Semiparametric estimation of the signal subspace*

***Denis Belomestny*[1], *Vladimir Panov*[2], *Vladimir Spokoiny*[3]**

1 — Laboratory for Structural Methods of Data Analysis in Predictive Modeling (MIPT) and University Duisburg-Essen, denis.belomestny@uni-due.de

2 — Laboratory for Structural Methods of Data Analysis in Predictive Modeling (MIPT) and University Duisburg-Essen, vladimir.panov@uni-due.de

3 — Laboratory for Structural Methods of Data Analysis in Predictive Modeling (MIPT) and Weierstrass Institute for Applied Analysis and Stochastics, spokoiny@wias-berlin.de

Let a high-dimensional random vector $\vec{X}$ be represented as a sum of two components — a signal $\vec{S}$ that belongs to some low-dimensional linear subspace $\mathcal{S}$, and a noise component $\vec{N}$. This paper presents a new approach for estimating the subspace $\mathcal{S}$ based on the ideas of the Non-Gaussian Component Analysis. Our approach avoids the technical difficulties that usually appear in similar methods — it requires neither the estimation of the inverse covariance matrix of $\vec{X}$ nor the estimation of the covariance matrix of $\vec{N}$.

Keywords: *dimension reduction, non-Gaussian components, signal subspace*

Introduction

Assume that a high-dimensional random variable $\vec{X} \in \mathbb{R}^d$ be represented as a sum of two independent components — a low-dimensional signal (which one can imagine as "a useful part" or "an information") and a noise component with a Normal distribution. More precisely,

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

where $\vec{S}$ belongs to some low-dimensional linear subspace $\mathcal{S}$, $\vec{N}$ is a normal vector with zero mean and unknown covariance matrix, and $\vec{S}$ is independent of $\vec{N}$. This structural assumption follows the observation that in applications the "useful part" is non-Gaussian while the "rest part" can be interpreted as a high-dimensional noise. For the sake of simplicity, we assume that the expectation of $\vec{X}$ vanishes and the covariance matrix of $\vec{X}$, which is denoted by Σ, is non-degenerated.

Denote the dimension of $\vec{S}$ by m. In this paper, m is fixed such that the representation (1) is unique; the existence of such m is proved under some mild assumptions by Theis and Kawanabe (2007). If $m = 0$, then our model reduces to the pure-parametric case; if $m = d$, then the model is pure-nonparametric. Obviously, the representation (1) links pure Gaussian (PCA) and pure non-Gaussian (ICA) modelling.

The aim of this paper is to estimate vectors from the subspace $\mathcal{S}$, which we call *the signal subspace*. A very related task, estimation of so called *the non-Gaussian subspace* $\mathcal{I}$ (the definition will be given below) has been extensively studied in the literature. The original method known as Non-Gaussian Component Analysis (NGCA) was proposed by Blanchard et al. (2006), and later improved by Dalalyan et al. (2007), Kawanabe et al. (2007), Diederichs et al. (2010).

In almost all papers mentioned above, the problem of estimation of the vectors from $\mathcal{S}$ is not considered in details; natural estimators require the estimation of the covariance matrix of the noise. Moreover, practical usage meets another technical problem — the estimation of the

The authors are partially supported by Laboratory for Structural Methods of Data Analysis in Predictive Modeling, MIPT, RF government grant, ag. 11.G34.31.0073.

inverse covariance matrix of $\vec{X}$. Each of these tasks is an obstacle in real-world applications of the method. In this article, we propose a new approach, which avoids the mentioned problems.

The paper is organized as follows. In the second section, we formulate and discuss the mathematical statements that yield a method for estimating the vectors from $\mathcal{S}$. We start with Theorem 1, which implies a special representation for the density function of $\vec{X}$. This representation is new and plays a central role in the validation of the method. The theoretical base of our approach is given in Theorem 2 and Lemma 3. Section 3 contains the full description of the algortihm. Next, we discuss different representations for the density function and point out the advantages of our formulation. The last section states what has been done before and what is the contribution of the present paper. All proofs are collected in Appendix A; some additional information about the NGCA methodology is given in Appendix B.

Theoretical base for the estimation of the signal subspace

This section presents the theoretical results that are needed for our purposes. The proofs are collected in Appendix A.

The first theorem gives the semiparametric density representation for the random vector $\vec{X}$. Such facts are known in the literature (see e.g. Blanchard et al., 2006) but the formulation given below principally differs from the previous versions, see the next section for discussion.

Theorem 1. Let the structural assumption (1) be fulfilled. Then the density function of the random vector $\vec{X}$ can be represented as follows:

$$p(\vec{x}) = \mathbf{g}(\mathbf{T}\vec{x})\; p^{\mathcal{N}}(\vec{x}, \Sigma), \tag{2}$$

where

— $\mathbf{T}:\ \mathbb{R}^d \to \Sigma^{-1/2}\mathcal{S}$ is the linear transformation defined as the projection of the vector $\Sigma^{-1/2}\vec{x}$ on the subspace $\Sigma^{-1/2}\mathcal{S}$, i.e.,

$$\mathbf{T}\vec{x} := \Pr{}_{\Sigma^{-1/2}\mathcal{S}}\{\Sigma^{-1/2}\vec{x}\}, \tag{3}$$

— $\mathbf{g}:\ \Sigma^{-1/2}\mathcal{S} \to \mathbb{R}$ is defined by

$$\mathbf{g}(\vec{t}\,) = \frac{q\,(\vec{t}\,)}{p^{\mathcal{N}}(\vec{t}, \mathbf{I}_m)}, \tag{4}$$

and $q(\cdot)$ is the density function of the random variable $\mathbf{T}\vec{X}$.

As it was mentioned in the introduction, our aim is to recover the subspace $\mathcal{S}$. This can be done thanks to the following theorem, which is a new result in the context of the Non-Gaussian Component Analysis.

Theorem 2. Let $\mathbf{T}$ be the linear transformation defined by (3). Then

$$\mathcal{S} = \Sigma\ (\operatorname{Ker}\mathbf{T})^{\perp}. \tag{5}$$

In Blanchard et al. (2006), a transformation $\mathcal{T}$ is considered instead of $\mathbf{T}$:

$$\mathcal{T}\vec{x} := \Pr{}_{\Gamma^{-1/2}\mathcal{S}}\{\Gamma^{-1/2}\vec{x}\}, \tag{6}$$

where by Γ we denote the covariance matrix of the noise component. In the paper by Blanchard et al., the subspace $(\operatorname{Ker}\mathcal{T})^{\perp}$ is called *the non-Gaussian subspace* and is in fact the main object of interest. We would like to stress here that $\mathcal{T} \neq \mathbf{T}$, and equalities like (5) are wrong for $\mathcal{T}$.

One of the main results about the NGCA approach gives the practical method for estimating vectors from $(\operatorname{Ker}\mathcal{T})^{\perp}$. Similar result can be formulated also for the subspace $(\operatorname{Ker}\mathbf{T})^{\perp}$.

Lemma 3. Assume that the structural assumption (1) is fulfilled. Then for any function $\psi \in \mathcal{C}^1(\mathbb{R}^d, \mathbb{R})$ there exists a vector $\beta \in (\text{Ker } \mathbf{T})^\perp$ such that

$$\mathbb{E}\left(\nabla\psi(\vec{X})\right) - \beta = \Sigma^{-1}\mathbb{E}\left(\vec{X}\psi(\vec{X})\right). \tag{7}$$

Corollary 4. Let the structural assumption (1) be fulfilled and let a function $\psi \in \mathcal{C}^1(\mathbb{R}^d, \mathbb{R})$ be such that $\mathbb{E}\left(\vec{X}\psi(\vec{X})\right) = 0$. Then

$$\mathbb{E}\left(\nabla\psi(\vec{X})\right) \in (\text{Ker } \mathbf{T})^\perp.$$

Theorem 2 and Lemma 3 yield a method for estimating vectors from the subspace $\mathcal{S}$ that is described in the next section.

Algorithm for the estimation of the signal subspace

The first step is to estimate vectors from the subspace $(\text{Ker } \mathbf{T})^\perp$ using Lemma 3. Theoretically, the best way for the estimation is to find a set of functions $\{\psi\}$ such that $\mathbb{E}\left(\vec{X}\psi(\vec{X})\right) = 0$ for any element of this set, and to estimate the vectors from $(\text{Ker } \mathbf{T})^\perp$ by $\mathbb{E}\left(\nabla\psi(\vec{X})\right)$, see Corollary 4. In practice, such set $\{\psi\}$ is unknown; usually it is more realistic to consider some ψ such that $\mathbb{E}\left(\vec{X}\psi(\vec{X})\right)$ is close to zero (but not exactly zero). In this case, according to Lemma 3, the vector $\mathbb{E}\left(\nabla\psi(\vec{X})\right)$ is close to some vector from the subspace $(\text{Ker } \mathbf{T})^\perp$.

We suggest to use the convex projection method, see Diederichs (2007) and Diederichs et al. (2010). Since the description of this method requires at least one page, and its discussion is not an objective of the article (and merits a separate publication), we put the explanation in the appendix of the paper.

The second step. Denote the vectors obtained on the first step by $\vec{\beta}_k$, $k = 1..K$. Now one can use Theorem 2 and estimate vectors from the signal subspace by $\vec{\gamma}_k := \hat{\Sigma}\vec{\beta}_k$, where $\hat{\Sigma}$ is an estimator of the matrix Σ,

$$\hat{\Sigma} := n^{-1}\sum_{i=1}^{n} \vec{X}_i\vec{X}_i^\top.$$

The third step. The next problem is how to utilize the set $\{\vec{\gamma}_k\}_{k=1}^K$ for recovering the target space. According to the method SAMM (structural adaptation via maximum minimization method, see Dalalyan et al. (2007)), the "optimal" projector on the target subspace is defined as follows:

$$\widehat{\Pi}_{NG} := \underset{\left\{\substack{\Pi:\Pi^*=\Pi \\ 0\leqslant\Pi\leqslant I \\ \mathrm{tr}(\Pi)\leqslant m}\right\}}{\arg\min} \max_{1\leqslant k\leqslant K} \vec{\gamma_k}^\top(I-\Pi)\vec{\gamma}_k. \tag{8}$$

Remark 1. Note that the inverse covariance matrix is included in the formula (7) but our approach doesn't require the estimation of it. In fact, Lemma 3 is used only for theoretical justification of the first step; practical method described above needs neither the estimation of Σ^{-1} nor the estimation of Γ. On the second step, one uses only the representation (5), which also allows to avoid the estimation of the inverse covariance matrix. The method SAMM doesn't require the estimation of the inverse covariance matrix and therefore can be used in this algorithm.

Representations for the density function of $\vec{X}$

The proofs of Theorem 2 and Lemma 3 are based on the special representation of the density function of $\vec{X}$ that is given in Theorem 1. Similar representations have been studied extensively in previous papers about NGCA. Such facts are usually stated in the following form: if structural assumption (1) is fulfilled, then the density function of a random vector $\vec{X} \in \mathbb{R}^d$ can be represented as

$$p(\vec{x}) = g(T\vec{x})\varphi_A(\vec{x}), \tag{9}$$

where $T : \mathbb{R}^d \to \mathcal{E}$ is a linear transformation ($\mathcal{E}$ — some subspace with $\dim\mathcal{E} = m$), $g : \mathcal{E} \to \mathbb{R}$ — a function, and A — a $d \times d$ symmetric positive matrix. In most papers, formula (9) is proven only for $A = \Gamma$, see e.g. Kawanabe et al. (2007). Another way is to start from the representation (9) without giving the motivation in the spirit of (1), see e.g. Blanchard et al. (2006). In this respect, the result of Theorem 1 can be briefly explained as follows: for any $\vec{X}$ in the form (1), one can find a function g such that (9) is fulfilled with $T = \mathbf{T}$ and $A = \Sigma$.

The existence of a representation in the form (9) can be easily shown in the following way. Note that the model (1) is equivalent to a linear mixing model

$$\vec{X} = A_S\vec{X}_S + A_N\vec{X}_N, \tag{10}$$

where $\vec{X}_S \in \mathbb{R}^m, \vec{X}_N \in \mathbb{R}^{d-m}$ are two random variables; $\vec{X}_N$ is a normal vector with unknown covariance matrix; $\vec{X}_S$ is independent of $\vec{X}_N$; $A_S \in \mathrm{Matr}(d \times m), A_N \in \mathrm{Matr}(d \times (d-m))$ are two deterministic matrices such that columns of these matrices are independent. In this formulation, the signal subspace is spanned by the columns of matrix A_S.

Therefore the vector X is a linear transformation of the vector $\vec{X}' := (\vec{X}_S; \vec{X}_N)$ (this notation means that $\vec{X}'$ is a concatenation of $\vec{X}_S$ and $\vec{X}_N$). This yields that $p(x) \propto g(\vec{X}_S)\varphi(\vec{X}_N)$, where by g we denote the density function of the m-dimensional non-Gaussian component, and by φ — the density function of the normally distributed random variable $\vec{X}_N$. Thus, the representation (9) is proven with some A and T.

As it was mentioned before, the precise formulation for the density of $\vec{X}$ that is needed for our purposes is given in Theorem 1.

Conclusion

The NGCA method has been already widely discussed in the literature (see the introduction). A substantial difference between the previous papers and this article lies in the object of estimation: Blanchard et al, (2006), Diederichs et al. (2010) and other authors concentrate on the estimation of the subspace $\mathcal{T}$ (defined by (6)), and we aim to estimate the subspace $\mathbf{T}$ (see (3)). The estimation of $\mathcal{T}$ also relies on Lemma 3, and this lemma can be found in previous articles. However, the second step is quite new for the NGCA approach. Our methodology (estimation of the signal subspace without estimation of the inverse covariance matrix) has not been previously discussed in the literature. The mathematical base for the second step is given in Theorem 2, which is also new.

Our method relies on the special representation for the density function, which is discussed in Section 4. I would like to stress here that the representation (2) with the operator $\mathbf{T}$ in the form (3) is also a new technical result.

Acknowledgments

The authors are partially supported by Laboratory for Structural Methods of Data Analysis in Predictive Modeling, MIPT, RF government grant, ag. 11.G34.31.0073.

Appendix A. Proofs of the main theoretical facts

Proof of Theorem 1

1. Denote

$$\vec{X}' := \Sigma^{-1/2}\vec{X} = \Sigma^{-1/2}\vec{S} + \Sigma^{-1/2}\vec{N}, \tag{11}$$

and introduce the notation $\vec{S}' = \Sigma^{-1/2}\vec{S}, \quad \vec{N}' = \Sigma^{-1/2}\vec{N}$.

The first component in (11) belongs to the subspace $\mathcal{S}' := \Sigma^{-1/2}\mathcal{S}$. Denote by $\mathcal{N}'$ the subspace that is orthogonal to $\mathcal{S}'$; one can show $\mathcal{N}'$ coincides with the subspace $\Sigma^{1/2}\mathcal{S}^{\perp}$ (Sugiyama et al, 2008).

Vector $\vec{N}'$ can be decomposed into the sum of two vectors, $\vec{N}' = \vec{N}_{\mathcal{S}'} + \vec{N}_{\mathcal{N}'}$, where $\vec{N}_{\mathcal{S}'} \in \mathcal{S}'$, $\vec{N}_{\mathcal{N}'} \in \mathcal{N}'$. This yields that $\vec{X}$ is represented by

$$\vec{X}' = \underbrace{\vec{S}' + \vec{N}_{\mathcal{S}'}}_{\in\,\mathcal{S}'} + \underbrace{\vec{N}_{\mathcal{N}'}}_{\in\,\mathcal{N}'}.$$

Let us choose the basis of $\mathbb{R}^d$ such that the first m vectors $\vec{v}_1, ..., \vec{v}_m$ belong to the subspace $\mathcal{S}'$, while $\vec{v}_{m+1}, ..., \vec{v}_d$ belong to $\mathcal{N}'$. Decompose the vectors $\vec{Z}' := \vec{S}' + \vec{N}_{\mathcal{S}'}$ and $\vec{N}_{\mathcal{N}'}$ into the basis $\vec{v}_1, .., \vec{v}_d$:

$$\vec{Z}' = \sum_{i=1}^{m} z_i \vec{v}_i, \quad \vec{N}_{\mathcal{N}'} = \sum_{i=m+1}^{d} n_i \vec{v}_i, \tag{12}$$

where all coefficients z_i and n_i are random. Since $\vec{X}'$ is a standardized vector,

$$\mathbf{I}_d = \mathbb{E}\left[\vec{X}'\vec{X}'^{\top}\right] = \sum_{i,j=1}^{m} \mathbb{E}\left[z_i z_j\right] \vec{v}_i \vec{v}_j^{\top} + \sum_{i=1}^{m}\sum_{j=m+1}^{d} \mathbb{E}\left[z_i n_j\right] \vec{v}_i \vec{v}_j^{\top} + \sum_{i,j=m+1}^{d} \mathbb{E}\left[n_i n_j\right] \vec{v}_i \vec{v}_j^{\top},$$

and we conclude that the vectors $\vec{Z}'$ and $\vec{N}_{\mathcal{N}'}$ are also standardized.

2. Denote by $F'(\vec{x}')$ and $p'(\vec{x}')$ the distribution function and the density function of the vector $\vec{X}'$,

$$F'(\vec{x}') = \mathbb{P}\left\{\vec{X}' \leqslant \vec{x}'\right\} = \mathbb{P}\left\{\vec{Z}' + \vec{N}_{\mathcal{N}'} \leqslant \vec{x}'\right\}. \tag{13}$$

Note that the vectors $\vec{S}' = \Sigma^{-1/2}\vec{S}$ and $\vec{N}' = \Sigma^{-1/2}\vec{N}$ are independent as the functions of the independent vectors $\vec{S}$ and $\vec{N}$. Therefore, the vectors $\vec{S}'$, $\vec{N}_{\mathcal{N}'}$ and $\vec{N}_{\mathcal{S}'}$ are mutually (not just pairwise) independent. Finally, $\vec{Z}'$ and $\vec{N}_{\mathcal{N}'}$ are independent as functions of independent variables.

Next, the choice of the basis allows to split the inequality

$$\vec{Z}' + \vec{N}_{\mathcal{N}'} \leqslant \vec{x}' = \sum_{i=1}^{d} x_i \vec{v}_i$$

into two

$$\vec{Z}' \leqslant \sum_{i=1}^{m} x_i \vec{v}_i =: \vec{x}_{\mathcal{S}'}, \quad \vec{N}_{\mathcal{N}'} \leqslant \sum_{i=m+1}^{d} x_i \vec{v}_i =: \vec{x}_{\mathcal{N}'},$$

and to represent the function F' as

$$F'(\vec{x}') = \mathbb{P}\left\{\vec{Z}' + \vec{N}_{\mathcal{N}'} \leqslant \vec{x}'\right\} = \mathbb{P}\left\{\vec{Z}' \leqslant \vec{x}_{\mathcal{S}'}, \vec{N}_{\mathcal{N}'} \leqslant \vec{x}_{\mathcal{N}'}\right\} = \mathbb{P}\left\{\vec{Z}' \leqslant \vec{x}_{\mathcal{S}'}\right\} \mathbb{P}\left\{\vec{N}_{\mathcal{N}'} \leqslant \vec{x}_{\mathcal{N}'}\right\}. \tag{14}$$

Taking the derivatives in (14), we arrive at the representation for the density function of $\vec{X}'$.

$$p'(\vec{x}') = q(\vec{x}_{\mathcal{S}'}) \cdot p^{\mathcal{N}}(\vec{x}_{\mathcal{N}'}, \mathbf{I}_{d-m}) = \frac{q(\vec{x}_{\mathcal{S}'})}{p^{\mathcal{N}}(\vec{x}_{\mathcal{S}'}, \mathbf{I}_m)} \cdot p^{\mathcal{N}}(\vec{x}', \mathbf{I}_d) = \frac{q(\Pr_{\mathcal{S}'}\vec{x}')}{p^{\mathcal{N}}(\Pr_{\mathcal{S}'}\vec{x}', \mathbf{I}_m)} \cdot p^{\mathcal{N}}(\vec{x}', \mathbf{I}_d),$$

where by $q(\cdot)$ we denote the density function of the random vector $\vec{Z}' = \vec{S}' + \vec{N}_{\mathcal{S}'} = \Pr_{\mathcal{S}'}\{\vec{X}'\}$.

3. We complete the proof by deriving representation for the density function of $\vec{X} = \Sigma^{1/2}\vec{X}'$ from the density function of $\vec{X}'$. According to the well-known formula for the density transformation,

$$p(\vec{x}) = \det\left(\Sigma^{-1/2}\right)\, p'(\Sigma^{-1/2}\vec{x}) = \det\left(\Sigma^{-1/2}\right) \cdot \frac{q\Big(\Pr_{\mathcal{S}'}\left\{\Sigma^{-1/2}\vec{x}\right\}\Big)}{p^{\mathcal{N}}\Big(\Pr_{\mathcal{S}'}\left\{\Sigma^{-1/2}\vec{x}\right\}, \mathbf{I}_m\Big)} \cdot p^{\mathcal{N}}(\Sigma^{-1/2}\vec{x}, \mathbf{I}_d).$$

The remark $p^{\mathcal{N}}(\Sigma^{-1/2}\vec{x}, \mathbf{I}_d) = \det\left(\Sigma^{1/2}\right) p^{\mathcal{N}}(\vec{x}, \Sigma)$ concludes the proof.

Proof of Lemma 3. We obtain a more general result:

Lemma 5. Assume that the density function of a random vector $\vec{X} \in \mathbb{R}^d$ be represented in the form (9), where $T : \mathbb{R}^d \to \mathcal{E}$ is any linear transformation ($\mathcal{E}$ — any linear space), $g : \mathcal{E} \to \mathbb{R}$ — any function, and A — any $d \times d$ symmetric positive-defined matrix. Then for any function $\psi \in \mathcal{C}^{(1)}(\mathbb{R}^d, \mathbb{R})$, the vector

$$\vec{\beta} := \mathbb{E}\left[\nabla\psi(\vec{X})\right] - A^{-1}\mathbb{E}\left[\vec{X}\psi(\vec{X})\right] \tag{15}$$

belongs to the subspace $(\mathrm{Ker}\ T)^{\perp}$.

Proof. Since for any function ψ and for any $u \in \mathbb{R}^d$,

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

we conclude that under some mild assumptions,

$$\mathbb{E}\left[\nabla\psi(\vec{X})\right] = \int \nabla\left[\psi(\vec{x})\right] p(\vec{x})dx = -\int \psi(\vec{x})\nabla\left[p(\vec{x})\right] d\vec{x}. \tag{16}$$

The gradient of the density function can be represented by the sum of two components:

$$\begin{aligned} \nabla p(\vec{x}) &= \nabla\left[\log p(\vec{x})\right] p(\vec{x}) \\ &= \nabla\left[\log g(T\vec{x})\right] p(\vec{x}) + \nabla\left[\log p^{\mathcal{N}}(\vec{x}, A)\right] p(\vec{x}). \end{aligned} \tag{17}$$

We have

$$\begin{aligned} \nabla\left[\log g(T\vec{x})\right] p(\vec{x}) &= \frac{\nabla g(T\vec{x})}{g(T\vec{x})} p(\vec{x}) = \nabla\left[g(T\vec{x})\right] p^{\mathcal{N}}(\vec{x}, A) \\ &= p^{\mathcal{N}}(\vec{x}, A) \cdot T^{\top}\, \nabla_{\vec{y}}\left[g(\vec{y})\right]|_{\vec{y}=T\vec{x}}, \\ \nabla\left[\log p^{\mathcal{N}}(\vec{x}, A)\right] p(\vec{x}) &= -A^{-1}\vec{x}p(\vec{x}). \end{aligned}$$

Denote

$$\vec{\Lambda} = -\int \psi(\vec{x}) \cdot p^{\mathcal{N}}(\vec{x}, A) \cdot \nabla_{\vec{y}}\left[g(\vec{y})\right]|_{\vec{y}=T\vec{x}}\, d\vec{x}.$$

Then (15) follows with $\vec{\beta} := T^\top \vec{\Lambda} \in (\mathrm{Ker}\ T)^\perp$ because (16) together with (17) yield

$$\mathbb{E}\left[\nabla\psi(\vec{X})\right] = T^\top \vec{\Lambda} + A^{-1}\mathbb{E}\left[\vec{X}\psi(\vec{X})\right].$$

Proof of Theorem 2

The proof is straightforward:

$$\begin{aligned}\mathrm{Ker}\,\mathbf{T} &= \left\{\vec{x}:\ \Sigma^{-1/2}\vec{x} \perp \Sigma^{-1/2}\mathcal{S}\right\} \\ &= \left\{\vec{x}:\ \exists \vec{s} \in \mathcal{S} \mid \vec{x}^\top \left(\Sigma^{-1/2}\right)^\top \Sigma^{-1/2}\vec{s} = 0\right\} = \left\{\vec{x}:\ \exists \vec{s} \in \mathcal{S} \mid \vec{x}^\top\ \Sigma^{-1}\vec{s} = 0\right\} \\ &= \left\{\vec{x}:\ \vec{x} \perp \Sigma^{-1}\mathcal{S}\right\}.\end{aligned}$$

Appendix B. Choice of ψ by the convex projection method

The appendix briefly explains the method of estimation of the functions $\{\psi\}$ called the convex projection method (Diederichs (2007) and Diederichs et al. (2010)). The method gives an algorithm for finding one function ψ; one can repeat the algorithm with different parameters and receive the whole set $\{\psi\}$.

The core of the method is the choice of the function ψ in the following form:

$$\psi(\vec{x}) := \sum_{j=1}^{J} c_j \psi_j(\vec{x}); \qquad \psi_j(\vec{x}) = f(\vec{\omega}_j^T \vec{x}) e^{-||\vec{x}||^2/2}, \tag{18}$$

where f can be any smooth function; for the numerical simulations, Diederichs uses $f(z) = = f_1(z)$ or $f(z) = (1+z^2)^{-1}e^z$; "directions"$\{\omega_j, j = 1..J\}$ are preliminary estimated through the Monte-Carlo sampling from the uniform distribution on the sphere; $\vec{c} = \{c_j, j = 1..J\}$ is a vector from the $\mathscr{L}_1$ - unit ball, which is the object of estimation.

Corollary 4 yields that for any function $\psi \in \mathcal{C}^{(1)}(\mathbb{R}^d, \mathbb{R})$ such that $\mathbb{E}\left[\vec{X}\psi(\vec{X})\right] = 0$, the vector $\mathbb{E}\left[\nabla\psi(\vec{X})\right]$ belongs to $(\mathrm{Ker}\ \mathbf{T})^\perp$. Changing the mathematical expectations in this statement by their empirical counterparts, we conclude that if the vector $\vec{c}$ is chosen such that

$$\gamma(\vec{c}) := \frac{1}{JN}\sum_{j=1}^{J} c_j \left(\sum_{i=1}^{N} \psi_j(\vec{X}_i)\vec{X}_i\right) \approx \vec{0}, \tag{19}$$

then

$$\vec{\beta}(\vec{c}) := \frac{1}{JN}\sum_{j=1}^{J} c_j \left(\sum_{i=1}^{N} \nabla\psi_j(\vec{X}_i)\right) \tag{20}$$

can be considered as the estimate of a vector from $(\mathrm{Ker}\ \mathbf{T})^\perp$. Diederichs (2007) proposes to estimate the coefficient vector $\vec{c}$ by solving the optimization problem

$$(\hat{c}_1, \hat{c}_2, ..., \hat{c}_J) := \arg\min_{\vec{c}}\left\{\|\vec{\xi} - \vec{\beta}(\vec{c})\|_2 \;\middle|\; \gamma(\vec{c}) = \vec{0},\ \|\vec{c}\|_1 \leqslant 1\right\}, \tag{21}$$

where $\vec{\xi}$ is a unit vector, which we call *a probe vector*. Afterwards the estimate of the function ψ is set to $\psi(\vec{x}) := \sum_{j=1}^{J} \hat{c}_j \psi_j(\vec{x})$.

See the article by Diederichs et al. (2007) for examples.

Another popular method for finding ψ was introduced by Blanchard et al. (2006). We don't present that method here because it uses the estimator of the inverse covariance matrix Σ^{-1} and contradicts the philosophy of this paper.

Литература

[1] Belomestny D., Spokoiny V. (2007). Spatial aggregation of local likelihood estimates with applications to classification. *Ann. Statist.* **35**, 2287–2311.

[2] Blanchard, G., Kawanabe, M., Sugiyama, M., Spokoiny, V., Müller, K.-R. (2006). In search of non-Gaussian components of a high-dimensional distribution. *J. Mach. Learn. Res.* **6**, 247–282.

[3] Cook, R.D. (1998). Principal hessian directions revised. *J. Am. Statist. Ass.* **93**, 85-100.

[4] Dalalyan, A., Juditsky, A., Spokoiny, V. (2007). A new algorithm for estimating the effective dimension — reduction subspace. *J. Mach. Learn. Res.* **9**, 1647–1678.

[5] Diaconis, P., and Friedman, D. (1984). Asymptotics of graphical projection pursuit. *Ann.Statist.* **12(3)**, 793–815.

[6] Diederichs, E. (2007). *Semi-parametric reduction of dimensionality.* Ph.D. thesis. Free University of Berlin.

[7] Diederichs, E., Juditsky, A., Spokoiny, V., Schütte, C. (2010). Sparse non-Gaussian component analysis. *IEEE Trans. Inf. Theory.* **15**, 3033–3047.

[8] Hall, P. (1989). Projection pursuit methods. *Ann. Statist.* **17**, 589–605.

[9] Hristache, M., Juditsky A., Polzehl, J., Spokoiny, V. (2001). Structure adaptive approach for dimension reduction. *Ann. Statist.* **6**, 1537–1566.

[10] Jolliffe, I.T. (2002). *Principal component analysis.* Springer series in statistics. Springer, Berlin and New York, 2nd edition.

[11] Hyvärinen, A., Karhunen, J., and Oja, E. (2001) *Independent Component Analysis.* Wiley, New York.

[12] Kawanabe, M., Sugiyama, M., Blanchard, G., Müller, K.-R. (2007). A new algorithm of non-Gaussian component analysis with radial kernel functions. *Ann. Inst. Stat. Math.* **59**, 57–75.

[13] Theis, F.J. and Kawanabe, M. (2007). Uniqueness of non-Gaussian subspace analysis. *Proc. ICA.* **4666**, 917–925, Springer, London.