Weierstraß-Institut
für Angewandte Analysis und Stochastik

im Forschungsverbund Berlin e.V.

Preprint ISSN 0946 – 8633

In Search of Non-Gaussian Components of a High-Dimensional Distribution

Gilles Blanchard[1] Motoaki Kawanabe[2], Masashi Sugiyama[3], Vladimir Spokoiny[4],

Klaus-Robert Müller[5]

submitted: 18th January 2006

[1] Fraunhofer FIRST.IDA
Kekuléstr. 7, 12489, Berlin, Germany and
CNRS, Université Paris-Sud, Orsay, France
E-Mail: `blanchar@first.fhg.de`

[2] Fraunhofer FIRST.IDA
Kekuléstr. 7, 12489, Berlin, Germany
E-Mail: `nabe@first.fhg.de`

[3] Fraunhofer FIRST.IDA, Kekuléstr. 7, 12489, Berlin, Germany and
Department of Computer Science, Tokyo Institute of Technology
2-12-1, O-okayama, Meguro-ku, Tokyo, 152-8552, Japan
E-Mail: `sugi@cs.titech.ac.jp`

[4] Weierstrass Institute for
Applied Analysis and Stochastics,
Mohrenstr. 39, 10117, Berlin, Germany
E-Mail: `spokoiny@wias-berlin.de`

[5] Fraunhofer FIRST.IDA
Kekuléstr. 7, 12489, Berlin, Germany and
Department of Computer Science
University of Potsdam
August-Bebel-Str.89, Haus 4,
14482 Potsdam, Germany
E-Mail: `klaus@first.fhg.de`

No. 1092
Berlin 2006

[4]Supported by the DFG Research Center MATHEON "Mathematics for key technologies" in Berlin

2000 *Mathematics Subject Classification.* 00A73, 62H30, 62H99, 68Q32 .

Key words and phrases. Non-Gaussian components, dimension reduction.

Edited by
Weierstraß-Institut für Angewandte Analysis und Stochastik (WIAS)
Mohrenstraße 39
10117 Berlin
Germany

Fax: + 49 30 2044975
E-Mail: preprint@wias-berlin.de
World Wide Web: http://www.wias-berlin.de/

Abstract

Finding non-Gaussian components of high-dimensional data is an important pre-processing step for efficient information processing. This article proposes a new *linear* method to identify the "non-Gaussian subspace" within a very general semi-parametric framework. Our proposed method, called NGCA (Non-Gaussian Component Analysis), is essentially based on a linear operator which, to any arbitrary nonlinear (smooth) function, associates a vector which belongs to the low dimensional non-Gaussian target subspace up to an estimation error. By applying this operator to a family of different nonlinear functions, one obtains a family of different vectors lying in a vicinity of the target space. As a final step, the target space itself is estimated by applying PCA to this family of vectors. We show that this procedure is consistent in the sense that the estimaton error tends to zero at a parametric rate, uniformly over the family, Numerical examples demonstrate the usefulness of our method.

1 Introduction

1.1 Setting and general principle

Suppose $\{X_i\}_{i=1}^n$ are i.i.d. samples in a high dimensional space $\mathbb{R}^d$ drawn from an unknown distribution with density $p(x)$. A general multivariate distribution is typically too complex to analyze from the data, thus dimensionality reduction is useful to decrease the complexity of the model [Cox and Cox, 1994, Schölkopf et al., 1998, Roweis and Saul, 2000, Tenenbaum et al., 2000, Belkin and Niyogi, 2003]. Here our point of departure is the following assumption: the high dimensional data includes low dimensional non-Gaussian components and the other components are Gaussian. This assumption follows the rationale that in most real-world applications the 'signal' or 'information' contained in the high-dimensional data is essentially non-Gaussian while the 'rest' can be interpreted as high dimensional Gaussian noise.

We want to emphasize from the beginning that we do *not* assume the Gaussian components to be of *smaller* order of magnitude than the signal components; all components are instead typically of *the same order*. This setting therefore excludes the use dimensionality reduction methods based on the assumption that the data lies, say, on a lower dimensional manifold, up to some small noise. In fact, this type of methods addresses a different kind of problem altogether.

Under our modeling assumption, therefore, the task is to recover the relevant *non-Gaussian* components. Once such components are identified and extracted, various tasks can be applied in the data analysis process, say, data visualization, clustering, denoising or classification.

If the number of Gaussian components is *at most one* and all the non-Gaussian components are mutually independent, *Independent Component Analysis (ICA)* techniques [e.g.

Comon, 1994, Hyvärinen et al., 2001] are relevant to identify the non-Gaussian subspace. Unfortunately, however, this is often a too strict assumption on the data.

The framework we consider is on the other hand very close to that of *Projection Pursuit (PP)* algorithms [Friedman and Tukey, 1974, Huber, 1985, Hyvärinen et al., 2001]. The goal of Projection Pursuit methods is to extract non-Gaussian components in a general setting, i.e., the number of Gaussian components can be more than one and the non-Gaussian components can be dependent.

Projection Pursuit methods typically proceed by fixing a *single* index which measures the non-Gaussianity (or 'interessingness') of a projection direction. This index is then optimized over all possible directions of projection; the procedure can be repeated iteratively (over directions orthogonal to the first ones already found) to find a higher dimensional projection of the data as needed.

However, it is known that some projection indices are suitable for finding super-Gaussian components (heavy-tailed distribution) while others are suited for identifying sub-Gaussian components (light-tailed distribution) [Hyvärinen et al., 2001]. Therefore, traditional PP algorithms may not work effectively if the data contains, say, both super- and sub-Gaussian components.

To summarize: existing methods for the setting we consider typically proceed by defining an appropriate interestingness index and then compute a projection that maximizes this index (projection pursuit methods, and some ICA methods). The philosophy that we would like to promote in this paper is in a sense different: in fact we do not specify what we are interested in, but we rather define what is *not interesting* [Jones and Sibson , 1987]. Clearly a multi-dimensional Gaussian subspace is a reasonable candidate for an undesired component (our idea could be generalized by defining, say, a Laplacian subspace to be uninformative). Having defined this uninteresting subspace, its (orthogonal) complement is by definition interesting: this therefore precisely defines our target space.

1.2 Presentation of the method

Technically, our new approach to identifying the non-Gaussian subspace uses a very general semi-parametric framework. The proposed method, called *Non-Gaussian Component Analysis (NGCA)*, is essentially based on a central property stating that there exists a linear mapping $h \mapsto \beta(h) \in \mathbb{R}^d$ which, to any *arbitrary* (smooth) nonlinear function $h : \mathbb{R}^d \to \mathbb{R}$, associates a vector β lying in the non-Gaussian target subspace. In practice, the vector $\beta(h)$ has to be estimated from the data, giving rise to an estimation error. However, our main consistency result shows that this estimation error vanishes at a rate $\sqrt{\log(n)/n}$ with the sample size n. Using a whole family of different nonlinear functions h then yields a family of different vectors $\widehat{\beta}(h)$ which all approximately lie in, and span, the non-Gaussian subspace. We finally perform PCA on this family of vectors to extract the principal directions and estimate the target space.

In practice, we consider functions of the particular form $h_{\omega,a}(x) = f_a(\langle \omega, x \rangle)$ where f is a function class parameterized, say, by a parameter a, and $\|\omega\| = 1$. Even for a fixed a, it is infeasible to compute values of $\beta(h_{\omega,a})$ for all possible values of ω (say on a discretized net of the unit sphere) because of the cardinality involved. In order to choose a relevant value for ω (still for fixed a), we then opt to use as a heuristic a well-known PP algorithm, FastICA [Hyvärinen, 1999]. This was suggested by the surprising observation that the mapping $\omega \to \beta(h_{\omega,a})$ is then *equivalent* to a *single* iteration of FastICA (although this algorithm was built using different theoretical considerations); hence in this special case FastICA is exactly the same as iterating our mapping. In short, we use a PP method as a proxy to select the most relevant direction ω for a fixed a. This results in a particular choice of ω_a, to which we apply the mapping once more, thus yielding $\beta_a = \beta(h_{\omega_a,a})$. Finally, we aggregate the different vectors β_a obtained when varying a by applying PCA as indicated previously.

Thus, apart from the conceptual point, defining uninterestingness as the point of departure instead of interestingness, another way to look at our method is to say that it allows the combination of information coming from different indices: here the above function f_a (for fixed a) plays a role similar to that of a non-Gaussianity index in PP, but we do combine a rich family of such functions (by varying a and even by considering several function classes at the same time). The important point here is while traditional Projection Pursuit does not provide a well-founded justification for combining directions from using different indices, our framework allows to do precisely this – thus implicitly selecting, in a given family of indices, the ones which are the most informative for the data at hand.

In the following section we will outline the theoretical cornerstone of the method, a novel semi-parametric theory for *linear* dimension reduction. Section 3 discusses the algorithmic procedure and is conluded with theoretical results establishing statistical consistence of the method. In Section 4, we study on simulated and real data examples the behavior of the algorithm. A brief conclusion is given in Section 5.

2 Theoretical framework

In this section, we give a theoretical basis for the non-Gaussian component search within a *semi-parametric* framework. We present a population analysis where expectations can in principle be calculated exactly in order to emphasize the main idea and how the algorithm is built. A more rigorous statistical study of the estimation error will be exposed later in section 3.5.

2.1 Motivation

Before introducing the semi-parametric density model which will be used as a foundation for developing our method, we motivate it by starting from elementary considerations. Suppose we are given a set of observations $X_i \in \mathbb{R}^d$, $(i = 1, \ldots, n)$ obtained as a sum of a signal S and an independent Gaussian noise component N :

$$X = S + N \,, \tag{1}$$

where $N \sim \mathcal{N}(0, \Gamma)$. Note that no particular structural assumption is made about the noise covariance matrix Γ.

Assume the signal S is contained in a lower-dimensional linear subspace E of dimension $m < d$. Loosely speaking, we would like to project X linearly so as to eliminate as much of the noise as possible while preserving the signal information. An important issue for the analysis of the model (1) is a suitable representation of the density of X which reflects the low dimensional structure of the non-Gaussian signal. The next lemma presents a generic representation of the density p for the model (1).

Lemma 2.1 *The density $p(x)$ for the model (1) with the m-dimensional signal S and an independent Gaussian noise N can be represented as*

$$p(x) = g(Tx)\phi_\Gamma(x)$$

where T is a linear operator from $\mathbb{R}^d$ to $\mathbb{R}^m$, $g(\cdot)$ is some function on $\mathbb{R}^m$ and $\phi_\Gamma(x)$ is the density of the Gaussian component.

The formal proof of this lemma is given in the Appendix. Note that the above density representation is not unique, as the parameters g, T, Γ are not identifable from the density p. However, the null suspace (kernel) $\mathfrak{K}(T)$ of the linear operator T is an identifiable parameter. In particular, is useful to notice that if the noise N is standard normal, then the operator T can be taken as the projector on the signal space E. In this case, $\mathfrak{K}(T)$ coincides with $E^\perp$, the orthogonal complementary subspace to E. In the general situation with "colored" Gaussian noise, the signal space E does not coincide with the orthogonal complementary of the kernel $\mathcal{I} = \mathfrak{K}(T)^\perp$ of the operator T. However, the density representation of Lemma 2.1 shows that the the subspace $\mathfrak{K}(T)$ is non-informative and contains only noise. The original data can then be projected orthogonally onto $\mathcal{I}$, which we call the *non-Gaussian subspace*, without loss of information. This way, we are preserving the totality of the signal information. This definition implements the general point of view outlined in the introduction, namely: we define what is considered *uninteresting*; the target space is then defined indirectly as the orthogonal of the uninteresting component.

2.2 Relation to ICA analysis

An equivalent view of the same model is to decompose the noise N appearing in Eq.(1) into a component N_1 belonging to the signal space E and an *independent* component N_2; it can then be shown that N_2 belongs to the subspace $\mathfrak{K}(T)$ defined above. In this view the space $\mathcal{I}$ is orthogonal to the independent noise component, and projecting the data onto $\mathcal{I}$ amounts to cancelling this independent noise component by an orthogonal projection.

In the present paper, we assume that we wish to project the data *orthogonally*, i.e., that the Euclidean geometry of the input space is meaningful for the data at hand, and that we want to respect it while projecting. An alternative point of view would be to disregard the input space geometry altogether, and to first map the data linearly to a reference space where it has covariance identity ("whitening" transform), which would be closer to a traditional ICA analysis. This would have on the one hand the advantage of resulting in an affine invariant procedure, but, on the other hand, the disadvantage of losing the information of the original space geometry.

2.3 Main model

Based on the above motivation, we assume to be dealing with an unknown probability density function $p(x)$ on $\mathbb{R}^d$ which can put under the form

$$p(x) = g(Tx)\phi_\Gamma(x), \tag{2}$$

where T is an unknown linear mapping from $\mathbb{R}^d$ to $\mathbb{R}^m$ with $m \le d$, g is an unknown function on $\mathbb{R}^m$, and ϕ_Γ is a centered[1] Gaussian density with covariance matrix Γ.

Note that the *semi-parametric* model (2) includes as particular cases both the pure parametric ($m = 0$) and purely non-parametric ($m = d$) models. For practical purposes however we are effectively interested in an intermediate case where d is large and m is relatively small. In what follows we denote by $\mathcal{I}$ the m-dimensional *linear* subspace in $\mathbb{R}^d$ generated by the adjoint operator T^*:

$$\mathcal{I} = \mathfrak{K}(T)^\perp = \Im(T^*)\,,$$

where $\Im(\cdot)$ denotes the range of an operator. We call $\mathcal{I}$ the *non-Gaussian subspace*.

The proposed goal is therefore to estimate $\mathcal{I}$ by some subspace $\widehat{\mathcal{I}}$ computed from an i.i.d. sample $\{X_i\}_{i=1}^n$ following the distribution with density $p(x)$. In this paper we assume the effective dimension m to be known or fixed *a priori* by the user. Note that we do *not* estimate Γ nor g when estimating $\mathcal{I}$. We measure the closeness of the two subspaces $\widehat{\mathcal{I}}$ and $\mathcal{I}$ by the following error function:

$$\mathcal{E}(\widehat{\mathcal{I}},\mathcal{I}) = (2m)^{-1}\left\|\Pi_{\mathcal{I}} - \Pi_{\widehat{\mathcal{I}}}\right\|^2_{Frob} = m^{-1}\sum_{i=1}^{m}\|(I_d - \Pi_{\widehat{\mathcal{I}}})v_i\|^2, \tag{3}$$

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

2.4 Key result

The main idea underlying our approach is summed up in the following Proposition (the proof is given in Appendix A.2). Whenever variable X has covariance[2] matrix identity, this result allows, from an *arbitrary* smooth real function h on $\mathbb{R}^d$, to find a vector $\beta(h) \in \mathcal{I}$.

Proposition 2.2 *Let X be a random variable whose density function $p(x)$ satisfies Eq.(2) 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 on h, the following vector $\beta(h)$ belongs to the target space $\mathcal{I}$:*

$$\beta(h) = \mathbb{E}\left[Xh(X) - \nabla h(X)\right]\,. \tag{4}$$

[1]It is possible to handle a more general situation where the Gaussian part has an unknown mean parameter θ in addition to the unknown covariance Γ. For simplicity of exposition we consider here only the case of $\theta = 0$.

[2]Here and in the sequel, with some abuse we call $\Sigma = \mathbb{E}\left[XX^\top\right]$ the *covariance matrix* even though we do not assume that the non-Gaussian part of the data is centered.

In the general case where the covariance matrix Σ is different from identity, provided it is nondegenerated, we can apply a whitening operation (also known as Mahalanobis transform). Namely, let us put $Y = \Sigma^{-\frac{1}{2}}X$ the "whitened" data; the covariance matrix of Y is then identity. Note that if the density function of X is of the form

$$p(x) = g(Tx)\phi_\Gamma(x),$$

then by change of variable the density function of $Z = AX$ is given by

$$q(z) = c_A g(TA^{-1}z)\phi_{A\Gamma A^\top}(z),$$

where c_A is a normalization constant depending on A.

This identity applied to $A = \Sigma^{-\frac{1}{2}}$ and the previous proposition allow to conclude that

$$\beta_Y(h) = \mathbb{E}\left[\nabla h(y) - yh(y)\right] \in \mathcal{J} = \Im(\Sigma^{\frac{1}{2}}T^*)$$

and therefore that

$$\gamma(h) = \Sigma^{-\frac{1}{2}}\beta_Y(h) \in \mathcal{I} = \Im(T^*)\ ,$$

where $\mathcal{I}$ is the non-Gaussian index space for the initial variable X, and $\mathcal{J} = \Sigma^{\frac{1}{2}}\mathcal{I}$ the transformed non-Gaussian space for the whitened variable Y.

3 Procedure

We now use the key proposition established in the previous section to design a practical algorithm in order to identify the non-Gaussian subspace. The first step is to apply the whitening transform to the data (where the true covariance matrix Σ is estimated by the empirical covariance $\widehat{\Sigma}$). We then estimate the "whitened" non-Gaussian space $\mathcal{J}$ by some $\widehat{\mathcal{J}}$ (this will be described next); this space is then finally pulled back in the original space by application of $\widehat{\Sigma}^{-\frac{1}{2}}$. To simplify the exposition, in this section we will forget about the whitening/dewhitening steps and always implicitly assume that we are dealing directly with the whitened data: every time we refer to the non-Gaussian space it is therefore to be understood that we refer to $\mathcal{J} = \Sigma^{\frac{1}{2}}\mathcal{I}$ corresponding to the whitened data Y.

3.1 Principle idea

In the previous section we have proved that for an arbitrary function h satisfying mild smoothness conditions, it is possible to construct a vector $\beta(h)$ which lies in the non-Gaussian subspace. However, since the unknown density $p(x)$ is used (via the expectation operator) to define β by Eq.(2.2), one cannot directly use this formula in practice: it is then natural to approximate it by replacing the true expectation by the empirical expectation. This gives rise to the estimated vector

$$\widehat{\beta}(h) = \frac{1}{n}\sum_{i=1}^{n} Y_i h(Y_i) - \nabla h(Y_i)\,, \tag{5}$$

which we expect to be close to the non-Gaussian subspace up to some estimation error. At this point, a natural step is to consider a whole family of functions $\{h_i\}_{i=1}^{n}$ giving

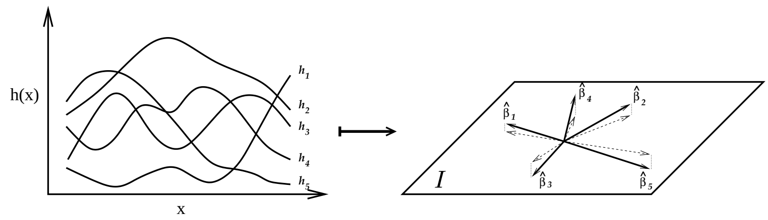


Figure 1: The NGCA principle idea: from a varied family of real functions, compute a family of vectors belonging to the target space up to small estimation error.

rise to an associated vector family of $\{\widehat{\beta}_i\}_{i=1}^L$ all lying close to the target subspace, where $\widehat{\beta}_i := \widehat{\beta}(h_i)$. The final step is to recover the non-Gaussian subspace from this set. For this purpose we suggest to use the principal directions of this family, i.e. to apply PCA (although other algorithmic options are certainly avalaible for this task). This general idea is illustrated on Figure 1.

3.2 Normalization of the vectors

When extracting information on the target subspace from the set of vectors $\{\widehat{\beta}_i\}_{i=1}^L$, attention should be paid to how the functions $\{h_i\}_{i=1}^L$ are normalized. As can be seen from its definition, the operator which maps a function h to $\beta(h)$ (and also its empirical counterpart $\widehat{\beta}(h)$) is linear. Therefore, if, for example, one of the functions $\{h_i\}_{i=1}^L$ is multiplied by an arbitrarily large scalar, the associated $\widehat{\beta}(h)$ could have an arbitrarily large norm: this is likely to influence heavily the procedure of principal direction extraction applied to the whole family.

To prevent this problem, the functions $\{h_i\}_{i=1}^L$ should be normalized in a reasonable way. Several possibilities can come to mind, like using the sup or L_2 norm of h or of ∇h. We argue here that a sensible way to normalize functions is such that the average squared deviation (estimation error) of $\widehat{\beta}(h)$ to its mean is of the same order for all functions h considered. This has a first direct intuitive interpretation in terms of making the length of each estimated vector proportional to its associated signal-to-noise ratio. We argue in more detail that the norm of $\widehat{\beta}(h)$ after normalization is directly linked to the amount of information brought by this vector about the target subspace.

Namely, if we measure the information that is brought by a certain vector $\widehat{\beta}(h)$ about the target space $\mathcal{J}$ through the angle $\theta(\widehat{\beta}(h))$ between the vector and the space, we have

$$\|\widehat{\beta}(h) - \beta(h)\| \geq dist(\widehat{\beta}(h), \mathcal{J}) = \sin(\theta(\widehat{\beta}(h)))\|\widehat{\beta}(h)\| \,. \tag{6}$$

Suppose we have ensured by renormalization that $\sigma(h)^2 = \mathbb{E}\left[\|\widehat{\beta}(h) - \beta(h)\|^2\right]$ is constant and independent of h, and assume that this results in $\|\widehat{\beta}(h) - \beta(h)\|^2$ being bounded by some constant with high probability. It entails that $\sin(\theta(\widehat{\beta}(h)))\|\widehat{\beta}(h)\|$ is bounded independently of h. We expect in this situation that the bigger $\|\widehat{\beta}\|$, the smaller is $\sin(\theta)$, and therefore the more reliable the information about $\mathcal{J}$. This intuition is illustrated in Figure 2, where the estimation error is represented by a confidence ball of equal size for all vectors[3].

[3]Of course, the situation depicted in Figure 2 is idealized: we actually expect (from the Central Limit

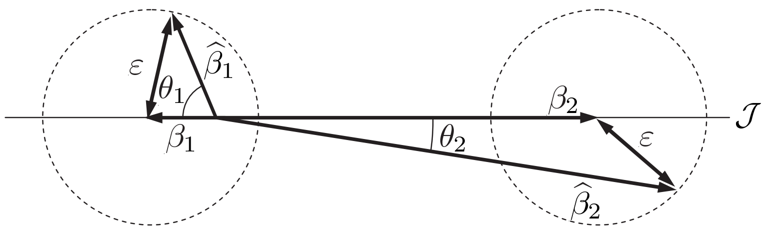


Figure 2: For the same estimation error represented as a confidence ball of radius ε, estimated vectors with higher norm give a more precise information about the true target space.

Therefore, at least at an intuitive level it appears appropriate to use $\sigma(h)$ as a renormalization. Note that this is just the square root of the trace of the covariance matrix of $\widehat{\beta}(h)$, and therefore easy to estimate in practice from its empirical counterpart. In section 3.5, we give actual theoretical confidence bounds for $\|\beta - \widehat{\beta}\|$ which justify this intuition in a more rigorous manner.

Finally, to confirm this idea on actual data, we plot in the top row Figure 3 the distribution of $\widehat{\beta}$ on an illustrative data set using the normalization scheme just described. In order to investigate the relation between the norm of the (normalized) $\widehat{\beta}$ and the amount of information on the non-Gaussian subspace brought by $\widehat{\beta}$, we plot in the right part of Figure 3 the relation between $\|\widehat{\beta}\|$ and $\|\Pi_{\mathcal{J}}\widehat{\beta}\|/\|\widehat{\beta}\| = \cos(\theta(\widehat{\beta}))$. As expected, the vectors $\widehat{\beta}$ with highest norm are indeed much closer to the non-Gaussian subspace in general. Furthermore, vectors $\widehat{\beta}$ with norm close to zero appear to bear almost no information about the non-Gaussian space, which is consistent with the setting depicted in Figure 2: whenever an estimated vector $\widehat{\beta}$ has norm smaller than the estimation error ε, its confidence ball contains the origin, which means that it brings no useable information about the non-Gaussian subspace.

These findings motivate two important points for the algorithm:

1. It should be beneficial to *actively search* for functions h which yield an estimated $\widehat{\beta}(h)$ with higher norm, since these are more informative about the target space $\mathcal{J}$;

2. The vectors $\widehat{\beta}$ with norm below a certain threshold ε can be discarded as they are non-informative. So far the theoretical bounds presented below in section 3.5 are not precise enough to give a canonical value for this threshold: we therefore recommend that it be determined by a calibration procedure. For this, we consider independent Gaussian data: in this case $\beta = 0$ for any h and thus $\|\widehat{\beta}\|$ represents pure estimation noise. A reasonable choice for the threshold is therefore the 95th percentile (say) of this distribution, which we expect to reject a large majority of the noninformative vectors.

Theorem) that $\beta - \widehat{\beta}$ has approximately a Gaussian distribution with some non-spherical variance, giving rise to a confidence ellipsoid rather than a confidence ball. To obtain a spherical error ball, we would have to apply a (linear) error whitening transform separately to each $\widehat{\beta}(h)$. However, in that case the error whitening transform would be different for each h, and the information of the vector family about the target subspace would then be lost. To preserve this information, only a scalar normalization is adequate, which is why we recommend the normalization scheme explained here.

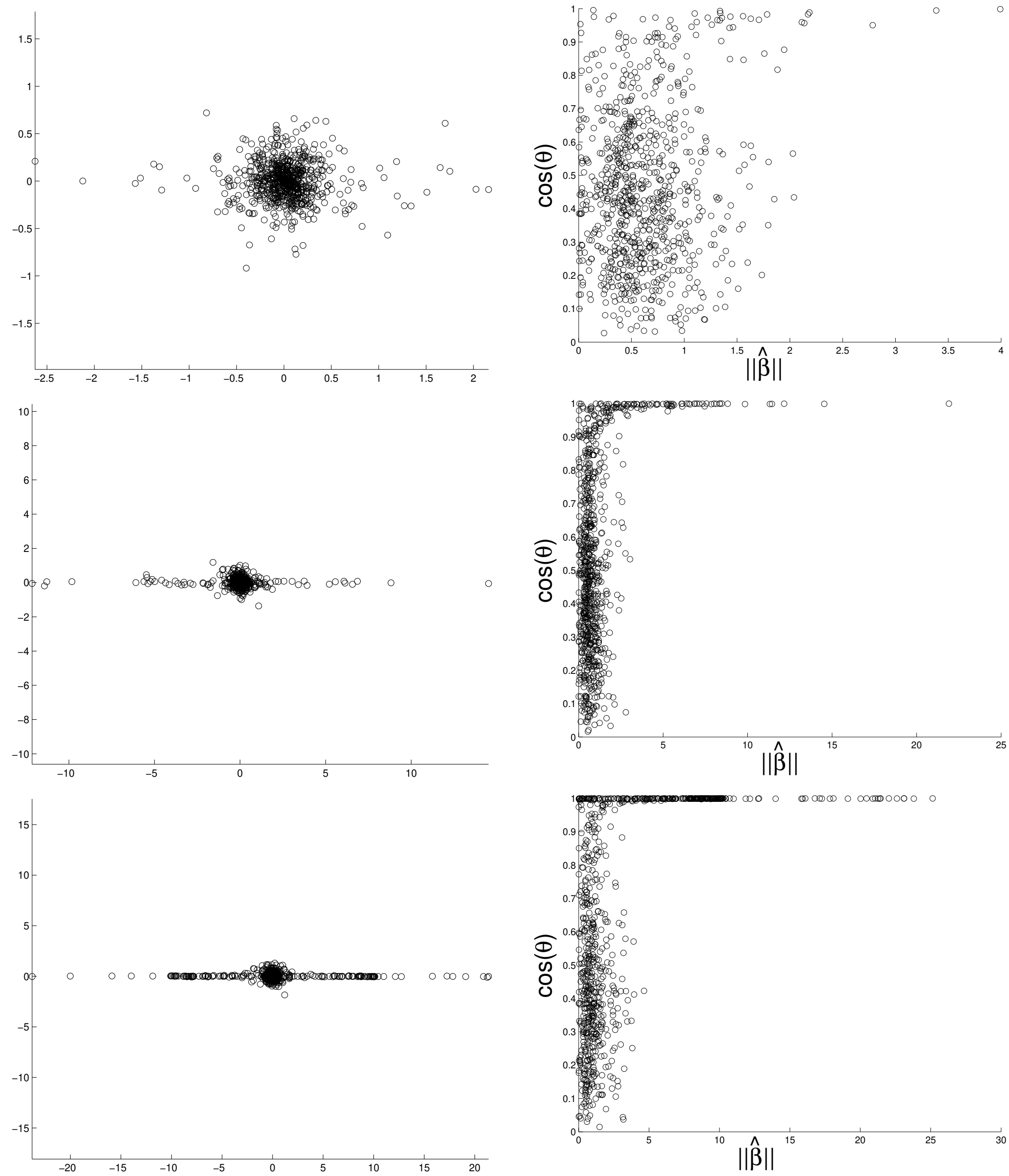


Figure 3: Illustrative plots of the method, applied to toy data of type **(A)** (See section 4.1). Left column: Distribution of $\widehat{\beta}$ projected on a direction belonging to the target space $\mathcal{J}$ (abscissa) and a direction orthogonal to it (ordinate). Right column: $\|\widehat{\beta}\|$ (after normalization) vs. $\cos(\theta(\widehat{\beta}, \mathcal{J}))$. From top to bottom rows: random draw of functions, after 1-step, and after 4-step of FastICA preprocessing. $\widehat{\beta}$'s are normalized as described in section 3.2.

3.3 Using FastICA as preprocessing to find promising functions

When considering a parametrized family of functions $\{h_\omega\}$, it is a desirable goal to search the parameter space to find indices ω such that $\widehat{\beta}(h_\omega)$ has a high norm, as proposed in the last section. From now on we will restrict our attention to functions of the form

$$h_\omega(x) = f(\langle \omega, x \rangle)\,, \tag{7}$$

where $\omega \in \mathbb{R}^d$, $\|\omega\| = 1$, and f is a smooth real function of a real variable. Clearly, it is not feasible to sample the entire parameter space for ω as soon as it has more than a few dimensions, and it is not obvious *a priori* to find parameters ω such that $\widehat{\beta}(h_\omega)$ has a high norm. Remember however that we do not need to find an *exact maximum* of this norm over the parameter space. We merely want to find parameters such that the associated norm is preferably high because they bring more information; this may also involve heuristics. Naturally, good heuristics should be able to find parameters giving rise to vectors with higher norm, bringing more information on the subspace and ultimately better practical results, nevertheless the underlying theoretical motivation stays unchanged regardless of the way the functions are picked.

A particularly relevant heuristic for choosing ω is suggested to us when we look at the form of $\widehat{\beta}$ given by Eq.(5) when we plug in functions of the specific form given by Eq.(7):

$$\widehat{\beta}(h_\omega) = \frac{1}{n} \sum_{i=1}^{n} \big(Y_i f(\langle \omega, Y_i \rangle) - f'(\langle \omega, Y_i \rangle)\omega \big)\ . \tag{8}$$

It is interesting and surprising to notice that this equation precisely coincides with *one* iteration of a well-known Projection Pursuit algorithm, FastICA [Hyvärinen, 1999]. More precisely, FastICA consists in iterating the following update rule to form a sequence $\omega_1, \ldots, \omega_T$:

$$\omega_{t+1} \propto \frac{1}{n} \sum_{i=1}^{n} \big(Y_i f(\langle \omega_t, Y_i \rangle) - f'(\langle \omega_t, Y_i \rangle)\omega_t \big) \tag{9}$$

where the sign $\propto$ indicates that vector ω_{t+1} is renormalized to be of unit norm.

Note that the FastICA procedure is derived from quite a different theoretical setting of what we considered here [see e.g. Hyvärinen et al., 2001]; its goal is in principle to optimize a non-Gaussianity measure $\mathbb{E}\left[F(\langle \omega, x \rangle)\right]$ (where F is such that F' formally coincides with our f above) and the solution is reached by an approximate Newton method giving rise to the update rule of Eq.(9), repeated until convergence.

This formal identity leads us to adopt the FastICA methodology as a heuristic for our method. Since finding an actual optimum point is not needed, convergence is not an issue, so that we only iterate the update rule of Eq.(9) for a fixed number of iterations T to find a relevant direction ω_T . Finally we apply Eq.(8) one more time to this choice of parameter, so that the procedure finally outputs $\widehat{\beta}(h_{\omega_T})$. On Figure 3, we plot the effect of a few iterations of this preprocessing for the method, applied on toy data and see that it leads to a significant improvement.

Paradoxically, if the convergence of this FastICA preprocessing is too good, there is in principle a risk that all vectors $\widehat{\beta}$ end up in the vicinity of one single "best" direction instead of spanning the whole target space: the preprocessing would then have the opposite effect of what is wished, namely impoverishing the vector family. One possible remedy against this is to apply so-called batch FastICA, which consists in iterating equation (9) on a m-dimensional system of vectors, which is orthonormalized anew before each new iteration. In our practical experiments we did not observe any significant change in the results when using this refinement, so we mention this possibility only as a matter of precaution. We suspect two mitigating factors against this possible unwished behavior are that (1) it is known that FastICA does not converge to a global maximum, so that we probably find vectors in the vicinity of different local optima and (2) the "optimal" directions depend on the function f used and we combine a large number of such functions.

In the next section, we will describe the full algorithm, which consists in applying the procedure just described to different choices of the function f. Since we are using Projection Pursuit as a heuristic to find suitable parameters ω for a fixed f, the theoretical setting proposed here can therefore also be seen as a suitable framework for combining Projection Pursuit results when using different index functions f.

3.4 Full procedure

The previous sections have been devoted to detailing some key points of the procedure. We now gather these points and describe the full algorithm. We previously considered the case of a basis function family $h_\omega(y) = f(\langle \omega, y \rangle)$. We now consider a finite family of possible choices $\{f_k\}_{k=1}^L$ which are then combined.

In the implementation tested, we have used the following forms of the functions f_k:

$$f_\sigma^{(1)}(z) = z^3 \exp\left(-\frac{z^2}{2\sigma^2}\right), \qquad \text{(Gauss-Pow3)}$$

$$f_b^{(2)}(z) = \tanh(bz), \qquad \text{(Hyperbolic Tangent)}$$

$$f_a^{(3)}(z) = \exp(iaz), \qquad \text{(Fourier[4])}$$

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). Note that using z^3 and Hyperbolic Tangent functions is inspired by the classical PP algorithms (including FastICA) where these indices are used. We multiplied z^3 by a Gaussian factor in order to satisfy the boundedness assumption needed to control the estimation error (see Theorem 3.1 and 3.2 below). Furthermore the introduction of the parameter σ^2 allows for a richer family. Finally, the Fourier functions were introduced as they constitute a rich and important family. A pseudocode for the NGCA algorithm is described in Figure 4.

3.5 Theoretical bounds on the statistical estimation error

In this section we tackle the question of controlling the estimation error when approximating the vectors $\beta(h)$ by their empirical estimations $\widehat{\beta}(h)$ from a rigorous theoretical point of view. These results were derived with the following goals in mind:

[4] In practice, separated into real and complex parts (sine and cosine).

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

Whitening.
The data X_i is recentered by subtracting the empirical mean.
Let $\widehat{\Sigma}$ denote 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 $t = 1, \ldots, T$: *[FastICA loop]*
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_k be the trace of the empirical covarariance matrix of $\widehat{\beta}_T$:
$$N_k = \frac{1}{n} \sum_{i=1}^{n} \left\| \widehat{Y}_i f_k(\langle \omega_{T-1}, \widehat{Y}_i \rangle) - f_k'(\langle \omega_{T-1}, \widehat{Y}_i \rangle) \omega_{T-1} \right\|^2 - \left\| \widehat{\beta}_T \right\|^2 .$$
Store $v^{(k)} \leftarrow \widehat{\beta}_T * \sqrt{n/N_k}$. *[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 $\widehat{\mathcal{J}}$ be the space spanned by the first m principal directions.
Pull back in original space.
$\widehat{\mathcal{I}} = \widehat{\Sigma}^{-\frac{1}{2}} \widehat{\mathcal{J}}$.

Output: $\widehat{\mathcal{I}}$.

Figure 4: Pseudocode of the NGCA algorithm.

- A cornerstone of the algorithm is that we consider a whole family $h_1, \ldots, h_L$ of functions and pick selected members from it. In order to justify this from a statistical point of view, we therefore need to control the estimation error not for a single function h and the associated $\widehat{\beta}(h)$, but instead uniformly over the function family. For this, a simple control of e.g. the averaged squared deviation $\mathbb{E}\left[\|\beta - \widehat{\beta}\|^2\right]$ for each individual h is not sufficient: we need a stronger result, namely an exponential control of the deviation probability. This allows, by the union bound, to obtain a uniform control over the whole family with a mild (logarithmic) dependence on the cardinality of the family.

- We aim at making the covariance trace $\widehat{\sigma}^2$ directly appear into the main bounding terms of our error control. This provides a more solid justification to the renormalization scheme developed in section 3.2 where we had used arguments based on a non rigorous intuition. The choice to involve directly the *empirical* covariance in the bound instead of the population one was made to emphasize that estimation error for the covariance itself is also taken into account for the bound.

- While the control of the deviation of an empirical average of the form given in Eq.(5) is a very classical problem, we want to explicitly take into account the effect of the empirical whitening/dewhitening using the empirical covariance matrix $\widehat{\Sigma}$. This complicates matters noticeably since this whitening is itself data-dependent.

- Our goal was *not* to obtain tight confidence intervals or even exact asymptotical behavior. There is a number of ways in which our results could be substantially refined, for example obtaining uniform bounds over continuous (instead of finite) families of functions using covering number arguments; showing asymptotical uniform central limit properties for a precise study of the typical deviations, etc. Here we tried to obtain simple, while still mathematically rigorous, results, covering essential statistical foundations of our method: consistency and order of the convergence rate.

In the sequel, for a matrix A, we denote $\|A\|$ its operator norm.

Analysis of the estimation error with exact whitening. We start by considering an idealized case where whitening is done using the true covariance matrix Σ: $Y = \Sigma^{-\frac{1}{2}} X$.

In this case we have the following control of the estimation error:

Theorem 3.1 *Let $\{h_k\}_{k=1}^L$ be a family of smooth functions from $\mathbb{R}^d$ to $\mathbb{R}$. Assume that $\sup_{y,k} \max\left(\|\nabla h_k(y)\|, \|h_k(y)\|\right) < B$ and that X has covariance matrix Σ with $\left\|\Sigma^{-1}\right\| \leq K^2$, and is such that for some $\lambda_0 > 0$ the following inequality holds:*

$$\mathbb{E}\left[\exp\left(\lambda_0 \|X\|\right)\right] \leq a_0 < \infty. \tag{10}$$

Denote $\widetilde{h}(y) = yh(y) - \nabla h(y)$. Suppose $X_1, \ldots, X_n$ are i.i.d. copies of X and let $Y_i = \Sigma^{-\frac{1}{2}} X_i$. If we define

$$\widehat{\beta}_Y(h) = \frac{1}{n}\sum_{i=1}^n \widetilde{h}(Y_i) = \frac{1}{n}\sum_{i=1}^n Y_i h(Y_i) - \nabla h(Y_i), \tag{11}$$

and

$$\widehat{\sigma}_Y^2(h) = \frac{1}{n}\sum_{i=1}^{n}\left\|\widetilde{h}(Y_i) - \widehat{\beta}_Y(h)\right\|^2 , \tag{12}$$

then for any $\delta < \frac{1}{4}$ *, with probability at least* $1-4\delta$ *the following bounds hold simultaneously for all* $k \in \{1,\ldots,L\}$ *:*

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

and

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

where $dist(\gamma, \mathcal{I})$ *denotes the distance between a vector* γ *and the subspace* $\mathcal{I}$ *, and* C, C' *are constants depending only on the parameters* $(d, \lambda_0, a_0, B, K)$ *.*

Comments.

1. The above inequality tells us that the rate of convergence of the estimated vectors to the target space is in this case of order $n^{-1/2}$ (classical "parametric" rate). Furthermore, the theorem gives us an estimation of the relative size of the estimation error for different functions h through the empirical factor $\widehat{\sigma}_Y(h)$ in the principal term of the bound. As announced in our initial goals, this therefore gives a rigorous foundation to the intuition exposed in section 3.2 for vector renormalization.
2. Also following our goals, we obtained a uniform control of the estimation error over a finite family with a logarithmic dependence in the cardinality. This does not correspond exactly to the continuous families we use in practice but comes close enough if we consider adequate parameter discretization. We will comment on this in more detail after the next theorem.

Whitening using empirical covariance. When Σ is unknown (which is in general the case), we use instead the empirical covariance matrix $\widehat{\Sigma}$. Here we will show that under a somewhat stronger assumption on the distribution of X and on the functions h, we are still able to obtain a convergence rate of order at most $\sqrt{\log(n)/n}$ towards the index space $\mathcal{I}$.

Let us denote $\widehat{Y}_i = \widehat{\Sigma}^{-\frac{1}{2}} X_i$ the empirically whitened datapoints, $\widetilde{h}(y) = yh(y) - \nabla h(y)$ as previously, and

$$\widehat{\beta}_{\widehat{Y}}(h) = \frac{1}{n}\sum_{i=1}^{n}\widetilde{h}(\widehat{Y}_i) = \frac{1}{n}\sum_{i=1}^{n}\widehat{Y}_i h(\widehat{Y}_i) - \nabla h(\widehat{Y}_i) ; \tag{13}$$

finally, let us denote

$$\widehat{\gamma}(h) = \widehat{\Sigma}^{-\frac{1}{2}}\widehat{\beta}_{\widehat{Y}}(h) , \qquad \text{and} \qquad \widehat{\sigma}_{\widehat{Y}}^2(h) = \frac{1}{n}\sum_{i=1}^{n}\left\|\widetilde{h}(\widehat{Y}_i) - \widehat{\beta}_{\widehat{Y}}(h)\right\|^2 .$$

We then have the following theorem:

Theorem 3.2 *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* $\left\|\Sigma^{-1}\right\| \leq K^2$;
(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* $k \in \{1, \dots, L\}$ *:*

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

and

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

where C_1, C_1' *are constants depending on parameters* $(\lambda_0, a_0, B, K, M)$ *only and* C_2, C_2' *on* $(d, \lambda_0, a_0, B, K, M)$ *.*

Comments.

1. Theorem 3.2 implies that the vectors $\widehat{\gamma}(h_k)$ obtained from any $h(x)$ converge to the unknown non-Gaussian subspace $\mathcal{I}$ *uniformly* at a rate of order $\sqrt{\log(n)/n}$.
2. The condition (i) is a restrictive assumption as it excludes some densities with heavy tails. We are considering weakening this assumption in future developments.
3. In the actual algorithm we consider a family of functions of the form $h_\omega(x) = f(\langle \omega, x\rangle)$, with ω on the unit sphere of $\mathbb{R}^d$. Suppose we approximate ω by its nearest neighbor $\widetilde{\omega}$ on a regular grid of scale ε . Then we only have to apply the bound to a discretized family of size $L = \mathcal{O}(\varepsilon^{1-d})$, giving rise only to an additional factor in the bound of order $\sqrt{d \log \varepsilon^{-1}}$. Taking for example $\varepsilon = 1/n$ (the fact that the function family depends on n is not a problem since the bounds are valid for any fixed n), this ensures convergence of the discretized functions to the initial continuous family while introducing only in an additional factor $\sqrt{d \log n}$ in the bound: this does not change fundamentally the order of the bound since there is already another $\sqrt{d \log n}$ term present.
4. For both Theorems 3.1 and 3.2 we have given bounds for estimation of both $\mathcal{I}$ and $\mathcal{J}$, that is, in terms of the initial data and of the "whitened" data. The result in terms of the initial data ensures the overall consistency of the approach, but the convergence in the whitened space is equally interesting since we use it as the main working space for the algorithm.
5. Comparing to Theorem 3.1 obtained for exact whitening, we see in the present case that there is an additional term of principal order in n coming from the estimation error of Σ , with a multiplicative factor which unfortunately is not known accurately.

This means that the renormalization scheme is not completely justified in this case, although we feel the idealized situation of Theorem 3.1 already provides some strong argument in this direction. However, the present result suggests that the accuracy of the normalization could probably be further improved.

4 Numerical results

Parameters used: *All* of the experiments presented where obtained with exactly the same set of parameters: $a \in [0,4]$ for the Fourier functions; $b \in [0,5]$ for the Hyperbolic Tangent functions; $\sigma^2 \in [0.5,5]$ for the Gauss-pow3 functions. Each of these ranges was divided into 1000 equispaced values, thus yielding a family $\{f_k\}$ of size 4000 (Fourier functions count twice because of the sine and cosine parts). The preliminary calibration procedure described in the end of section 3.2 suggested to take $\varepsilon = 1.5$ as the threshold under which vectors are not informative (strictly speaking, the threshold should be calibrated separately for each function f but we opted here for a single threshold for simplicity). Finally we fixed the number of FastICA iterations $T = 10$. With this choice of parameters and 1000 data points in the sample, the computation time is typically of the order of less than 10 seconds on a modern PC under our Matlab implementation.

4.1 Tests in a controlled setting

For testing our algorithm and comparing it with PP, we performed numerical experiments using various synthetic data. Here we report exemplary results using the following 4 data sets. Each data set includes 1000 samples in 10 dimension and each sample consists of 8-dimensional independent standard Gaussian. Other 2 non-Gaussian components are as follows.

(A) Simple Gaussian Mixture: 2-dimensional independent Gaussian mixtures with density of each component given by

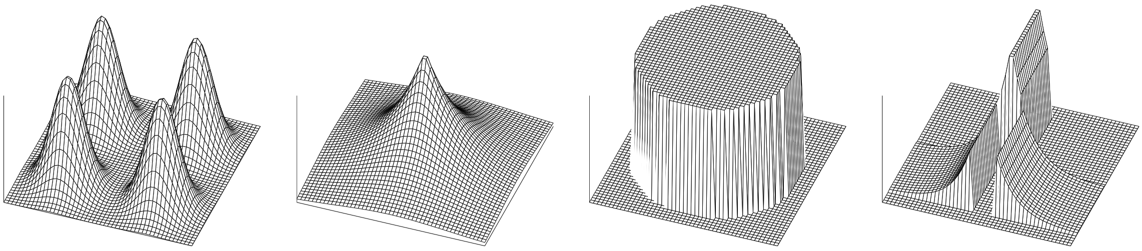
$$\frac{1}{2}\phi_{-3,1}(x) + \frac{1}{2}\phi_{3,1}(x), \tag{14}$$

(B) Dependent super-Gaussian: 2-dimensional isotropic distribution with density proportional to $\exp(-\|x\|)$.

(C) Dependent sub-Gaussian: 2-dimensional isotropic uniform with constant positive density for $\|x\| \leq 1$ and 0 otherwise.

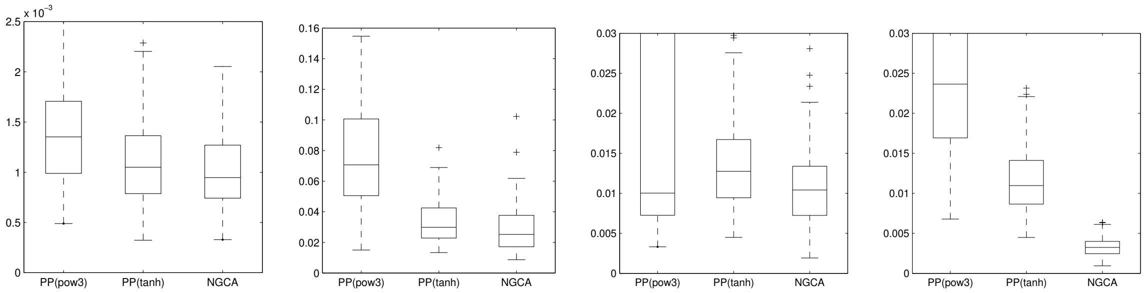
(D) Dependent super- and sub-Gaussian: 1-dimensional Laplacian with density proportional to $\exp(-|x_{Lap}|)$ and 1-dimensional dependent uniform $U(c, c+1)$, where $c = 0$ for $|x_{Lap}| \leq \log 2$ and $c = -1$ otherwise.

For each of these situations, the non-Gaussian components are additionally rescaled coordinatewise by a fixed factor so that each coordinate has unit variance. The profiles of the density functions of the non-Gaussian components in the above data sets are described in Figure 5.



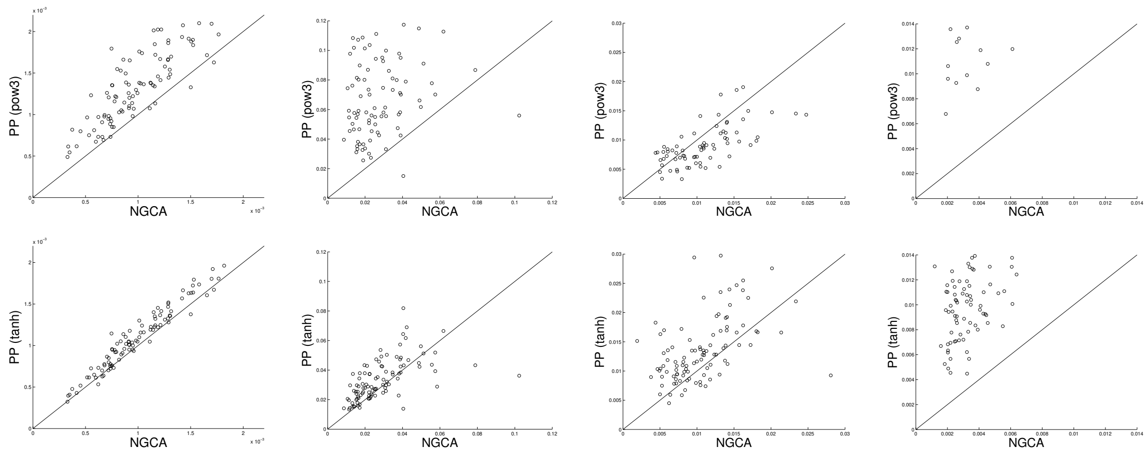
(A) (B) (C) (D)

Figure 5: Densities of non-Gaussian components. The data sets are: (a) 2D independent Gaussian mixtures, (b) 2D isotropic super-Gaussian, (c) 2D isotropic uniform and (d) dependent 1D Laplacian + 1D uniform.



(A) (B) (C) (D)

Figure 6: Boxplots of the error criterion $\mathcal{E}(\widehat{\mathcal{I}}, \mathcal{I})$.



(A) (B) (C) (D)

Figure 7: Performance comparison plots (for error criterion $\mathcal{E}(\widehat{\mathcal{I}}, \mathcal{I})$) of NGCA versus FastICA; top: versus pow3 index; bottom: versus tanh index.

We compare the following three methods in the experiments: PP with 'pow3' or 'tanh' index[5] (denoted by PP(pow3) and PP(tanh), respectively), and the proposed NGCA.

Figure 6 shows boxplots of the error criterion $\mathcal{E}(\widehat{\mathcal{I}}, \mathcal{I})$ defined in Eq.(3) obtained from 100 runs. Figure 7 shows comparison of the errors obtained by different methods for each individual trial. Because PP tends to get trapped into local optima of the index function it optimizes, we restarted it 10 times with random starting points and took the subspace obtaining the best index value. However, even when it is restarted 10 times, PP (especially with the 'pow3' index) still gets caught in local optima in a small percentage of cases (we also tried up to 500 restarts but it led to negligible improvement).

For the simplest data set (A), NGCA is comparable or slightly better than PP methods. It is known that PP(tanh) is suitable for finding super-Gaussian components (heavy-tailed distribution) while PP(pow3) is suitable for finding sub-Gaussian components (light-tailed distribution) [Hyvärinen et al., 2001]. This can be observed in the data sets (B) and (C): PP(tanh) works well for the data set (B) and PP(pow3) works well for the data set (C), although the upper-quantile is very large for the data set (C) (because of PP getting trapped in local minima). The plots of Figure 7 confirm that NGCA is on average on par or slightly better than PP with 'correct' non-Gaussianity index without having to prefix such a non-Gaussianity index. For the data set (C), NGCA appears to be very slightly worse than PP(pow3) (excluding those cases where PP fails due to local minima: the corresponding points are outside the range of the figure), but this appears hardly significant. The superiority of the index adaptation feature of NGCA can be clearly observed in the data set (D), which includes both sub- and super-Gaussian components. Because of this composition, there is no single best non-Gaussianity index for this data set, and the proposed NGCA gives significantly lower error than that of either PP method.

Failure modes. We now try to explore the limits of the method and the conditions under which estimation of the target space will fail. First, we study the behaviour of NGCA again compared with PP as the total dimension of the data increases. We use the same synthetic data sets with 2-dimensional non-Gaussian components, while the number of Gaussian components increases. The averaged errors over 100 experiments are depicted in Figure 8. In all cases we seem to observe a sharp phase transition between a good behaviour regime and a failure mode where the procedure is unable to estimate the correct subspace. In 3 out of 4 cases however we observe that the phase transition to the failure mode occurs for a higher dimension for NGCA than for the PP methods, which indicates better robustness of NGCA.

In the synthetic data sets used so far, the data was always generated with a covariance matrix equal to identity. Another interesting setting to study is the robustness with respect to bad conditioning of the covariance matrix. We consider again a fixed-dimension setting, with 2 non-Gaussian and 8 gaussian dimensions.

While the non-Gaussian coordinates always have variance unity, the standard deviation of the 8 Gaussian dimensions now follows the geometrical progression $10^{-r}, 10^{-r+2r/7}, \ldots, 10^{r}$. Thus, the higher r, the worse conditioned is the total covariance

[5] We used the deflation mode of the FastICA code [Hyvärinen et al., 2001] as an implementation of PP. The 'pow3' flavor is equivalent to a kurtosis based index, i.e. in this case, FastICA iteratively maximizes the kurtosis. On the other hand, the 'tanh' flavor uses a robust index which is appropriate in particular for heavy-tailed data.

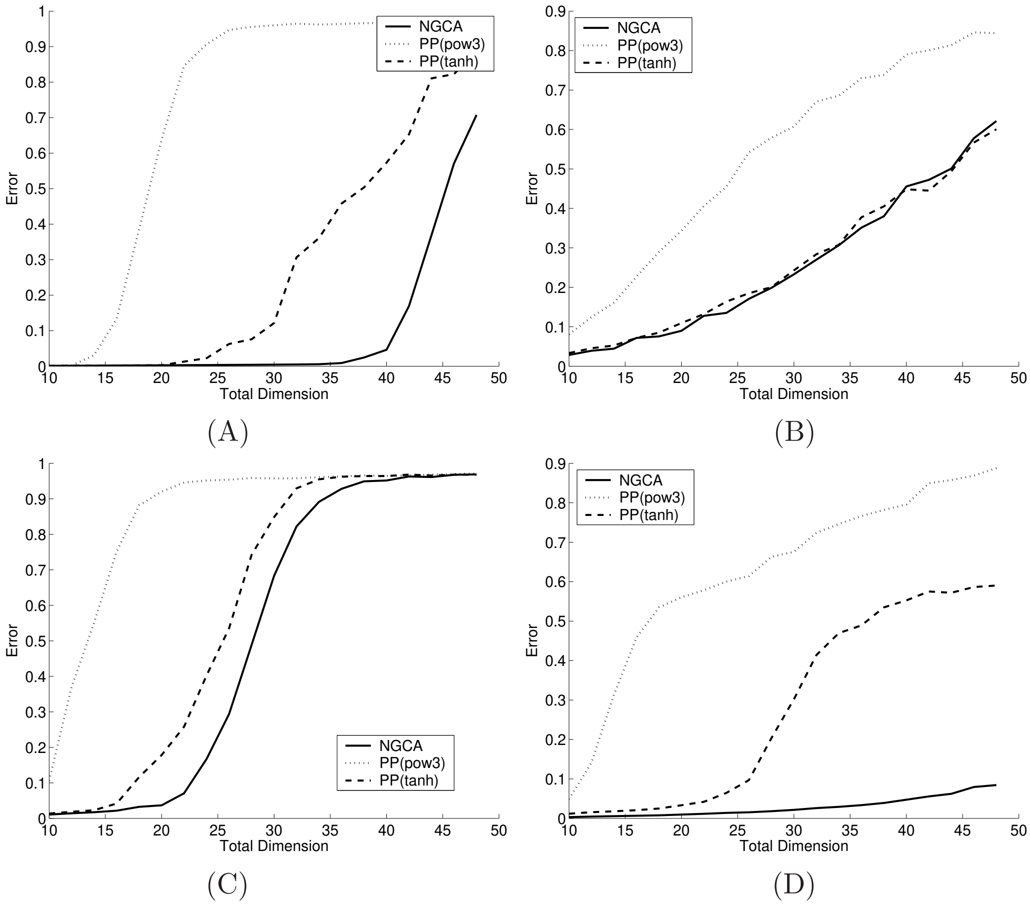


Figure 8: Results when the total dimension of the data increases.

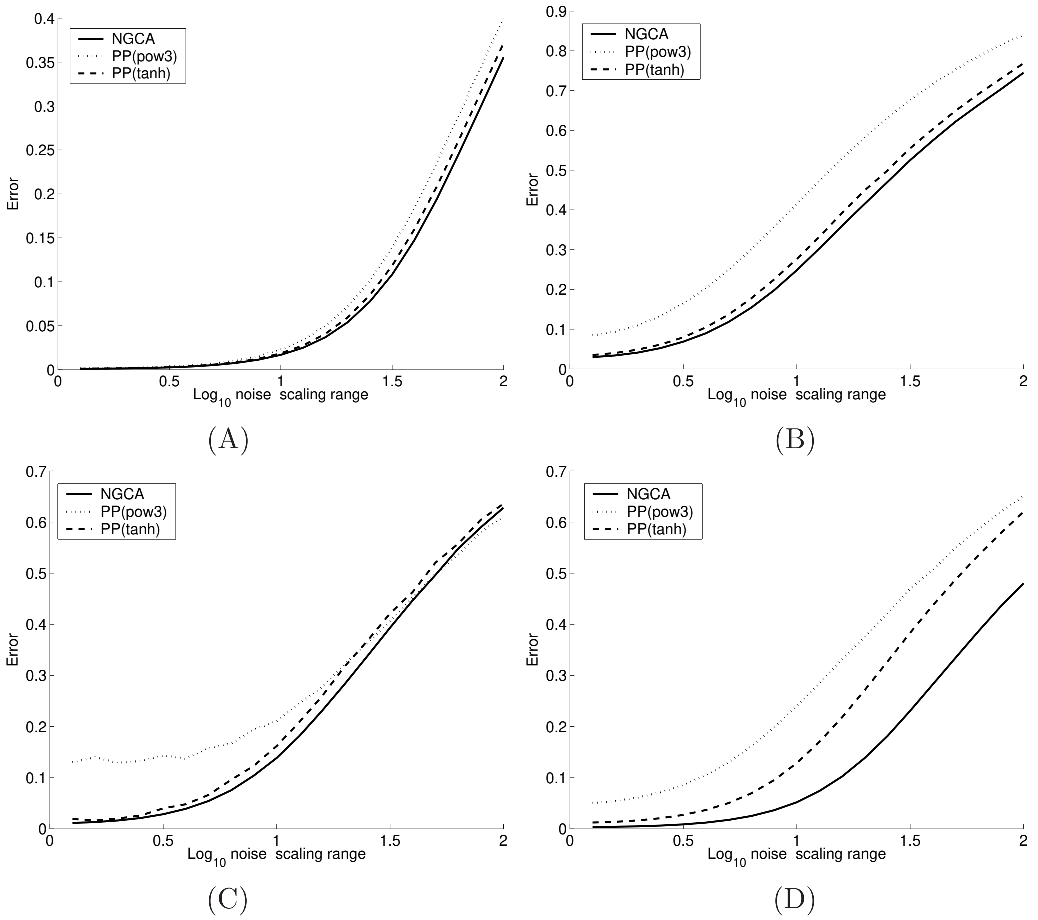


Figure 9: Results when the Gaussian (noise) components have different scales (the standard deviations follow a geometrical progression on $[10^{-r}, 10^{r}]$ where r is the parameter on the abscissa).

matrix.

The results are depicted in Figure 9, where we observe again a transition to a failure mode when the covariance matrix is too badly conditioned. Although NGCA still appears as the best method, we observe that on 3 out of 4 data sets the transition to failure mode seems to happen roughly at the same point as for PP methods. This suggests that there is no or only little added robustness of NGCA with respect to PP in this regard. However, this result is not entirely surprising, as we expect this type of failure mode to be caused by a too large estimation error in the covariance matrix and therefore in the whitening/dewhitening steps. Since these steps are common to NGCA and the PP algorithms, it seems logical to expect a parallel evolution of their errors.

4.2 Example of application for realistic data: visualization and clustering

We now give an example of application of our methodology to visualization and clustering of realistic data. We consider here "oil flow" data which has been obtained by numerical simulation of a complex physical model. This data was already used before for testing techniques of dimension reduction [Bishop et al., 1998]. The data is 12-dimensional and it is not known a priori if some dimensions are more relevant. Here our goal is to visualize the data and possibly exhibit a clustered structure. Furthermore it is known that the data is divided into 3 classes. We show classes in different colors but the class information is not used in finding the directions (i.e. the process is unsupervised).

We compare the NGCA methodology described in the previous section, and PP ("vanilla" FastICA) using the tanh or the pow3 index. The results are shown on Figure 10. A 3D projection of the data was computed using these methods, which was in turn projected in 2D to draw the figure; this last projection was chosen manually so as to make the cluster structure as visible as possible in each case.

We see that the NGCA methodology gives a much more relevant projection than PP using either tanh or pow3 alone: we can distinguish 10-11 clusters versus at most 5 for the PP methods. Furthermore the classes are clearly separated only on the NGCA projection, on the other ones they are partially confounded in one single cluster. Finally, we confirm, by applying the projection found to held-out test data (i.e. data not used to determine the projection), that the cluster structure is relevant and not due to some overfitting artifact. This, in passing, shows one advantage of a linear projection method, namely that it can be extended to new data in a straightforward way.

Presumably, an important difference between the NGCA projection and the others comes from the Fourier functions, since they are not present in either of the PP methods. It can be confirmed by looking at the vector norms that Fourier functions are more relevant for this data set; they gave rise to estimated vectors with generally higher norms and had consequently a sizable influence of the choice of the projection. One could object that we have been merely lucky for this specific data because Fourier functions happened to be more relevant, and neither PP method uses this index. One could suggest, for fair comparison, to use the PP algorithm with a Fourier index. However, beside the fact that this index is not generally used in classical PP methods, the results would be highly dependent of the specific frequency parameter chosen, so we did not make experiments in that

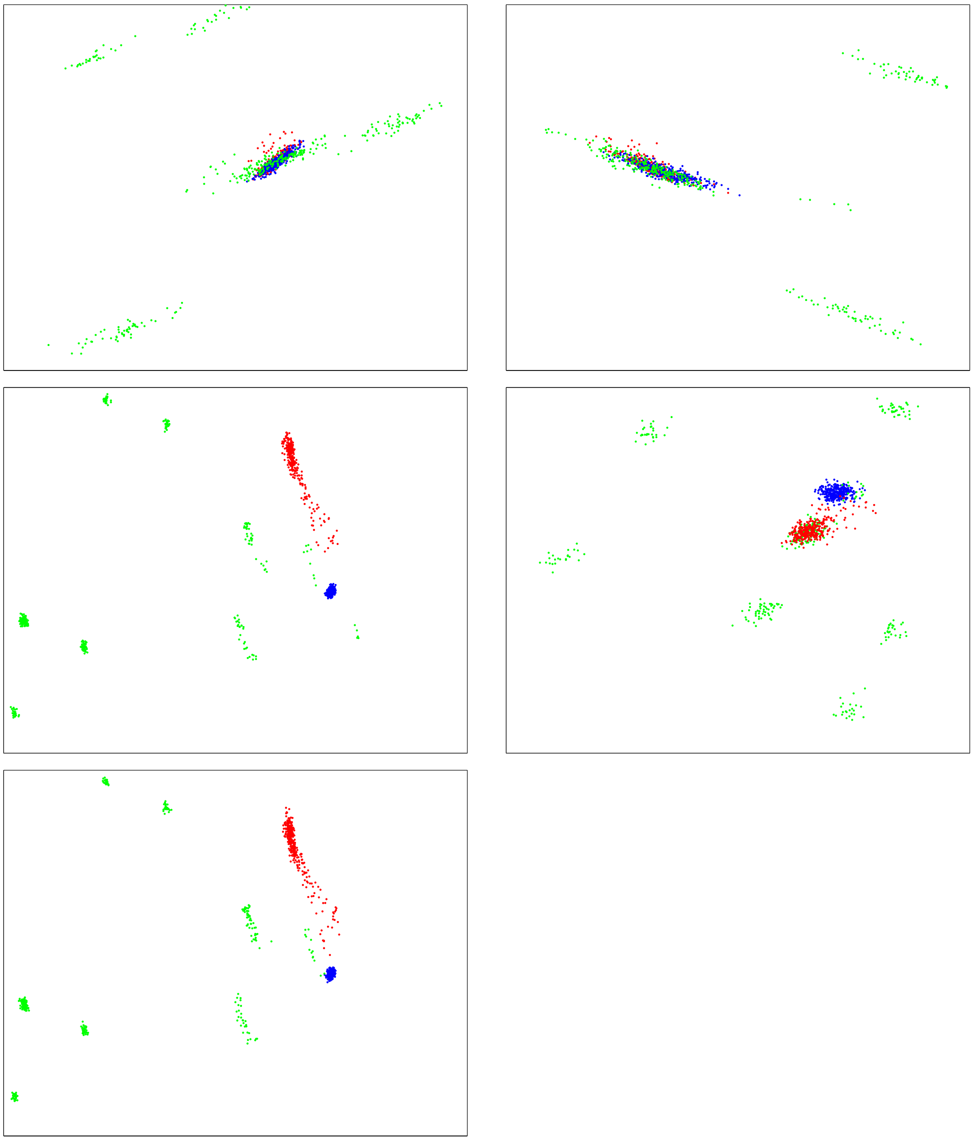


Figure 10: 2D projection of the "oil flow" (12-dimensional) data obtained by different algorithms. Top: vanilla FastICA methods (left: tanh index, right: pow3 index). Middle: on the left, projection obtained by NGCA (using a combination of Fourier, tanh, Gauss-pow3 indices). On the right is the projection obtained by applying NGCA without the Fourier index. Finally, the bottom panel shows the projection of additional test data using the projection found by NGCA on the middle-left panel. In each case (except the bottom panel), the data was first projected in 3D using the respective methods (without class information), from which a 2D projection was chosen visually so as to yield the clearer cluster structure. For the bottom panel, the 2D projection found from the middle-left panel was used to visualize additional test data.

direction (by contrast, the NGCA methodology allows to combine vectors obtained from different frequencies). On the other hand, another route to investigate the relevance of this objection is to look at the results obtained by the NGCA method if Fourier functions are *not* used – thus only considering Gauss-pow3 and tanh. In this case we still hope an improvement over PP because NGCA is combining indices (as well as combining over the parameters ranges σ^2 and b). This is confirmed in Figure 10: even without the relevant Fourier functions, NGCA yields a projection where 8 clusters can be distinguished, and the classes are much more clearly separated than with PP methods. Finally, a visual comparison with the results obtained by Bishop et al. [1998] demonstrated that the projection found by our algorithm exhibits a clearer clustered structure; moreover ours is a purely *linear* projection whereas the latter reference was a nonlinear data representation

Further analysis on clustering performance with additional data sets are given in the Appendix and underline the usefulness of our method.

5 Conclusion

We proposed a new semi-parametric framework for constructing a linear projection to separate an uninteresting multivariate Gaussian 'noise' subspace of possibly of large amplitude from the 'signal-of-interest' subspace. Our theory provides generic consistency results on how well the non-Gaussian directions can be identified (Theorem 3.2). To estimate the non-Gaussian subspace from the set of vectors obtained, PCA is finally performed after suitable renormalization and elimination of uninformative vectors. The key ingredient of our NGCA method is to make use of the *gradient* computed for the nonlinear basis function $h(x)$ in Eq.(11) after data whitening. Once the low-dimensional 'signal' part is extracted, we can use it for a variety of applications such as data visualization, clustering, denoising or classification.

Numerically we found comparable or superior performance to, e.g., FastICA in deflation mode as a generic representative of the family of PP algorithms. Note that in general PP methods need to pre-specify a projection index with which they search non-Gaussian components. By contrast, an important advantage of our method is that we are able to simultaneously use several families of nonlinear functions; moreover, also inside a same function family we are able to use an entire range of parameters (such as frequency for Fourier functions). Thus, our new method provides higher flexibility, and less restricting assumptions *a priori* on the data. In a sense, the functional indices that are the most relevant for the data at hand are automatically selected.

Future research will adapt the theory to simultaneously estimate the dimension of the non-Gaussian subspace. Extending the proposed framework to non-linear projection scenarios [Cox and Cox, 1994, Schölkopf et al., 1998, Roweis and Saul, 2000, Tenenbaum et al., 2000, Belkin and Niyogi, 2003, Harmeling et al., 2003] and to finding the most discriminative directions using labels are examples for which the current theory could be taken as a basis.

Acknowledgements: The authors would like to thank Stefan Harmeling for discussions and J.-F. Cardoso for suggesting us the pre-whitening step for increased efficiency and robustness. We would also like to thank anonymous reviewers for many insightful comments, in particular pointing out the ICA interpretation. We acknowledge partial financial support by DFG, BMBF and the EU NOE PASCAL (EU # 506778). GB and MS also thank Alexander von Humboldt foundation for partial financial support.

A Proofs of the theorems

A.1 Proof of Lemma 2.1

Suppose first that the noise N is standard normal. Denote by Π_E the projector from $\mathbb{R}^d$ to $\mathbb{R}^m$ which corresponds to the subspace E. Let also $E^\perp$ be the subspace complementary to E and $\Pi_{E^\perp}$ mean the projector on $E^\perp$. The standard normal noise can be decomposed as $N = N_1 \dotplus N_2$ where $N_1 = \Pi_E N$ and $N_2 = \Pi_{E^\perp} N$ are independent noise components. Similarly, the signal X can be decomposed as

$$X = (\Pi_E S + N_1) \dotplus N_2$$

where we have used the model assumption that the signal S is concentrated in E and it is independent of N. It is clear that the density of $\Pi_E S + N_1$ in $\mathbb{R}^m$ can be represented as the product $g(x_1)\phi(x_1)$ for some function g and the standard normal density $\phi(x_1)$, $x_1 \in \mathbb{R}^m$. The independence of N_1 and N_2 yields the in the similar way for $x = (x_1, x_2)$ with $x_1 = \Pi_E x$ and $x_2 = \Pi_{E^\perp} x$ that $p(x) = g(x_1)\phi(x_1)\phi(x_2) = g(x_1)\phi(x)$. Note that for the linear mapping $T = \Pi_E$ characterizes the signal subspace E. Namely, E is the image $\Im(T^*)$ of the dual operator T^* while $E^\perp$ is the null subspace (kernel) of T: $E^\perp = \mathfrak{K}(T)$.

Next we drop the assumption of the standard normal noise and assume only that the covariance matrix Γ of the noise is nondegenerated. Multiplying the both sides of the equation (1) by the matrix $\Gamma^{-1/2}$ leads to $\Gamma^{-1/2}X = \Gamma^{-1/2}S + \widetilde{N}$ where $\widetilde{N} = \Gamma^{-1/2}N$ is standard normal. The transformed signal $\widetilde{X} = \Gamma^{-1/2}S$ belongs to the subspace $\widetilde{E} = \Gamma^{-1/2}E$. Therefore, the density of $\widetilde{X}$ can be represented as $p(\widetilde{x}) = \widetilde{g}(\Pi_{\widetilde{E}}\widetilde{x})\phi(\widetilde{x})$ where $\Pi_{\widetilde{E}}$ is the projector corresponding to $\widetilde{E}$. Coming back the variable x yields the density of X in the form $p(x) = g(Tx)\phi(\Gamma^{-1/2}x)$ where $T = \Pi_{\widetilde{E}}\Gamma^{-1/2}$.

A.2 Proof of Proposition 2.2

For any function $\psi(x)$, it holds that

$$\int \psi(x+u)p(x)dx = \int \psi(x)p(x-u)dx,$$

if the integrals exists. Under mild regularity conditions on $p(x)$ and $\psi(x)$ allowing differentiation under the integral sign, differentiating this with respect to u gives

$$\int \nabla\psi(x)p(x)dx = -\int \psi(x)\nabla p(x)dx. \tag{15}$$

Let us take the following function

$$\psi_h(x) := h(x) - x^\top \mathbb{E}\left[Xh(X)\right],$$

whose gradient is

$$\nabla\psi_h(x) = \nabla h(x) - \mathbb{E}\left[Xh(X)\right].$$

The vector $\beta(h)$ is the expectation of $-\nabla\psi_h$. From Eq.(15) and using $\nabla p(x) = \nabla \log p(x)\, p(x)$, we have

$$\beta(h) = \int \psi_h(x)\nabla\log p(x)\, p(x)dx.$$

Applying Eq.(2) to the above equation yields

$$\begin{aligned}
\beta(h) &= \int \psi_h(x)\nabla\log g(Tx)\; p(x)dx - \int \psi_h(x)\Gamma^{-1}x\, p(x)dx \\
&= T^* \int \psi_h(x)\nabla g(Tx)\phi_{\theta,\Gamma}(x)dx - \Gamma^{-1}\!\!\int x\psi_h(x)p(x)dx. \qquad (16)
\end{aligned}$$

Under the assumption $\mathbb{E}\left[XX^\top\right] = I_d$, we get

$$\mathbb{E}\left[X\psi_h(X)\right] = \mathbb{E}\left[Xh(X)\right] - \mathbb{E}\left[XX^\top\right]\mathbb{E}\left[Xh(X)\right] = 0,$$

that is, the second term of Eq.(16) vanishes. Since the first term of Eq.(16) belongs to $\mathcal{I}$ by the definition of $\mathcal{I}$, we finally have $\beta(h) \in \mathcal{I}$.

A.3 Proof of Theorem 3.1

For a fixed function h, we will essentially apply Lemma A.1 stated below for each coordinate of $\beta_Y(h)$. Denoting the k-th coordinate of a vector v by $v^{(k)}$, and $y = \Sigma^{-\frac{1}{2}}x$, we have

$$\widetilde{h}^{(k)}(x) = \left|\left[\nabla h(y) - yh(y)\right]^{(k)}\right| \le B(1 + \|y\|) \le B\left(1 + K\|x\|\right).$$

It follows that $\widetilde{h}^{(k)}(x)$ is such that

$$\mathbb{E}\left[\exp\left(\frac{\lambda_0}{BK}\widetilde{h}^{(k)}(x)\right)\right] \le a_0 \exp\left(\frac{\lambda_0}{K}\right),$$

and hence satisfies the assumption of Lemma A.1. Denoting by $\widehat{\sigma}_k^2$ the sample variance of $\widetilde{h}^{(k)}$, we apply the lemma with $\delta' = \delta/d$, obtaining by the union bound that with probability at least $1-4\delta$, for all $1 \le k \le d$:

$$\left(\left[\beta_Y - \widehat{\beta}_Y\right]^{(k)}\right)^2 \le 4\widehat{\sigma}_k^2 \frac{\log\left(d\delta^{-1}\right)}{n} + C_1(\lambda_0, a_0, B, d, K)\frac{\log^2(n\delta^{-1})\log^2\delta^{-1}}{n^{\frac{3}{2}}},$$

where we have used the inequality $(a+b)^2 \le 2(a^2+b^2)$ and C_1 denotes some function depending only on the indicated quantities. Now summing over the coordinates, taking the square root and using $\sqrt{a+b} \le \sqrt{a} + \sqrt{b}$ leads to:

$$\left\|\beta_Y - \widehat{\beta}_Y\right\| \le 2\sqrt{\widehat{\sigma}_Y^2(h)\frac{\log\delta^{-1} + \log d}{n}} + C_2(\lambda_0, a_0, B, d, K)\left(\frac{\log(n\delta^{-1})\log\delta^{-1}}{n^{\frac{3}{4}}}\right), \qquad (17)$$

with probability at least $1-4\delta$. To turn this into a uniform bound over the family $\{h_k\}_{k=1}^L$, we simply apply this inequality separately to each function in the family with $\delta'' = \delta/L$. This leads to the first announced inequality of theorem. We obtain the second one by multiplying the first by $\Sigma^{-\frac{1}{2}}$ to the left and using the assumption on $\left\|\Sigma^{-1}\right\|$.

Lemma A.1 *Let X be a real random variable such that for some $\lambda_0 > 0$:*

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

Let $X_1, \ldots, X_n$ denote an i.i.d. sequence of copies of X. Let $\mu = \mathbb{E}[X]$, $\widehat{\mu} = \frac{1}{n}\sum_{i=1}^n X_i$ and $\widehat{\sigma}^2 = \frac{1}{2n(n-1)}\sum_{i\neq j}(X_i - X_j)^2$ be the sample variance.

Then for any $\delta < \frac{1}{4}$ the following holds with probability at least $1-4\delta$, where c is a universal constant:

$$|\mu - \widehat{\mu}| \le \sqrt{\frac{2\widehat{\sigma}^2 \log\delta^{-1}}{n}} + c\lambda_0^{-1}\max\left(1, \log\left(na_0\delta^{-1}\right)\right)\left(\left(\frac{\log\delta^{-1}}{n}\right)^{\frac{3}{4}} + \frac{\log\delta^{-1}}{n}\right).$$

Proof. For $A > 0$ denote $X^A = X\mathbf{1}\{|X| \le A\}$. We decompose

$$\left|\frac{1}{n}\sum_{i=1}^n X_i - \mu\right| \le \left|\frac{1}{n}\sum_{i=1}^n \left(X_i - X_i^A\right)\right| + \left|\frac{1}{n}\sum_{i=1}^n X_i^A - \mathbb{E}\left[X^A\right]\right| + \left|\mathbb{E}\left[X - X^A\right]\right|;$$

these three terms will be denoted by T_1, T_2, T_3. By Markov's inequality, it holds that

$$\mathbb{P}\left[|X| > t\right] \le a_0 \exp\left(-\lambda_0 t\right),$$

Fixing $A = \log\left(n\delta^{-1}a_0\right)/\lambda_0$ for the rest of the proof, it follows by taking $t = A$ in the above inequality that for any $1 \le i \le n$:

$$\mathbb{P}\left[X_i^A \neq X_i\right] \le \frac{\delta}{n}.$$

By the union bound, we then have $X_i^A = X_i$ for all i, and therefore $T_1 = 0$, except for a set Ω_A of probability bounded by δ.

We now deal with the third term: we have

$$\begin{aligned}
T_3 = \left|\mathbb{E}\left[X\mathbf{1}\{|X| > A\}\right]\right| \le \mathbb{E}\left[X\mathbf{1}\{X > A\}\right] &= \int_0^\infty \mathbb{P}\left[X\mathbf{1}\{X > A\} > t\right] dt \\
&\le A\mathbb{P}\left[X > A\right] + \int_A^\infty a_0 \exp\left(-\lambda_0 t\right) dt \\
&\le a_0\left(A + \lambda_0^{-1}\right)\exp\left(-\lambda_0 A\right) \\
&= \frac{\delta}{n\lambda_0}\left(1 + \log\left(n\delta^{-1}a_0\right)\right).
\end{aligned}$$

Finally, for the second term, since $\left|X^A\right| \le A = \lambda_0^{-1}\log\left(n\delta^{-1}a_0\right)$, Bernstein's inequality ensures that with probability as least $1-2\delta$ the following holds:

$$\left|\frac{1}{n}\sum_{i=1}^n X_i^A - \mathbb{E}\left[X^A\right]\right| \le \sqrt{\frac{2\mathrm{Var}\left[X^A\right]\log\delta^{-1}}{n}} + 2\,\frac{\log\left(n\delta^{-1}a_0\right)\log\delta^{-1}}{\lambda_0 n}.$$

We finally turn to the estimation of $\mathrm{Var}\left[X^A\right]$. The sample variance of X^A is given by

$$(\widehat{\sigma}^A)^2 = \frac{1}{2n(n-1)} \sum_{i \neq j} \left(X_i^A - X_j^A\right)^2 ;$$

Note that $(\widehat{\sigma}^A)^2$ is an unbiased estimator of $\mathrm{Var}\left[X^A\right]$. Furthermore, replacing X_i^A by $X_i'^A$ in the above expression changes this quantity at most of $4A^2/n$ since X_i^A appears only in $2(n-1)$ terms. Therefore, application of the bounded difference (a.k.a. McDiarmid's) inequality [McDiarmid , 1989] to the random variable $\widehat{\sigma}^A$ yields that with probability $1-\delta$ we have

$$\left|(\widehat{\sigma}^A)^2 - \mathrm{Var}\left[X^A\right]\right| \leq 4A^2 \sqrt{\frac{\log \delta^{-1}}{n}} ;$$

finally, except for samples in the set Ω_A which we have already excluded above, we have $\widehat{\sigma}^A = \widehat{\sigma}$. Gathering these inequalities lead to the conclusion.

A.4 Proof of Theorem 3.2

In this proof we will denote by $C(\cdot)$ a factor depending only on the quantities inside the parentheses, and whose exact value can vary from line to line.

From Lemmas B.4 and B.5 below, we conclude that for big enough n, the following inequality is satisfied with probability $1-2/n$:

$$\left\| \Sigma^{-\frac{1}{2}} - \widehat{\Sigma}^{-\frac{1}{2}} \right\| \leq C(a_0, \lambda_0, K) \sqrt{\frac{d \log n}{n}} ; \tag{18}$$

also, it is a a weaker consequence of Lemmas B.2 and B.3 that the following inequalities hold with probability at least $1-1/n$ each (again for n big enough):

$$\frac{1}{n} \sum_{i=1}^{n} \|X_i\| \leq C(a_0, \lambda_0) , \tag{19}$$

$$\frac{1}{n} \sum_{i=1}^{n} \|X_i\|^2 \leq C(a_0, \lambda_0) . \tag{20}$$

Let us denote Ω the set of samples where (18), (19) and (20) are satisfied simultaneously; from the above we conclude that for large enough n the set Ω contains the sample with probability at least $1-4/n$. For the remainder of the proof we suppose that this condition is satisfied.

For any function h, we have

$$\left\| \widehat{\beta}_{\widehat{Y}} - \beta_Y \right\| \leq \left\| \widehat{\beta}_{\widehat{Y}} - \widehat{\beta}_Y \right\| + \left\| \widehat{\beta}_Y - \beta_Y \right\|$$

Note that (up to some changes in the constants) the assumption on the Laplace transform is stronger than the assumption of Theorem 3.1; hence equation (17) in the proof of this theorem holds and we have with probability at least $1-4\delta$, for any function in the family $\{h_k\}_{k=1}^{L}$:

$$\left\|\beta_Y - \widehat{\beta}_Y\right\| \leq 2\sqrt{\widehat{\sigma}^2_Y(h)\frac{\log\left(L\delta^{-1}\right)+\log d}{n}} + C(\lambda_0, a_0, B, d, K)\left(\frac{\log(nL\delta^{-1})\log\left(L\delta^{-1}\right)}{n^{\frac{3}{4}}}\right). \tag{21}$$

On the other hand, conditions (18) and (19) imply that for any function h in the family,

$$\begin{aligned}
\left\|\widehat{\beta}_{\widehat{Y}} - \widehat{\beta}_Y\right\| = \left\|\frac{1}{n}\sum_{i=1}^{n}\left(\widetilde{h}(\widehat{Y}_i) - \widetilde{h}(Y_i)\right)\right\| &\leq \frac{M}{n}\sum_{i=1}^{n}\left\|\widehat{Y}_i - Y_i\right\| \\
&\leq \frac{M}{n}\left\|\Sigma^{-\frac{1}{2}} - \widehat{\Sigma}^{-\frac{1}{2}}\right\|\sum_{i=1}^{n}\|X_i\| \\
&\leq C(a_0, \lambda_0, K)M\sqrt{\frac{d\log n}{n}}.
\end{aligned}$$

where in the first inequality, we have used the Lispchitz assumption on the function h.

One remaining technicality is to replace the term $\widehat{\sigma}_Y(h)$ (which cannot be evaluated from the data since it depends on the exactly whitened data Y_i) in (21) by $\widehat{\sigma}_{\widehat{Y}}(h)$ which can be evaluated from the data. For this use the following, holding for any function h in the family:

$$\left|\widehat{\sigma}^2_Y(h) - \widehat{\sigma}^2_{\widehat{Y}}(h)\right| = \frac{1}{2n(n-1)}\left|\sum_{i\neq j}\left\|\widetilde{h}(Y_i) - \widetilde{h}(Y_j)\right\|^2 - \left\|\widetilde{h}(\widehat{Y}_i) - \widetilde{h}(\widehat{Y}_j)\right\|^2\right|;$$

let us now focus on one term of the above sum:

$$\begin{aligned}
&\left\|\widetilde{h}(Y_i) - \widetilde{h}(Y_j)\right\|^2 - \left\|\widetilde{h}(\widehat{Y}_i) - \widetilde{h}(\widehat{Y}_j)\right\|^2 \\
&\qquad = \left(\widetilde{h}(Y_i) - \widetilde{h}(\widehat{Y}_i) - \widetilde{h}(Y_j) + \widetilde{h}(\widehat{Y}_j)\right)^\top\left(\widetilde{h}(Y_i) - \widetilde{h}(Y_j) + \widetilde{h}(\widehat{Y}_i) - \widetilde{h}(\widehat{Y}_j)\right) \\
&\qquad \leq M^2\left(\left\|Y_i - \widehat{Y}_i\right\| + \left\|Y_j - \widehat{Y}_j\right\|\right)\left(\|Y_i - Y_j\| + \left\|\widehat{Y}_i - \widehat{Y}_j\right\|\right) \\
&\qquad \leq M^2\left\|\Sigma^{-\frac{1}{2}} - \widehat{\Sigma}^{-\frac{1}{2}}\right\|\left(\left\|\Sigma^{-\frac{1}{2}}\right\| + \left\|\widehat{\Sigma}^{-\frac{1}{2}}\right\|\right)(\|X_i\| + \|X_j\|)^2 \\
&\qquad \leq M^2 C(a_0, \lambda_0, K)\sqrt{\frac{d\log n}{n}}\left(\|X_i\|^2 + \|X_j\|^2\right),
\end{aligned}$$

where we have used the Cauchy-Schwarz inequality, the triangular inequality and the Lipschitz assumption on $\widetilde{h}$ at the third line. Summing over $i \neq j$ this expression and using condition (20) we obtain

$$\left|\widehat{\sigma}^2_Y(h) - \widehat{\sigma}^2_{\widehat{Y}}(h)\right| \leq M^2 C(a_0, \lambda_0, K)\sqrt{\frac{d\log n}{n}},$$

so that we can effectively replace $\widehat{\sigma}_Y$ by $\widehat{\sigma}_{\widehat{Y}}$ in (21) up to additional lower-order terms. This concludes the proof of the first inequality in the theorem.

For the second inequality, we additionally write

$$\begin{aligned}
dist(\widehat{\gamma}(h), \mathcal{I}) &\leq \left\|\widehat{\Sigma}^{-\frac{1}{2}}\widehat{\beta}_{\widehat{Y}} - \Sigma^{-\frac{1}{2}}\beta_Y\right\| \\
&\leq \left\|\Sigma^{-\frac{1}{2}} - \widehat{\Sigma}^{-\frac{1}{2}}\right\|\|\beta_Y\| + \left\|\Sigma^{-\frac{1}{2}}\right\|\left\|\widehat{\beta}_{\widehat{Y}} - \beta_Y\right\| + \left\|\Sigma^{-\frac{1}{2}} - \widehat{\Sigma}^{-\frac{1}{2}}\right\|\left\|\widehat{\beta}_{\widehat{Y}} - \beta_Y\right\|;
\end{aligned}$$

we now conclude using (18), the previous inequalities controlling $\left\|\widehat{\beta}_{\widehat{Y}} - \beta_Y\right\|$, the assumption on $\left\|\Sigma^{-\frac{1}{2}}\right\|$ and the fact that

$$\|\beta_Y\| = \|\mathbb{E}\left[Xh(X) - \nabla h(X)\right]\| \leq B(1 + \mathbb{E}\left[\|x\|\right]) \leq C(a_0, \lambda_0, B)\,.$$

B Additional proofs and results

We have used Bernstein's inequality which we recall here for completeness under the following form:

Theorem B.1 *Suppose $X_1, \ldots, X_n$ are i.i.d. real random variables such that $|X| \leq b$ and $VarX = \sigma^2$. Then*

$$\mathbb{P}\left[\left|n^{-1}\sum_i X_i - E(X_i)\right| > \sqrt{2\sigma^2\frac{x}{n}} + 2b\frac{x}{n}\right] \leq 2\exp(-x).$$

The following results concern the estimation of $\Sigma^{-\frac{1}{2}}$, needed in the proof of Theorem 3.2. We divide this into 4 lemmas.

Lemma B.2 *Let $\xi_1, \ldots, \xi_n$ be i.i.d. with $\mathbb{E}\left[\xi_1\right] = m$ and assume $\log \mathbb{E}\left[\exp \mu(\xi_1 - m)\right] \leq c\mu^2/2$ holds for all $\mu \leq \mu_0$, for some positive constants c and μ_0. Then for sufficiently large n*

$$\mathbb{P}\left[n^{-1/2}\sum_{i=1}^{n}(\xi_i - m) > z\right] \leq e^{-c^{-1}z^2/2}.$$

Proof. This is an application of Chernoff's bounding method:

$$\begin{aligned} R_n &:= \log \mathbb{P}\left[n^{-1/2}\sum_{i=1}^{n}(\xi_i - m) > z\right] \\ &\leq -\mu z\sqrt{n} + \log \mathbb{E}\left[\exp \sum_{i=1}^{n} \mu(\xi_i - m)\right] \\ &= -\mu z\sqrt{n} + n \log \mathbb{E}\left[\exp \mu(\xi_1 - m)\right]\,, \end{aligned}$$

where the above inequality is Markov's. We select $\mu = zn^{-1/2}c^{-1}$. For n sufficiently large, it holds that $\mu \leq \mu_0$ and by the lemma condition

$$R_n \leq -\mu z\sqrt{n} + nc\mu^2/2 = -z^2c^{-1}/2.$$

The goal of the following Lemma is merely to replace the assumption about the Laplace transform (in the previous Lemma) by a simpler assumption (existence of some exponential moment). This allows a simpler statement – as far as we are not really interested in the precise constants involved.

Lemma B.3 *Let X be a real random variable such that for some $\mu_0 > 0$:*

$$\mathbb{E}\left[\exp(\mu_0 |X|)\right] = e_0 < \infty.$$

Then there exists $c > 0$ (depending only on μ_0 and e_0) such that

$$\forall \mu \in \mathbb{R} \qquad |\mu| \leq \mu_0/2 \Rightarrow \log \mathbb{E}\left[\exp\left(\mu\left(X - \mathbb{E}\left[X\right]\right)\right)\right] \leq c\mu^2/2.$$

Proof. Note that X has finite expectation since $|X| \leq \mu_0^{-1} \exp \mu_0 |X|$. Taylor's expansion gives that

$$\forall x \in \mathbb{R},\ \forall \mu \in \mathbb{R},\ |\mu| < \mu_0/2 \Rightarrow \exp(\mu x) \leq 1 + \mu x + \frac{\mu^2}{2} x^2 \exp(|\mu_0|\,|x|/2). \tag{22}$$

There exists some constant $c > 0$ (depending on μ_0) such that

$$\forall x \in \mathbb{R},\ x^2 \exp(|\mu_0 x|/2) \leq c\left(\exp(|\mu_0 x|)\right).$$

Using this and the assumption, taking expectation in (22) yields that for $c' = \frac{1}{2} c e_0 > 0$

$$\forall \mu \in \mathbb{R},\ |\mu| < \mu_0/2 \Rightarrow \mathbb{E}\left[\exp(\mu X)\right] \leq 1 + \mu \mathbb{E}\left[X\right] + c'\mu^2 \leq \exp\left(\mu \mathbb{E}\left[X\right] + c'\mu^2\right),$$

implying

$$\mathbb{E}\left[\exp\left(\mu\left(X - \mathbb{E}\left[X\right]\right)\right)\right] \leq \exp\left(c'\mu^2\right);$$

taking logarithms on both sides yields the conclusion.

The next two Lemmas, once combined, provide the confidence bound on $\left\|\Sigma^{-\frac{1}{2}} - \widehat{\Sigma}^{-\frac{1}{2}}\right\|$ which we need for the proof of Theorem 3.2.

Lemma B.4 *Let $X_1, \ldots, X_n$ be i.i.d. vectors in $\mathbb{R}^d$. Assume that, for some $\mu_0 > 0$,*

$$\mathbb{E}\left[\exp\left(\mu_0 \|X\|^2\right)\right] = e_0 < \infty; \tag{23}$$

denote $\Sigma = \mathbb{E}\left[XX^\top\right]$ and $\widehat{\Sigma}$ it empirical counterpart. Then for some constant κ depending only on (μ_0, e_0), and for big enough n,

$$R_n^* := \mathbb{P}\left[\left\|\Sigma - \widehat{\Sigma}\right\| > \sqrt{\frac{\kappa d \log n}{n}}\right] \leq \frac{2}{n}.$$

Proof. Along this proof C, c will denote constants depending only on μ_0, e_0; their exact value can change from line to line. Note that by definition of Σ and $\widehat{\Sigma}$,

$$\left\|\Sigma - \widehat{\Sigma}\right\| = \sup_{\theta \in \mathcal{B}_d} \frac{1}{n} \sum_{i=1}^{n} \left((X_i^\top \theta)^2 - \mathbb{E}\left[\left(X^\top \theta\right)^2\right]\right),$$

where $\mathcal{B}_d$ denotes the unit ball of $\mathbb{R}^d$. For $\varepsilon < 1$, let $\mathcal{B}_{d,\varepsilon}$ denote a ε-packing set of $\mathcal{B}_d$, that is, a discrete ε-separated set of points of $\mathcal{B}_d$ of maximum cardinality. By the maximality assumption and the triangle inequality, $\mathcal{B}_{d,\varepsilon}$ is also a 2ε-covering net of $\mathcal{B}_d$. On the other hand, the ε-balls centered on these points are disjoint and their union is included in the ball of radius $(1+\varepsilon)$, so that a volume comparison allows us to conclude that $\#(B_{d,\varepsilon})\varepsilon^d \leq (1+\varepsilon)^d \leq 2^d$. This shows that $\mathcal{B}_{d,2\varepsilon}$ is a 4ε-covering set of $\mathcal{B}_d$ of cardinality bounded by ε^{-d}.

Now, if $\theta, \theta' \in \mathcal{B}_d$ are such that $\|\theta - \theta'\| \leq 4\varepsilon$, then we have

$$\begin{aligned}\left|\sum_{i=1}^n (X_i^\top \theta)^2 - \sum_{i=1}^n (X_i^\top \theta')^2\right| &= \left|\sum_{i=1}^n (X_i^\top(\theta-\theta'))(X_i^\top(\theta+\theta'))\right| \\ &\leq 8\varepsilon \sum_{i=1}^n \|X_i\|^2 ,\end{aligned}$$

where we have applied the Cauchy-Schwarz inequality at the last line.

Now application of Lemmas B.2 and B.3 yields that for n large enough, with probability at least $1-1/n$,

$$n^{-1}\sum_{i=1}^n \|X_i\|^2 \leq \mathbb{E}\left[\|X\|^2\right] + \sqrt{\frac{c\log n}{n}} \leq C .$$

The above implies that with probability at least $1-1/n$,

$$\sup_{\theta,\theta'\in\mathcal{B}_d:\|\theta-\theta'\|\leq 2\varepsilon} n^{-1/2}\left|\sum_{i=1}^n (X_i^\top \theta)^2 - \sum_{i=1}^n (X_i^\top \theta')^2\right| \leq C\varepsilon\sqrt{n} .$$

We can also show a similar inequality about the corresponding expectation

$$\sup_{\theta,\theta'\in\mathcal{B}_d:\|\theta-\theta'\|\leq 2\varepsilon} n^{-1/2}\left|\mathbb{E}\left[(X^\top\theta)^2\right] - \mathbb{E}\left[(X^\top\theta')^2\right]\right| \leq C\varepsilon\sqrt{n} .$$

We now select $\varepsilon = n^{-\frac{1}{2}}$. Therefore, approximating any $\theta \in \mathcal{B}_d$ by its nearest neighbour in $\mathcal{B}_{d,2\varepsilon}$ and using the above inequalitites, we obtain that

$$\begin{aligned} R_n^* &\leq \frac{1}{n} + \mathbb{P}\left[\sup_{\theta\in\mathcal{B}_{d,2\varepsilon}} n^{-1/2}\sum_{i=1}^n\left((X_i^\top\theta)^2 - \mathbb{E}\left[\left(X^\top\theta\right)^2\right]\right) > \sqrt{\kappa d\log n} - C\right] \\ &\leq \frac{1}{n} + \sum_{\theta\in B_{d,2\varepsilon}} \mathbb{P}\left[n^{-1/2}\sum_{i=1}^n\left((X_i^\top\theta)^2 - \mathbb{E}\left[\left(X^\top\theta\right)^2\right]\right) > \sqrt{(\kappa - C)d\log n}\right] \\ &\leq \frac{1}{n} + \#(B_{d,2\varepsilon})\exp\{-0.5c^{-1}(\kappa-C)d\log n\} \leq \frac{2}{n}\end{aligned}$$

provided that κ is chosen so that $c^{-1}(\kappa - C)d/2 > d/2+1$. Here we have again used Lemmas B.2 and B.3, noting that for any $\theta \in \mathcal{B}_d$ it holds that $\mathbb{E}\left[\exp\mu_0\left|\theta^\top X\right|\right] \leq \mathbb{E}\left[\exp\mu_0\|X\|\right] < \exp(\mu_0) + e_0$ by assumption.

Lemma B.5 *Let A, B be two real positive definite symmetric matrices satisfying $\|A-B\| \leq \varepsilon$ with $\varepsilon \leq (2\left\|A^{-1}\right\|)^{-1}$. Then there exists a constant C such that*

$$\left\|A^{-\frac{1}{2}} - B^{-\frac{1}{2}}\right\| \leq C\left\|A^{-1}\right\|^{\frac{3}{2}}\varepsilon .$$

Proof.

Note that for $\|M\| < 1$ it holds that

$$(I-M)^{-\frac{1}{2}} = \sum_{k\geq 0} \gamma_k M^k \,,$$

with $(\gamma_k) \geq 0$ the coefficients of the power series development of the function $1/\sqrt{1-x}$.

Denote $\lambda_{max}(M), \lambda_{min}(M)$ the biggest and smallest eigenvalue of a matrix M. Put $K = \|A\| = \lambda_{max}(A)$ and $L = \left\|A^{-1}\right\| = \lambda_{min}(A)^{-1}$. Note that $LK \geq 1$. Put $A' = A/K, B' = B/K$. All eigenvalues of A' belong to $(0,1]$ and therefore

$$\left\|I - A'\right\| = \lambda_{\max}(I - A') = 1 - \lambda_{min}(A') = 1 - (LK)^{-1} \,.$$

By the assumption that $\varepsilon \leq (2L)^{-1}$, it holds that

$$\lambda_{max}(B') = K^{-1}\|B\| \leq K^{-1}(\|A\| + \varepsilon) \leq 1 + (2LK)^{-1} \leq \frac{3}{2} \,,$$

and that

$$\lambda_{min}(B') \geq K^{-1}(\lambda_{min}(A) - \varepsilon) \geq (2KL)^{-1} \,,$$

from this we deduce that

$$\left\|I - B'\right\| = \max(\lambda_{max}(B') - 1, 1 - \lambda_{min}(B')) \leq \max\left(\frac{1}{2}, 1 - (2LK)^{-1}\right) = 1 - (2LK)^{-1} \,.$$

Putting $\overline{A} = I - A', \overline{B} = I - B'$, we have ensured that $\left\|\overline{A}\right\| < 1$ and $\left\|\overline{B}\right\| < 1$; we can thus write

$$\begin{aligned} A'^{-\frac{1}{2}} - B'^{-\frac{1}{2}} &= \left(I - \overline{A}\right)^{-\frac{1}{2}} - \left(I - \overline{B}\right)^{-\frac{1}{2}} \\ &= \sum_{k\geq 1} \gamma_k (\overline{A}^k - \overline{B}^k) \,. \end{aligned}$$

Noticing that

$$\left\|\overline{A}^k - \overline{B}^k\right\| = \left\|\sum_{i=0}^{k-1} \overline{A}^i (\overline{A} - \overline{B}) \overline{B}^{k-1-i}\right\| \leq k \max\left(\left\|\overline{A}\right\|, \left\|\overline{B}\right\|\right)^{k-1} \left\|A' - B'\right\| \,,$$

we obtain

$$\begin{aligned} \left\|A'^{-\frac{1}{2}} - B'^{-\frac{1}{2}}\right\| &\leq \left\|A' - B'\right\| \sum_{k\geq 1} k\gamma_k \left(1 - (2LK)^{-1}\right)^{k-1} \\ &= \frac{\varepsilon}{K} \frac{1}{2} (2LK)^{\frac{3}{2}} = CL^{\frac{3}{2}} K^{\frac{1}{2}} \varepsilon \,. \end{aligned}$$

From this we deduce that

$$\left\|A^{-\frac{1}{2}} - B^{-\frac{1}{2}}\right\| = K^{-\frac{1}{2}} \left\|A'^{-\frac{1}{2}} - B'^{-\frac{1}{2}}\right\| \leq CL^{\frac{3}{2}} \varepsilon.$$

C Clustering results

The goal of NGCA is to discover interesting structure in the data. It is naturally a difficult task to quantify this property precisely. In this appendix we try to make this apparent using clustering techniques. We apply a mean distance linkage clustering algorithm to data projected in lower dimension using various techniques: NGCA, FastICA, PCA, LLE, Isomap.

There is no single well-defined performance measure for the performance of clustering. Here we resort to indirect criteria that should however allow a comparative study. We consider the two following criteria:

(1) Label cross-information. We apply clustering to benchmark data for which label information Y is available. Although this information is not used in determining the clustering, we will use it as a yardstick to measure whether the clustering gives rise to relevant structure discovery. We measure this by the scaled mutual information $I(C,Y)/H(Y)$, where C is the cluster labelling and the normalization ensures that the quantity lies between 0 and 1. Note that there is *a priori* no mathematical reason why clustering should be related to label information, but this is often the case for real data, so this can be a relevant criterion of structure discovery. A higher score indicates a better match between discovered cluster structure and label structure.

(2) Stability. Recent attempts at formalizing criteria for clustering have proposed that clustering stability should be a relevant criterion for data clustering (e.g. [Meinecke et al. , 2002, Lange et al., 2004]). Again, this is only an indirect criterion as, for example, a trivial clustering algorithm dividing the space without actually looking at the data would be very stable. But with this caveat in mind, it provides a relevant diagnostic tool. Here we measured stability in the following way: the data is divided randomly into 2 groups of equal size on which we apply clustering. Then the cluster labels obtained on group 1 are extended to group 2 by the nearest-neighbor rule and vice-versa. This thus gives rise to two different cluster labellings C_1, C_2 of the whole data and we measure their agreement through relative mutual information $I(C_1,C_2)/H(C_1,C_2)$. Again, this score lies in the interval $[0,1]$ and a high score indicates better stability.

We consider the "oil flow" data already presented in section 4.2, and additional data sets from the UCI classification repository, for which the features all take continuous values. (When there are features taking only discrete values, NGCA is inappropriate since these will generally be picked up as strongly non-Gaussian). Size and dimension of these data sets are given in Table 1.

Table 1: Description of data sets

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	10

The results are depicted in Figure 11. On the Oil data set, NGCA works very well for both criteria (as was expected from the good visualization results of section 4.2). On the

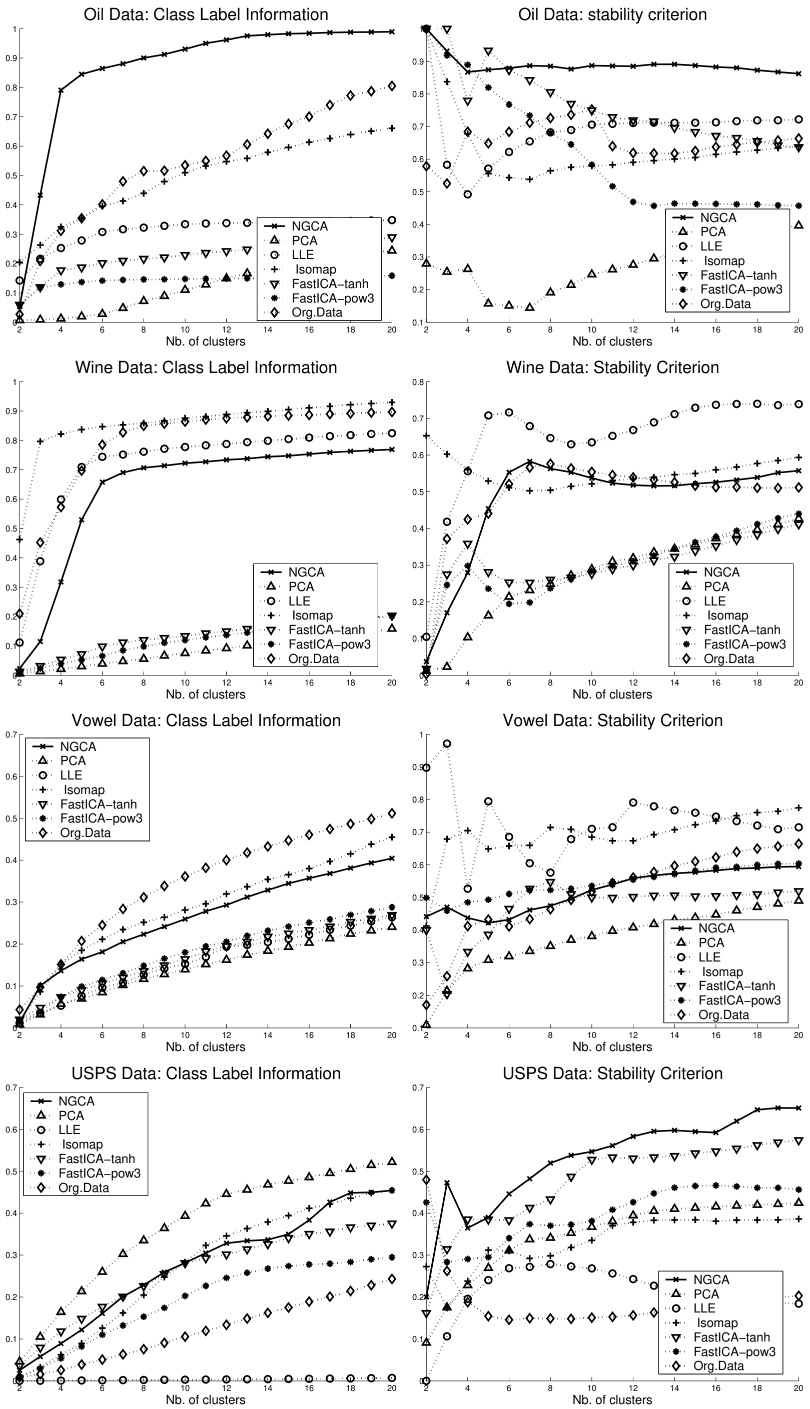


Figure 11: Clustering results

Wine data set, the different algorithms appear to be divided in two clear groups, with the performance in the first group (NGCA, Isomap, LLE) noticeably better than in the second (PCA, FastICA). NGCA belongs to the better group although the best methods appear to be the non-linear projections LLE and Isomap. The results of the Vowel data set are probably the most difficult to interpret as most methods appear to be relatively close. Isomap appear as the winner method in this case with NGCA quite close in terms of label cross-information and in the middle range for stability. Finally, for the USPS data set we used the 30 first principal components obtained by Kernel-PCA and a polynomial kernel of degree 3. In this case PCA gives better results in terms of label cross-information with NGCA a close second, while NGCA is the clear winner in terms of stability.

To summarize: NGCA performed very well in 2 of the 4 data sets tried (Oil data and USPS), and was in the best group of methods for the Wine Data and had average performance on the last data set. Even when NGCA is outperformed by nonlinear methods LLE and Isomap, it generally achieves a comparable performance though being a linear method, which has other advantages such as clearer geometrical interpretation, direct extension to additional data if needed, possible assessment of variable importance in original space.

References

M. Belkin and P. Niyogi. Laplacian eigenmaps for dimensionality reduction and data representation. *Neural Computation*, 15(6):1373–1396, 2003.

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

C.M. Bishop and G.D. James. Analysis of multiphase flow using dual-energy gamma densitometry and neural networks. *Nuclear Instruments and Methods in Physics Research*, A327:580–593, 1993.

P. Comon. Independent component analysis—a new concept? *Signal Processing*, 36:287–314, 1994.

T.F. Cox and M.A.A. Cox. *Multidimensional Scaling*. Chapman & Hall, London, 2001.

L. Devroye, L. Györfi, and G. Lugosi. *A Probabilistic Theory of Pattern Recognition*, Springer, 1996.

B. Efron. Bootstrap methods: Another look at the jackknife. *The Annals of Statistics*, 7(1):1–26, 1979.

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

S. Harmeling, A. Ziehe, M. Kawanabe and K.-R. Müller. Kernel-based nonlinear blind source separation. *Neural Computation*, 15(5):1089–1124, 2003.

P.J. Huber. Projection pursuit. *The Annals of Statistics*, 13:435–475, 1985.

A. Hyvärinen. Fast and robust fixed-point algorithms for independent component analysis. *IEEE Transactions on Neural Networks*, 10(3):626–634, 1999.

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

M. C. Jones and R. Sibson. What is projection pursuit? *Journal of the Royal Statistical Society, series A*, 150:1–36, 1987.

C. McDiarmid. On the method of bounded differences, *Surveys in Combinatorics*, London Math. Soc. Lecture Notes Series 141:148–188, 1989.

F. Meinecke, A. Ziehe, M. Kawanabe, and K.-R. Müller. A resampling approach to estimate the stability of one-dimensional or multidimensional independent components. *IEEE Transactions on Biomedical Engineering*, 49:1514–1525, 2002.

S. Roweis and L. Saul. Nonlinear dimensionality reduction by locally linear embedding. *Science*, 290(5500):2323–2326, 2000.

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

B. Schölkopf, A.J. Smola and K.–R. Müller. Nonlinear component analysis as a kernel eigenvalue problem. *Neural Computation*, 10(5):1299–1319, 1998.

J.B. Tenenbaum, V. de Silva and J.C. Langford. A global geometric framework for nonlinear dimensionality reduction. *Science*, 290(5500):2319–2323, 2000.