The Development of NonGaussian Component Analysis

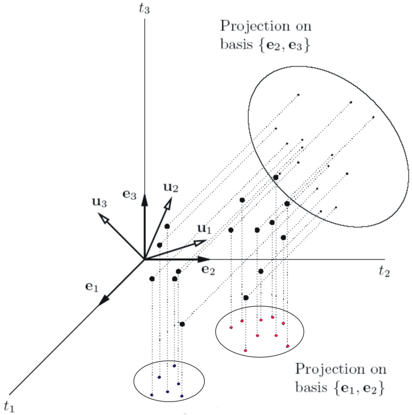
Elmar Diederichs[2]

joint work with Anatoli Juditsky[3] and Vladimir Spokoiny[1]

[1]Weierstrass-Institute and Humbolt University, Berlin
[2]Moscow Institute for Physics and Technology, Moscow
[3]U. Joseph Fourier, Grenoble

International Workshop on Statistical Learning
Moscow, 26.-27.05.2013

Structural Analysis



Design of Unsupervised Feature Extraction

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.

Problem: most of all random projections $X^\top \omega$ are almost approximately normal

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

Requirements for an acceptable statistical method:

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 or unrealistic assumptions on the data density as found e.g. in ICA.

The Semi-Parametric Model

Semiparametric structural assumption:

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

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

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

goal: Estimate a projector without estimating the model parameter.

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

Recovery of Target Space

Lemma

Assume that $\rho(x)$ follows the semiparametric assumption with $\boldsymbol{E}[X] = 0$. Then for $\psi(x) \in \mathcal{C}^1(\mathbb{R}^d, \mathbb{R})$ and

$$\beta(\psi) \stackrel{\text{def}}{=} \boldsymbol{E}\left[\nabla\psi(x)\right] \tag{2}$$

there is $\beta \in \mathcal{I}$ such that there is a uniform error bound

$$\|\beta(\psi) - \beta\|_2 \leq \left\|\Sigma^{-1}\boldsymbol{E}\left[x\psi(x)\right]\right\|_2 \tag{3}$$

of dist$(\beta(\psi), \mathcal{I})$. Moreover if $\boldsymbol{E}\left[x\psi(x)\right] = 0$, then $\beta(\psi) \in \mathcal{I}$.

Unsupervised Feature Extraction Using Projections

1) linear approach: $\psi(x) = \sum_\ell^L c_\ell h_\ell(x)$
2) test functions: $h_\ell(x) \stackrel{\text{def}}{=} h(\omega_\ell^\top x) e^{-\lambda \|x\|^2/2}$
3) define:

$$\gamma_\ell \stackrel{\text{def}}{=} \boldsymbol{E}[X h_\ell(X)], \;\; \eta_\ell \stackrel{\text{def}}{=} \boldsymbol{E}[\nabla h_\ell(X)],$$

and let $\widehat{\gamma}_\ell$ and $\widehat{\eta}_\ell$ be their "empirical counterparts" that can be computed, such that for a set A of probability at least $1-\epsilon$ it holds $\max_{|c|_1 \le 1} \sum_\ell |\widehat{\eta}_\ell - \eta_\ell|_2 \le \delta_N$ and $\max_{|c|_1 \le 1} \sum_\ell |\widehat{\gamma}_\ell - \gamma_\ell|_2 \le \nu_N$, .

Lemma

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

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

NonGaussian Component Analysis

Approach of Blanchard et al.(JMLR, 2007):

Consider $\psi(x) = h(x) - \alpha^T x$ and select α s.t.

$$\boldsymbol{E}\left[X\psi(X)\right] = \boldsymbol{E}\left[Xh(X)\right] - \boldsymbol{E}XX^T\alpha = 0.$$

in order to suppress the noise when estimating elements from $\mathcal{I}$.

Then $\beta(\psi) \stackrel{\text{def}}{=} \boldsymbol{E}[\nabla\psi(x)]$ leads to:

$$\widehat{\beta}_\ell = \widehat{\eta}_\ell - \widehat{\Sigma}^{-1}\widehat{\gamma}_\ell.$$

drawbacks: requires to compute and study $\widehat{\Sigma}^{-1}$, m cannot be estimated.

Sparse NonGaussian Component Analysis

Approach of Diederichs et al. (IEEE Trans. Inf. Theo, 2009):

Set $\psi(x) = \sum_{\ell}^{L} c_\ell h_\ell(x)$ and consider the convex projection problem

$$\widehat{c} = \arg\min_c \left\{ \|\xi - \sum_\ell c_\ell \widehat{\eta}_\ell\|_2 \;\middle|\; \sum_\ell c_\ell \widehat{\gamma}_\ell = 0,\ \|c\|_1 \le 1 \right\}$$

and define $\widehat{\beta}_\ell = \sum_\ell \widehat{c}_\ell \widehat{\eta}_\ell$. Then under some regularity conditions there is $\mathcal{C} \stackrel{\text{def}}{=} \{\|c\|_1 \le 1, \sum_\ell c_\ell \widehat{\gamma}_\ell\}$

$$\|(\mathbf{1}_d - \Pi^*)\widehat{\beta}\|_2 \le \sqrt{d}\, C(d, N^{-1/2})(1 + \|\Sigma^{-1}\|_2)$$

drawbacks: choice of informative probe vector ξ, m cannot be estimated

Both approaches require the solution of the Reduced Rank Regression problem: given m, recover $\mathcal{I}$ or the projector $\Pi_{\mathcal{I}}$ from $\widehat{\beta}_1, \ldots, \widehat{\beta}_L$.

Dimension Reduction Step: the RRR problem

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

$$\|(1_d - \Pi_{\mathcal{I}})\widehat{\beta}_\ell\| \leq \epsilon$$

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

PCA solution:

$$\widehat{\mathcal{I}} = \operatorname*{arg\,min}_{dim(\mathcal{I})=m} \sum_\ell \|(\mathbf{1}_d - \Pi_{\mathcal{I}})\widehat{\beta}_\ell\|^2 = \langle \text{first } m \text{ eigenvectors of } \sum_\ell \widehat{\beta}_\ell \widehat{\beta}_\ell^T \rangle.$$

Requires that $\lambda_m\left(\sum_\ell \beta_\ell \beta_\ell^T\right) \geq L\epsilon^2$. Works poorly if most of the $\widehat{\beta}_\ell$'s are non-informative.

Rounding Ellipsoid: (see Yu.Nesterov, 2004) Define the convex set

$$\mathcal{S} := \{\widehat{\beta}_1, -\widehat{\beta}_1, \widehat{\beta}_2, -\widehat{\beta}_2, \ldots\}.$$

For $\mathcal{S}$ a centered ellipsoid $\mathcal{E}$ of minimum volume that encloses $\mathcal{S}$ always exists and recovering $\mathcal{I}$ from the principal axis of $\mathcal{E}$ comes with the accuracy

$$\|\Pi_{\mathcal{I}} - \Pi_{\widehat{\mathcal{I}}}\|_2^2 \leq C(\lambda_m, \epsilon^2) d\sqrt{d}.$$

Numerical Illustration of Progress

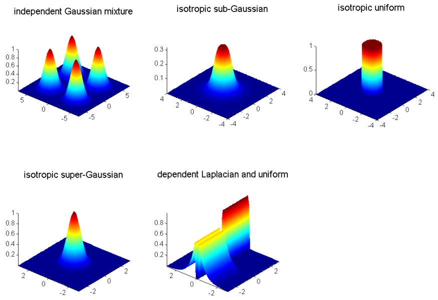


Figure : (A) NonGaussian test densities: 2 d independent Gaussian mixtures, (B) 2 d isotropic super-Gaussian, (C) 2 d isotropic uniform and (D) dependent 1 d Laplacian with additive 1 d uniform with $N = 1000$ respectively.

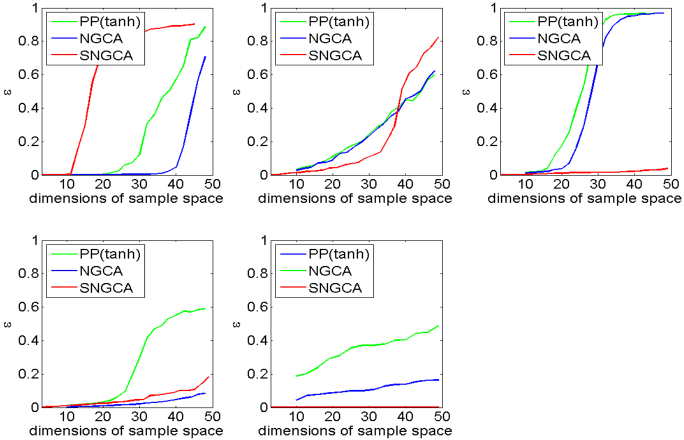
Error Measure

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

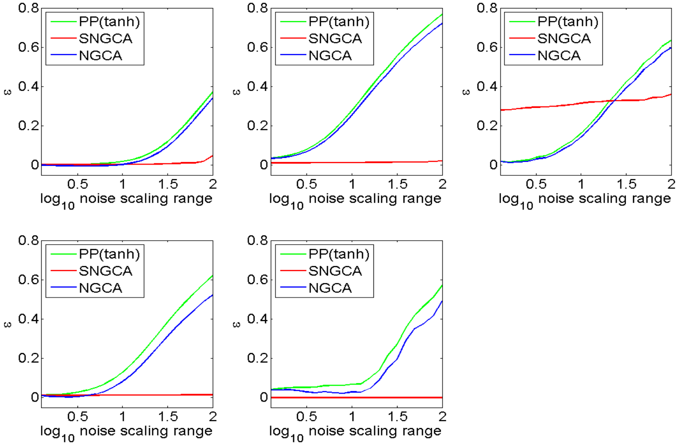
$$\mathcal{E}(\widehat{\mathcal{I}}, \mathcal{I}) = \frac{1}{m} \sum_{i=1}^{m} \|(\mathbf{1}_d - \Pi) v_i\|_2^2 \qquad (4)$$

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.

Estimation Error for Increasing Dimensionality



Comparison of Methods Cont'd: Noise



Design of Semidefinite Structural Analysis

first step of approach of Juditsky et al. (JML, 2012): Choose randomly a family of directions $\{\omega_\ell\}$, $\ell = 1, \ldots, L$ and compute

$$\widehat{\gamma}_\ell \stackrel{\text{def}}{=} \boldsymbol{E}_N[X h_\ell(X)], \qquad \widehat{\eta}_\ell \stackrel{\text{def}}{=} \boldsymbol{E}_N[\nabla h_\ell(X)]$$

second step: avoid any probe vectors and the RRR problem by solving the semidefinite problem

$$\Pi^* = \min_{\Pi} \max_{c} \left\{ \left\| (I - \Pi) U c \right\|_2^2 \;\middle|\; \begin{array}{c} \Pi \text{ is a projector on a} \\ m\text{-dimensional subspace of } \mathbb{R}^d \\ c \in \mathbb{R}^L,\ Gc = 0 \end{array} \right\}. \qquad (5)$$

where $U = [\eta_1, \ldots, \eta_L] \in \mathbb{R}^{d \times L}$, $G = [\gamma_1, \ldots, \gamma_L] \in \mathbb{R}^{d \times L}$ and Π^* is the Euclidean projector on $\mathcal{I}$.

Design of Semiparametric Structural Analysis cont'd

third step: structural adaptation idea (Hristache, Juditsky, Polzehl and Spokoiny, 2003):

> use the estimated projector $\widehat{\Pi}_{k-1}$ as a prior information for the directional sampling to improve the quality of estimation at iteration k of SD-NGCA.

This leads to a sequential procedure: alternate two steps

a) sample some directions ω_ℓ from the space spanned by the m principal directions of $\widehat{\Pi}_{k-1}$
b) estimate the structure $\widehat{\Pi}_k$

This ensures that a certain fraction of $\widehat{\gamma}_\ell$ and $\widehat{\eta}_\ell$ is informative.

Relaxation of the hard problem in SD-NGCA

idea: drop constraints to get convexity and solve an approximating problem

i) Use: $\|(I-\Pi)\widehat{U}c\|_2^2 = \mathrm{Tr}\left[\widehat{U}(I-\Pi)\widehat{U}cc^T\right]$.

ii) **Linearization**: 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 $\mathrm{Tr}[\widehat{G}X\widehat{G}] \leq \varrho^2$.

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

Then we arrive at the relaxed and constrained saddle point problem:

$$\min_P \max_X \left\{ \mathrm{Tr}\left[\widehat{U}(I-P)\widehat{U}X\right] \;\middle|\; \begin{array}{c} 0 \preceq P \preceq I,\ \mathrm{Tr}[P] = m, \\ X \succeq 0,\ |X|_1 \leq 1,\ \mathrm{Tr}[\widehat{G}X\widehat{G}] \leq \varrho^2 \end{array} \right\}. \qquad (6)$$

where $|X|_1 := \sum_{ij} |X_{ij}|$. We call P - defined as above - a subprojector.

Accuracy of the estimated projector

Lemma

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

i) Π^* *on* $\mathcal{I}$ *is a convex combination of rank-one matrices* Ucc^TU^T

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

Then it holds of $\widehat{\Pi}$*, spanned by the first* m *eigenvectors of* $\widehat{P}$*, with probability* $\geq (1-\epsilon)$*;*

$$
\begin{aligned}
\left\|(\mathbf{1}_d - \widehat{\Pi})Uc\right\|_2 &\leq C_1\sqrt{m+1}((\varrho+\nu_N)\lambda_{\min}^{-1}(\Sigma) + 2\delta_N) \\
\mathrm{Tr}[(\mathbf{1}_d - \widehat{\Pi})\Pi^*] &\leq C_2\left[(\varrho+\nu_N)\lambda_{\min}^{-1}(\Sigma) + 2\delta_N\right]^2 \\
\|\widehat{\Pi} - \Pi^*\|_{Frob}^2 &\leq C_3(m+1)\left[(\varrho+\nu_N)\lambda_{\min}^{-1}(\Sigma) + 2\delta_N\right]^2
\end{aligned}
$$

where $C_i = C_i(h)$ *with* $i = 1, 2, 3$ *does not depend on* d *or* L.

Linear Constraints

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

$$Tr(\widehat{G}^T\widehat{G}X) = 0 \quad \Rightarrow \quad X = QZQ^T \tag{7}$$

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 $\widehat{G}^T\widehat{G}$.

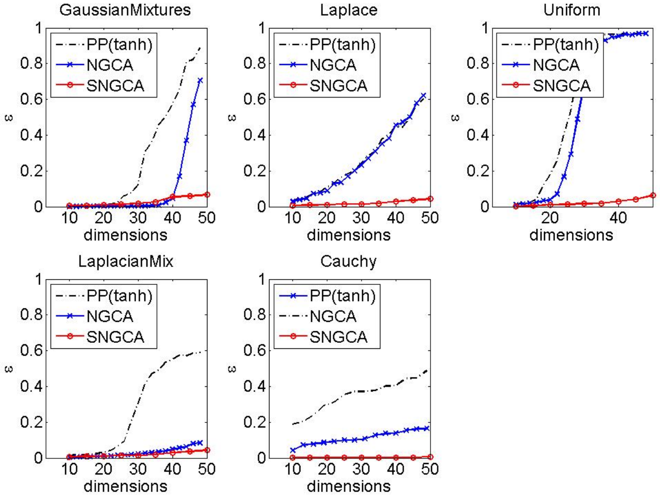
Let $V = \widehat{G}Q$. Than we get a regularized and hence unconstrained convex reformulation of the relaxed problem:

$$\min_{\Pi,W}\left[\max_{Z\in\mathcal{Z},Y} \mathrm{Tr}[V^T(I-\Pi_{\widehat{\mathcal{I}}})VZ] + \mathrm{Tr}[W(QZQ^T - Y)]\right] \tag{8}$$

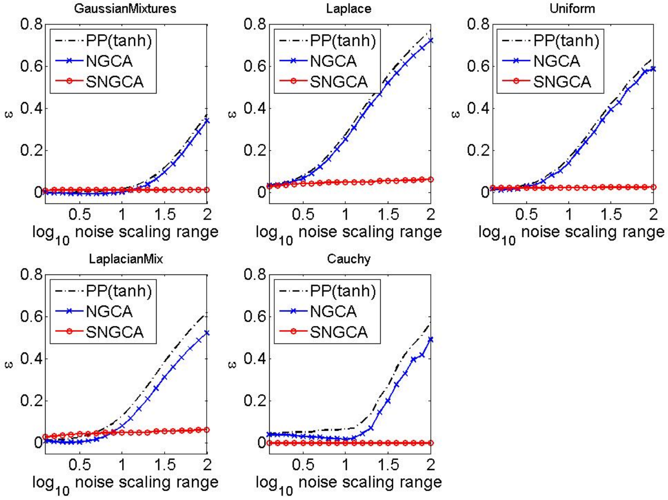
where $Z \in \mathcal{Z}$ and $\mathcal{Z} := \{Z \in \mathcal{S}_{L-d} \mid Z \succeq 0, Tr(Z) \leq 1\}$.

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

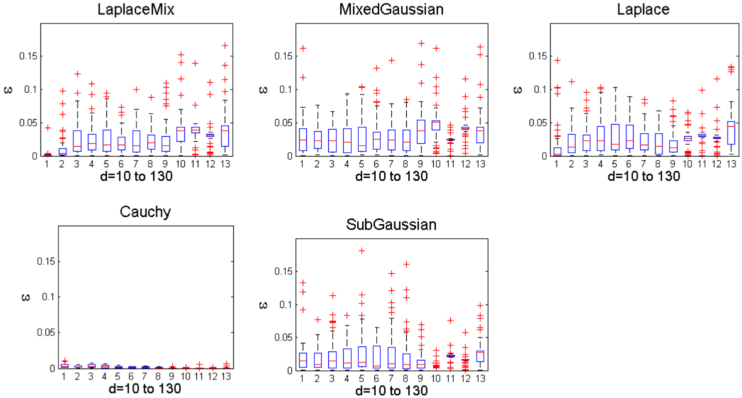
Estimation Error for Increasing Dimensionality



Comparison of Methods Cont'd: Noise



Numerical Performance



Final Slide

Thank you for your attention!

Dual Extrapolation Algorithm:

Consider the linear matrix game

$$\min_{x\in\mathcal{A}} \max_{y\in\mathcal{B}} \langle x, Ay\rangle + \langle a, x\rangle + \langle b, y\rangle. \tag{9}$$

with $\mathcal{A} \subset \mathbb{E}^n$ and $\mathcal{B} \subset \mathbb{E}^m$ be closed and convex sets.

motivation: size $L^2 \sim 10^6$ of the variable X rules out the possibility of using state-of-the-art interior point methods

idea: use a subgradient method to get low complexity

Scheme of Dual Extrapolation Algorithm

a) vector field of descend-ascend directions: $F(z) = (-A^T y - a, Ax + b)$

b) $z = (x, y)$, $z_k, z_k^+ \in \mathcal{A} \times \mathcal{B}$ and $s_k \in E^*$ at the k-th iteration

c) minimizer of a distance-generating function over $\mathcal{A} \times \mathcal{B}$: $\bar{z}$

update:

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

where $\lambda_k > 0$ is the current (adaptively chosen) stepsize.

approximate solution:

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

Choice of Prox-Transform

a) **distance-generating function** $d(z) = d_x(x) + d_y(y)$:

α-strongly convex and differentiable on $\mathcal{A} \times \mathcal{B}$

b) **prox-function** V on $\mathcal{A} \times \mathcal{B}$ in $z_0 = (x_0, y_0)$:

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

c) **prox-transform** $T(z_0, s)$ of $s = (s_x, s_y)$

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