Inhalt

1 Motivation, Data and Problem
2 Semi-parametric framework
3 NGCA procedures
4 Numerical Experiments
Motivation, Data and Problem

- Examples: EEG and microarray data
- Robust risk management
- High dimensional clustering
- Conformational Changes of Biomolecules
Motivating Example: Highdimensional EEG Data

noisy mixed signal of 26 sensors;
top: raw; middle: $\delta$-waves (sleep); bottom: $\alpha$-waves
Robust risk management

\(R_t\), a vector of observed log-returns for a big portfolio.

**Problem:**

- Assumption of normal returns do not capture large losses caused by external shocks.
- Requires to include non-normal heavy tailed components in the distribution of the portfolio returns.


Clustering in high dimension

Given a sample $X_1, \ldots, X_n$ from a measure $\mathcal{P}$ on $\mathbb{R}^d$, identify the clustering (multimodality) structure of $\mathcal{P}$.

Curse of dimensionality: non-parametric methods poor for $d$ large.

NGCA approach: Gaussian component does not contribute to clustering, focus on non-Gaussian part (with a clear multimodal structure).
Example: folding states of 12-alanine

Most probable large scale shapes of 12-alanine, $\alpha$-helix and $\beta$-sheet
Observation of Different Time Scales in the Dynamics

a. *small and fast* variations around stable geometric mean due to random perturbations of the molecule from the solvent

b. rare flipping between *long-living* geometric mean configurations of a molecule, called *conformations*

conformational changes of 12-alanine as transition in the landscape of potential energy
General Picture of Dimension Reduction for Biomolecules

**Observation:** In conformational dynamics the detection of rare folding events coincides with structural data analysis.

**Aim:** find a linear combination of rotational angles (dieder angles) spanning a low dimensional conformational subspace.
Example: cluster structure in reduced data (12-alanine)

multimodal component of 12-alanine
Limitations of other Methods in this field

Drawbacks of standard methods to detect the cluster structure:

1. Conformational changes are realized by small variations s.t. PCA fails to detect them.

2. Direct Perron Cluster Analysis is unreliable for \( 10 \leq d \).

3. Fitting of HMM via EM-algorithm is computationally very expensive for \( 35 \leq d \) and the EM-algorithm has only local convergence.
Strategic for metastability analysis of high-dimensional biomolecules:

a. Reduce high-dimensional data with **SDNGCA**.

b. Fit a **HMM with Gaussian via EM** output and sufficient high number $M$ of hidden states to the data.

c. Consider the resulting Viterbi path, describing the macroscopic dynamics as a realization of a **Markov jump process**.

d. Perform **Perron Cluster Analysis** (PCCA) to detect the metastable states.
The trajectory of a 15-AT B-DNA oligonucleotide is simulated by AMBER with explicit water in \( d = 84 \) dieder angles contains \( T = 1 \cdot 10^5 \) time steps with each time step of 100 fs length and covers 1 ns at \( T = 300 K \).
Metastability of reduced B-DNA: 7 states

SDNGCA returns a $9d$ target space with $5d$ multimodal subspace. For illustration we show only a $3$ dimensional subspace of the target space with $7$ metastable states.

Reduced Gaussian target space of 12-alanine
Time Series of reduced B-DNA: 7 states

First five most multimodal components from the target space
The trajectory simulated by AMBER with implicit water in $d = 11$ dieder angles contains $T = 2 \cdot 10^4$ time steps with each time step of 50 $fs$ length and covers 0.5 $ns$ at $T = 300K$. 

Structure of Phenylalanyl-Glycyl-Glycine Tripetide
Metastability of reduced PGGT data: 9 states

SDNGCA returns a $4d$ target space with $3d$ multimodal subspace containing 9 metastable states.

Reduced Gaussian target space of 12-alanine
Time Series of reduced PGGT data: 9 states

first 3 most multimodal components from the target space
Time Series of remaining component

Projection of the data onto the components 4-6.
2 Semi-parametric framework
   ■ Model
   ■ The key idea
Data and problem

Data $X_1, \ldots, X_n \in \mathbb{R}^d$ i.i.d., $d$ large. For simplicity $\mathbb{E}X_i = 0$.

Goal: structural analysis.

Basic observation: high dimensional data tends to be normal:

A random projection $X^\top \omega$ is approximately normal for the most of directions $\omega$.

Gaussian component of the data is usually uninformative (noise).

Approach: project the data on the Non-Gaussian component.
Semiparametric Structural Assumption

Let $X_i$ i.i.d. with a density $\rho(\cdot)$. Suppose

$$\rho(x) = \phi_{0,\Sigma}(x)q(Tx) \quad (1)$$

- $\phi_{\mu,\Sigma}$, the normal density with parameter $(\mu, \Sigma)$
- $T : \mathbb{R}^d \to \mathbb{R}^m$ is a linear operator with $\mathcal{J} = \text{Ker}(T)^\perp$.
- $q : \mathbb{R}^m \to \mathbb{R}$, $m \leq d$, a smooth nonlinear function.

$\mathcal{J}$ is the target non-Gaussian subspace, $m$ is the non-Gaussian dimension.

Interpretation: $X = Z + \varepsilon$ where $\varepsilon$ is an independent Gaussian noise, $Z$, a signal.

(1) links pure Gaussian (PCA) and pure non-Gaussian (ICA) modeling.

Aim: recover $\mathcal{J}$ and possibly $m$. 
**Lemma**

Assume that $\rho(x)$ is the structured density according to (1). If $\psi(x) \in C^1(\mathbb{R}^d, \mathbb{R})$ fulfills

$$IE[X\psi(X)] = 0$$

then

$$\beta(\psi) \overset{\text{def}}{=} IE[\nabla \psi(X)] \in J.$$  

Moreover, if $IE[X\psi(X)] = \Delta \neq 0$, then there exists $\beta \in J$ s.t.

$$\|\beta - \beta(\psi)\|_2 \leq \|\Sigma^{-1}\Delta\|_2.$$
NGCA: Translation to RRR problem

Two big steps:

- **Sampling** Build some functions $\psi_1, \ldots, \psi_M$ such that

\[
\mathbb{E}_n \{X \psi_j(X)\} = n^{-1} \sum_{i=1}^n X_i \psi_j(X_i) = 0.
\]

Then every vector

\[
\hat{\beta}_j = \mathbb{E}_n \nabla \psi_j(X) = n^{-1} \sum_{i=1}^n \nabla \psi_j(X_i)
\]

belongs $J$ up to the empirical errors $(\mathbb{E} - \mathbb{E}_n) \nabla \psi_j(X)$ and $\Sigma^{-1} (\mathbb{I} - \mathbb{I}_n) \{X \psi_j(X)\}$.

- **Reduced Rank Regression** problem: Utilize $\{\hat{\beta}_j\}$ for recovering the target $m$-dimensional non-Gaussian subspace.
NGCA procedures

- NGCA 1G: Linear projection + PCA
- NGCA 2G: Convex projection
- NGCA 3G: SD Relaxation

(1): take any $h(\cdot)$ and consider $\psi(x) = h(x) - \alpha^\top x$.

Select $\alpha$ s.t. $\mathbb{E}_n \{X\psi(X)\} = \mathbb{E}_n \{Xh(X)\} - \mathbb{E}_n XX^\top \alpha = 0$.

Problem: requires to compute and study the inverse of the empirical covariance matrix.

(2): use PCA to recover the non-Gaussian subspace from the $\hat{\beta}_j$’s.

Problem: most of vectors $\hat{\beta}_j$ are uninformative, PCA often fails in dimensions over 10.

Given functions $h_1, \ldots, h_L$ compute

\[
\hat{\gamma}_\ell \overset{\text{def}}{=} \mathbb{E}_n[X h_\ell(X)] \approx \gamma_\ell \overset{\text{def}}{=} \mathbb{E}[X h_\ell(X)] \\
\hat{\eta}_\ell \overset{\text{def}}{=} \mathbb{E}_n[\nabla h_\ell(X)] \approx \eta_\ell \overset{\text{def}}{=} \mathbb{E}[
abla h_\ell(X)].
\]

**Convex projection:** given a direction $\xi \in \mathbb{R}^d$, solve

\[
\hat{c} = \arg\min_{c \in \mathbb{R}^L} \left\| \xi - \sum_\ell c_\ell \hat{\eta}_\ell \right\|_2 \quad \text{subject to} \quad \|c\|_1 \overset{\text{def}}{=} \sum_\ell |c_\ell| \leq 1, \quad \sum_\ell c_\ell \hat{\gamma}_\ell = 0,
\]

Define

\[
\hat{\beta} = \hat{\beta}(\hat{c}) = \sum_\ell \hat{c}_\ell \hat{\eta}_\ell.
\]
Consider the functions of the form

\[ h_\omega(x) \overset{\text{def}}{=} h(\omega^\top x)e^{-\lambda \|x\|^2/2} \]

with a given function \( h \) and a vector \( \omega \in \mathcal{B}_d \).

- Choose randomly a set of directions \( \{\xi_j\}, \ j = 1, \ldots, M \) and for every \( j \) a family of directions \( \{\omega_{j\ell}\}, \ \ell = 1, \ldots, L \).
- Compute \( \hat{\gamma}_{\ell,j} = IE_n X h_{\omega_{j\ell}}(X) \) and \( \hat{\eta}_{\ell,j} = IE_n [\nabla h_{\omega_{j\ell}}(X)] \).

- Solve for every \( j = 1, \ldots, M \)

\[
\{\hat{c}_{\ell,j}\} = \arg\min_{c \in \mathbb{R}^L} \left\| \xi_j - \sum_\ell c_\ell \hat{\eta}_{\ell,j} \right\|_2, \quad \text{subject to} \quad \sum_\ell c_\ell \hat{\gamma}_{\ell,j} = 0, \ |c|_1 \leq 1
\]

leading to

\[
\hat{\beta}_j = \sum_{\ell=1} \hat{c}_{\ell,j} \hat{\eta}_{\omega_{\ell,j}}
\]
Lemma

Let $h(\cdot)$ be bounded and continuously differentiable. For a fixed constant $C = C(h)$, it holds

$$\mathbb{E} \max_{\ell} |\hat{\gamma}_\ell - \gamma_\ell|^2 + |\hat{\eta}_\ell - \eta_\ell|^2 \leq C(h)n^{-1}\min\{d, \log L\} =: \varepsilon^2.$$
Suppose to be given the vectors $\hat{\beta}_1, \ldots, \hat{\beta}_M$ such that

$$\|(I - \Pi_I)\hat{\beta}_j\| \leq \epsilon$$

where $\Pi_I$ is a projector on a $m$-dimensional subspace.

Reduced Rank Regression problem: given $m$, recover $I$ (or $\Pi_I$) from $\hat{\beta}_1, \ldots, \hat{\beta}_M$.

More challenging: recover $m$ and $I$. 

NGCA 2G: Fritz-John rounding ellipsoid

PCA solution:

$$\hat{J} = \arg\min_{\text{dim}(J) = m} \sum_{j} \| (I - \Pi_{\hat{J}}) \hat{\beta}_{j} \|^{2} = \langle \text{first } m \text{ eigenvectors of } \sum_{j} \hat{\beta}_{j} \hat{\beta}_{j}^{\top} \rangle.$$ 

Requires that $\lambda_{m}(\sum_{j} \hat{\beta}_{j} \hat{\beta}_{j}^{\top}) \geq M\varepsilon^{2}$. Works poorly if most of the $\hat{\beta}_{j}$’s are non-informative.

Rounding ellipsoid approach: (see Yu.Nesterov, 2004) Define the set

$$\mathcal{A} \overset{\text{def}}{=} \{ \hat{\beta}_{1}, -\hat{\beta}_{1}, \hat{\beta}_{2}, -\hat{\beta}_{2}, \ldots \}.$$ 

and a centered ellipsoid of minimum volume that encloses $\mathcal{A}$. Recover $J$ from $\mathcal{E}$.

Leads to the accuracy $\| \Pi_{J} - \Pi_{\hat{J}} \|$ of order $d^{1/2}\varepsilon$. 
Structural adaptation idea (Hristache, Juditsky, Polzehl and Sp., 2003):

use the estimated ellipsoid $E_{k-1}$ as a prior information to improve the quality of estimation in the next step.

Leads to sequential procedure: alternate two steps

- estimate the model (vectors $\beta_j$) using the given structure
- estimate the structure (ellipsoid $E$)

Method: sample the directions $\xi_j$ and the vectors $\omega_{\ell,j}$ due to length of semiaxis of $E_{k-1}$.

This ensures that a certain fraction of $\xi_j$, $\gamma_{\ell,j}$ and $\eta_{\ell,j}$ is informative and hence, the corresponding solutions $\hat{\beta}_j$ are informative as well.
Pros:

- Convex projection helps preserves the individual estimation error;
- Rounding ellipsoid approach is more robust than PCA.

Open questions: choice of informative $\xi$, estimation of $m$.

Drawbacks:

- Computation of $\hat{\beta}_j$ using randomly chosen directions $\{\xi_j\}$ is expensive.
- Computation of Fritz-John ellipsoid of the set $S := \{\hat{\beta}_1, -\hat{\beta}_1, \hat{\beta}_2, -\hat{\beta}_2, \ldots\}$ always requires use of the inverse covariance matrix $\hat{\Sigma}^{-1}$.
- Structural adaptation does not work in high dimensions.
NGCA 3G: Minimax Approach

Aims at a direct estimation of the projector $\Pi$ on the target space $I$ from the data by solving a single semidefinite optimization problem.

Problem: given: $\omega_1, \ldots, \omega_L \in S_d$

▶ suppress the noise via the constraint

$$\sum_{\ell=1} \hat{c}_\ell \hat{\gamma}_\ell = \sum_{\ell=1} \hat{c}_\ell \mathbb{E}_n [Xh(\omega_\ell^\top X)] = 0.$$  

▶ access $I$ via $\sum_{\ell} c_\ell \hat{\eta}_\ell = \sum_{\ell} c_\ell \mathbb{E}_n \nabla h(\omega_\ell^\top X).$
NGCA 3G: Minimax Approach

Notation: $\hat{U} \overset{\text{def}}{=} [\hat{\eta}_1, \ldots, \hat{\eta}_L] \in \mathbb{R}^{d \times L}$, $\hat{G} \overset{\text{def}}{=} [\hat{\gamma}_1, \ldots, \hat{\gamma}_L] \in \mathbb{R}^{d \times L}$.

Minimax Approach (cf. Dalalyan, Juditsky, Sp 2009, JMLR): Solve the problem

$$\widehat{\Pi} = \arg\min_{\Pi} \max_c \left\{ \left\| (I - \Pi)\hat{U}c \right\|_2^2 \right\}$$

where $\Pi$ is a projector on a \(m\)-dimensional subspace of $\mathbb{R}^d$.

\[ c \in \mathbb{R}^L, \hat{G}c = 0, \left\| c \right\|_1 = 1 \]

where $\Pi$ is a Euclidean projector in $\mathbb{R}^d$.

- **Advantage**: shortcut of point estimation and target space reconstruction.

- **Problem**: (2) is a non-convex, non-smooth, hard optimization problem.
NGCA 3G: Relaxation

Aim: reduce the original problem to an approximate, convex-concave and smooth problem with an acceptable complexity.

Idea: drop non-convex constraints and solve an approximating semidefinite problem.
Joint with A. Nemirovsky:

i. Use positive semidefinite matrix $X = cc^\top$ as ”new variable”:

$$\|(I - \Pi)\hat{U}c\|_2^2 = \text{tr}\left[\hat{U}(I - \Pi)\hat{U}X\right].$$

ii. Relax rank $X = 1$ to $|X|_1 \overset{\text{def}}{=} \sum |X_{ij}| \leq 1$.

iii. Relax $\hat{G}c = 0$ to $\text{tr}[\hat{G}X\hat{G}] \leq \rho^2$.

iv. Relax rank $\Pi = m$ to $\text{tr} \Pi = m$, $0 \preceq \Pi \preceq I$.

Leads to a relaxed saddle point convex-concave problem:

$$\min_{P} \max_{X} \left\{ \text{tr}\left[\hat{U}(I - P)\hat{U}X\right] \right\} \begin{array}{c} 0 \preceq P \preceq I, \text{tr}[P] = m, \ \ X \succeq 0, \ |X|_1 \leq 1, \ \text{tr}[\hat{G}X\hat{G}] \leq \rho^2 \end{array}.$$
Solving the relaxed convex-concave SD problem:

▶ For large $L > 10^3$, interior point methods are too expensive;

▶ Adopt a subgradient descent-ascent method, e.g. dual extrapolation method (*Nesterov* 2007);

▶ complexity of one step $O(d \log d)$;

▶ precision $O\left(\frac{1}{k}\right)$, where $k$ is the number of steps.
Theorem

Let $\hat{P}$ be an optimal solution of the relaxed SDP and assume that

i. $\Pi^*$ on $J$ is a convex combination of rank-one matrices $U c c^T U^T$

ii. $c$ satisfies $Gc = 0$ and $\|c\|_1 \leq 1$.

Then it holds of $\hat{\Pi}$, spanned by $m$ eigenvectors of $\hat{P}$:

\[
\| (I - \hat{\Pi}) U c \|_2 \leq C_1 \sqrt{m + 1} (\rho + \lambda_{\min}^{-1}(\Sigma) + \epsilon)
\]

\[
\| \hat{\Pi} - \Pi^* \|_2^2 \leq C_2 (m + 1) [(\rho + \epsilon) \lambda_{\min}^{-1}(\Sigma)]^2.
\]
Aim: improve the estimation error of $\Pi$.

Approach:

i. **directional sampling**: choose $L$ directions $\omega_\ell$ uniform from $S_d$ to compute

$$\hat{U} = [\hat{\eta}_1, \ldots, \hat{\eta}_L] \in \mathbb{R}^{d \times L} \text{ and } \hat{G} = [\hat{\gamma}_1, \ldots, \hat{\gamma}_L] \in \mathbb{R}^{d \times L}$$

ii. **use result** $\hat{P}_k$ to get a "better" initial guess for the directions $\omega_\ell$ in iteration $k + 1$.

Definition of final projector $\hat{\Pi}$:

$$\hat{P}_{k+1} := [h_1, \ldots, h_d]^T \Lambda [h_1, \ldots, h_d] \text{ and } \hat{\Pi} := [h_1, \ldots, h_m].$$
4 Numerical Experiments

- Artificial Distributions
Test Distributions

Densities of the non-Gaussian components
Error criterion

The closeness of \( I \) and its estimate \( \hat{I} \) measured by

\[
\mathcal{E}(\hat{I}, I) \overset{\text{def}}{=} \frac{1}{2m} \| \Pi_I - \Pi_{\hat{I}} \|_{\text{Frob}}^2 = \frac{1}{m} \sum_{i=1}^{m} \| (1_d - \Pi_{\hat{I}}) h_i \|_2^2
\]

(3)

where \( \Pi_J \) denotes the orthogonal projection onto \( J \), \( \| \cdot \|_{\text{Frob}} \) is the Frobenius norm, \( \{ h_i \}_{i=1}^{m} \) is an orthonormal basis of \( \hat{J} \) and \( 1_d \) denotes the identity matrix.
Performance in $\mathbb{R}^{10}$

Comparison of PP, NGCA and SDNGCA by estimation error in 10 dimensions.
Comparison of Methods Cont’d: Increase of Dimension I

Comparison of PP, NGCA and SNGCA by estimation error for increasing dimensionality.
Comparison of Methods Cont’d: Increase of Dimension II

Comparison of PP, SNGCA, SDNGCA by estimation error for increasing dimensionality.
Effect of Numerical Condition on $\Sigma$

Comparison of PP, NGCA, SDNGCA by estimation error for increasing numerical condition of $\Sigma^{-1}$.
Motivating Example: Metastable analysis of biomolecules

Folding states of 12-alanine:

Most probable large scale shapes of 12-alanine, $\alpha$-helix and $\beta$-sheet
Observation of Different Time Scales in the Dynamics

a. **small and fast** variations around stable geometric mean due to random perturbations of the molecule from the solvent

b. rare flipping between **long-living** geometric mean configurations of a molecule, called **conformations**

conformational changes of 12-alanine as transition in the landscape of potential energy
Observation: In conformational dynamics the detection of rare folding events coincides with structural data analysis.

Aim: find a linear combination of rotational angles (dieder angles) spanning a low dimensional conformational subspace.
Example: cluster structure in reduced data (12-alanine)

multimodal component of 12-alanine
Limitations of other Methods in this field

Drawbacks of standard methods to detect the cluster structure:

1. Conformational changes are realized by small variations s.t. PCA fails to detect them.

2. Direct Perron Cluster Cluster Analysis is unreliable for $10 \leq d$.

3. Fitting of HMM via EM-algorithm is computationally very expensive for $35 \leq d$ and the EM-algorithm has only local convergence.
Determination of Metastable Subspace

i. Let \( \hat{P} := [h_1, \ldots, h_d]^T \Lambda [h_1, \ldots, h_d] \) and \( \hat{P}_J := [h_1, \ldots, h_m] \), where \( \hat{P} \) the solution of the relaxed SDP.

ii. Project the data \( X \) on \( [h_1, \ldots, h_m] \).

iii. Compute the well-known dip-index, that is significant to multimodality of every projected data \( h_i^\top X \).

iv. Take the subspace \( J_{multi} \subseteq J \) as final target space where the projected data with highest dip-index is located.
Structure of DNA Oligonucleotide

The trajectory of a 15-AT B-DNA oligonucleotide is simulated by AMBER with explicit water in $d = 84$ dieder angles contains $T = 1 \cdot 10^5$ time steps with each time step of $100 fs$ length and covers $1 ns$ at $T = 300 K$. 
SDNGCA returns a $9d$ target space with $5d$ multimodal subspace. For illustration we show only a $3$ dimensional subspace of the target space with $7$ metastable states.

Reduced Gaussian target space of 12-alanine
Time Series of reduced B-DNA: 7 states

First five most multimodal components from the target space
The trajectory simulated by AMBER with implicit water in $d = 11$ dieder angles contains $T = 2 \cdot 10^4$ time steps with each time step of $50 fs$ length and covers $0.5 ns$ at $T = 300K$.

Structure of Phenylalanyl-Glycyl-Glycine Tripeptide
Metastability of reduced PGGT data: 9 states

SDNGCA returns a 4$d$ target space with 3$d$ multimodal subspace containing 9 metastable states.

Reduced Gaussian target space of 12-alanine
Time Series of reduced PGGT data: 9 states

first 3 most multimodal components from the target space
Time Series of remaining component

Projection of the data onto the components 4-6.
1. Structural data analysis based on the non-Gaussian vs. Gaussian distinction is effective and computational not too expansive.

2. The Algorithm is independent from any use of $\hat{\Sigma}$.

3. Semidefinite relaxation leads to a statistically more sensitive and structural analysis with not too large complexity $O(kn^2 + L \log L)$.

4. Convergence rate of the estimation error: $O((m+1)[\rho \sqrt{\frac{d}{N}} \lambda_{\min}^{-1}(\Sigma)]^2)$.

5. The stochastic reduction of dimensionality works also with stochastic dynamical systems like large biomolecules.
Outlook

1. Estimation of the reduced dimension $m$ inside of the SDP-approach.

2. Development of criterion to check the new approach in the setting of biomolecules.

3. Development of code with very high performance.