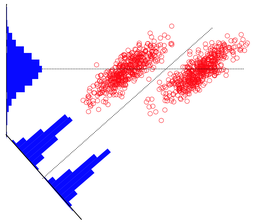
Non-Gaussian Component Analysis: a Semiparametric Framework for Linear Dimension Reduction

G.Blanchard[1], M.Sugiyama[1,2], M.Kawanabe[1], V.Spokoiny[3], K.-R.Müller[1,4]

[1] Fraunhofer FIRST.IDA, Germany
[2] Dept. of Computer Science, Tokyo Institute of Technology, Japan
[3] Weierstrass Institute and Humboldt University, Berlin, Germany
[4] Dept. of Computer Science, University of Potsdam, Germany

Introduction

- We want to perform dimensionality reduction on high-dimensional data.
- Manifold learning and other non-linear methods generally assume that the data is concentrated around a lower-dimensional structure (manifold)
- We consider here a different problem: assume the data can be decomposed into different linear components:
  - A purely Gaussian, "uninteresting", part
  - A non-Gaussian, "interesting", component



- Important: we do not assume that the Gaussian component is of lower order. This excludes manifold learning methods.
- Our goal is to recover the non-Gaussian component.

The Projection Pursuit method

(Friedman & Tukey, 1975)
Principle: find a direction $w \in \mathbb{R}^d$ maximaxing a "non-Gaussianity" measure:

$$\max_{\|w\|=1} \left|\mathbb{E}\left[G(\langle w,x\rangle)\right] - E_\nu\left[G(\nu)\right]\right|,$$

where $\nu \sim \mathcal{N}(0,1)$.
Popular choices for G:

$$G_1(\eta) = \eta^4 \qquad G_2(\eta) = b^{-1}\log\cosh(b\eta)\,.$$

The FastICA algorithm

FastICA (Hyvärinen & Oja 1997): efficient instantiation of Projection Pursuit. After data whitening the following equations are iterated:

$$(i) \quad w \leftarrow \mathbb{E}\left[xg(\langle w,x\rangle)\right] - \mathbb{E}\left[g'(\langle w,x\rangle)\right] w$$
$$(ii) \quad w \leftarrow w/\left\|w\right\|$$

where $g \propto G'$, G being the Projection Pursuit criterion, e.g.,

$$g_1(x) = x^3 \qquad g_2(x) = \tanh(bx)\,.$$

▶ Drawbacks: g_1 is appropriate to find sub-Gaussian (light-tailed) distributions, g_2 to find super-Gaussian (heavy-tailed) distributions.
However, one does not always know *a priori* which index is the more appropriate. The choice of the parameter b also affects performance.

▶ Interesting goal: to be able to combine information from different non-Gaussianity indices.

A new semiparametric framework

We adopt the following model for the density of the observations:

$$\boldsymbol{p(x) = g(Tx)\phi_\Gamma(x),} \qquad (1)$$

where:

- T: unknown linear mapping $\mathbb{R}^d \to \mathbb{R}^m$.
- g: unknown function $\mathbb{R}^m \to \mathbb{R}$.
- ϕ_Γ: centered Gaussian density, unknown covariance matrix Γ.

Target space: the goal is to recover the non-Gaussian subspace defined as

$$\mathcal{I} = Ker(T)^\perp = Range(T^\top)\,.$$

Note that we do *not* estimate Γ, g, and T when estimating $\mathcal{I}$.

Independent Components Interpretation

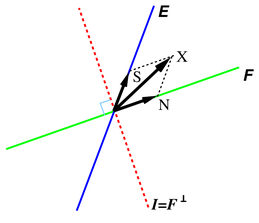
The model (1) has the following equivalent interpretation: the data $X \in \mathbb{R}^d$ can be decomposed as

$$X = S \dotplus N\,,$$

where:

- $S \in E$ is a non-Gaussian signal belonging to a lower-dimensional subspace;
- $N \in F$ is a Gaussian component belonging to a subspace F in direct sum with E, and independent of S.

In this representation $\mathcal{I} = F^\top$.



Key Property

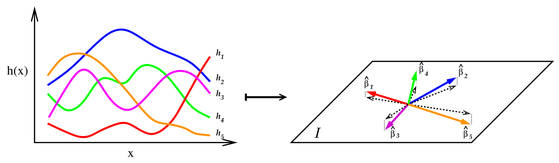
Proposition:

Let X be a random variable whose density function $p(x)$ satisfies (1) and suppose that $h(x)$ is a smooth real function on $\mathbb{R}^d$. Assume furthermore that $\Sigma = \mathbb{E}\left[XX^\top\right] = I_d$. Then under mild regularity conditions the following vector belongs to the target space $\mathcal{I}$:

$$\boldsymbol{\beta(h) = \mathbb{E}\left[Xh(X) - \nabla h\right]\,.} \qquad (2)$$

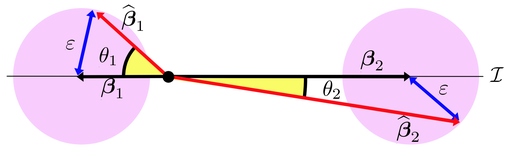
Main roadmap of the algorithm

1. Apply whitening to the data so that it has covariance identity.
2. Consider a family of smooth functions (h_i). Apply (2), replacing the true expectation by the empirical expectation over the training sample. This yields a family of estimated vectors $(\widehat{\beta}_i)$.
3. Apply PCA to the family of $(\widehat{\beta}_i)$ to recover m principal directions.
4. Pull back in original (non-whitened) data space.



Vector Normalization

- The mapping $h \mapsto \beta(h)$ is linear, implying that $\|\beta(h)\|$ can be arbitrary if h has an arbitrarily scaling.
- We would like $\|\beta\|$ to be representative of the information brought forth by this vector about the target space $\mathcal{I}$: this will justify applying PCA to the estimated vector family.
- It is necessary to introduce a suitable normalization.
- We propose a renormalization by an estimated value of $\mathbb{E}\left[\left\|\widehat{\beta}(h) - \beta(h)\right\|^2\right]^{\frac{1}{2}}$: the scaled vector norm is then proportional to its signal-to-noise ratio. Equivalently, this ensures that the estimation error is of the same order for all vectors.



Searching for informative vectors

- After renormalization, vectors $\widehat{\beta}(h)$ with a larger norm are more informative.
- We would like to search a potentially large function family $\mathcal{H}$ for the functions giving rise to more informative vectors.
- Observation: if we consider functions of the form $h(x) = f(\langle w,x\rangle)$, with $f : \mathbb{R} \to \mathbb{R}$, $w \in \mathbb{R}^d$, then equation (2) is equivalent to step (i) of FastICA.
- We use FastICA iterations as a proxy to find good candidates values of w for a fixed f. Note: since the Key Property is valid for any w, convergence is not an issue. We use a fixed number of iterations.
- This is repeated over a collection of different choices for f.
- This makes the NGCA algorithm comparable to a multi-index FastICA.

The NGCA algorithm

Input: Data points $(X_i) \in \mathbb{R}^d$, dimension m of target subspace.
Parameters: Number $T_{\max}$ of FastICA iterations; threshold ϵ; family of real functions (f_k).

Whitening.
The data X_i is recentered by subtracting the empirical mean.
Let $\widehat{\Sigma}$ be the empirical covariance matrix of the data sample (X_i).
Put $\widehat{Y}_i = \widehat{\Sigma}^{-\frac{1}{2}} X_i$ the empirically whitened data.
Main Procedure.
Loop on $k = 1, \ldots, L$:
Draw ω_0 at random on the unit sphere of $\mathbb{R}^d$.
Loop on $t = 1, \ldots, T_{\max}$: *[FastICA loop]*

$$\text{Put } \widehat{\beta}_t \leftarrow \frac{1}{n}\sum_{i=1}^n \left(\widehat{Y}_i f_k(\langle\omega_{t-1}, \widehat{Y}_i\rangle) - f_k'(\langle\omega_{t-1}, \widehat{Y}_i\rangle)\omega_{t-1}\right).$$

Put $\omega_t \leftarrow \widehat{\beta}_t / \|\widehat{\beta}_t\|$.
End Loop on t
Let N_i be the trace of the empirical covariance matrix of $\widehat{\beta}_{T_{\max}}$:

$$N_i = \frac{1}{n}\sum_{i=1}^n \left\|\widehat{Y}_i f_k(\langle\omega_{T_{\max}-1}, \widehat{Y}_i\rangle) - f_k'(\langle\omega_{T_{\max}-1}, \widehat{Y}_i\rangle)\omega_{T_{\max}-1}\right\|^2 - \left\|\widehat{\beta}_{T_{\max}}\right\|^2.$$

Store $v^{(k)} \leftarrow \widehat{\beta}_{T_{\max}} * \sqrt{n/N_i}$. *[Normalization]*
End Loop on k
Thresholding.
From the family $v^{(k)}$, throw away vectors having norm smaller than threshold ϵ.
PCA step.
Perform PCA on the set of remaining $v^{(k)}$.
Let V_m be the space spanned by the first m principal directions.
Pull back in original space.
Output: $W_m = \widehat{\Sigma}^{-\frac{1}{2}} V_m$.

Families of functions

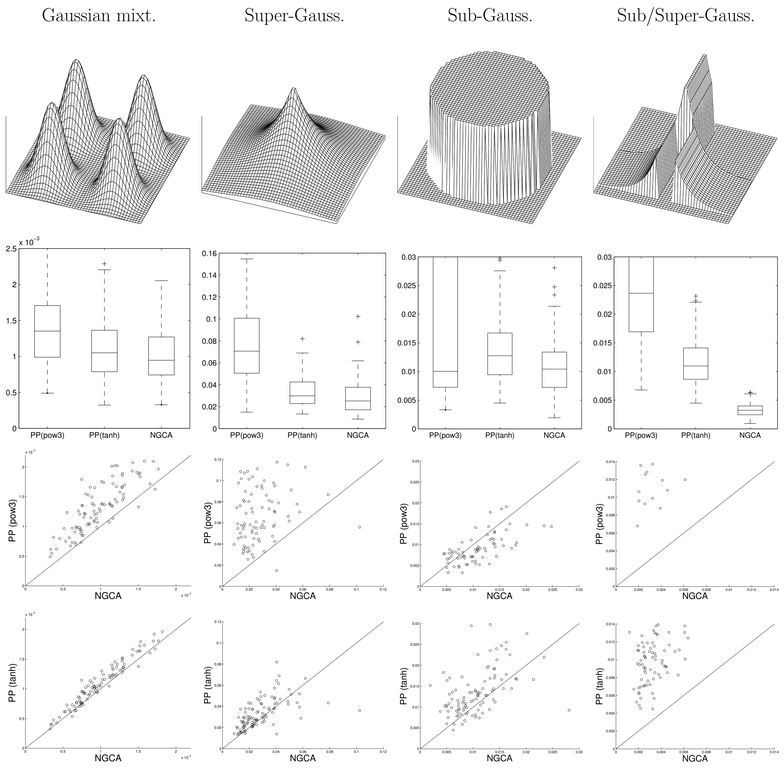
We have used the following forms of the functions f_k:

$$f_\sigma^{(1)}(z) = z^3 \exp\left(-\frac{z^2}{2\sigma^2}\right), \qquad \text{(Gauss-Pow3)}$$
$$f_b^{(2)}(z) = \tanh(bz), \qquad \text{(Hyperbolic Tangent)}$$
$$f_a^{(3)}(z) = \exp(iaz)\,, \qquad \text{(Fourier)}$$

More precisely, we consider discretized ranges for $\sigma \in [\sigma_{\min}, \sigma_{\max}]$, $b \in [0, B]$, and $a \in [0, A]$, giving rise to a finite collection $\{f_k\}$ (which therefore includes *simultaneously* functions of the three different above families).

Numerical Examples

Synthetic data



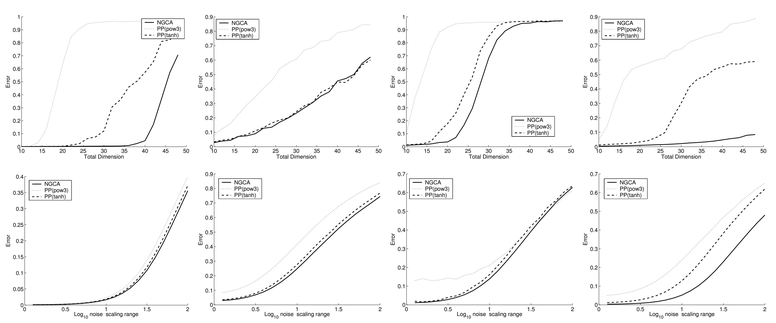
- Each dataset has 2 non-Gaussian components + 8 Gaussian components; 1000 data points per trial.
- The same parameters (defining the function families) have been used in all experiments. No dataset-specific tuning. Parameter values are discretized so that each family gives rise to a collection of 1000 functions.
- The criterion for measuring performance is

$$\mathcal{E}(\widehat{\mathcal{I}},\mathcal{I}) = (2m)^{-1}\left\|\Pi_{\mathcal{I}} - \Pi_{\widehat{\mathcal{I}}}\right\|_{Frob}^2 = m^{-1}\sum_{i=1}^{m}\|(I_d - \Pi_{\widehat{\mathcal{I}}})v_i\|^2,$$

where Π_V denotes the orthogonal projection on V, and $\{v_i\}_{i=1}^m$ is an orthonormal basis of $\mathcal{I}$.

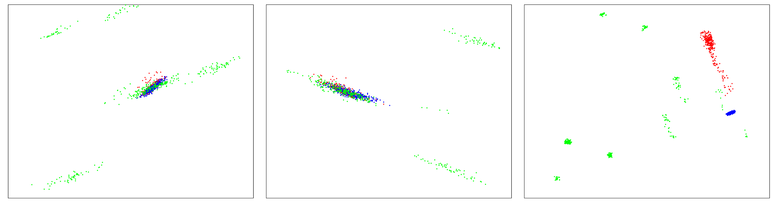
Robustness / Failure modes

We investigate the limit conditions under which NGCA correctly estimates the target space $\mathcal{I}$.



- Top row shows the error criterion vs. increasing data dimensionality (all other parameters being equal).
- Bottom rows shows the error criterion for increasingly worse conditioning of the covariance matrix of the Gaussian part.

Application to the "oil flow" dataset



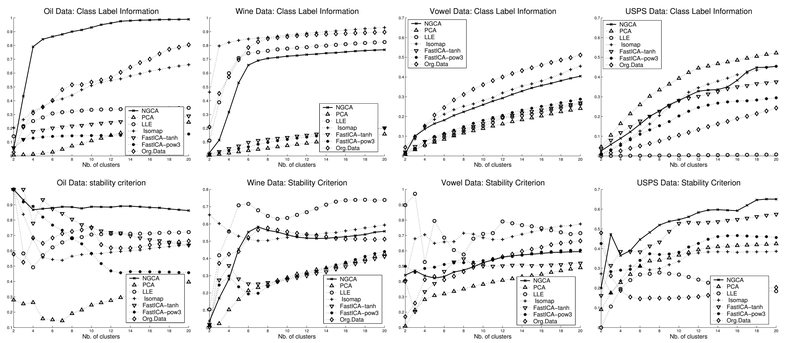
Proj. Pursuit (tanh) Proj. Pursuit (power3) NGCA

- 12-dimensional data coming from a complex simulated model of oil flow, used as an example for other methods of dimensionality reduction (Bishop et al. 1998).
- Clustered structure is expected. The data is divided into 3 classes but this information is not used here.
- We used the different methods to yield a 3D projection out of which the best 2D representation was chosen visually to exhibit the clearer cluster structure.

Application to clustering

Data set	Nb. of Classes	Nb. of samples	Total dimension	Projection Dim.
Oil	3	2000	12	3
Wine	3	178	13	3
Vowel	11	528	10	3
USPS	10	7291	30 (KPCA)	10

- We study the influence of different dimensionality reduction methods on clustering on several datasets.
- There is no single well-defined performance measure for the performance of clustering. Here we use the following criteria:
  - **Label cross-information.** The datasets used have label information Y not used for the dimensionality reduction. If C is the cluster labelling, we compute the scaled mutual information $I(C,Y)/H(Y)$ as a measure of relevance of the cluster structure found.
  - **Stability.** Inspired by recent work on clustering (Lange et al. 2004), we measure the stability of the cluster structure found by applying the clustering algorithm on two separate subsamples of equal size and compute $I(C_1,C_2)/H(C_1,C_2)$.



Theoretical guarantees on the estimation error

Goals of a theoretical control of the estimation error:

- Obtain a uniform control of the estimation error over a whole family of functions, since considering many functions simultaneously plays a crucial role in the method.
- Make the variance dependence appear explicilty in the error control to justify the renormalization procedure.
- Take into account explicitly the effect of empirical whitening/dewhitening.

Theorem 1 (identity covariance matrix case).

Let $\{h_k\}_{k=1}^L$ be a family of smooth functions. Assume that $\sup_{k,y}\max\left(\|\nabla h_k(y)\|, \|h_k(y)\|\right) < B$, that X has covariance matrix $\mathbb{E}\left[XX^\top\right] = I_d$, and is such that for some $\lambda_0 > 0$:

$$\mathbb{E}\left[\exp\left(\lambda_0\|X\|\right)\right] \le a_0 < \infty.$$

Denote $\widetilde{h}(x) = \nabla h(x) - xh(x)$. Suppose $X_1,\dots,X_n$ are i.i.d. copies of X and define

$$\widehat{\beta}(h) = \frac{1}{n}\sum_{i=1}^n \widetilde{h}(X_i)\,, \text{ and } \widehat{\sigma}(h) = \frac{1}{n}\sum_{i=1}^n \left\|\widetilde{h}(X_i) - \widehat{\beta}(h)\right\|^2;$$

then with probability $1-4\delta$ the following holds simultaneously for all $k\in\{1,\dots,L\}$:

$$dist\left(\widehat{\beta}(h_k),\mathcal{I}\right) \le 2\sqrt{\widehat{\sigma}^2(h_k)\frac{\log(L\delta^{-1})+\log d}{n}} + C\left(\frac{\log(nL\delta^{-1})\log(L\delta^{-1})}{n^{\frac{3}{4}}}\right),$$

where C depends only on the parameters (d,λ_0,a_0,B,K).

Theorem 2 (general case).

Let us assume the following :

(i) There exists $\lambda_0>0, a_0>0$ such that

$$\mathbb{E}\left[\exp\left(\lambda_0\|X\|^2\right)\right] = a_0 < \infty;$$

(ii) The covariance matrix Σ of X is such that $\left\|\Sigma^{-1}\right\| \le K^2$;

(iii) $\sup_{k,y}\max\left(\|\nabla h_k(y)\|, \|h_k(y)\|\right) < B$;

(iv) The functions $\widetilde{h}_k(y) = \nabla h_k(y) - y h_k(y)$ are all Lipschitz with constant M.

Then for big enough n, with probability at least $1-\frac{4}{n}-4\delta$ the following bounds hold true simultaneously for all $k\in\{1,\dots,L\}$:

$$dist(\widehat{\beta}(h_k),\mathcal{I}) \le C_1\sqrt{\frac{d\log n}{n}} + 2K\sqrt{\widehat{\sigma}^2_{\widehat{Y}}(h_k)\frac{\log(L\delta^{-1})+\log d}{n}} + C_2\frac{\log(nL\delta^{-1})\log(L\delta^{-1})}{n^{\frac{3}{4}}},$$

where C_1 depends on parameters (λ_0,a_0,B,K,M) only and C_2 on (d,λ_0,a_0,B,K,M). Here $\widehat{\beta}(h)$ is obtained as in the actual algorithm: equation (2) is applied using the (empirically) whitened data, and the resulting vector is pulled back again in original data space by application of $\widehat{\Sigma}^{-\frac{1}{2}}$.

J.H. Friedman and J.W. Tukey. *A projection pursuit algorithm for exploratory data analysis.* IEEE Transactions on Computers, 23 (9):881–890, 1975.

A. Hyvärinen, J. Karhunen and E. Oja. *Independent Component Analysis.* Wiley, 2001.

T. Lange, V. Roth, M. L. Braun and J. M. Buhmann. *Stability-based validation of clustering solutions.* Neural Computation, 16(6):1299-1323, 2004.

C.M. Bishop, M. Svensen and C.K.I. Wiliams. *GTM: The generative topographic mapping.* Neural Computation, 10 (1): 215–234, 1998.