Weierstraß-Institut für
Angewandte Analysis und Stochastik

Non-Gaussian Component Analysis using Semi Definite Relaxation

Vladimir Spokoiny

joint with A.Juditsky and E. Diederichs

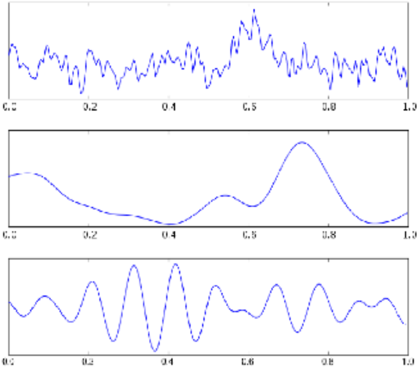
Inhalt

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

1 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: δ-waves (sleep); bottom: α-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.

- Chen, Härdle, VS (2010) GHICA – Risk analysis with GH distributions and independent components, J. Empirical Finance, 17 (2010) pp. 255–269.
- Chen, Härdle, VS (2007) Portfolio value at risk based on independent components analysis, J. Comp. Appl. Math., 205 (2007) pp. 594–607.

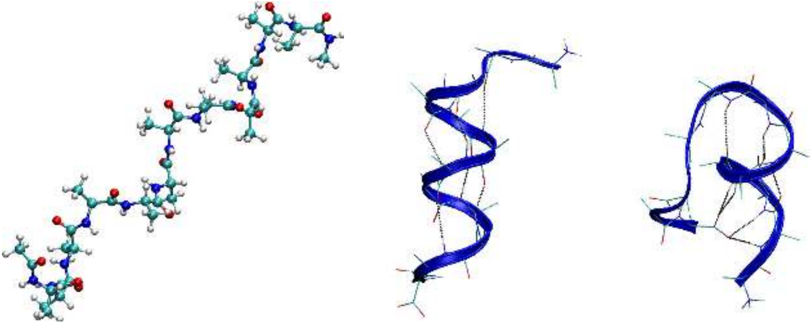
Clustering in high dimension

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

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

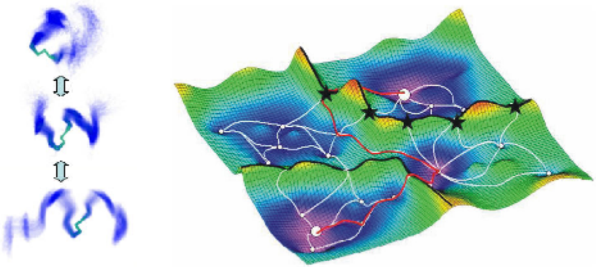


Most probable large scale shapes of 12-alanine, α-helix and β-sheet

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

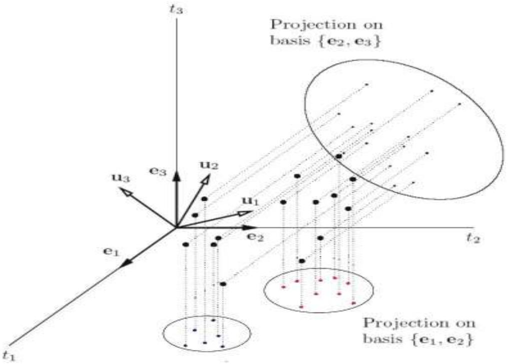
b. rare flipping between long-living geometric mean configurations of a molecule, called conformations



conformational changes of 12-alanine as transition in the landscape of potential energy

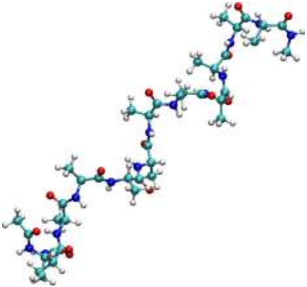
General Picture of Dimension Reduction for Biomolecules

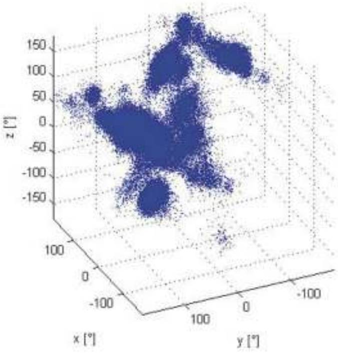
Observation: In conformational dynamics the detection of rare folding events coincides with structural data analysis.


t_3
Projection on basis $\{\mathbf{e}_2, \mathbf{e}_3\}$
$\mathbf{u}_2$
$\mathbf{u}_3$
$\mathbf{e}_3$
$\mathbf{u}_1$
$\mathbf{e}_2$
t_2
$\mathbf{e}_1$
Projection on basis $\{\mathbf{e}_1, \mathbf{e}_2\}$
t_1

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.

Metastability Analysis of Reduced Biomolecular Data

Strategy for metastability analysis of highdimensional biomulecules:

a. Reduce highdimensional data with SDNGCA.

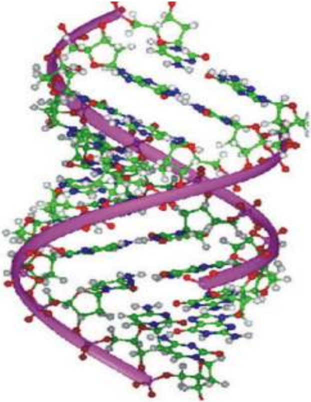
b. Fit a HMM with Gaussian via EM output and sufficient high number M of hidden states to the data

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.

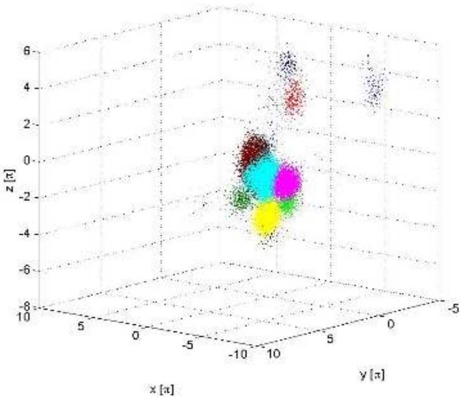
Structure of DNA Oligonucleotide

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 $100fs$ length and covers $1ns$ at $T = 300K$.



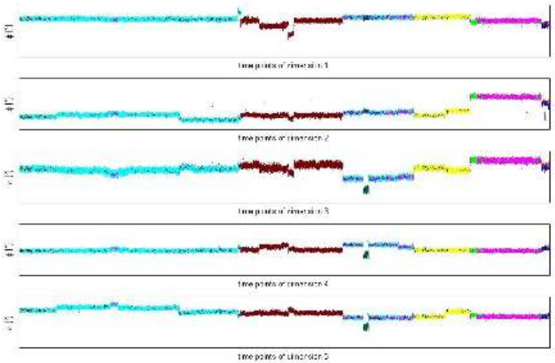
Metastability of reduced B-DNA: 7 states

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

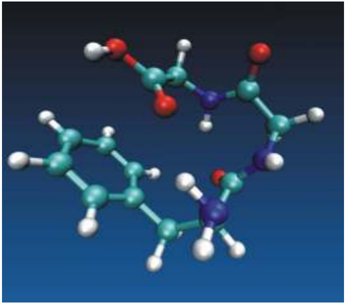
Time Series of reduced B-DNA: 7 states



First five most multimodal components from the target space

Structure of Phenylalanyl-Glycyl-Glycine Tripetide

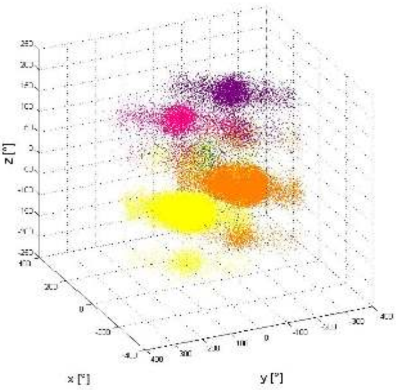
The trajectory simulated by AMBER with implicit water in $d = 11$ dieder angles contains $T = 2 \cdot 10^4$ time steps with each time step of $50fs$ length and covers $0.5ns$ at $T = 300K$.



Structure of Phenylalanyl-Glycyl-Glycine Tripetide

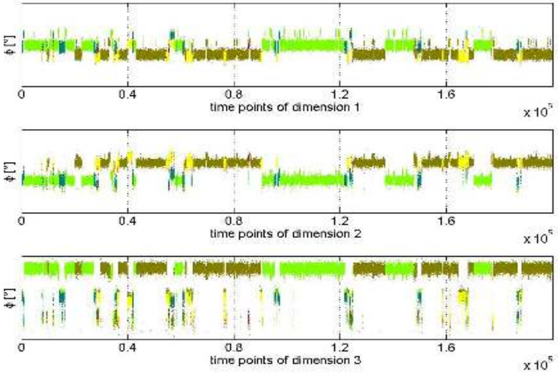
Metastability of reduced PGGT data: 9 states

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



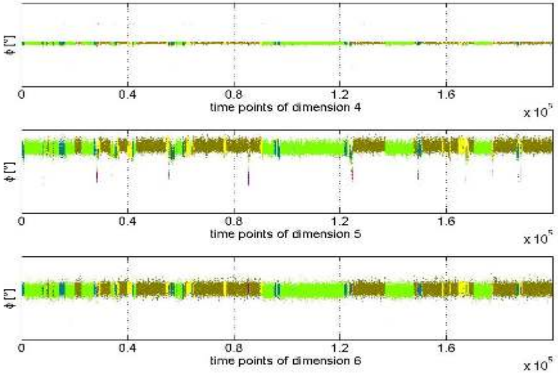
Reduced Gaussian target space of 12-alanine

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 $X_1, \dots, X_n \in I\!R^d$ i.i.d., d large. For simplicity $I\!E X_i = 0$.

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 ω.

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

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

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

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

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

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

Aim: recover $\mathcal{I}$ and possibly m.

Basic Lemma

Lemma

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

$$\mathbb{E}\left[X\psi(X)\right] = 0$$

then

$$\beta(\psi) \overset{\text{def}}{=} \mathbb{E}\left[\nabla\psi(X)\right] \in \mathcal{I}\,.$$

Moreover, if $\mathbb{E}\left[X\psi(X)\right] = \Delta \neq 0$, *then there exists* $\beta \in \mathcal{I}$ *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

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

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

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

3 NGCA procedures

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

NGCA 1G: Linear projection + PCA

Blanchard, Kawanabe, Sugiyama, Sp, K.-R. Müller, In search of non-Gaussian components of a high-dimensional distribution, J. Mach. Learn. Res., 7 (2006) pp. 247–282.

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

Select α 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 $\widehat{\beta}_j$'s.

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

NGCA 2G: Convex projection

Diederichs, Juditsky, Sp, Schütte (2010). Sparse Non-Gaussian Component Analysis. to appear IEEE of Inf. Theory, 2010.

Given functions $h_1, \ldots, h_L$ compute

$$\begin{array}{ccccccc} \widehat{\gamma}_\ell & \stackrel{\text{def}}{=} & I\!\!E_n\big[X h_\ell(X)\big] & \approx & \gamma_\ell & \stackrel{\text{def}}{=} & I\!\!E\big[X h_\ell(X)\big] \\ \widehat{\eta}_\ell & \stackrel{\text{def}}{=} & I\!\!E_n\big[\nabla h_\ell(X)\big] & \approx & \eta_\ell & \stackrel{\text{def}}{=} & I\!\!E\big[\nabla h_\ell(X)\big]. \end{array}$$

Convex projection: given a direction $\xi \in I\!\!R^d$, solve

$$\widehat{c} = \operatorname*{argmin}_{c \in I\!\!R^L} \Big\| \xi - \sum_\ell c_\ell \widehat{\eta}_\ell \Big\|_2 \quad \text{subject to} \quad \|c\|_1 \stackrel{\text{def}}{=} \sum_\ell |c_\ell| \le 1, \quad \sum_\ell c_\ell \widehat{\gamma}_\ell = 0,$$

Define

$$\widehat{\beta} = \widehat{\beta}(\widehat{c}) = \sum_\ell \widehat{c}_\ell \widehat{\eta}_\ell.$$

NGCA 2G: Directional sampling

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 $\widehat{\gamma}_{\ell,j} = I\!E_n X h_{\omega_{\ell,j}}(X)$ and $\widehat{\eta}_{\ell,j} = I\!E_n \left[\nabla h_{\omega_{\ell,j}}(X)\right]$.
- Solve for every $j = 1, \ldots, M$

  $$\{\widehat{c}_{\ell,j}\} = \operatorname*{argmin}_{c \in I\!R^L} \left\| \xi_j - \sum_\ell c_\ell \widehat{\eta}_{\ell,j} \right\|_2, \qquad \text{subject to} \qquad \sum_{\ell=1} c_\ell \widehat{\gamma}_{\ell,j} = 0, \ \|c\|_1 \le 1$$

  leading to

  $$\widehat{\beta}_j = \sum_{\ell=1} \widehat{c}_{\ell,j} \widehat{\eta}_{\omega_{\ell,j}}$$

NGCA 2G: Estimation error

Lemma

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

$$I\!\!E \max_{\ell} \bigl|\widehat{\gamma}_\ell - \gamma_\ell\bigr|^2 + \bigl|\widehat{\eta}_\ell - \eta_\ell\bigr|^2 \le C(h) n^{-1} \min\{d, \log L\} =: \varepsilon^2 .$$

NGCA 2G: Dimension Reduction Step

Suppose to be given the vectors $\widehat{\beta}_1, \ldots, \widehat{\beta}_M$ such that

$$\|(I - \Pi_{\mathcal{J}})\widehat{\beta}_j\| \le \varepsilon$$

where $\Pi_{\mathcal{J}}$ is a projector on a m-dimensional subspace.

Reduced Rank Regression problem: given m, recover $\mathcal{J}$ (or $\Pi_{\mathcal{J}}$) from $\widehat{\beta}_1, \ldots, \widehat{\beta}_M$.

More challenging: recover m and $\mathcal{J}$.

NGCA 2G: Fritz-John rounding ellipsoid

PCA solution:

$$\widehat{\mathcal{I}} = \operatorname*{argmin}_{dim(\mathcal{I})=m} \sum_j \|(I - \Pi_{\mathcal{I}})\widehat{\beta}_j\|^2 = \langle \text{first } m \text{ eigenvectors of } \sum_j \widehat{\beta}_j \widehat{\beta}_j^\top \rangle.$$

Requires that $\lambda_m\left(\sum_j \beta_j \beta_j^\top\right) \geq M\varepsilon^2$. Works poorly if most of the $\widehat{\beta}_j$'s are non-informative.

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

$$\mathcal{A} \overset{\text{def}}{=} \{\widehat{\beta}_1, -\widehat{\beta}_1, \widehat{\beta}_2, -\widehat{\beta}_2, \ldots\}.$$

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

Leads to the accuracy $\|\Pi_{\mathcal{I}} - \Pi_{\widehat{\mathcal{I}}}\|$ of order $d^{1/2}\varepsilon$.

NGCA 2G: Structural adaptation

Structural adaptation idea (Hristache, Juditsky, Polzehl and Sp., 2003):

use the estimated ellipsoid $\mathcal{E}_{k-1}$ 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 β_j) using the given structure
- estimate the structure (ellipsoid $\mathcal{E}$)

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

This ensures that a certain fraction of ξ_j, $\widehat{\gamma}_{\ell,j}$ and $\widehat{\eta}_{\ell,j}$ is informative and hence, the corresponding solutions $\widehat{\beta}_j$ are informative as well.

NGCA 2G: Pros and Cons

Pros:

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

Open questions: choice of informative ξ, estimation of m.

Drawbacks:

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

NGCA 3G: Minimax Approach

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

Problem: given: $\omega_1, \ldots, \omega_L \in \mathcal{S}_d$

- suppress the noise via the constraint

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

- access $\mathcal{I}$ via $\sum_\ell c_\ell \widehat{\eta}_\ell = \sum_\ell c_\ell \mathbb{E}_n \nabla h(\omega_\ell^\top X)$.

NGCA 3G: Minimax Approach

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

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

$$\widehat{\Pi} = \underset{\Pi}{\operatorname{argmin}} \max_{c} \left\{ \left\| (I - \Pi)\widehat{U}c \right\|_2^2 \,\middle|\, \begin{array}{c} \Pi \text{ is a projector on a} \\ m\text{-dimensional subspace of } \mathbb{R}^d \\ c \in \mathbb{R}^L,\ \widehat{G}c = 0,\ \|c\|_1 = 1 \end{array} \right\} \tag{2}$$

where Π 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.

NGCA 3G: Relaxation: Main steps

Joint with A. Nemirovsky:

i. Use positive semidefinite matrix $X = cc^\top$ as "new variable":
$\|(I - \Pi)\widehat{U} c\|_2^2 = \operatorname{tr}\big[\widehat{U}(I - \Pi)\widehat{U} X\big]$.

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

iii. Relax $\widehat{G} c = 0$ to $\operatorname{tr}[\widehat{G} X \widehat{G}] \le \rho^2$.

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

Leads to a relaxed saddle point convex-concave problem:

$$\min_P \max_X \left\{ \operatorname{tr}\left[\widehat{U}(I - P)\widehat{U} X\right] \;\middle|\; \begin{array}{c} 0 \preceq P \preceq I,\ \operatorname{tr}[P] = m, \\ X \succeq 0,\ |X|_1 \le 1,\ \operatorname{tr}[\widehat{G} X \widehat{G}] \le \rho^2 \end{array} \right\} .$$

NGCA 3G: Prize of Relaxation

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 $\mathcal{O}(d \log d)$;
- precision $\mathcal{O}(\frac{1}{k})$, where k is the number of steps.

NGCA 3G: Quality of estimation

Theorem

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

i. *Π^* on $\mathfrak{I}$ is a convex combination of rank-one matrices $Ucc^\top U^\top$*

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

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

$$\begin{aligned}\left\|(I-\widehat{\Pi})Uc\right\|_2 &\le C_1\sqrt{m+1}(\rho+\lambda_{\min}^{-1}(\Sigma)+\varepsilon)\\ \|\widehat{\Pi}-\Pi^*\|_2^2 &\le C_2(m+1)\left[(\rho+\varepsilon)\,\lambda_{\min}^{-1}(\Sigma)\right]^2.\end{aligned}$$

NGCA 3G: Structural Adaptation

Aim: improve the estimation error of Π.

Approach:

i. **directional sampling**: choose L directions ω_ℓ uniform from $\mathcal{S}_d$ to compute $\widehat{U} = [\widehat{\eta}_1, \ldots, \widehat{\eta}_L] \in \mathbb{R}^{d \times L}$ and $\widehat{G} = [\widehat{\gamma}_1, ..., \widehat{\gamma}_L] \in \mathbb{R}^{d \times L}$

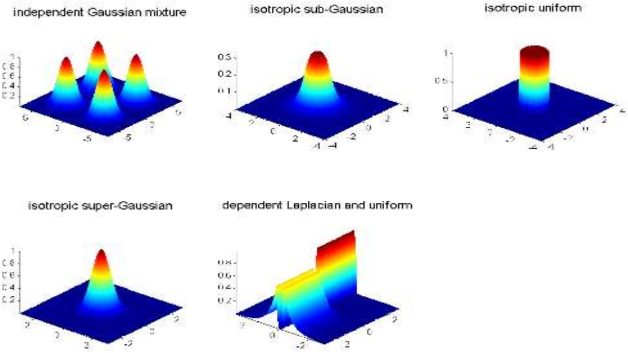
ii. **use result** $\widehat{P}_k$ to get a ”better” initial guess for the directions ω_ℓ in iteration $k+1$.

Definition of final projector $\widehat{\Pi}$: $\widehat{P}_{k+1} := [h_1, \ldots, h_d]^T \Lambda [h_1, \ldots, h_d]$ and $\widehat{\Pi} := [h_1, \ldots, h_m]$.

4 Numerical Experiments

- Artificial Distributions

Test Distributions


independent Gaussian mixture

isotropic sub-Gaussian

isotropic uniform

isotropic super-Gaussian

dependent Laplacian and uniform

Densities of the non-Gaussian components

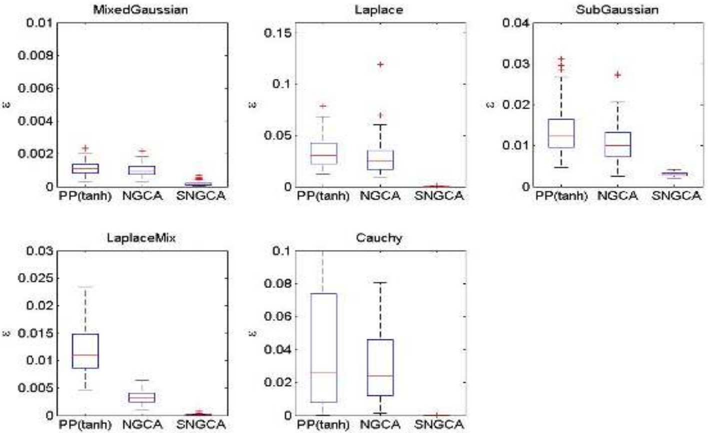
Error criterion

The closeness of $\mathcal{I}$ and its estimate $\widehat{\mathcal{I}}$ measured by

$$\mathcal{E}(\widehat{\mathcal{I}},\mathcal{I}) \overset{\text{def}}{=} \frac{1}{2m}\|\Pi_{\mathcal{I}} - \Pi_{\widehat{\mathcal{I}}}\|^2_{Frob} = \frac{1}{m}\sum_{i=1}^{m}\|(\mathbf{1}_d - \Pi_{\widehat{\mathcal{I}}})h_i\|^2 \tag{3}$$

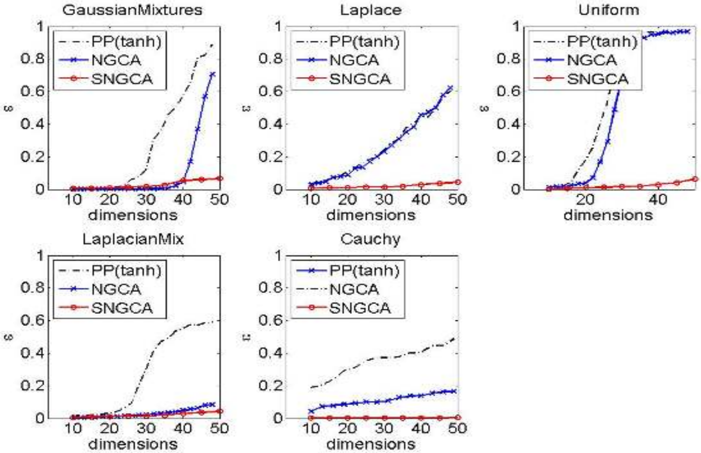
where $\Pi_{\mathcal{I}}$ denotes the orthogonal projection onto $\mathcal{I}$, $\|\cdot\|_{Frob}$ is the Frobenius norm, $\{h_i\}_{i=1}^m$ is an orthonormal basis of $\widehat{\mathcal{I}}$ and $\mathbf{1}_d$ denotes the identity matrix.

Performance in $\mathbb{R}^{10}$



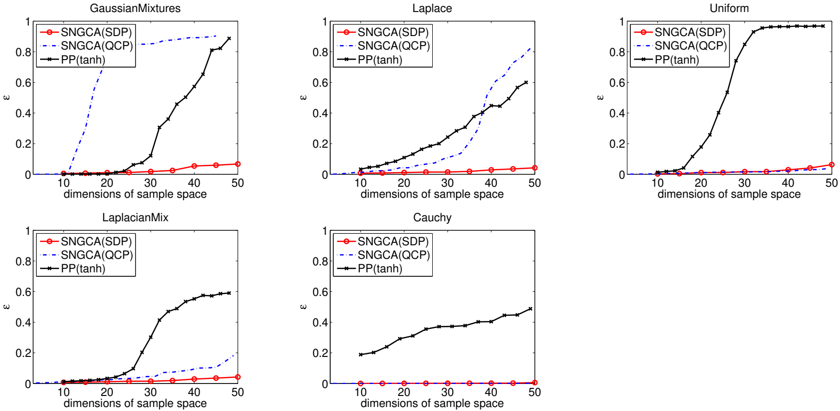
Comparison of PP, NGCA and SDNGCA by estimation error in 10 dimensions.

Comparison of Methods Cont'd: Increase of Dimension I



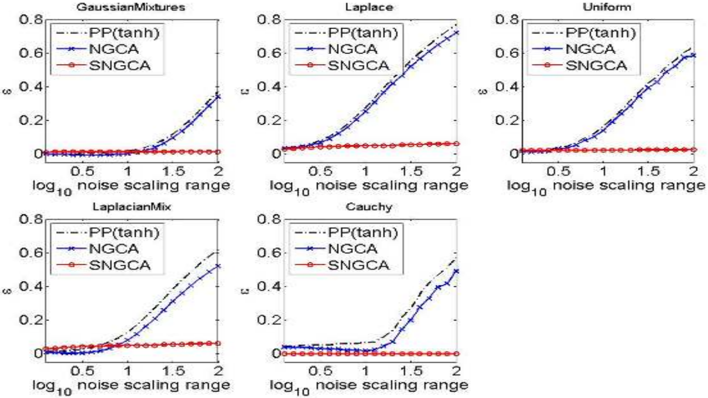
Comparison of PP, NGCA and SNGCA by estimation error for increasing dimensionality .

Comparison of Methods Cont'd: Increase of Dimension II



Comparison of PP, SNGCA, SDNGCA by estimation error for increasing dimensionality.

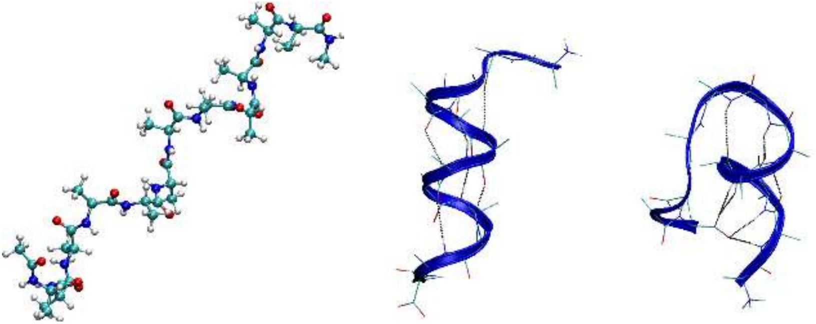
Effect of Numerical Condition on Σ



Comparison of PP, NGCA, SDNGCA by estimation error for increasing numerical condition of Σ^{-1}.

Motivating Example: Metastable analysis of biomoleculs

Folding states of 12-alanine:

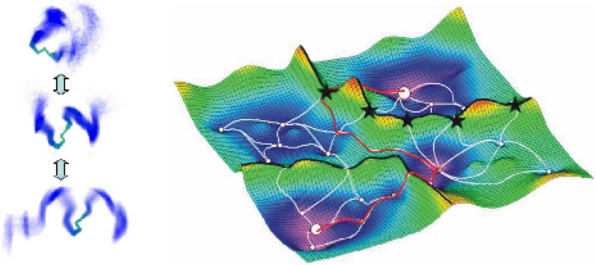


Most probable large scale shapes of 12-alanine, α-helix and β-sheet

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

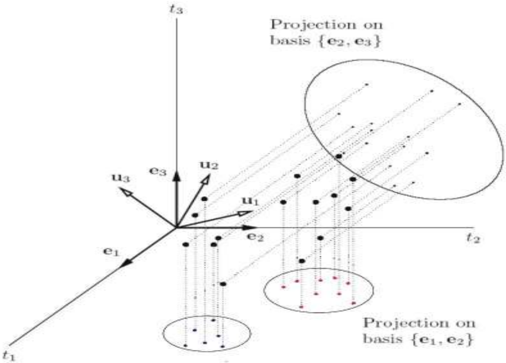
b. rare flipping between long-living geometric mean configurations of a molecule, called conformations



conformational changes of 12-alanine as transition in the landscape of potential energy

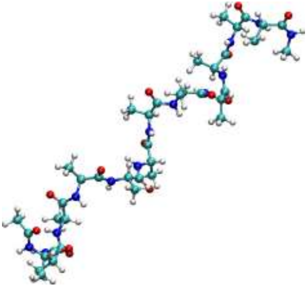
General Picture of Dimension Reduction for Biomolecules

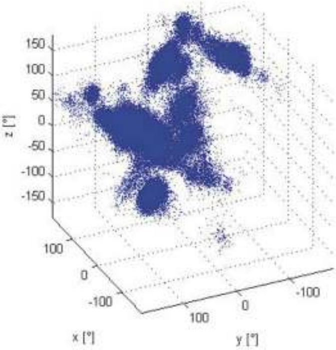
Observation: In conformational dynamics the detection of rare folding events coincides with structural data analysis.


t_3
Projection on basis $\{\mathbf{e}_2, \mathbf{e}_3\}$
$\mathbf{u}_3$
$\mathbf{e}_3$
$\mathbf{u}_2$
$\mathbf{u}_1$
$\mathbf{e}_1$
$\mathbf{e}_2$
t_2
t_1
Projection on basis $\{\mathbf{e}_1, \mathbf{e}_2\}$

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 $\widehat{P} := [h_1, \dots, h_d]^T \Lambda [h_1, \dots, h_d]$ and $\widehat{P}_{\mathcal{J}} := [h_1, \dots, h_m]$, where $\widehat{P}$ the solution of the relaxed SDP.

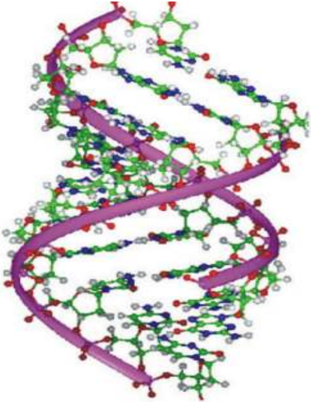
ii. Project the data X on $[h_1, \dots, h_m]$.

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

iv. Take the subspace $\mathcal{J}_{multi} \subseteq \mathcal{J}$ as final target space where the projected data with highest dip-index is located.

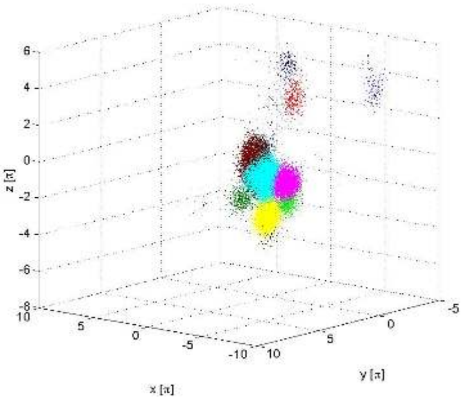
Structure of DNA Oligonucleotide

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 $100fs$ length and covers $1ns$ at $T = 300K$.



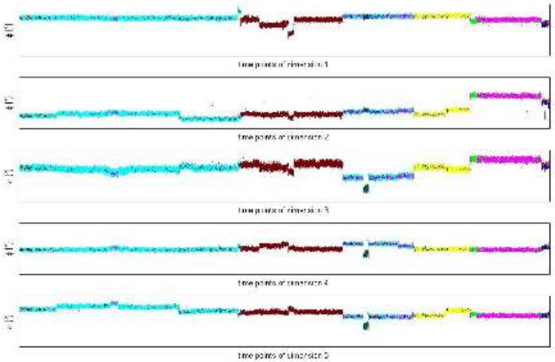
Metastability of reduced B-DNA: 7 states

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

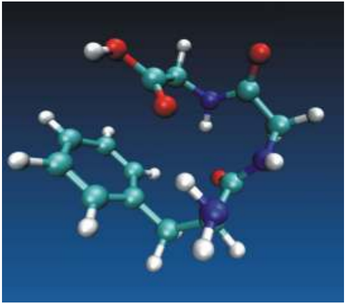
Time Series of reduced B-DNA: 7 states



First five most multimodal components from the target space

Structure of Phenylalanyl-Glycyl-Glycine Tripetide

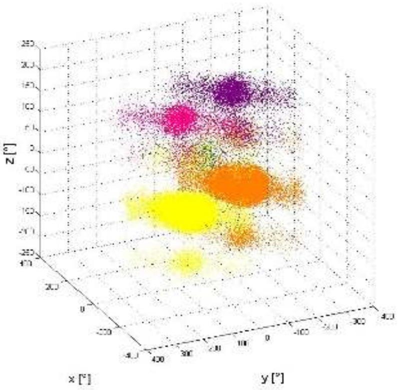
The trajectory simulated by AMBER with implicit water in $d = 11$ dieder angles contains $T = 2 \cdot 10^4$ time steps with each time step of $50fs$ length and covers $0.5ns$ at $T = 300K$.



Structure of Phenylalanyl-Glycyl-Glycine Tripetide

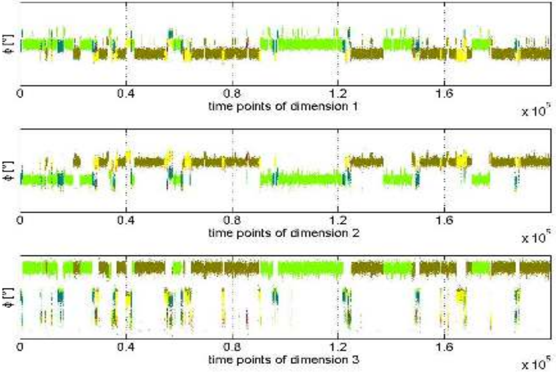
Metastability of reduced PGGT data: 9 states

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



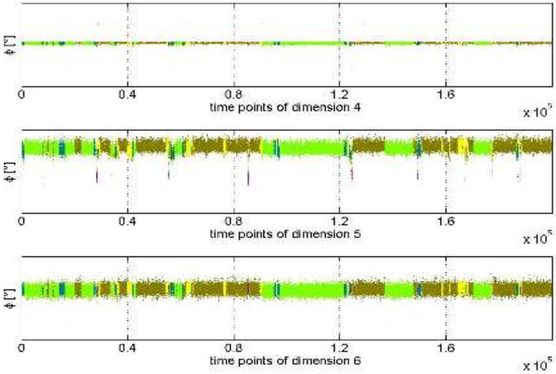
Reduced Gaussian target space of 12-alanine

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.

Summary

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 $\widehat{\Sigma}$.

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

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

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.