Sparse NonGaussian Component Analysis with Applications to Conformation Dynamics of Biomolecular Systems

E. Diederichs[1]

joint work with A.Juditsky[3] and V. Spokoiny[2]

[1]Free University of Berlin

[2]WIAS and Humboldt University

[3]Laboratoire Jean Kuntzmann, Université Joseph Fourier, Grenoble

June 29, 2009

Outline

Outline

Outline

Outline

Outline

Motivation for Structural Data Analysis

Under physical constraints of constant volume and temperature we observe:

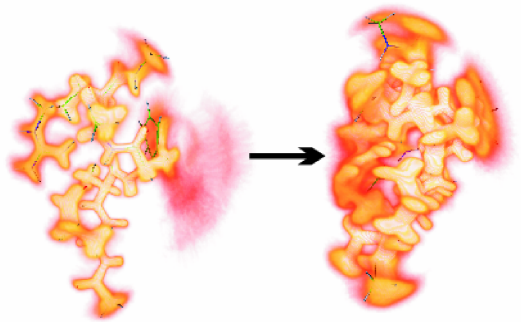


Figure: Changes between different conformations of a biological active molecule.

Observe that small variations around stable geometric mean configurations of a molecule, called conformations, correspond to connected set of the state space.

Motive: The large scale geometry of a molecular system determines its biological function.

Different Time Scales in the Dynamics

Observation: Changes of geometric large scale configurations of a molecule have life times much longer than the time scale of the internal interactions between the atoms and the random perturbations of the molecule from the solvent.

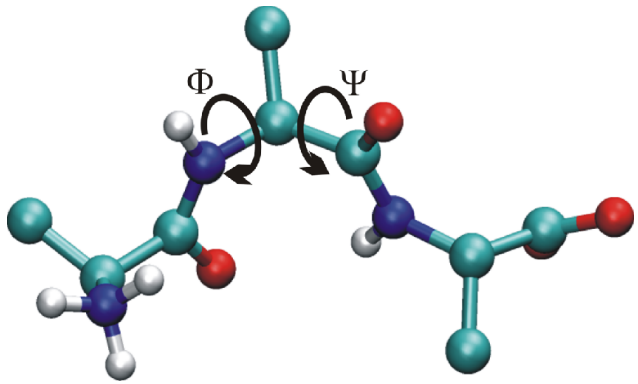


Figure: Backbone of alanine-dipeptid with dihedral-angels (Φ, Ψ).

The rotational degrees of freedom (Φ, Ψ) allow to observe the rare macroscopic folding events of a biomolecule as a change of the geometric configuration of the backbone.

Detection of Rare Events in High Dimensional Time Series

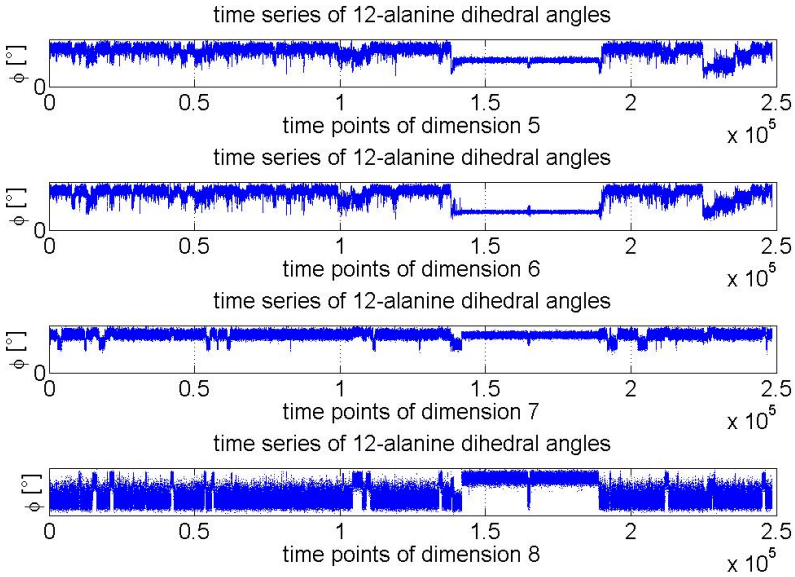


Figure: Selected dihedral angels of 12-alanine obtained from MD-simulations.

Curse of dimensionality: Due to the inherent sparsity of high-dimensional data statistical analysis is typically unreliable and prohibitively time consuming.

General Picture of Dimension Reduction for Biomolecules

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

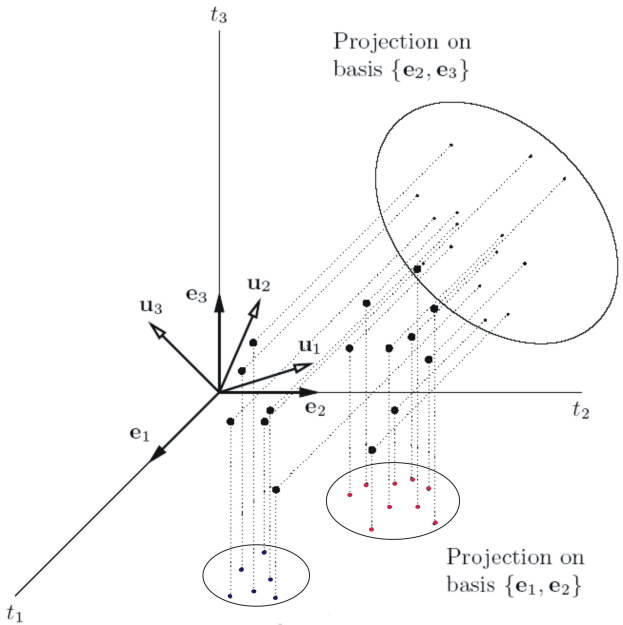


Figure: Aim: find a linear combination of dihedrals s.t. the rare folding events can be observed in a low dimensional subspace.

Unsupervised Feature Extraction Using Projections

Data $X_1, \ldots, X_n \in \mathbb{R}^d$ i.i.d., d large. For simplicity let $\boldsymbol{E}[X_i] = 0$ for all i.

Basic Observation: High dimensional data tends to be normal.

Problem: a random projection $X^\top \omega$ is almost approximately normal for most of the arbitrary directions $\omega \in \mathcal{B}_d$, where $\mathcal{B}_d$ is the d-dimensional unit ball.

Approach: Gaussian component of the data is entropy-maximizing and hence uninformative (noise). Project the data on the non-Gaussian components.

Requirements:

i) No apriori knowledge about the data density is used.

ii) No dependency on the magnitude of second moments of Gaussian and non-Gaussian components as found e.g. in PCA.

iii) No unrealistic assumptions on the whole data density as found e.g. in ICA.

The Semi-Parametric Model

Let $X_1, \ldots, X_N \in \mathbb{R}^d$ be i.i.d. random observable, distributed according to the structured and stationary density

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

This links pure Gaussian Analysis (PCA) and pure NonGaussian Analysis (ICA).

$q : \mathbb{R}^m \to \mathbb{R}$, $m \leq d$ is a smooth nonlinear function.
$T : \mathbb{R}^d \to \mathbb{R}^m$ is a linear operator with $\mathcal{I} = Ker(T)^{\perp}$.
$\mathcal{I}$ is the linear subspace of the non-Gaussian components.

goal: Estimate a projector without estimating the model parameter q and covariance matrix Σ.

interpretation: (1) lead to the stationary data model $X = Z + \zeta$ where ζ represents independent Gaussian noise components and Z the signal.

Estimation Procedure

Lemma

Assume that $\rho(x)$ is the structured density according to (1) with $\mu = 0$. If $\psi(x) \in \mathcal{C}^1(\mathbb{R}^d, \mathbb{R})$ has the property

$$\boldsymbol{E}\Big[x\psi(x)\Big] = 0 \tag{2}$$

then one can show that

$$\beta(\psi) = \boldsymbol{E}\Big[\nabla\psi(x)\Big] \quad \in \quad \mathcal{I} \tag{3}$$

Moreover, if (2) is not fulfilled, then there exists a vector $\beta \in \mathcal{I}$ s.t.

$$\|\beta - \beta(\psi)\|_2 \leq \|\Sigma^{-1} \int (x-\mu)\psi(x)\rho(x)\, dx\|_2 = \epsilon \tag{4}$$

i.e. dist$(\beta(\psi), \mathcal{I})$ is uniformly bounded.

Algorithmic Realization of the Lemma

idea: Compute $\psi(x)$ from the data using the linear approach:

$$\psi_{h,c}(x) = \sum_{l=1}^{L} c_l h_\omega(x) \tag{5}$$

Let N be the sample size. If we find coefficients $\{c_l\}_{l=1}^{L}$ such that

$$\boldsymbol{E}\Big[x\psi_{h,c}(x)\Big] \approx \frac{1}{N}\sum_{n=1}^{N} X_n\psi_{h,c}(X_n) = \frac{1}{N}\sum_{n=1}^{N}\sum_{l=1}^{L} c_l X_n h_{\omega_l}(X_n) = 0$$

it follows that $\beta \in \mathcal{I}$ with

$$\beta = \boldsymbol{E}\Big[\nabla\psi_{h,c}(x)\Big] \approx \frac{1}{N}\sum_{n=1}^{N} \nabla\psi_{h,c}(X_n) = \frac{1}{N}\sum_{n=1}^{N}\sum_{l=1}^{L} c_l \nabla h_{\omega_l}(X_n)$$

By the right choice "test functions" $h_\omega(x) \in \mathcal{C}^{1,1}(\mathbb{R}^d \times \mathbb{R}^d, \mathbb{R})$ are informative with respect to non-Gaussianity.

Algorithmic Realization of the Lemma cont'd

Remaining tasks:

a) Sampling of the data space using an appropriate function $h_\omega(x)$.

b) Find "good" coefficients $\{c_l\}_{l=1}^{L}$ with low computational effort.

c) Construct an ONB for the estimated target space $\widehat{\mathcal{I}}$.

d) Determine the reduced dimension m.

Note that the use of the semi-parametric framework combined with the Lemma is not unique:

(A) iterative approach: utilize $\{\widehat{\beta}_j^{(k)}\}_{j=1}^{J}$ for recovering a sequence of target spaces $\widehat{\mathcal{I}}^{(k)}$.

(B) non-iterative approach: direct estimation of the projector Π onto the target space $\widehat{\mathcal{I}}$.

Directional Sampling (both approaches)

Consider the functions of the form

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

with a smooth function h and a vector $\omega \in \mathcal{B}_d$, where $\mathcal{B}_d$ denotes the unit ball in $\mathbb{R}^d$. Define also

$$\begin{array}{lllllll} \widehat{\gamma}_\omega & := & N^{-1}\sum_i X_i h_\omega(X_i) & \approx & \gamma_\omega & := & \boldsymbol{E}\big[X h_\omega(X)\big] \\ \widehat{\eta}_\omega & := & N^{-1}\sum_i \nabla h_\omega(X_i) & \approx & \eta_\omega & := & \boldsymbol{E}\big[\nabla h_\omega(X)\big]. \end{array}$$

Then for the estimation accuracy it holds

Theorem

Let h_ω be bounded and continuously differentiable. Then there is $C = C(h)$ s.t.

$$\boldsymbol{E} \sup_{\omega \in \mathcal{B}_d} \left|\widehat{\gamma}_\omega - \gamma_\omega\right|^2 + \left|\widehat{\eta}_\omega - \eta_\omega\right|^2 \le C N^{-1} d^2 =: \epsilon^2.$$

Iterative "Convex Projection" Approach

Idea: For a given set $\{\omega_1, \ldots, \omega_L\}$ construct ψ as convex combinations of the $h_{\omega_\ell}(\cdot)$: $\psi(\cdot) = \sum_\ell c_\ell h_{\omega_\ell}(\cdot)$.

Convex optimization: given an arbitrary probe vector $\xi \in \mathcal{B}_d$, solve the non-smooth, convex problem

$$\{\widehat{c}_\ell\} = \underset{\|c\|_1 \le 1}{\arg\min} \left\| \xi - \sum_\ell c_\ell \widehat{\eta}_{\omega_\ell} \right\|_2^2 \quad \text{subject to} \quad \sum_\ell c_\ell \widehat{\gamma}_{\omega_\ell} = 0.$$

Then define an estimator $\widehat{\beta}$ of $\beta \in \mathcal{I}$ as

$$\widehat{\beta} \stackrel{\text{def}}{=} \sum_\ell \widehat{c}_\ell \widehat{\eta}_{\omega_\ell}.$$

and utilize $\{\widehat{\beta}_j\}_{j=1}^J$ for recovering the m-dimensional non-Gaussian target space $\mathcal{I}$.

Accuracy of the "Convex Projection"-Approach

"Ideal" vs. empirical projection:

$$\begin{aligned} \{c_\ell^*\} &= \underset{\|c\|_1 \le 1}{\arg\min} \left\| \xi - \textstyle\sum_\ell c_\ell \eta_{\omega_\ell} \right\|_2 \quad \text{s.t.} \quad \textstyle\sum_\ell c_\ell \gamma_{\omega_\ell} = 0, \\ \{\widehat{c}_\ell\} &= \underset{\|c\|_1 \le 1}{\arg\min} \left\| \xi - \textstyle\sum_\ell c_\ell \widehat{\eta}_{\omega_\ell} \right\|_2 \quad \text{s.t.} \quad \| \textstyle\sum_\ell c_\ell \widehat{\gamma}_{\omega_\ell} \| \le \epsilon \end{aligned}$$

and define:

$$\beta^* = \sum_\ell c_\ell^* \eta_{\omega_\ell} \qquad \widehat{\beta} = \sum_\ell \widehat{c}_\ell \widehat{\eta}_{\omega_\ell} \tag{6}$$

The the "convex projection"-approach is associated with the accuracy result:

Theorem

Let $h_\omega(x) \in \mathcal{C}^{1,1}$ have bounded variance in both arguments and let $\widehat{\beta}$ be defined as in (6). Then there is a set A of probability at least $1-\epsilon$, that

$$\left\| (I - \Pi^*)\widehat{\beta} \right\|_2 \le \sqrt{d}\, \delta_N \left(1 + \|\Sigma^{-1}\|_2\right),$$

where $\delta_N = \mathcal{O}(N^{-1}d)$.

Translation to Reduced Rank Regression Problem

Let the vectors $\widehat{\beta}_1, \ldots, \widehat{\beta}_L$ be given s.t.

$$\|(I - \Pi)\widehat{\beta}_j\|_2 \leq \epsilon$$

where Π is a projector on a m-dimensional target space.

Reduced Rank Regression problem: for given m, recover Π.

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

First guess to RRR: use PCA

$$\widehat{\mathcal{I}} = \underset{dim(\mathcal{I})=m}{\arg\min} \sum_j \|(I - \Pi)\widehat{\beta}_j\|^2 = \langle \text{first } m \text{ eigenvectors of } \sum_j \widehat{\beta}_j \widehat{\beta}_j^\top \rangle.$$

However it turns out numerically that this works poorly if most of the $\widehat{\beta}_j$'s are non-informative.

Reduced Rank Regression using Rounding Ellipsoids

Next guess: use the rounding ellipsoid of the symmetrized convex set

$$\mathcal{S} \stackrel{\text{def}}{=} \langle \widehat{\beta}_1, -\widehat{\beta}_1, \widehat{\beta}_2, -\widehat{\beta}_2, \ldots \rangle.$$

$\mathcal{E}(B) \equiv \mathcal{E}_1(B)$ is α-*rounding* ellipsoid for $\mathcal{S}$ if

$$\mathcal{E}_{1/\alpha}(B) \subseteq \mathcal{S} \subseteq \mathcal{E}(B), \qquad \alpha \leq 1,$$

where $\mathcal{E}_r(B) \stackrel{\text{def}}{=} \{x \in \mathbb{R}^d | \; x^\top B x \leq r^2\}$, .

Theorem (F. John, 1985; Nesterov, 2004)

For any convex $\mathcal{S} \subset \mathbb{R}^d$, there exists a rounding ellipsoid with $\alpha = d^{-1/2}$.

Advantage: To recover $\mathcal{I}$ compute the principal axis of $\mathcal{E}(\mathcal{B})$ with complexity $\mathcal{O}(d^2 J \log J)$ and select some of them according to a criterion of multimodality.

Accuracy of the "Rounding Ellipsoid" Solution

Theorem

1. *For any unit vector* $v \perp \mathcal{I}$,

$$v^\top B^{-1} v \leq \delta^2.$$

2. *If there is* $w \in \mathbb{R}^J$ *with* $w_j \geq 0$ *and* $\sum_j w_j = 1$ *such that*

$$\lambda_m \left(\sum_j w_j \beta_j \beta_j^\top \right) > 2\delta^2,$$

and $\widehat{\Pi}$ *projects on the* m *principal eigenvectors of* B^{-1}, *then*

$$\|\widehat{\Pi} - \Pi^*\|_2^2 \leq C(\delta^2)\mathcal{O}(d\sqrt{d}).$$

Iteration allows for Structural Adaptation

Use the estimated ellipsoid $\mathcal{E}_{k-1}$ as a prior information to improve the quality of estimation.

This leads to sequential procedure: alternate two steps

i) estimate the model vector β_j using a given structure

ii) estimate the structure, i.e. the rounding ellipsoid $\mathcal{E}$

Method: sample some of the probe vectors ξ_j and some vectors $\omega_{\ell,j}$ due to identified semi-axis 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.

Computability in High Dimensions

The iterative approach leads to one quadratic, constrained optimization problem (QCP) for each $\beta_j \in \mathcal{T}$.

However about fast interior-point-methods (IPM) to high accuracy we know:

1) Assembling and solving a $L \times L$ Newton system of linear equations takes $\mathcal{O}(L^3)$ operations unless the matrix of the system is highly sparse with favourable patterns.
2) SNGCA leads to optimization problems with dense Newton systems.

In the context of the "convex projection"-approach $\mathcal{O}(JLN^2 + (16L)^3)$ operations are needed for the k^{th} iteration of SNGCA.

"Semidefinite Programming"-Approach

Some notations: let

i) $G \in \mathbb{R}^{d \times L}$ be a matrix of averaged gradients of test functions h_ω with columns η_l

ii) $U \in \mathbb{R}^{d \times L}$ a matrix of averaged functions xh_ω with columns γ_l.

and let $\widehat{G} \in \mathbb{R}^{d \times L}$ and $\widehat{U} \in \mathbb{R}^{d \times L}$ from the data counterparts respectively s.t.

$$\|G - \widehat{G}\|_2 \leq \epsilon \quad \text{and} \quad \|U - \widehat{U}\|_2 \leq \epsilon.$$

Then solve the non-convex, non-smooth contrained problem

$$\min_{\Pi} \max_{c} \left\{ \|(I - \Pi)\widehat{U}c\|_2^2 \;\middle|\; \begin{array}{c} 0 \preceq \Pi \preceq I,\ \mathrm{Tr}[\Pi] = m,\ rank\Pi = m; \\ c \in \mathbb{R}^L,\ \|c\|_1 \leq 1,\ \|\widehat{G}c\|_2 \leq \delta \end{array} \right\}. \tag{7}$$

Recipe of Semidefinite Relaxation

idea: drop constraints to get convexity and than solve.

i) Use the identity:

$$\|(I-\Pi)\widehat{U}c\|_2^2 = \mathrm{Tr}\left[\widehat{U}(I-\Pi)\widehat{U}X\right]. \tag{8}$$

ii) Linearization: consider the positive semidefinite matrix $X = cc^T$ with $rankX = 1$ as "new variable".

iii) Set $|X|_1 \stackrel{\text{def}}{=} \sum_{i,j=1}^{L} |X_{ij}|$ and transform $\|\widehat{G}c\|_2 \leq \delta$ into $\mathrm{Tr}[\widehat{G}X\widehat{G}] \leq \delta^2$.

iv) Drop the non-convex constraints $rankX = 1$ and $rank\Pi = m$.

Then we arrive at the relaxed semidefinite constrained problem:

$$\min_P \max_X \left\{ \mathrm{Tr}\left[\widehat{U}(I-P)\widehat{U}X\right] \;\middle|\; \begin{array}{c} 0 \preceq P \preceq I,\ \mathrm{Tr}[P] = m, \\ X \succeq 0,\ |X|_1 \leq 1,\ \mathrm{Tr}[\widehat{G}X\widehat{G}] \leq \delta^2 \end{array} \right\}. \tag{9}$$

Bounds for Relaxation Error

Theorem

Suppose that the projector Π^* is μ^* times a convex combination of rank-one matrices $Ucc^T U^T$ where c satisfies the constraints $Gc = 0$ and $\|c\|_1 \leq 1$, i.e.

$$\Pi^* \preceq \sum_{k=1}^{\overline{m}} \mu^k U c_k c_k^T U^T. \tag{10}$$

Then an optimal solution $\widehat{P}$ of the relaxed problem satisfies

$$\mathrm{Tr}\left[(I - \widehat{P})\Pi^*\right] \leq 4\mu^* \epsilon^2 (\lambda_{\min}^{-1}(\Sigma) + 1)^2. \tag{11}$$

Further, if $\widehat{\Pi}$ is the projector onto the subspace spanned by m principal eigenvectors of $\widehat{P}$, then

$$\|\widehat{\Pi} - \Pi^*\|_2^2 \leq \frac{8\mu^* \epsilon^2 (\lambda_{\min}^{-1}(\Sigma) + 1)^2}{1 - 4\mu^* \epsilon^2 (\lambda_{\min}^{-1}(\Sigma) + 1)^2} \tag{12}$$

How to Include the Constraints

Observe that $\widehat{G}^T\widehat{G} = \Gamma\Lambda\Gamma^T$ and X are symmetric and positive. Hence:

$$Tr(\widehat{G}^T\widehat{G}X) = 0 \quad \Rightarrow \quad X = QZQ^T \tag{13}$$

where $Z \in \mathcal{S}^{L-d}$ and $Q \in \mathcal{S}^{L\times(L-d)}$ is a submatrix of columns of Γ corresponding to the vanishing eigenvalues of $\widehat{G}^T\widehat{G}$.

Let $V = \widehat{G}Q$. Than we get a regularized and hence unconstrained convex reformulation of the relaxed problem:

$$\min_{\Pi,W}\left[\max_{Z\in\mathcal{Z},Y} \mathrm{Tr}[V^T(I-\Pi_{\widehat{\mathcal{I}}})VZ] + \mathrm{Tr}[W(QZQ^T - Y)]\right] \tag{14}$$

where $Z \in \mathcal{Z}$ and $\mathcal{Z} := \{Z \in \mathcal{S}_{L-d} \mid Z \succeq 0, Tr(Z) \leq 1\}$.

The latter problem can be solved using a gradient-type method with complexity $\mathcal{O}(d\log d)$ and $\mathcal{O}(\epsilon^{-1})$ iterations (Nesterov 2007).

Non-gaussian Components of Test Densities

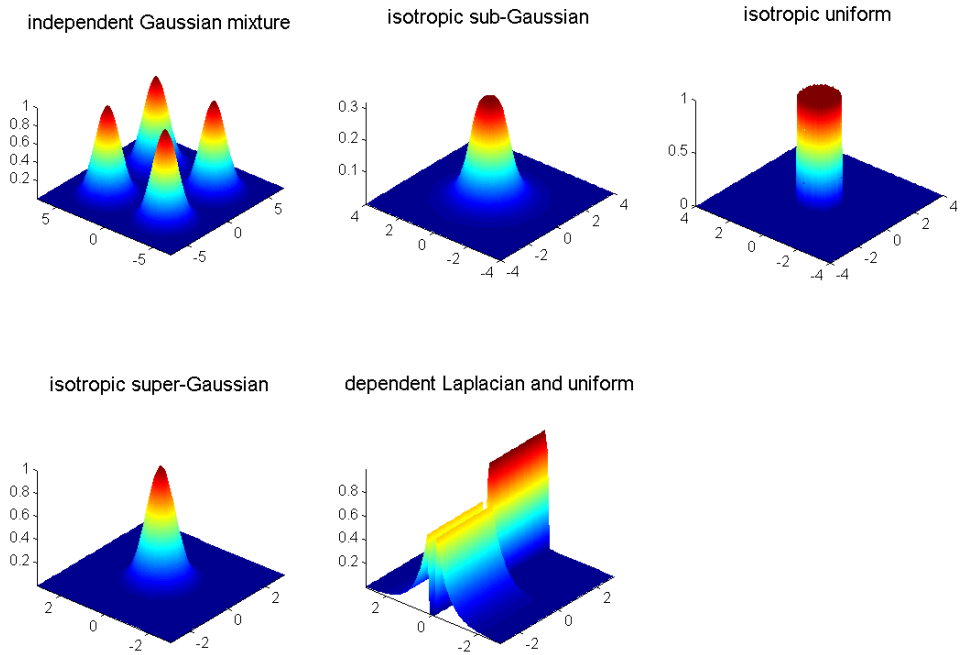


Figure: (A) 2 d independent Gaussian mixtures, (B) 2 d isotropic super-Gaussian, (C) 2 d isotropic uniform and (D) dependent 1 d Laplacian with additive 1 d uniform with $N = 1000$ respectively.

One Step Improvement of the Iterative Approach

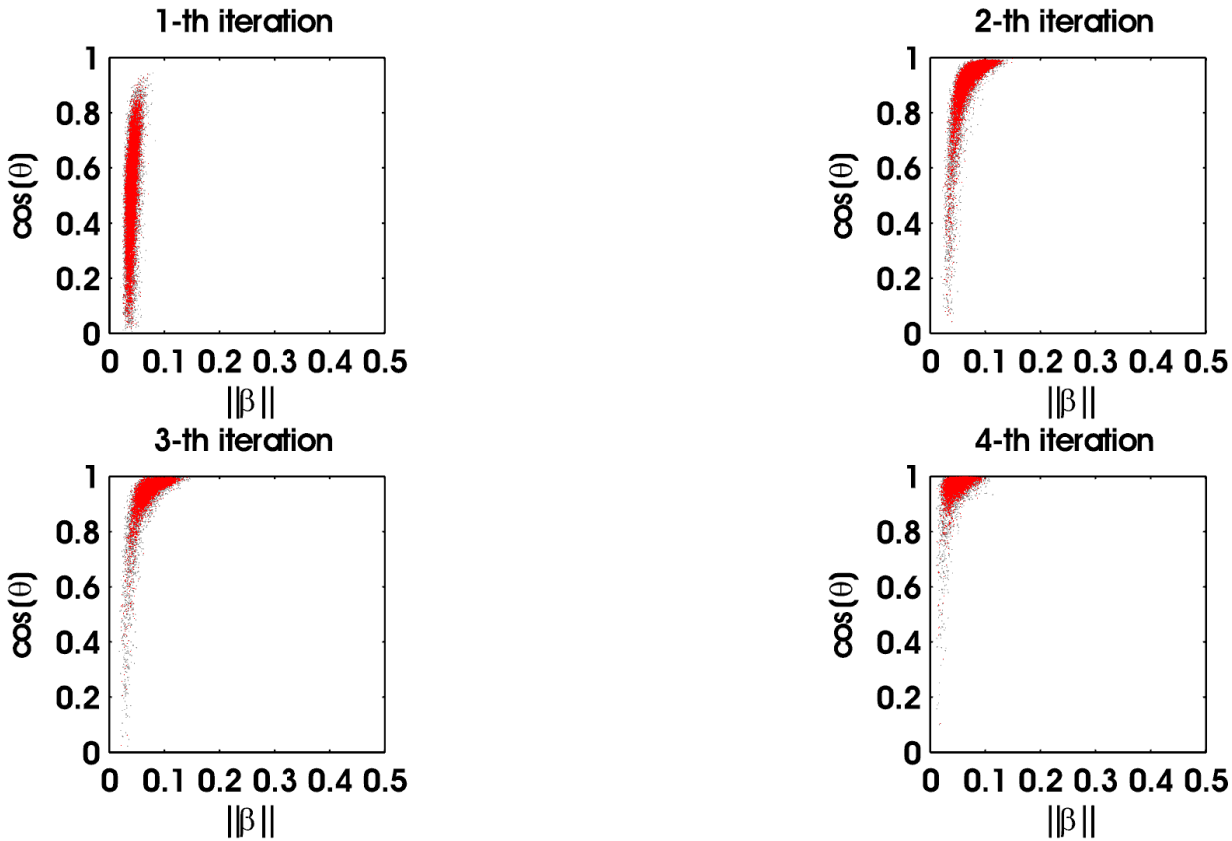


Figure: Sub-Gaussian density with 2 components in $\mathbb{R}^{20}$

Error Criterion

The closeness of the subspaces $\mathcal{I}$ and its estimate $\widehat{\mathcal{I}}$ can be measured by the error function

$$\mathcal{E}(\widehat{\mathcal{I}}, \mathcal{I}) \stackrel{\text{def}}{=} \frac{1}{m} \sum_{i=1}^{m} \|(I - \Pi) v_i\|^2 \tag{15}$$

where Π denotes the orthogonal projection onto $\widehat{\mathcal{I}}$, $\{v_i\}_{i=1}^m$ is an orthonormal basis of $\widehat{\mathcal{I}}$ and I denotes the identity matrix.

Comparison dimension

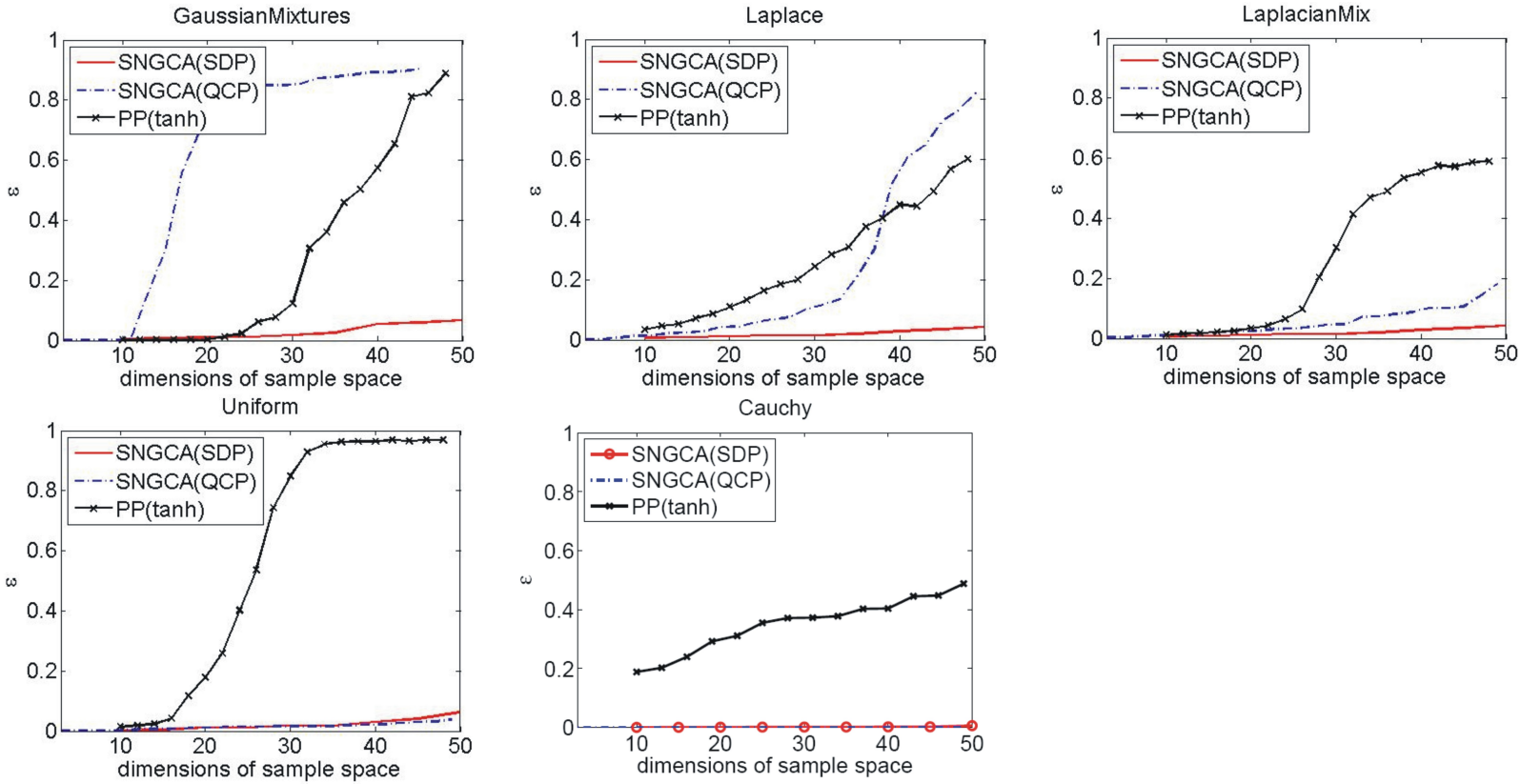


Figure: Comparison of PP, iterative and non-iterative SNGCA by estimation error for increasing dimensionality .

Comparison noise

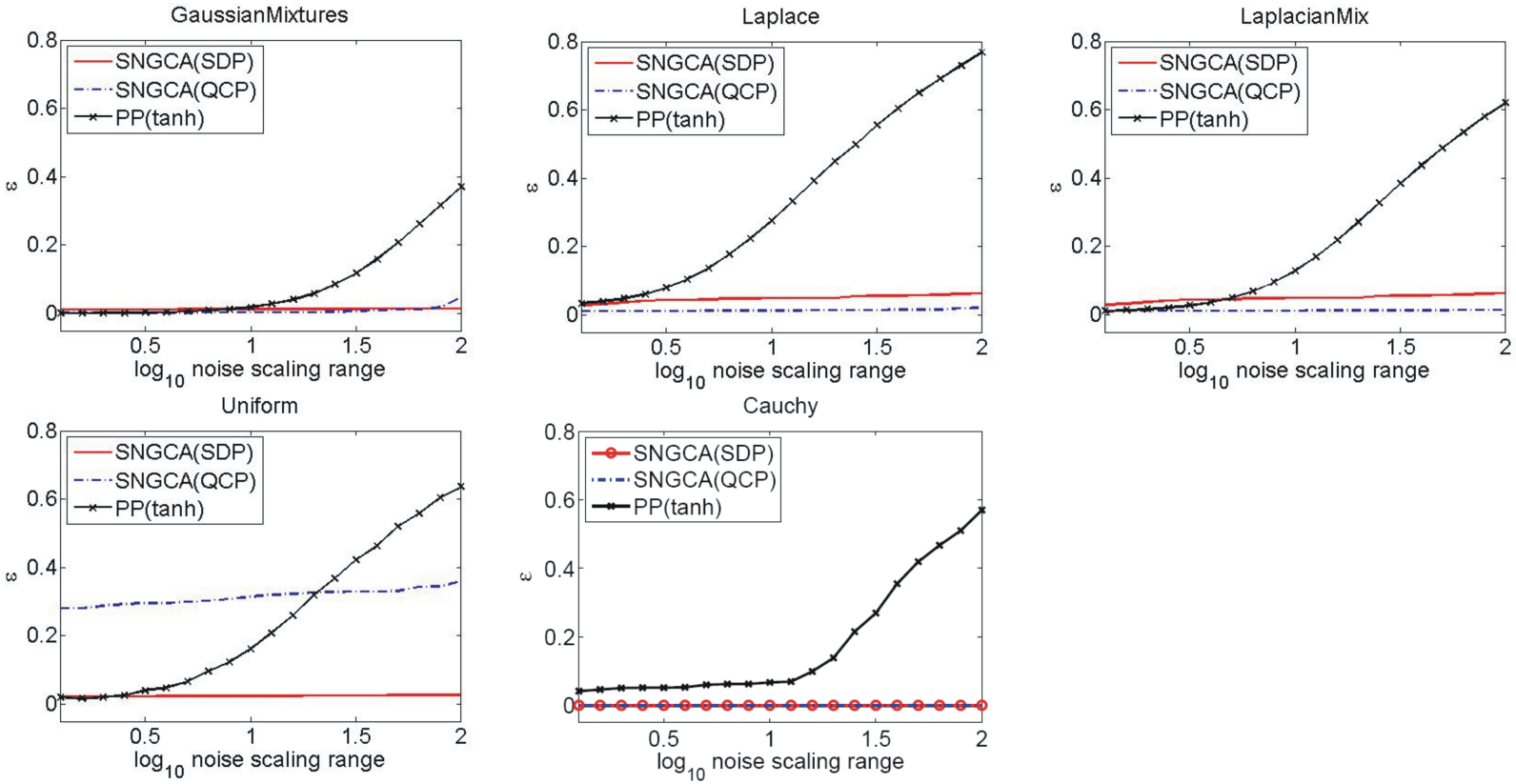


Figure: Comparison of PP, iterative and non-iterative SNGCA for increasing numerical condition for Σ^{-1} .

Application to Protein Study

The molecule was simulated using CHARMM with an implicit water environment at $300K$. We analyzed a 1ìs long simulation with 2fs time steps observing the 33 backbone torsion angles.

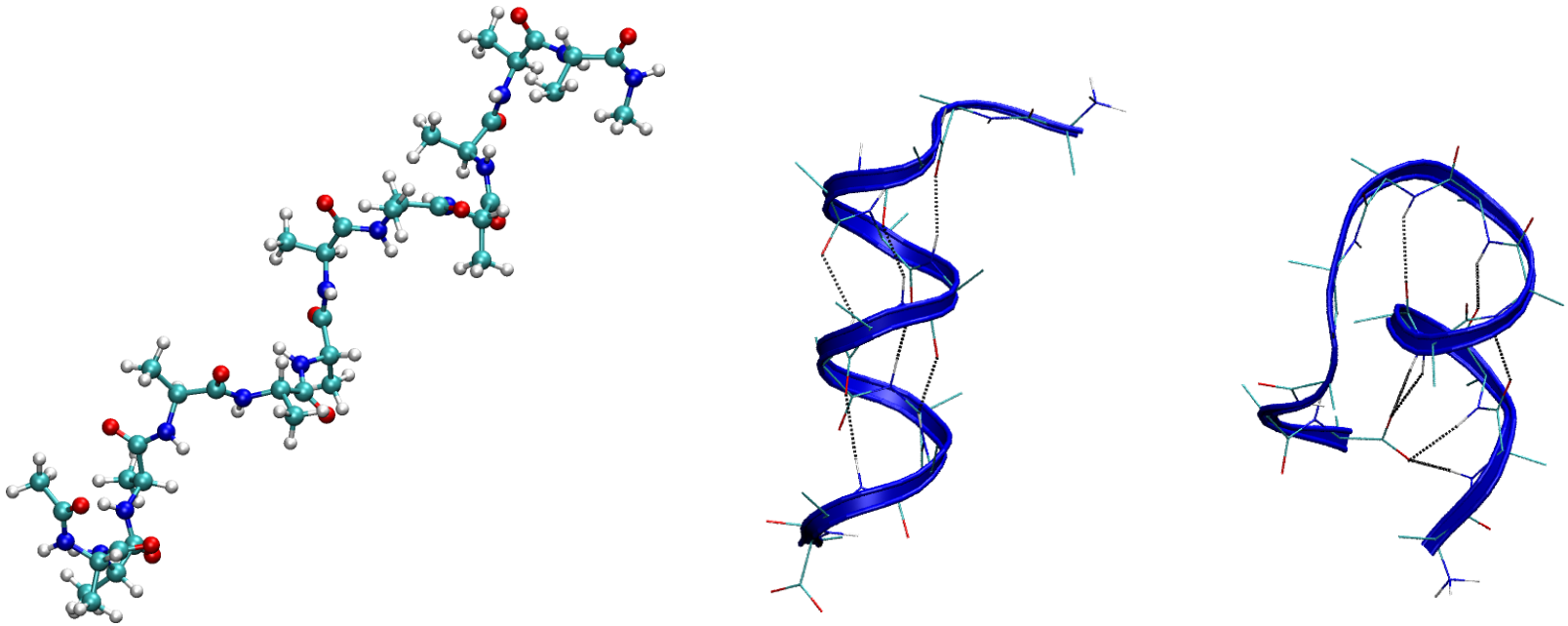
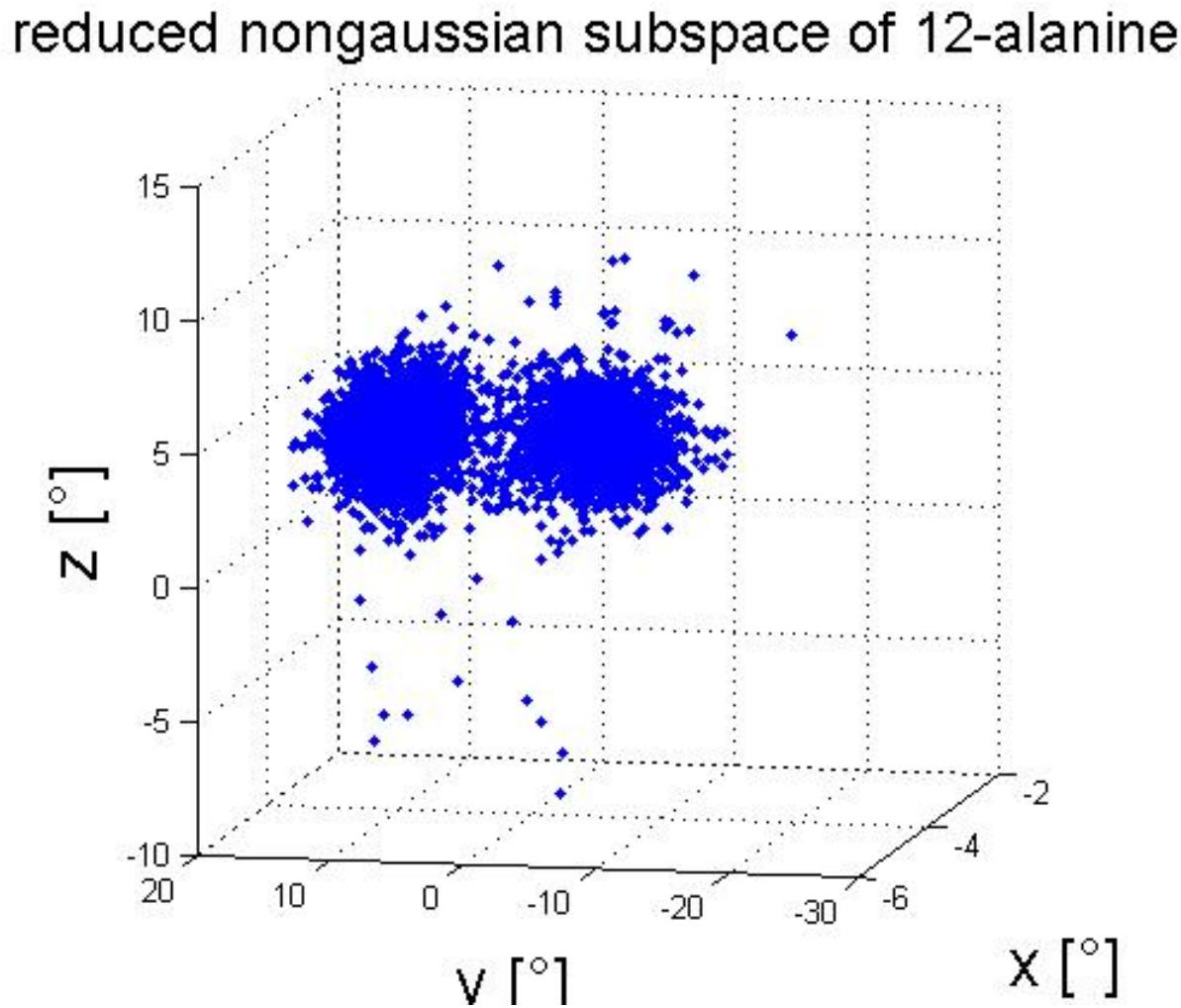


Figure: most probable conformations of 12-alanine, α-helix and β-sheet

SNGCA-result of 12-alanine



Summary

1) Structural data analysis based on the non-Gaussian vs. Gaussian distinction is effective and computational not too expansive.

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

3) The stochastic reduction of dimensionality works also with stochastic dynamical systems like large biomolecules.

Summary

1) Structural data analysis based on the non-Gaussian vs. Gaussian distinction is effective and computational not too expansive.

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

3) The stochastic reduction of dimensionality works also with stochastic dynamical systems like large biomolecules.

Summary

1) Structural data analysis based on the non-Gaussian vs. Gaussian distinction is effective and computational not too expansive.

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

3) The stochastic reduction of dimensionality works also with stochastic dynamical systems like large biomolecules.

Final Slide

Thank you for your attention!