal of Λ . Then solve the “vector” problem

$$p^* = \arg \max_{0 \leq p \leq 1, \sum p \leq m} \lambda^T p - \frac{\beta_P}{m} \sum_{i=1}^d p_i \ln(p_i/m) \quad (30)$$

and compose

$$T_{\beta}(P, S) = \Gamma \operatorname{diag}(y^*) \Gamma^T.$$

Now, the solution of (30) can be obtained by simple bisection: indeed, using Lagrange duality we conclude that the components of y^* satisfies

$$p_i^* = \exp \left(\frac{\lambda_i}{\beta} - \nu \right) \wedge 1, \quad i = 1, \dots, d,$$

and the Lagrange multiplier ν is to be set to obtain $\sum p_i^* = m$, what can be done by bisection in ν . When the solution is obtained, the optimal value of (29) can be easily computed.

Prox-transform on \mathcal{W} The distance-generating function of \mathcal{W} is $\beta_W \operatorname{trace}[W^2]/2 = \|W\|_2^2/2$ so that we have to solve for $S \in \mathbf{S}^L$

$$T_{\beta}(W_0, S) = \arg \max_{\|W\|_2 \leq 1} \left\{ \operatorname{trace}[S(W - W_0)] - \frac{\beta_W}{2} \frac{\|W - W_0\|_2^2}{2} \right\}. \quad (31)$$

The optimal solution to (31) can be easily computed

$$T_{\beta}(W_0, S) = \begin{cases} W_0 + S/\beta_W & \text{if } \|W_0 + S/\beta_W\|_2 \leq 1, \\ (W_0 + S/\beta_W)/\|W_0 + S/\beta_W\|_2 & \text{if } \|W_0 + S/\beta_W\|_2 > 1. \end{cases}$$

Prox-transform on \mathcal{Z} The prox-function of \mathcal{Z} is the matrix entropy and we have to solve for $S \in \mathbf{S}^{L-d}$

$$\begin{aligned} T_\beta(Z, S) &= \arg \max_{Z \in \mathcal{Z}} \text{trace}[S(Z - Z_0)] - \beta_Z \text{trace}[Z(\ln(Z) - \ln(Z_0))] \\ &= \arg \max_{Z \in \mathcal{Z}} \text{trace}[(S + \beta_Z \ln(Z_0))Z] - \beta_Z \text{trace}[Z \ln Z]. \end{aligned}$$

Once again, in the basis of eigenvectors of $S + \beta_Z \ln(Z_0)$ the problem reduces to

$$z^* = \arg \max_{z \geq 0, \sum z \leq 1} \lambda^T z - \beta_Z \sum_{i=1}^d z_i \ln(z_i),$$

where λ is the diagonal of Λ with $S + \beta_Z \ln Z_0 = \Gamma \Lambda \Gamma^T$. In this case

$$z_i^* = \frac{\exp(\frac{\lambda_i}{\beta})}{\sum_{j=1}^L \exp(\frac{\lambda_j}{\beta})}, \quad i = 1, \dots, L - d.$$

Prox-transform on \mathcal{Y} The distance generating function for the domain \mathcal{Y} is defined as follows:

$$\begin{aligned} d(Y) = \min \left\{ \sum_{i,j=1}^L (u_{ij} \ln[u_{ij}] + v_{ij} \ln[v_{ij}]) : \sum_{i=1}^n (u_{ij} + v_{ij}) = 1, \right. \\ \left. Y_{ij} = u_{ij} - v_{ij}, u_{ij} \geq 0, v_{ij} \geq 0, 1 \leq i, j \leq L \right\}. \end{aligned}$$

In other words, the element $Y \in \mathcal{Y}$ is decomposed according to $Y = u - v$, where (u, v) is an element of the $2L^2$ -dimensional simplex $\Delta = \{x \in \mathbb{R}^{2L^2}, x \geq 0, \sum_i x_i = 1\}$. To find the Y -component of the prox-transform amounts to find for $S \in \mathbf{S}^L$

$$\begin{aligned} T_{\beta_Y}(Y_0, S) &= T_{\beta_Y}(u^0, v^0, S) \\ &= \arg \max_{u,v \in \Delta} \text{trace}[S(u - v)] - \beta_Y \sum_{ij} \left[u_{ij} \ln\left(\frac{u_{ij}}{u_{ij}^0}\right) + v_{ij} \ln\left(\frac{v_{ij}}{v_{ij}^0}\right) \right]. \end{aligned} \tag{32}$$

One can easily obtain an explicit solution to (32): let

$$a_{ij} = u_{ij}^0 \exp\left(\frac{S_{ij}}{\beta_Y}\right), \quad b_{ij} = v_{ij}^0 \exp\left(-\frac{S_{ij}}{\beta_Y}\right).$$

Then $T_{\beta_Y}(Y_0, S) = u^* - v^*$, where

$$u_{ij}^* = \frac{a_{ij}}{\sum_{ij} (a_{ij} + b_{ij})}, \quad v_{ij}^* = \frac{b_{ij}}{\sum_{ij} (a_{ij} + b_{ij})}.$$

References

- Amadei, A., Linssen, A. B., & Berendsen, H. J. (1993). Essential dynamics of proteins. *Proteins*, 17(4), 412–425.
- Belkin, M., & Niyogi, P. (2009). Laplacian eigenmaps for dimensionality reduction and data representation. *Neural Computation*, 15(6), 1373–1396.
- Ben Tal, A., & Nemirovski, A. (2001). *MPS/SIAM series on optimization: Vol. 1. Lectures on modern convex optimization. Analysis, algorithms and engineering applications*. Philadelphia: SIAM.
- Blanchard, G., Kawanabe, M., Sugiyama, M., Spokoiny, V., & Müller, K.-R. (2006). In search of non-Gaussian components of a high-dimensional distribution. *Journal of Machine Learning Research*, 247–282.
- Candès, E. (2006). Compressive sampling. In *Int. congress of mathematics*, Madrid, Spain (Vol. 3, pp. 1433–1452).
- d'Aspremont, A., El Ghaoui, L., Jordan, M. I., & Lanckriet, G. R. G. (2007). A direct formulation for sparse PCA using semidefinite programming. *SIAM Review*, 49(3), 434–448.
- d'Aspremont, A., Bach, F., & El Ghaoui, L. (2008). Optimal solutions for sparse principal component analysis. *Journal of Machine Learning Research*, 9, 1269–1294.
- Diaconis, P., & Freedman, D. (1984). Asymptotics of graphical projection pursuit. *The Annals of Statistics*, 12, 793–815.
- Diederichs, E., Juditsky, A., Spokoiny, V., & Schütte, C. (2009). Sparse nonGaussian component analysis. *IEEE Transactions on Information Theory*, 15(7), 5249–5262.
- Frauenfelder, H., & McMahon, B. H. (2000). Energy landscape and fluctuations in proteins. *Annalen der Physik*, 9, 655–667.
- Guyon, I., & Elisseeff, A. (2003). An introduction to variable and feature selection. *Journal of Machine Learning Research*, 3, 1157–1182.
- Hastie, T. J., Tibshirani, R., & Friedman, J. (2001). *Springer series in statistics. The elements of statistical learning*. New York: Springer.
- Horenko, I., & Schütte, Ch. (2008). Likelihood-based estimation of multidimensional Langevin models and its application to biomolecular dynamics. *Multiscale Modeling & Simulation. SIAM Interdisciplinary Journal*, 7(2), 731–773.
- Horn, R. A., & Johnson, C. R. (1985). *Matrix analysis*. Cambridge: Cambridge University Press.
- Hyvärinen, A. (1999). Survey on independent component analysis. *Neural Computing Surveys*, 2, 94–128.
- Kawanabe, M., Sugiyama, M., Blanchard, G., & Müller, K.-R. (2007). A new algorithm of non-Gaussian component analysis with radial kernel functions. *Annals of the Institute of Statistical Mathematics*, 59(1), 57–75.
- Lu, Z., Monteiro, R., & Nemirovski, A. (2007). Large-scale semidefinite programming via saddle point mirror-prox algorithm. *Mathematical Programming*, 109(2–3), 211–237.
- Mizuta, M. (2004). Dimension reduction methods. In J. E. Gentle, W. Härdle, & Y. Mori (Eds.), *Handbook of computational statistics* (pp. 566–589).
- Nemirovski, A. (2004). Prox-method with rate of convergence $\mathcal{O}(1/t)$ for variational inequalities with Lipschitz continuous monotone operators and smooth convex-concave saddle point problems. *SIAM Journal on Optimization*, 15, 229–251.
- Nemirovski, A., & Yudin, D. (1983). *Problem complexity and method efficiency in optimization*. New York: Wiley.
- Nesterov, Yu. E. (2005). Smooth minimization of non-smooth functions. *Mathematical Programming. Series A and B*, 103(1), 127–152.
- Nesterov, Yu. E. (2007). Dual extrapolation and its applications for solving variational inequalities and related problems. *Mathematical Programming. Series A and B*, 109(2), 319–344.
- Pillardy, J., & Piela, L. (1995). Molecular dynamics on deformed energy hypersurfaces. *Journal of Physical Chemistry*, 99, 11805–11812.
- Reha, D., Valdes, H., Vondrasek, J., Hobza, P., Abu-Riziq, A., Crews, B., & de Vries, M. S. (2005). Structure an IR spectrum of phenylalanyl-glycyl-glycine tripeptide in the gas-phase. *European Journal of Chemistry*, 11, 6083–6817.
- Roweis, S., & Saul, L. (2000). Nonlinear dimensionality reduction by locally linear embedding. *Science*, 290, 2323–2326.
- Schaefer, M., & Karplus, M. (1996). A comprehensive analytical treatment of continuum electrostatics. *Journal of Chemical Physics*, 100, 1578–1599.
- Schütte, C., & Huisinga, W. (2003). Biomolecular conformations can be identified as metastable sets of molecular dynamics. In *Computational chemistry, handbook of numerical analysis* (pp. 699–744).

- Spokoiny, V. (2009). A penalized exponential risk bound in parametric estimation. <http://arxiv.org/abs/0903.1721>.
- Tenenbaum, de Silva, V., & Langford, J. C. (2000). A global geometric framework for nonlinear dimensionality reduction. *Science*, *290*, 2319–2323.
- van der Vaart, A., & Wellner, J. A. (1996) *Springer series in statistics. Weak convergence and empirical processes*. New York: Springer.
- Wasserman, L. (2006). *Springer texts in statistics. All of nonparametric statistics*. New York: Springer.