A Novel Dimension Reduction Procedure for Searching Non-Gaussian Subspaces

Motoaki Kawanabe[1], Gilles Blanchard[1], Masashi Sugiyama[2],
Vladimir Spokoiny[3], and Klaus-Robert Müller[1,4]

[1] Fraunhofer FIRST.IDA, Germany
[2] Department of Computer Science, Tokyo Institute of Technology, Japan
[3] Weierstrass Institute and Humboldt University, Germany
[4] Department of Computer Science, University of Potsdam, Germany
{blanchar, nabe}@first.fhg.de, sugi@cs.titech.ac.jp,
spokoiny@wias-berlin.de, klaus@first.fhg.de

Abstract. In this article, we consider high-dimensional data which contains a low-dimensional non-Gaussian structure contaminated with Gaussian noise and propose a new *linear* method to identify the non-Gaussian subspace. Our method NGCA (Non-Gaussian Component Analysis) is based on a very general semi-parametric framework and has a theoretical guarantee that the estimation error of finding the non-Gaussian components tends to zero at a parametric rate. NGCA can be used not only as preprocessing for ICA, but also for extracting and visualizing more general structures like clusters. A numerical study demonstrates the usefulness of our method.

1 Introduction

Suppose that we are given a set of i.i.d. observations $\boldsymbol{x}_i \in \mathbb{R}^d$, $(i = 1, \ldots, n)$ obtained as a sum of a signal $\boldsymbol{s} \in \mathbb{R}^m$ $(m \leq d)$ with an unknown non-Gaussian distribution and an independent Gaussian noise component $\boldsymbol{n} \in \mathbb{R}^d$:

$$\boldsymbol{x} = A\boldsymbol{s} + \boldsymbol{n}, \tag{1}$$

where A is a $d \times m$ matrix and $\boldsymbol{n} \sim N(\boldsymbol{0}, \Gamma)$. The rationale behind this model is 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 that we do *not* assume the Gaussian components to be of *smaller* order of magnitude than the signal components. This setting therefore excludes the use of common (nonlinear) dimensionality reduction methods such as PCA, Isomap [12] and LLE [11] that are based on the assumption that the data lies, say, on a lower dimensional manifold, up to some small noise distortion.

If the non-Gaussian components s_i's are mutually independent, the model turns out to be the under-complete noisy ICA [9]. Although some algorithms have been proposed, combinations of dimension reduction like PCA or Factor Analysis and noise-free ICA methods are often used, when the number m of the sources is relatively small. However, the classical methods for dimension reduction are based on second order statistics

J. Rosca et al. (Eds.): ICA 2006, LNCS 3889, pp. 149–156, 2006.

and do not consider non-Gaussianity of the sources. In this research, we will construct a dimension reduction procedure called NGCA (Non-Gaussian Component Analysis) which extracts the non-Gaussian subspace by higher order statistics. Since mutual independence of the sources is not assumed, our NGCA method can be used not only as preprocessing for ICA, but also for searching more general and dependent non-Gaussian structures (cf. [10]).

The NGCA approach is built upon a very general semi-parametric framework where the density of the sources is not specified at all. We will present an implementation here which is close in spirit to *Projection Pursuit (PP)* [5, 7, 8, 9] for visualization of interesting structures in high-dimensional data. However, 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*. To be more precise, in PP methods, a *single* index which measures the non-Gaussianity (or 'interestingness') of a projection direction has to be fixed and optimized , while NGCA takes many various indices into account at the same time. Therefore it can outperform PP algorithms, if the data contains say, both super- and sub-Gaussian components.

In the following section we will outline a novel semi-parametric theory for *linear* dimension reduction and theoretical guarantees of the NGCA procedure. The algorithm will be presented in Section 3 and simulation results underline the usefulness of NGCA; finally a brief conclusion is given.

2 Theoretical Framework

The probability density function $p(\boldsymbol{x})$ of the observations defined by the mixing model (1) can be put under the following semi-parametric form:

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

where T is an unknown linear mapping from $\mathbb{R}^d$ to another subspace $\mathbb{R}^m$, g is an unknown function on $\mathbb{R}^m$ related to the distribution of the source $\boldsymbol{s}$ and ϕ_{Γ} is a centered Gaussian density with unknown covariance matrix Γ. The model (2) includes as particular cases both the pure parametric ($m = 0$) and pure non-parametric ($m = d$) models. In practice we are interested in an intermediate case where d is large and m is rather small.

Note that the decomposition (2) is non-unique, but we will show that the following m-dimensional *linear* subspace $\mathcal{I}$ of $\mathbb{R}^d$ is *identifiable*:

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

We call $\mathcal{I}$ the *non-Gaussian index space*. Its geometrical meaning is the following: in the model (1), the noise term can be decomposed into two components, $\boldsymbol{n} = \boldsymbol{n}_1 + \boldsymbol{n}_2$, where $\boldsymbol{n}_1 = A\boldsymbol{\eta} \in \mathrm{Range}(A)$ and $\boldsymbol{n}_2$ is restricted in the $(d-m)$-dimensional complementary subspace s.t. $\mathrm{Cov}(\boldsymbol{n}_1, \boldsymbol{n}_2) = 0$ (i.e. $\boldsymbol{n}_1$ and $\boldsymbol{n}_2$ are independent). Thus, we have the representation

$$\boldsymbol{x} = A\widetilde{\boldsymbol{s}} + \boldsymbol{n}_2, \tag{3}$$

where $\widetilde{\boldsymbol{s}} := \boldsymbol{s} + \boldsymbol{\eta}$ and the noise term $\boldsymbol{n}_2$ distributes with a $(d-m)$-dimensional degenerated Gaussian independent of $\widetilde{\boldsymbol{s}}$. The subspace $\mathcal{I}$ is then the orthogonal complement

of the $(d-m)$-dimensional subspace containing the independent Gaussian component $\boldsymbol{n}_2$. Once we can estimate the index space $\mathcal{I}$, we can project out the noise $\boldsymbol{n}_2$ by projecting the data $\boldsymbol{x}$ onto $\mathcal{I}$. In the representation (2) we can assume that $TA = I_m$ and $T\boldsymbol{x} = \widetilde{\boldsymbol{s}}$ without loss of generality, in which case T corresponds to the demixing matrix in under-complete ICA, but here we are not interested in the individual directions of the components $\widetilde{\boldsymbol{s}}_i$ (which are not assumed to be independent).

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

Proposition 1. *Let $\boldsymbol{x}$ be a random variable whose density function $p(\boldsymbol{x})$ satisfies* (2) *and suppose that $h(\boldsymbol{x})$ is a smooth real function on $\mathbb{R}^d$. Assume furthermore that $\Sigma = \mathbb{E}\left[\boldsymbol{x}\boldsymbol{x}^\top\right] = I_d$. Then under mild regularity conditions the following vector belongs to the target space $\mathcal{I}$:*

$$\boldsymbol{\beta}(h) = \mathbb{E}_{\boldsymbol{x}}\left[\nabla h(\boldsymbol{x}) - \boldsymbol{x}h(\boldsymbol{x})\right] . \tag{4}$$

Since an expectation over the unknown density $p(\boldsymbol{x})$ is used to define $\boldsymbol{\beta}$ by Eq.(4), in practice, it must be approximated using empirical expectation over the available data:

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

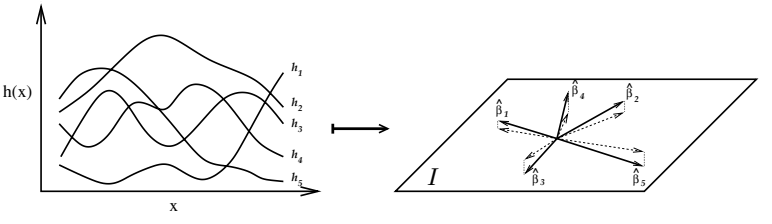

Fig. 1. The NGCA principle idea: from a varied family of real functions h, compute a family of vectors $\widehat{\boldsymbol{\beta}}$ belonging to the target space up to small estimation error

In the extended version of this paper [3], we show a probabilistic confidence bound of estimation error of our NGCA method under certain regularity conditions.

- If we assume $\mathbb{E}\left[\boldsymbol{x}\boldsymbol{x}^\top\right] = I_d$, the empirical estimator $\widehat{\boldsymbol{\beta}}(h)$ converges at a rate $\mathcal{O}(n^{-1/2})$ to a vector in the index space $\mathcal{I}$.
- In the general case where $\mathbb{E}\left[\boldsymbol{x}\boldsymbol{x}^\top\right]$ is an arbitrary positive definite matrix, we consider a "whitening" step, computing $\widehat{\boldsymbol{y}}_i = \widehat{\Sigma}^{-1/2}\boldsymbol{x}_i$ beforehand, where $\widehat{\Sigma} := \frac{1}{n}\sum_{i=1}^{n}\boldsymbol{x}_i\boldsymbol{x}_i^\top$. Taking into account the extra error introduced by this step, we can bound the the convergence rate of $\boldsymbol{\gamma}(h) := \widehat{\Sigma}^{-1/2}\widehat{\boldsymbol{\beta}}_{\boldsymbol{y}}(h)$ to the index space $\mathcal{I}$ by $\mathcal{O}(\sqrt{d\log n/n})$.
- The entire index space $\mathcal{I}$ can be estimated from a family of vectors $\widehat{\boldsymbol{\beta}}_k$ (see Fig. 1) for a large set of functions $\{h_k\}_{k=1}^{L}$ and applying PCA to the set $\{\widehat{\boldsymbol{\beta}}_k\}_{k=1}^{L}$.

- Thanks to an exponential deviation inequality for the convergence rate of single functions, a union bound over L functions leads to a *uniform* convergence bound over the whole set of functions with rate of order $\mathcal{O}(\sqrt{d \log n/n} + \sqrt{\log L/n})$. Therefore, taking, e.g., $L = O(n^d)$ we still have insurance that convergence holds.

3 The NGCA Algorithm

As is briefly mentioned in the last section, in our NGCA procedure, basically we calculate a family of vectors $\widehat{\boldsymbol{\beta}}_k$ for a large family of such functions $\{h_k\}_{k=1}^L$ and apply PCA to the set $\{\widehat{\boldsymbol{\beta}}_k\}_{k=1}^L$ to find out the m-dimensional subspace $\widehat{\mathcal{I}}$ which gives the least approximation error. Although the principle of NGCA is very simple, there are some implementation issues.

- The theoretical results guarantee that the convergence order is achieved for any smooth functions $\{h_k\}_{k=1}^L$ with mild regularity conditions. However, in practice, it is important to find out good functions which provide a lot of information on the index space $\mathcal{I}$ and make the estimator $\widehat{\mathcal{I}}$ more accurate, because there exist many uninformative functions.
- Since the mapping $h \mapsto \boldsymbol{\beta}(h)$ is linear, we need an appropriate renormalization of h or $\boldsymbol{\beta}(h)$, otherwise it is meaningless to combine many vectors $\{\boldsymbol{\beta}_k\}$ from various functions $\{h_k\}$ by PCA. Here we propose renormalizing by the trace of the variance $\mathrm{Var}\{\widehat{\boldsymbol{\beta}}(h)\}$. Under this condition the norm of each vector is proportional to its signal-to-noise ratio so that longer vectors are more informative, while vectors with too small a norm are uninformative and can be discarded.

In the proposed algorithm we will restrict our attention to functions of the form $h_{f,\boldsymbol{\omega}}(\boldsymbol{x}) = f(\langle \boldsymbol{\omega}, \boldsymbol{x} \rangle)$, where $\boldsymbol{\omega} \in \mathbb{R}^d$, $\|\boldsymbol{\omega}\| = 1$, and f belongs to a finite family $\mathcal{F}$ of smooth real functions of real variable. Our theoretical setting allows to ensure that the approximation error remains small uniformly over $\mathcal{F}$ and $\boldsymbol{\omega}$. However it is not feasible in practice to sample the whole parameter space for $\boldsymbol{\omega}$ as soon as it has more than a few dimensions. To overcome this difficulty we advocate using a well-known PP algorithm, FastICA [8], as a heuristic to find good candidates for $\boldsymbol{\omega}_f$ for a fixed f. We remark that FastICA, as a standalone procedure, requires to fix the "index function" f beforehand. The new point of our method is that we provide a theoretical setting and a methodology which allows to *combine* the results of this Projection Pursuit method when used over a possibly large spectrum of arbitrary index functions f.

Summing up, the NGCA algorithm then consists of the following steps: (1) Data whitening, (2) Applying FastICA to each function $f \in \mathcal{F}$ to find a promising candidate value for $\boldsymbol{\omega}_f$, (3) Computing the corresponding family of vectors $(\widehat{\boldsymbol{\beta}}(h_{f,\boldsymbol{\omega}_f}))_{f\in\mathcal{F}}$ (using Eq. (5)), (4) Normalize the vectors appropriately; threshold and throw out uninformative ones, (5) apply PCA, (6) Pull back in original space (cf. Pseudocode). Note that the PCA step could be replaced by other, more refined principal directions extraction methods. In the implementation tested, we have used the following forms of the functions f_k: $f_\sigma^{(1)}(z) = z^3 \exp(-z^2/2\sigma^2)$ (Gauss-Pow3), $f_b^{(2)}(z) = \tanh(bz)$ (Hyperbolic Tangent), $f_a^{(3)}(z) = \{\sin, \cos\}(az)$ (Fourier). More precisely, we consider

PSEUDOCODE FOR THE NGCA ALGORITHM

Input: Data points $(\boldsymbol{x}_i) \in \mathbb{R}^d$, dimension m of target subspace.

Parameters: Number $T_{\max}$ of FastICA iterations; threshold ε; family of real functions (f_k).

Whitening.

The data $\boldsymbol{x}_i$ is recentered by subtracting the empirical mean.

Let $\widehat{\Sigma}$ denote the empirical covariance matrix of the data sample $(\boldsymbol{x}_i)$;

put $\widehat{\boldsymbol{y}}_i = \widehat{\Sigma}^{-\frac{1}{2}} \boldsymbol{x}_i$ the empirically whitened data.

Main Procedure.

Loop on $k = 1, \ldots, L$:

Draw $\boldsymbol{\omega}_0$ at random on the unit sphere of $\mathbb{R}^d$.

Loop on $t = 1, \ldots, T_{\max}$: *[FastICA loop]*

Put $\widehat{\boldsymbol{\beta}}_t \leftarrow \dfrac{1}{n} \displaystyle\sum_{i=1}^{n} \big(\widehat{\boldsymbol{y}}_i f_k(\langle \boldsymbol{\omega}_{t-1}, \widehat{\boldsymbol{y}}_i \rangle) - f_k'(\langle \boldsymbol{\omega}_{t-1}, \widehat{\boldsymbol{y}}_i \rangle) \boldsymbol{\omega}_{t-1} \big)$.

Put $\boldsymbol{\omega}_t \leftarrow \widehat{\boldsymbol{\beta}}_t / \|\widehat{\boldsymbol{\beta}}_t\|$.

End Loop on t

Let N_i be the trace of the empirical covariance matrix of $\widehat{\boldsymbol{\beta}}_{T_{\max}}$:

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

Store $v^{(k)} \leftarrow \widehat{\boldsymbol{\beta}}_{T_{\max}} * \sqrt{n/N_i}$. *[Normalization]*

End Loop on k

Thresholding.

From the family $\boldsymbol{v}^{(k)}$, throw away vectors having norm smaller than threshold ε.

PCA step.

Perform PCA on the set of remaining $\boldsymbol{v}^{(k)}$.

Let V_m be the space spanned by the first m principal directions.

Pull back in original space.

Output: $W_m = \widehat{\Sigma}^{-\frac{1}{2}} V_m$.

discretized ranges for $a \in [0, A], b \in [0, B], \sigma \in [\sigma_{\min}, \sigma_{\max}]$, which gives rise to a finite family (f_k) (which includes *simultaneously* functions of the three different above families).

4 Numerical Results

All 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). Some preliminary calibration suggested to take $\varepsilon = 1.5$ as the threshold under which vectors are not informative. Finally we fixed the number of FastICA iterations $T_{\max} = 10$. With this choice of parameters, with 1000 points of data the computation time is typically of the order of 10 seconds on a modern PC under a Matlab implementation.

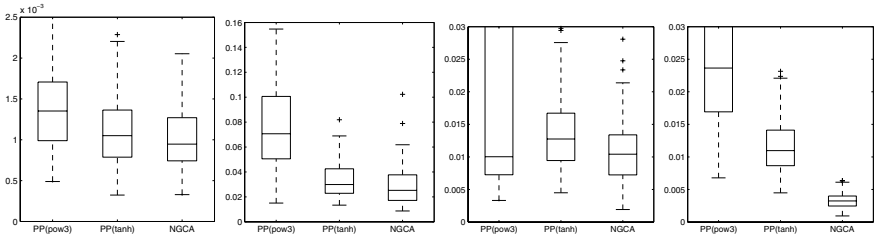

Fig. 2. Boxplots of the error criterion $\mathcal{E}(\widehat{\mathcal{I}}, \mathcal{I})$ over 100 training samples of size 1000

Tests in a controlled setting. We performed numerical experiments using various synthetic data. We report exemplary results using 4 data sets. Each data set includes 1000 samples in 10 dimensions, and consists of 8-dimensional independent standard Gaussian and 2 non-Gaussian components as follows:

(A) Simple Gaussian Mixture: 2-dimensional independent bimodal Gaussian mixtures;
(B) Dependent super-Gaussian: 2-dimensional density is proportional to $\exp(-\|x\|)$;
(C) Dependent sub-Gaussian: 2-dimensional uniform on the unit circle;
(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.

We compare the NGCA method against standalone FastICA with two different index functions. Figure 2 shows boxplots, over 100 samples, of the error criterion $\mathcal{E}(\widehat{\mathcal{I}}, \mathcal{I}) = m^{-1} \sum_{i=1}^{m} \|(I_d - \Pi_{\mathcal{I}})\widehat{\boldsymbol{v}}_i\|^2$, where $\{\widehat{\boldsymbol{v}}_i\}_{i=1}^m$ is an orthonormal basis of $\widehat{\mathcal{I}}$, I_d is the identity matrix, and $\Pi_{\mathcal{I}}$ denotes the orthogonal projection on $\mathcal{I}$. In datasets (A),(B),(C), NGCA appears to be on par with the best FastICA method. As expected the best index for FastICA is data-dependent: the 'tanh' index is more suited to the super-Gaussian data (B) while the 'pow3' index works best with the sub-Gaussian data (C) (although in this case FastICA with this index has a tendency to get caught in local minima, leading to a disastrous result for about 25% of the samples. Note that NGCA does *not* suffer from this problem). Finally, the advantage of the implicit index adaptation feature of NGCA can be clearly observed in the data set (D), which includes both sub- and super-Gaussian components. In this case neither of the two FastICA index functions taken alone does well and NGCA gives significantly lower error than either FastICA flavor.

Example of application for realistic data: visualization and clustering. We now give an example of application of NGCA 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 [2]. The data is 12-dimensional and our goal is to visualize the data and possibly exhibit a clustered structure. We compared results obtained with the NGCA methodology, regular PCA, FastICA with tanh index and Isomap. The results are shown on Figure 3. A 3D projection of the data was first computed using these methods, which was in turn projected in 2D to draw the figure; this last projection

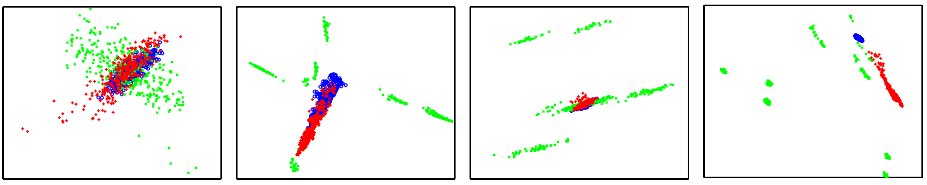

Fig. 3. 2D projection of the "oil flow" (12-dimensional) data obtained by different algorithms, from left two right: PCA, Isomap, FastICA (tanh index), NGCA. In each case, the data was first projected in 3D using the respective methods, from which a 2D projection was chosen visually so as to yield the clearest cluster structure. Colors indicate label information (*not used to determine the projections*).

was chosen manually so as to make the cluster structure as visible as possible in each case. The NGCA result appears better with a clearer clustered structure appearing. This structure is only partly visible in the Isomap result; the NGCA method additionally has the advantage of a clear geometrical interpretation (linear orthogonal projection). Finally, datapoints in this dataset are distributed in 3 classes. This information was not used in the different procedures, but we can see *a posteriori* that only NGCA clearly separates the classes in distinct clusters.

5 Conclusion

We proposed a new semi-parametric framework for constructing a linear projection to separate an uninteresting, possibly of large amplitude multivariate Gaussian 'noise' subspace from the 'signal-of-interest' subspace. We also provided generic consistency results on how well the non-Gaussian directions can be identified (an extended version of this paper). 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 ICA/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, NGCA 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 [11, 12, 1, 6] and to finding the most discriminative directions using labels are examples for which the current theory could be taken as a basis.

Acknowledgements. This work was supported in part by the IST Programme of the European Community, under the PASCAL Network of Excellence, IST-2002-506778. This publication only reflects the authors' views.

References

1. M. Belkin and P. Niyogi. Laplacian eigenmaps for dimensionality reduction and data representation. *Neural Computation*, 15(6):1373–1396, 2003.
2. C.M. Bishop, M. Svensen and C.K.I. Wiliams. GTM: The generative topographic mapping. *Neural Computation*, 10(1):215–234, 1998.
3. G. Blanchard, M. Kawanabe, M. Sugiyama, V. Spokoiny and K.-R. Müller. In search of non-Gaussian components of a high-dimensional distribution. (Preprint available at http://www.cs.titech.ac.jp/)
4. P. Comon. Independent component analysis—a new concept? *Signal Processing*, 36:287–314, 1994.
5. J.H. Friedman and J.W. Tukey. A projection pursuit algorithm for exploratory data analysis. *IEEE Transactions on Computers*, 23(9):881–890, 1975.
6. S. Harmeling, A. Ziehe, M. Kawanabe and K.-R. Müller. Kernel-based nonlinear blind source separation. *Neural Computation*, 15(5):1089–1124, 2003.
7. P.J. Huber. Projection pursuit. *The Annals of Statistics*, 13:435–475, 1985.
8. A. Hyvärinen. Fast and robust fixed-point algorithms for independent component analysis. *IEEE Transactions on Neural Networks*, 10(3):626–634, 1999.
9. A. Hyvärinen, J. Karhunen and E. Oja. *Independent Component Analysis*. Wiley, 2001.
10. M. Kawanabe and K.-R. Müller. Estimating functions for blind separation when sources have variance dependencies. *Journal of Machine Learning Research*, 6:453–482, 2005.
11. S. Roweis and L. Saul. Nonlinear dimensionality reduction by locally linear embedding. *Science*, 290(5500):2323–2326, 2000.
12. J.B. Tenenbaum, V. de Silva and J.C. Langford. A global geometric framework for nonlinear dimensionality reduction. *Science*, 290(5500):2319–2323, 2000.

Proof of Proposition 1

Put $\boldsymbol{\alpha} = \mathbb{E}_{\boldsymbol{x}}[\boldsymbol{x}h(\boldsymbol{x})]$ and $\psi(\boldsymbol{x}) = h(\boldsymbol{x}) - \boldsymbol{\alpha}^\top \boldsymbol{x}$. Note that $\nabla\psi = \nabla h - \boldsymbol{\alpha}$, hence $\beta(h) = \mathbb{E}_{\boldsymbol{x}}[\nabla\psi(\boldsymbol{x})]$. Furthermore, it holds by change of variable that

$$\int \psi(\boldsymbol{x}+\boldsymbol{u})p(\boldsymbol{x})d\boldsymbol{x} = \int \psi(\boldsymbol{x})p(\boldsymbol{x}-\boldsymbol{u})d\boldsymbol{x}.$$

Under mild regularity conditions on $p(\boldsymbol{x})$ and $h(\boldsymbol{x})$, differentiating this with respect to u gives

$$\mathbb{E}_{\boldsymbol{x}}[\nabla\psi(\boldsymbol{x})] = \int \nabla\psi(\boldsymbol{x})p(\boldsymbol{x})d\boldsymbol{x} = -\int \psi(\boldsymbol{x})\nabla p(\boldsymbol{x})d\boldsymbol{x} = -\mathbb{E}_{\boldsymbol{x}}[\psi(\boldsymbol{x})\nabla \log p(\boldsymbol{x})]\ ,$$

where we have used $\nabla p(\boldsymbol{x}) = \nabla \log p(\boldsymbol{x})\, p(\boldsymbol{x})$. Eq.(2) now implies $\nabla \log p(\boldsymbol{x}) = \nabla \log g(T\boldsymbol{x}) - \Gamma^{-1}\boldsymbol{x}$, hence

$$\begin{aligned}\beta(\psi) &= -\mathbb{E}_{\boldsymbol{x}}[\psi(\boldsymbol{x})\nabla \log g(T\boldsymbol{x})] + \mathbb{E}_{\boldsymbol{x}}\left[\psi(\boldsymbol{x})\Gamma^{-1}\boldsymbol{x}\right] \\ &= -T^\top \mathbb{E}_{\boldsymbol{x}}[\psi(\boldsymbol{x})\nabla g(T\boldsymbol{x})/g(T\boldsymbol{x})] + \Gamma^{-1}\mathbb{E}_{\boldsymbol{x}}\left[\boldsymbol{x}h(\boldsymbol{x}) - \boldsymbol{x}\boldsymbol{x}^\top \mathbb{E}[\boldsymbol{x}h(\boldsymbol{x})]\right]\ .\end{aligned}$$

The last term above vanishes because we assumed $\mathbb{E}_{\boldsymbol{x}}\left[\boldsymbol{x}\boldsymbol{x}^\top\right] = Id$. The first term belongs to $\mathcal{I}$ by definition. This concludes the proof. □