Sparse Non Gaussian Component Analysis by Semidefinite Programming

Elmar Diederichs*
Anatoli Juditsky**
Arkadi Nemirovski***
Vladimir Spokoiny*

* Weierstrass Institute (WIAS) Berlin, Germany
** Université J. Fourier Grenoble, France
*** Georgia Institute of Technology Atlanta, USA

BERLIN

ECONOMIC RISK

SFB 649

Sparse Non Gaussian Component Analysis by Semidefinite Programming

Elmar Diederichs [*], **Anatoli Juditsky**,
Arkadi Nemirovski, Vladimir Spokoiny [†],

Abstract Sparse non-Gaussian component analysis (SNGCA) is an unsupervised method of extracting a linear structure from a high dimensional data based on estimating a low-dimensional non-Gaussian data component. In this paper we discuss a new approach to direct estimation of the projector on the target space based on semidefinite programming which improves the method sensitivity to a broad variety of deviations from normality.

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

Mathematical Subject Classification: 62G07, 62H25, 62H30, 90C48, 90C90

JEL-Classification: C14

1 Introduction

Numerous statistical applications are confronted today with the so-called curse of dimensionality (cf. [13, 31]). 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

Elmar Diederichs
Weierstrass Institute Mohrenstr. 39, 10117 Berlin, Germany E-mail: diederic@wias-berlin.de

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

Arkadi Nemirovski
ISyE, Georgia Institute of Technology Atlanta, Georgia 30332, USA E-mail: nemirovs@isye.gatech.edu

Vladimir Spokoiny
Weierstrass Institute and Humboldt University, Mohrenstr. 39, 10117 Berlin, Germany E-mail: spokoiny@wias-berlin.de

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 [17,25]. In this study we consider the case where the phenomenon of interest is (approximately) located in a linear subspace $\mathcal{I}$. When compared to other approaches which involve construction of nonlinear mappings from the original data space onto the "subspace of interest", such that isomaps [29], local-linear embedding [13] or Laplacian eigenmaps [4], a linear mapping appears attractive due to its simplicity — it may be identified with a simple object, the projector Π^* from $\mathbb{R}^d$ onto $\mathcal{I}$. To find the structure of interest a statistician may seek for the non-Gaussian components of the data distribution, while its Gaussian components, as usual in the statistical literature may be treated as non-informative noise.

Several techniques of estimating the "non-Gaussian subspace" have been proposed recently. In particular, NGCA (for *Non-Gaussian Component Analysis*) procedure, introduced in [6], and then developed into SNGCA (for *Sparse NGCA*) in [10], is based on the decomposition the problem of dimension reduction into two tasks: the first one is to extract from the data a set $\{\widehat{\beta}_j\}$ of candidate vectors $\widehat{\beta}_j$ which are "close" to $\mathcal{I}$. The second is to recover an estimation $\widehat{\Pi}$ of the projector Π^* on $\mathcal{I}$ from $\{\widehat{\beta}_j\}$. In this paper we discuss a new method of SNGCA based on Semidefinite Relaxation of a nonconvex minmax problem which allows for a direct recovery of Π^*. When compared to previous implementations of the SNGCA in [6,7,10], the new approach "shortcuts" the intermediary stages and makes the best 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 $\mathcal{I}$.

The paper is organized as follows: in Section 2 we present the setup of SNGCA and briefly review some existing techniques. Then in Section 3 we introduce the new approach to recovery of the non-Gaussian subspace and analyze its accuracy. Further we provide a simulation study in Section 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. [6]. Let $X_1, ..., X_N$ 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:

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

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 $q : \mathbb{R}^m \to \mathbb{R}$ 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. Section 2 of [6]): 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 [6]) 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 [6]: suppose that the function q is smooth, then for any smooth function $\psi : \mathbb{R}^d \to \mathbb{R}$ the assumptions of (1) and $\mathbf{E}[X] = 0$ ensure that for

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

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

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

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

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

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 heuristics has been first used to estimate $\mathcal{I}$ in [6]. To be more precise, the construction implemented in [6] can be summarized as follows: let for a family $\{h_\ell\}$, $\ell = 1, ..., L$ of smooth bounded (test) functions on $\mathbb{R}^d$

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

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 [6] is as follows: $\widehat{\beta}_\ell = \widehat{\eta}_\ell - \widehat{\Sigma}^{-1}\widehat{\gamma}_\ell$, $\ell = 1, ..., L$, 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, $j = 1, ..., m$, are m principal eigenvectors of the matrix $\sum_{\ell=1}^L \widehat{\beta}_\ell \widehat{\beta}_\ell^T$. A numerical study, provided in [6], 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 [6] a different estimation procedure has been proposed in [10]. 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

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

Observe that $\beta(c) \in \mathcal{I}$ conditioned that $\gamma(c) = 0$. Indeed, if $\psi(x) = \sum_\ell c^\ell h_\ell(x)$, then $\sum_\ell c^\ell \mathbf{E}[X h_\ell(X)] = 0$, 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 [10] vectors $\{\widehat{c}_j\}$ are computed as follows: let

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

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

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

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 [10]), 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 [10,28]). 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 [10]: 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 $U = [\eta_1, ..., \eta_L] \in \mathbb{R}^{d \times L}$, $G = [\gamma_1, ..., \gamma_L] \in \mathbb{R}^{d \times L}$. 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

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

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

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

Indeed, it is well known (cf., e.g., Lemma 1 in [10] or [30]) that if functions $h_\ell(x) = f(x, \omega_\ell)$, $\ell = 1, ..., L$, 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

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

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:

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

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

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

Here we have substituted the constraint $Gc = 0$ with the inequality constraint $|\widehat{G}c|_2 \le \varrho$ 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 $\varrho \ge \nu_N$).

As we will see in a moment, when c runs the ν_N-neighborhood of intersection C_N of the standard hyperoctahedron $\{c \in \mathbb{R}^L, \; |c|_1 \le 1\}$ 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 $C_N = \{c \in \mathbb{R}^L, |c|_1 \le 1, |\widehat{G}c|_2 \le \nu_N\}$, 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. [18]).

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., [5, Chapter 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):

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

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

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

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

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

$$\min_P \max_X \left\{ \mathrm{trace}\left[\widehat{U}^T(I - P)\widehat{U}X\right] \;\middle|\; \begin{array}{c} 0 \preceq P \preceq I, \; \mathrm{trace}\,P = m, \\ X \succeq 0, \; |X|_1 \le 1, \; \mathrm{trace}[\widehat{G}X\widehat{G}^T] \le \varrho^2 \end{array} \right\} \tag{11}$$

(we write $P \preceq Q$ 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* $c_1, ..., c_{\overline{m}}$, $m \leq \overline{m} \leq L$ *such that* $|c_k|_1 \leq 1$ *and* $Gc_k = 0$, $k = 1, ..., \overline{m}$, *and non-negative constants* $\mu^1, \ldots, \mu^{\overline{m}}$ *such that*

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

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

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

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

and

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

(here $\|A\|_2 = \left(\sum_{i,j} A_{ij}^2\right)^{1/2} = \left(\text{trace}\,[A^T A]\right)^{1/2}$ 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 $1 - \epsilon$):

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

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

$$\min_{P,t} \left\{ t \left| \begin{array}{c} \operatorname{trace} P \leq t, \ \max_X \operatorname{trace} \left[\widehat{U}^T(I-P)\widehat{U}X\right] \leq \rho^2, \ 0 \preceq P \preceq I, \\ X \succeq 0, \ |X|_1 \leq 1, \ \operatorname{trace}[\widehat{G}X\widehat{G}^T] \leq \varrho^2 \end{array} \right. \right\} \tag{15}$$

The problem (15) is closely related to the ℓ_1-recovery estimator of sparse signals (see, e.g., the tutorial [8] and the references therein) and the trace minimization heuristics widely used in the Sparse Principal Component Analysis (SPCA) (cf. [2,3]). 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 $\widehat{t} = \operatorname{trace} \widehat{P}$ 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*

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

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

(i)

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

(ii) furthermore, if Assumption 1 hold then

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

and

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

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

---

[1] Here $\lfloor a \rfloor$ 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

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

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 [22] 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:

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

where

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

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

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

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 [19–22]. 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 [22] 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

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

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

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

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 $\overline{z}$ 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

$$z_{k+1} = T(\overline{z}, s_k),$$
$$z_{k+1}^+ = T(z_{k+1}, \lambda_k F(z_{k+1})),$$
$$s_{k+1} = s_k + \lambda_k F(z_{k+1}^+),$$

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-generaing 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 $\mathcal{P}$ 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.

---

[2] Recall 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 $x, y \in \mathcal{F}$.

4.2 Modified problem

We act as follows: first we eliminate the linear equality constraint which, taken along with $X \succeq 0$, says that $X = Q^T Z Q$ with $Z \succeq 0$ 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 $QQ^T = I$ (the orthogonal basis of the kernel of $\widehat{G}$). Note that from the constraints on X it follows that $\mathrm{trace}\,[X] \le 1$, whence

$$\mathrm{trace}\,[Q^T Z Q] = \mathrm{trace}\,[ZQQ^T] = \mathrm{trace}\,[Z] \le 1.$$

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

$$\mathcal{Z} = \{Z \in \mathbf{S}^{L-d}, \ Z \succeq 0, \ \mathrm{trace}\,[Z] \le 1\}.$$

Now can rewrite our problem equivalently as follows:

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

Let, further,

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

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

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

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*

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

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

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

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

$$\overline{G}(\widehat{P}) - \underline{G}(\widetilde{Z}) \le \epsilon,$$

where

$$\overline{G}(P) = \max_{Z \in \mathcal{Z}, \, |Q^T Z Q|_1 \le 1} \mathrm{trace}\,[\widehat{U}^T(I - P)\widehat{U}Q^T Z Q]; \ \ \underline{G}(Z)$$
$$= \min_{P \in \mathcal{P}} \mathrm{trace}\,[\widehat{U}^T(I - P)\widehat{U}Q^T Z Q].$$

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 ω_ℓ, $l = 1, ..., L$ are unit vectors in $\mathbb{R}^d$.

We implement here a multi-stage variant of the SNGCA (cf [10]). At the first stage of the SNGCA(SDP) algorithm 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 present for the estimation procedure with three stages. The size of the set of test function is set to $L = 10 \, d$, and the target accuracy of solving the problem (11) is set to $1e - 4$.

We can summarize the SNGCA(SDP) algorithm as follows:

Algorithm 1: SNGCA (SDP)

`% Initialization:`

The data $(X_i)_{i=1}^N$ are re-centered. Let $\sigma = (\sigma_1, \ldots \sigma_d)$ be the standard deviations of the components of X_i. We denote $Y_i = \mathrm{diag}(\sigma^{-1}) X_i$ 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 `η_ℓ` and `γ_ℓ

    for $\ell=1$ **to** L **do**

        $\widehat{\eta}_\ell^{(i)} = \frac{1}{N} \sum_{j=1}^N \nabla h_{\omega_\ell^{(i)}}(Y_j);$

        $\widehat{\gamma}_\ell^{(i)} = \frac{1}{N} \sum_{j=1}^N Y_j h_{\omega_l^{(i)}}(Y_j);$

    end

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

end

5.2 Experiment description

Each simulated data set $X^N = [X_1, ..., X_N]$ of size $N = 1000$ represents N 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 r.v.. 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: 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 plots of the densities of the non-Gaussian components on Figure 1.
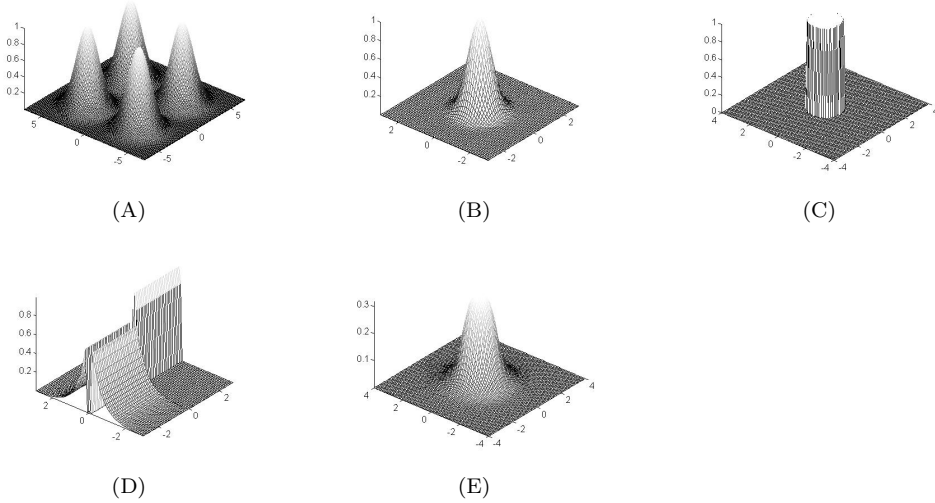
Fig. 1 (A) independent Gaussian mixtures, (B) isotropic super-Gaussian, (C) isotropic uniform and (D) dependent 1d Laplacian with additive 1d uniform, (E) isotropic sub-Gaussian

We start with comparing the presented algorithm with Projection Pursuit (PP) method [16] and the NGCA for $d = 10$. The results are presented on Figure 2 (the corresponding results for PP and NGCA has been already reported in [10] and [6]).
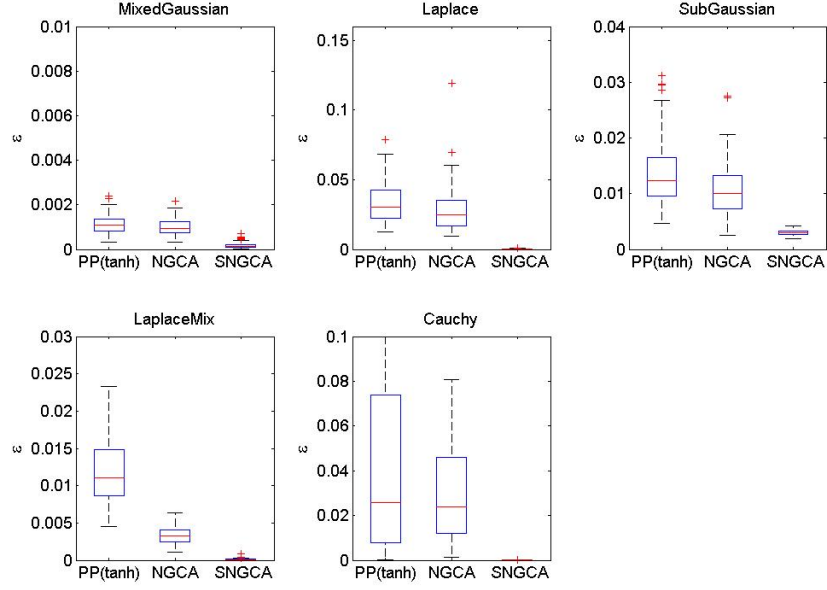
Fig. 2 Comparison of PP, NGCA and SNGCA(SDP)

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 Figure 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 $10^{-r}, 10^{-r+2r/7}, \ldots, 10^{r}$ where $r = 1, \ldots, 8$.
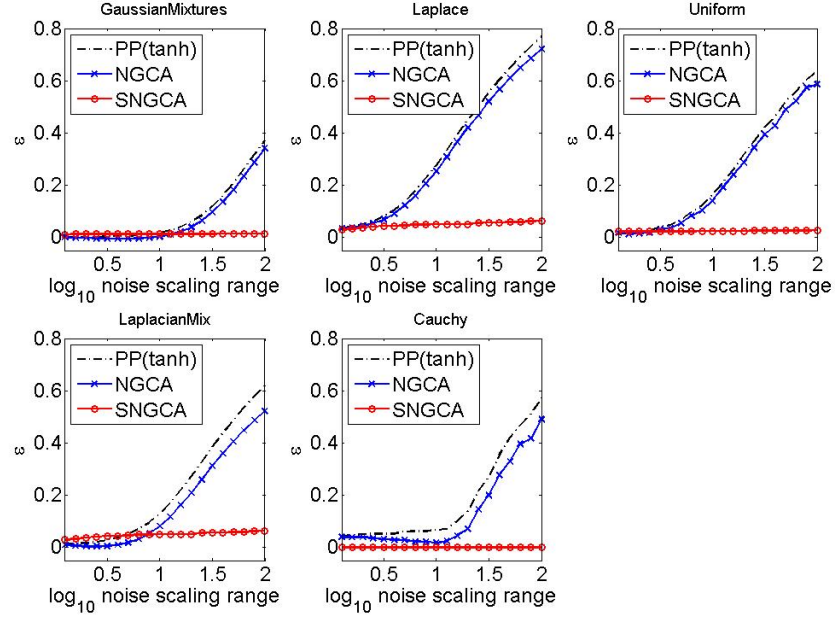
Fig. 3 estimation error with respect to the standard deviation of Gaussian components following a geometrical progression on $[10^{-r}, 10^{r}]$ where r is the parameter on the abscissa

The conditioning of the covariance matrix heavily influences the estimation error of PP(tanh) and NGCA, but not that of SNGCA(SDP). The latter method appears to be insensitive to the differences in the noise variance along different direction in all test cases.

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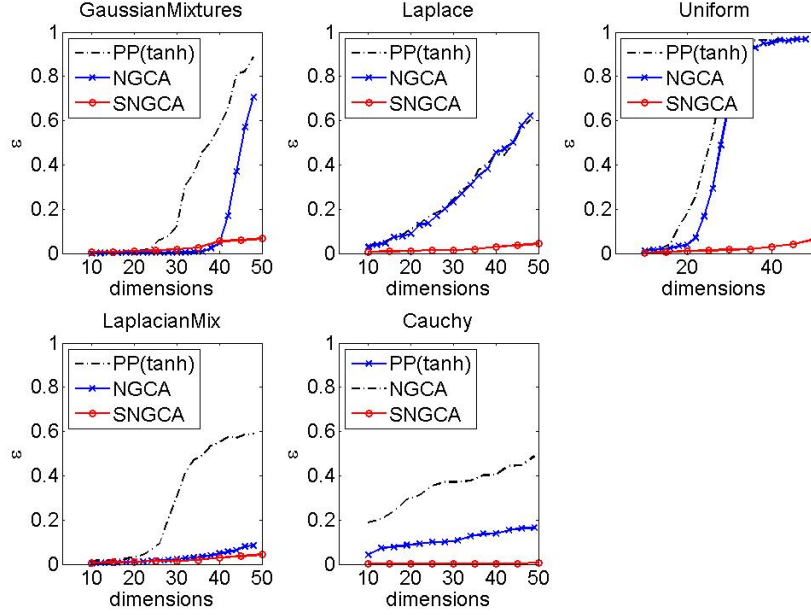
Fig. 4 mean-square estimation error vs problem dimension d

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 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 that the time scale of the local atomic vibrations, the structure is called metastable conformation [27]. 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 particulary deep wells each containing many local minima [23, 12]. Such wells represent different almost invariant geometrical large scale structures [1]. 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 [14]. 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.

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 14-dimensional time series consists of the cyclic data set of all backbone torsion angles. The simulation using CHARMM was done at $T = 300K$ with implicit water by means of the solvent model ACE2 [26]. A symplectic Verlet integrator with integration step of $1fs$ was used; the total trajectory length was $4\mu s$ and every $\tau = 50fs$ a set of coordinates was recorded.

The dimension reduction reported in the next figure was obtained using SNGCA(SDP) with for a given dimension $m = 5$ of the target space containing the multimodal component.
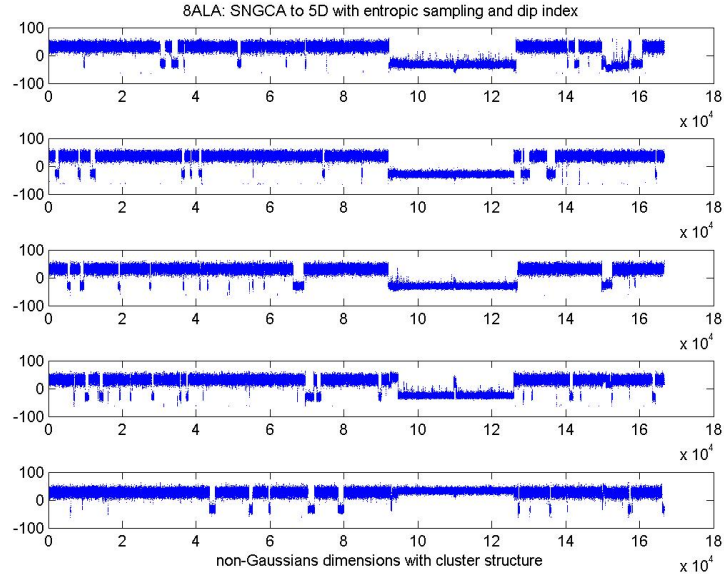
Fig. 5 low dimensional multimodal component of 8-alanine

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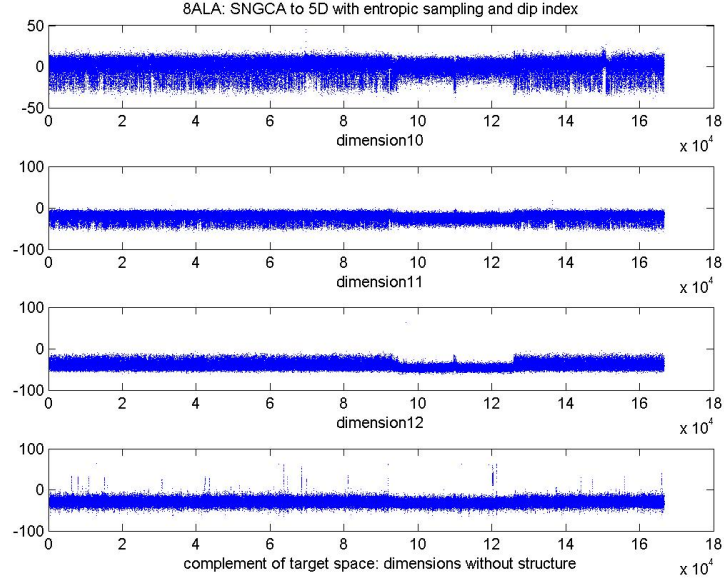
Fig. 6 Gaussian noise in the complement of the SNGCA target space.

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 [24]. The simulation is done using GROMACS at $T = 300K$ with implicit water. An integration step of a symplectic Verlet integrator is set to $2fs$, and every $\tau = 50fs$ a set of 31 diedre 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).
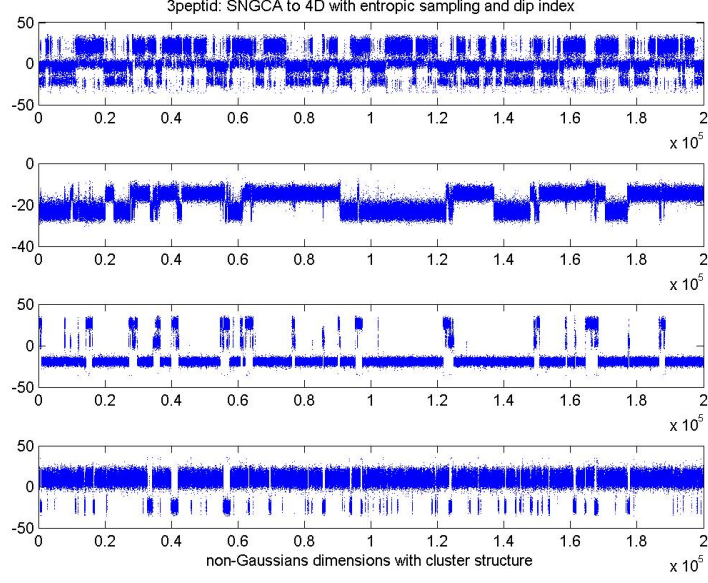
Fig. 7 low dimensional multimodal component of 3-peptide

6 Conclusions

We have studied a new procedure of non-Gaussian component analysis. The suggested method, same as the techniques proposed in [6, 10], has two stages: on the first stage certain linear functionals of unknown distribution are computed, then 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.

On the other hand, 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 the main practical limitation of the proposed approach.

20

References

1. A. Amadei & A. B. Linssen & H. J. Berendsen (1993) Essential dynamics of proteins. *Proteins*, **17**(4):412-425.

2. A. d'Aspremont, L. El Ghaoui, M.I. Jordan, and G. R. G. Lanckriet (2007) A direct formulation for sparse PCA using semidefinite programming. *SIAM Review*, **49**(3):434-448.

3. A. d'Aspremont, F. Bach and L. El Ghaoui (2008) Optimal solutions for sparse principal component analysis. *Journal of Machine Learning Research*, 9:1269-1294.

4. M. Belkin, P. Niyogi (2009) Laplacian Eigenmaps for dimensionality reduction and data representation *Neural Computation*, **15**(6):1373-1396.

5. A. Ben Tal & A. Nemirovski (2001) *Lectures on Modern Convex Optimization. Analysis, Algorithms and Engineering Applications*, Volume 1 of MPS/ SIAM Series on Optimization, SIAM, Philadelphia.

6. G. Blanchard, M. Kawanabe, M. Sugiyama, V. Spokoiny, K.-R. Müller (2006) In Search of Non-Gaussian Components of a High-Dimensional Distribution. *J. of Machine Learning Research*, p. 247-282.

7. M.Kawanabe, M.Sugiyama, G.Blanchard and K.-R.M'uller" (2007) A new algorithm of non-Gaussian component analysis with radial kernel functions, *Annals of the Institute of Statistical Mathematics*, vol. 59(1), 57-75.

8. E. Candès (2006) Compressive sampling. *Int. Congress of Mathematics, Madrid, Spain*, **3**:1433-1452.

9. P. Diaconis & D. Freedman (1984) Asymptotics of graphical projection pursuit *Annals of Statistics*, **12**:793-815.

10. E. Diederichs & A. Juditsky & V. Spokoiny & C. Schütte (2009) Sparse NonGaussian Component Analysis. *IEEE Transactions on Information Theory*, **15**(7):5249-5262.

11. I. Guyon & A. Elisseeff (2003) An Introduction to Variable and Feature Selection. *Journal of Machine Learning Research*, **3**:1157-1182.

12. H. Frauenfelder & B.H. McMahon (2000) Energy Landscape and fluctuations in proteins. *Ann. Phys. (Leipzig)*,**9**:655-667.

13. T. J. Hastie & R. Tibshirani & J. Friedman (2001) *The Elements of Statistical Learning*. New York, Springer Series in Statistcs.

14. I. Horenko & Ch. Schütte (2008) Likelihood-Based Estimation of Multidimensional Langevin Models and its Application to Biomolecular Dynamics. *Mult. Mod. Sim.*, **7**(2):731-773.

15. R. A. Horn and C. R. Johnson (1985) *Matrix Analysis*. Cambridge University Press.

16. A. Hyvärinen. (1999) Survey on independent component analysis. *Neural Computing Surveys*, **2**:94-128.

17. M. Mizuta (2004) Dimension Reduction Methods. In J.E. Gentle & W. Härdle, and Y. Mori (eds.): *Handbook of Computational Statistics* pp. 566-89.

18. A. Nemirovski & D. Yudin (1983) *Problem Complexity and Method Efficiency in Optimization*. New York, J. Wiley and Sons

19. A. Nemirovski (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

20. Z. Lu & R. Monteiro &A. Nemirovski (2007) Large-Scale Semidefinite Programming via Saddle Point Mirror-Prox Algorithm. *Mathematical Programming*, **109**:2-3, 211-237.

21. Yu. E. Nesterov (2005) Smooth minimization of non-smooth functions. *Mathematical Programming: Series A and B*, **103**(1):127-152.

22. Yu. E. Nesterov (2007) Dual extrapolation and its applications for solving variational inequalities and related problems. *Mathematical Programming: Series A and B*, **109**(2):319-344.

23. J. Pillardy & L. Piela (1995) Molecular dynamics on deformed energy hypersurfaces. *J.Phys.Chem.*, **99**:11805-11812.

24. D. Reha & H. Valdes & J.Vondrasek & P. Hobza & A. Abu-Riziq & B. Crews & M.S. de Vries (2005) Structure an IR Spectrum of Phenylalanyl-Glycyl-Glycine Tripeptide in the Gas-Phase. *Cem. Eur. J.*, **11**:6083-6817.

25. S. Roweis & L. Saul (2000) Nonlinear dimensionality reduction by locally linear embedding. *Science*, **290**:2323-2326.

26. M. Schaefer & M. Karplus (1996) A Comprehensive Analytical Treatment of Continuum Electrostatics. *J. Chem. Phys.*, **100**:1578-1599.

27. C. Schütte & W. Huisinga (2003) Biomolecular Conformations can be identified as metastable sets of melcular dynamics. *Computational Chemistry, Handbook of Numerical Analysis*, 699-744.

28. V. Spokoiny (2009) *A penalized exponential risk bound in parametric estimation.* `http://arxiv.org/abs/0903.1721`.

29. Tenenbaum & V. de Silva & J.C. Langford (2000) A global geometric framework for nonlinear dimensionality reduction. *Science*, **290**:2319-2323.

30. A. van der Vaart & J.A. Wellner (1996) *Weak Convergence and Empirical Proccesses.* Springer Series in Statistics - New York

31. L. Wasserman (2006) *All of Nonparametric Statistics.* New York, Springer Texts in Statistcss

A 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, $i = 1, .., L$ 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

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

(Recall that for a matrix $A \in \mathbb{R}^{d \times L}$ with columns a_i, $i = 1, ..., L$, $|A|_2$ 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

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

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, $\ell = 1, ..., L$ 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*

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

Proof. We write:

$$\max_c \left\{ |(I - \widehat{P})^{1/2} U c|_2 \mid |c|_1 \leq 1, \; Gc = 0 \right\}$$
$$\leq \max_Y \left\{ \|(I - \widehat{P})^{1/2} U Y\|_2 \mid |Y^2|_1 \leq 1, \; GY = 0 \right\}$$
$$\leq \max_Y \left\{ \|(I - \widehat{P})^{1/2} \widehat{U} Y\|_2 \mid |Y^2|_1 \leq 1, \; GY = 0 \right\}$$
$$+ \max_Y \left\{ \|(I - \widehat{P})^{1/2} (\widehat{U} - U) Y\|_2 \mid |Y^2|_1 \leq 1, \; GY = 0 \right\}$$
$$\leq \max_Y \left\{ \|(I - \widehat{P})^{1/2} \widehat{U} Y\|_2 \mid |Y^2|_1 \leq 1, \; \|\widehat{G} Y\|_2 \leq \varrho \right\}$$
$$\text{(by (24))} \quad + |(I - \widehat{P})^{1/2} (\widehat{U} - U)|_2$$
$$\text{(due to } 0 \preceq I - \widehat{P} \preceq I) = \|(I - \widehat{P})^{1/2} \widehat{U} \widehat{Y}\|_2 + \delta_N \leq \|(I - \Pi^*)^{1/2} \widehat{U} \widehat{Y}\|_2 + \delta_N$$
$$\text{(again by (24))} \leq \|(I - \Pi^*)^{1/2} U \widehat{Y}\|_2 + 2\delta_N.$$

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

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

and by (3),

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

This implies (25). $\qquad \square$

We now come back to the proof of the theorem. Let $\widehat{\lambda}_j$ and $\widehat{\theta}_j$, $j = 1, \ldots, d$ be respectively the eigenvalues and the

eigenvectors of $\widehat{P}$. Assume that $\widehat{\lambda}_1 \geq \widehat{\lambda}_2 \geq \ldots \geq \widehat{\lambda}_d$. 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):

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

(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 [15]), thus by (26),

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

On the other hand,

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

and we conclude that

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

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

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

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

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

(27)

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

$$\max_X \left\{ \text{trace}\left[\widehat{U}^T(I-\Pi^*)\widehat{U}X\right] \,\Big|\, X \succeq 0,\ |X|_1 \leq 1,\ \text{trace}\left[\widehat{G}X\widehat{G}^T\right] \leq \varrho^2 \right\}$$

$$= \max_Y \left\{ \|(I-\Pi^*)\widehat{U}Y\|_2^2 \,\Big|\, |Y^2|_1 \leq 1,\ \|\widehat{G}Y\|_2 \leq \varrho \right\}$$

$$\leq \max_Y \left\{ (\|(I-\Pi^*)UY\|_2 + \delta_N)^2 \,\Big|\, |Y^2|_1 \leq 1,\ \|\widehat{G}Y\|_2 \leq \varrho \right\}$$

$$\leq \left(\lambda_{min}^{-1}(\Sigma)(\varrho + \nu_N) + \delta_N \right)^2.$$

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

$$\max_c \left\{ |(I-\widehat{P})^{1/2}Uc|_2 \mid |c|_1 \leq 1,\ Gc = 0 \right\}$$

$$\leq \max_Y \left\{ \|(I-\widehat{P})^{1/2}UY\|_2 \mid |Y^2|_1 \leq 1,\ GY = 0 \right\}$$

$$\leq \max_Y \left\{ \|(I-\widehat{P})^{1/2}\widehat{U}Y\|_2 \mid |Y^2|_1 \leq 1,\ GY = 0 \right\} + |(I-\widehat{P})^{1/2}(\widehat{U}-U)|_2$$

$$\leq \max_Y \left\{ \|(I-\widehat{P})^{1/2}\widehat{U}Y\|_2 \mid |Y^2|_1 \leq 1,\ \|\widehat{G}Y\|_2 \leq \varrho \right\} + \delta_N \leq \rho + \delta_N$$

because of the feasibility of $\widehat{P}$. $\square$

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

$$\underline{F}(\widehat{Z},\widehat{Y}) = \min_{(P,W)\in\mathcal{P}\times\mathcal{W}} \left\{ \text{trace}\left[B^T(I-P)BQ^T\widehat{Z}Q\right] + \lambda\,\text{trace}\left[W(Q^T\widehat{Z}Q - \widehat{Y})\right] \right\}$$

$$= \min_{P\in\mathcal{P}} \left\{ \text{trace}\left[B^T(I-P)BQ^T\widehat{Z}Q\right] - \lambda\|Q^T\widehat{Z}Q - \widehat{Y}\|_2 \right\}$$

$$\leq \min_{P\in\mathcal{P}} \left\{ \text{trace}\left[B^T(I-P)BQ^T\widehat{Z}Q\right] \right\} = \underline{G}(\widehat{Z}); \tag{28}$$

and

$$\overline{F}(\widehat{P},\widehat{W}) = \max_{(Z,Y)\in\mathcal{Z}\times\mathcal{Y}} \left\{ \text{trace}\left[B^T(I-\widehat{P})BQ^TZQ\right] + \lambda\,\text{trace}\left[\widehat{W}(Q^TZQ - Y)\right] \right\}$$

$$\geq \max_{Z\in\mathcal{Z},\ |Q^TZQ|_1 \leq 1,\ Y=Q^TZQ} \left\{ \text{trace}\left[B^T(I-\widehat{P})BQ^TZQ\right] + \lambda\,\text{trace}\left[\widehat{W}(Q^TZQ - Y)\right] \right\}$$

$$= \overline{G}(\widehat{P}):$$

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 Â¸ 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:

$$\underline{F}(\widehat{Z},\widehat{Y}) = \min_{P\in\mathcal{P}} \left\{ \text{trace}\left(B^T(I-P)BQ^T\widehat{Z}Q\right) - \lambda\|Q^T\widehat{Z}Q - \widehat{Y}\|_2 \right\}$$

$$\leq \min_{P\in\mathcal{P}} \left\{ \text{trace}\left[B^T(I-P)BQ^T\widehat{Z}Q\right] - \frac{\lambda}{L}|Q^T\widehat{Z}Q - \widehat{Y}|_1 \right\}$$

$$\leq \min_{P\in\mathcal{P}} \left\{ \text{trace}\left[B^T(I-P)BQ^T\widehat{Z}Q\right] - \frac{\lambda}{L}(s-1) \right\}$$

$$= \min_{P\in\mathcal{P}} \left\{ s\,\text{trace}\left[B^T(I-P)BQ^T\widetilde{Z}Q\right] - \frac{\lambda}{L}(s-1) \right\}$$

$$\leq \min_{P\in\mathcal{P}} \left\{ \text{trace}\left[B^T(I-P)BQ^T\widetilde{Z}Q\right] + \underbrace{(s-1)|B^T(I-P)B|_\infty|Q^T\widetilde{Z}Q|_1}_{\leq (s-1)|B^TB|_\infty = (s-1)|B|_2^2} - \frac{\lambda}{L}(s-1) \right\}$$

$$\leq \min_{P\in\mathcal{P}} \left\{ \text{trace}\left[B^T(I-P)BQ^T\widetilde{Z}Q\right] \right\} = \underline{G}(\widetilde{Z}),$$

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. □

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:

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

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

$$T_\beta(P_0, S) = \arg \max_{P \in \mathcal{P}} \left\{ \text{trace} \left[S(P - P_0) \right] - \beta_P \, \text{trace} \left[\frac{P}{m} \left(\ln \left(\frac{P}{m} \right) - \ln \left(\frac{P_0}{m} \right) \right) \right] \right\} \tag{29}$$

$$= \arg \max_{P \in \mathcal{P}} \left\{ \text{trace} \left[\left(S + \frac{\beta_P}{m} \ln \left(\frac{P_0}{m} \right) \right) P \right] - \beta_P \, \text{trace} \left[\frac{P}{m} \ln \left(\frac{P}{m} \right) \right] \right\}.$$

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

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

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

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

and compose

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

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

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

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 $\beta_W \text{trace} \left[W^2 \right]/2 = \|W\|_2^2/2$ so that we have to solve for $S \in \mathbf{S}^L$

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

The optimal solution to (31) can be easily computed

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

---

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

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

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

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

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

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

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

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 $\Delta = \left\{ x \in \mathbb{R}^{2L^2}, \, x \geq 0, \, \sum_i x_i = 1 \right\}$. To find the Y-component of the prox-transform amounts to find for $S \in \mathbf{S}^L$

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

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

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

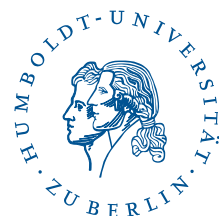# SFB 649 Discussion Paper Series 2011

For a complete list of Discussion Papers published by the SFB 649,
please visit http://sfb649.wiwi.hu-berlin.de.

001   "Localising temperature risk" by Wolfgang Karl Härdle, Brenda López Cabrera, Ostap Okhrin and Weining Wang, January 2011.

002   "A Confidence Corridor for Sparse Longitudinal Data Curves" by Shuzhuan Zheng, Lijian Yang and Wolfgang Karl Härdle, January 2011.

003   "Mean Volatility Regressions" by Lu Lin, Feng Li, Lixing Zhu and Wolfgang Karl Härdle, January 2011.

004   "A Confidence Corridor for Expectile Functions" by Esra Akdeniz Duran, Mengmeng Guo and Wolfgang Karl Härdle, January 2011.

005   "Local Quantile Regression" by Wolfgang Karl Härdle, Vladimir Spokoiny and Weining Wang, January 2011.

006   "Sticky Information and Determinacy" by Alexander Meyer-Gohde, January 2011.

007   "Mean-Variance Cointegration and the Expectations Hypothesis" by Till Strohsal and Enzo Weber, February 2011.

008   "Monetary Policy, Trend Inflation and Inflation Persistence" by Fang Yao, February 2011.

009   "Exclusion in the All-Pay Auction: An Experimental Investigation" by Dietmar Fehr and Julia Schmid, February 2011.

010   "Unwillingness to Pay for Privacy: A Field Experiment" by Alastair R. Beresford, Dorothea Kübler and Sören Preibusch, February 2011.

011   "Human Capital Formation on Skill-Specific Labor Markets" by Runli Xie, February 2011.

012   "A strategic mediator who is biased into the same direction as the expert can improve information transmission" by Lydia Mechtenberg and Johannes Münster, March 2011.

013   "Spatial Risk Premium on Weather Derivatives and Hedging Weather Exposure in Electricity" by Wolfgang Karl Härdle and Maria Osipenko, March 2011.

014   "Difference based Ridge and Liu type Estimators in Semiparametric Regression Models" by Esra Akdeniz Duran, Wolfgang Karl Härdle and Maria Osipenko, March 2011.

015   "Short-Term Herding of Institutional Traders: New Evidence from the German Stock Market" by Stephanie Kremer and Dieter Nautz, March 2011.

016   "Oracally Efficient Two-Step Estimation of Generalized Additive Model" by Rong Liu, Lijian Yang and Wolfgang Karl Härdle, March 2011.

017   "The Law of Attraction: Bilateral Search and Horizontal Heterogeneity" by Dirk Hofmann and Salmai Qari, March 2011.

018   "Can crop yield risk be globally diversified?" by Xiaoliang Liu, Wei Xu and Martin Odening, March 2011.

019   "What Drives the Relationship Between Inflation and Price Dispersion? Market Power vs. Price Rigidity" by Sascha Becker, March 2011.

020   "How Computational Statistics Became the Backbone of Modern Data Science" by James E. Gentle, Wolfgang Härdle and Yuichi Mori, May 2011.

021   "Customer Reactions in Out-of-Stock Situations – Do promotion-induced phantom positions alleviate the similarity substitution hypothesis?" by Jana Luisa Diels and Nicole Wiebach, May 2011.

SFB 649 Discussion Paper Series 2011

For a complete list of Discussion Papers published by the SFB 649,
please visit http://sfb649.wiwi.hu-berlin.de.

SFB 649 Discussion Paper Series 2011

For a complete list of Discussion Papers published by the SFB 649,
please visit http://sfb649.wiwi.hu-berlin.de.

044 "Predicting Bid-Ask Spreads Using Long Memory Autoregressive Conditional Poisson Models" by Axel Groß-Klußmann and Nikolaus Hautsch, July 2011.

045 "Bayesian Networks and Sex-related Homicides" by Stephan Stahlschmidt, Helmut Tausendteufel and Wolfgang K. Härdle, July 2011.

046 "The Regulation of Interdependent Markets", by Raffaele Fiocco and Carlo Scarpa, July 2011.

047 "Bargaining and Collusion in a Regulatory Model", by Raffaele Fiocco and Mario Gilli, July 2011.

048 "Large Vector Auto Regressions", by Song Song and Peter J. Bickel, August 2011.

049 "Monetary Policy, Determinacy, and the Natural Rate Hypothesis", by Alexander Meyer-Gohde, August 2011.

050 "The impact of context and promotion on consumer responses and preferences in out-of-stock situations", by Nicole Wiebach and Jana L. Diels, August 2011.

051 "A Network Model of Financial System Resilience", by Kartik Anand, Prasanna Gai, Sujit Kapadia, Simon Brennan and Matthew Willison, August 2011.

052 "Rollover risk, network structure and systemic financial crises", by Kartik Anand, Prasanna Gai and Matteo Marsili, August 2011.

053 "When to Cross the Spread: Curve Following with Singular Control" by Felix Naujokat and Ulrich Horst, August 2011.

054 "TVICA - Time Varying Independent Component Analysis and Its Application to Financial Data" by Ray-Bing Chen, Ying Chen and Wolfgang K. Härdle, August 2011.

055 "Pricing Chinese rain: a multi-site multi-period equilibrium pricing model for rainfall derivatives" by Wolfgang K. Härdle and Maria Osipenko, August 2011.

056 "Limit Order Flow, Market Impact and Optimal Order Sizes: Evidence from NASDAQ TotalView-ITCH Data" by Nikolaus Hautsch and Ruihong Huang, August 2011.

057 "Optimal Display of Iceberg Orders" by Gökhan Cebiroğlu and Ulrich Horst, August 2011.

058 "Optimal liquidation in dark pools" by Peter Kratz and Torsten Schöneborn, September 2011.

059 "The Merit of High-Frequency Data in Portfolio Allocation" by Nikolaus Hautsch, Lada M. Kyj and Peter Malec, September 2011.

060 "On the Continuation of the Great Moderation: New evidence from G7 Countries" by Wenjuan Chen, September 2011.

061 "Forward-backward systems for expected utility maximization" by Ulrich Horst, Ying Hu, Peter Imkeller, Anthony Réveillac and Jianing Zhang.

062 "On heterogeneous latent class models with applications to the analysis of rating scores" by Aurélie Bertrand and Christian M. Hafner, October 2011.

063 "Multivariate Volatility Modeling of Electricity Futures" by Luc Bauwens, Christian Hafner and Diane Pierret, October 2011.

SFB 649 Discussion Paper Series 2011

For a complete list of Discussion Papers published by the SFB 649,
please visit http://sfb649.wiwi.hu-berlin.de.