OBTAINING THE BEST LINEAR UNBIASED ESTIMATOR OF NOISY SIGNALS BY NON-GAUSSIAN COMPONENT ANALYSIS

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

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

ABSTRACT

Obtaining the best linear unbiased estimator (BLUE) of noisy signals is a traditional but powerful approach to noise reduction. Explicitly computing BLUE usually requires the prior knowledge of the subspace to which the true signal belongs and the noise covariance matrix. However, such prior knowledge is often unavailable in reality, which prevents us from applying BLUE to real-world problems. In this paper, we therefore give a method for obtaining BLUE *without* such prior knowledge. Our additional assumption is that the true signal follows a non-Gaussian distribution while the noise is Gaussian.

1. INTRODUCTION

One of the most fundamental and important problems in signal processing is reduction of the noise contained in observed signals. In this paper, we focus on the cases where the observed signal is the sum of a low-dimensional true signal and full-dimensional noise.

A traditional but useful approach to noise reduction in this setting is to obtain the *best linear unbiased estimator* (BLUE) of the observed signals [1], i.e., the linear unbiased estimator of the true signal that has the minimum variance among all linear unbiased estimators. BLUE can be obtained by linearly projecting the observed signal onto the subspace $\mathcal{S}$ of the true signal along a (not necessarily orthogonal) complementary subspace $\mathcal{T}$ determined by the noise covariance matrix $\boldsymbol{Q}$. In order to explicitly compute BLUE, the signal subspace $\mathcal{S}$ and the noise covariance matrix $\boldsymbol{Q}$ are required. Unfortunately, however, they are typically unknown in practice, so the application of BLUE to real-world problems has been rather limited so far. The purpose of this paper is to theoretically show a possibility of obtaining BLUE *without* the prior knowledge of $\mathcal{S}$ and $\boldsymbol{Q}$.

We first show that BLUE can be obtained by using the data covariance matrix $\boldsymbol{\Sigma}$, *without* using the noise covariance matrix $\boldsymbol{Q}$. This finding is practically meaningful because estimating the data covariance matrix $\boldsymbol{\Sigma}$ can be directly carried out in a consistent manner using the data samples, while estimating the noise covariance matrix $\boldsymbol{Q}$ is not a straightforward task in general. Thanks to this result, we only need to estimate the signal subspace $\mathcal{S}$ for obtaining BLUE.

Then we outline our new method named *non-Gaussian component analysis* (NGCA)[2], which enables us to identify the desired signal subspace $\mathcal{S}$ under the assumption that the signal components follow a non-Gaussian distribution while the noise is Gaussian. It is shown experimentally that by NGCA, $\mathcal{S}$ can be successfully identified, and thus our contribution in this paper opens a possibility of obtaining BLUE *without* the prior knowledge of $\mathcal{S}$ and $\boldsymbol{Q}$.

2. FORMULATION

Let $\boldsymbol{x} \in \mathbb{R}^d$ be the observed noisy signal, which is composed of an unknown true signal $\boldsymbol{s}$ and noise $\boldsymbol{n}$.

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

We treat $\boldsymbol{s}$ and $\boldsymbol{n}$ as random variables (thus $\boldsymbol{x}$ also), and assume $\boldsymbol{s}$ and $\boldsymbol{n}$ are statistically independent. We further suppose that the true signal $\boldsymbol{s}$ lies in a subspace $\mathcal{S} \subset \mathbb{R}^d$ of *known* dimension $m = \dim(\mathcal{S})$, where $1 \leq m < d$. On the other hand, the noise $\boldsymbol{n}$ spreads out over the entire space $\mathbb{R}^d$ and is assumed to be mean zero. Following this generative model, we are given a set of i.i.d. observations $\{\boldsymbol{x}_i\}_{i=1}^n$. Our goal is to obtain a set of denoised signals $\{\widehat{\boldsymbol{s}}_i\}_{i=1}^n$ that are *close* to the true signals $\{\boldsymbol{s}_i\}_{i=1}^n$.

A standard approach to noise reduction in this setting is to project the noisy signal $\boldsymbol{x}$ onto the true signal subspace $\mathcal{S}$, by which the noise is reduced while the signal component $\boldsymbol{s}$ is still preserved. Here the projection does not have to be orthogonal, thus we may want to optimize the projection direction so that the maximum amount of noise can be removed.

In statistics, the linear estimator which fulfills the above requirement is called the *best linear unbiased estimator (BLUE)* [1]. BLUE has the minimum variance among all linear unbiased estimators. More precisely, in the current set-

ting, BLUE of $\boldsymbol{s}$ denoted by $\widehat{\boldsymbol{s}}$ is defined by

$$\widehat{\boldsymbol{s}} = \boldsymbol{H}\boldsymbol{x}, \tag{2}$$

where, with $\mathbb{E}_{\boldsymbol{n}|\boldsymbol{s}}$ being the conditional expectation over the noise $\boldsymbol{n}$ given signal $\boldsymbol{s}$,

$$\boldsymbol{H} = \operatorname*{argmin}_{\widetilde{\boldsymbol{H}} \in \mathbb{R}^{d \times d}} \mathbb{E}_{\boldsymbol{n}|\boldsymbol{s}} \|\widetilde{\boldsymbol{H}}\boldsymbol{x} - \mathbb{E}_{\boldsymbol{n}|\boldsymbol{s}}\widetilde{\boldsymbol{H}}\boldsymbol{x}\|^2$$
$$\text{subject to} \quad \mathbb{E}_{\boldsymbol{n}|\boldsymbol{s}}[\widetilde{\boldsymbol{H}}\boldsymbol{x}] = \boldsymbol{s}. \tag{3}$$

Let $\boldsymbol{Q}$ be the noise covariance matrix:

$$\boldsymbol{Q} = \mathbb{E}_{\boldsymbol{n}}[\boldsymbol{n}\boldsymbol{n}^\top], \tag{4}$$

which we assume non-degenerated. Let $\boldsymbol{P}$ be the orthogonal projection matrix onto the subspace $\mathcal{S}$. Then, the estimation matrix $\boldsymbol{H}$ is given by (see e.g., [1, 3])

$$\boldsymbol{H} = (\boldsymbol{P}\boldsymbol{Q}^{-1}\boldsymbol{P})^\dagger \boldsymbol{Q}^{-1}, \tag{5}$$

where † denotes the Moore-Penrose generalized inverse. Let

$$\mathcal{T} = \boldsymbol{Q}\mathcal{S}^\perp, \tag{6}$$

where $\mathcal{S}^\perp$ is the orthogonal complement of $\mathcal{S}$. Then it can be confirmed that $\boldsymbol{H}$ is an *oblique* projection onto $\mathcal{S}$ along $\mathcal{T}$ (see e.g., [3]):

$$\boldsymbol{H}\boldsymbol{x} = \begin{cases} \boldsymbol{x} & \text{if } \boldsymbol{x} \in \mathcal{S}, \\ \boldsymbol{0} & \text{if } \boldsymbol{x} \in \mathcal{T}. \end{cases} \tag{7}$$

This is illustrated in Figure 1.

When calculating BLUE by Eqs.(2) and (5), the signal subspace $\mathcal{S}$ and the noise covariance matrix $\boldsymbol{Q}$ should be known. However, $\mathcal{S}$ and $\boldsymbol{Q}$ are often unknown in practice, and estimating them from data samples $\{\boldsymbol{x}_i\}_{i=1}^n$ is not generally a straightforward task. For this reason, the applicability of BLUE to real-world problems has been rather limited so far.

In this paper, we therefore propose a new algorithm which opens a possibility to obtain BLUE even in the *absence* of the prior knowledge of $\mathcal{S}$ and $\boldsymbol{Q}$.

3. OBTAINING BLUE WITHOUT Q

In this section, we show that $\boldsymbol{Q}$ is not needed in computing BLUE.

With some abuse, we call the following matrix $\boldsymbol{\Sigma}$ the *data covariance matrix*, although we do not assume $\boldsymbol{x}$ has mean zero:

$$\boldsymbol{\Sigma} = \mathbb{E}_{\boldsymbol{x}}[\boldsymbol{x}\boldsymbol{x}^\top]. \tag{8}$$

Then we have the following lemmas.

Lemma 1 *The subspace $\mathcal{T}$ is expressed as*

$$\mathcal{T} = \boldsymbol{\Sigma}\mathcal{S}^\perp. \tag{9}$$

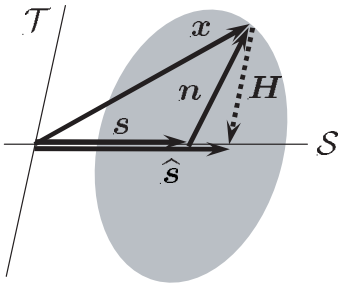

Fig. 1. Illustration of BLUE.

(Proof) Since $\boldsymbol{s}$ and $\boldsymbol{n}$ are independent and $\mathbb{E}_{\boldsymbol{n}}[\boldsymbol{n}] = \boldsymbol{0}$,

$$\begin{aligned}\boldsymbol{\Sigma}\mathcal{S}^\perp &= \mathbb{E}_{\boldsymbol{s}}[\boldsymbol{s}\boldsymbol{s}^\top\mathcal{S}^\perp] + \mathbb{E}_{\boldsymbol{x}}[\boldsymbol{s}\boldsymbol{n}^\top + \boldsymbol{n}\boldsymbol{s}^\top]\mathcal{S}^\perp + \mathbb{E}_{\boldsymbol{n}}[\boldsymbol{n}\boldsymbol{n}^\top]\mathcal{S}^\perp \\ &= \boldsymbol{Q}\mathcal{S}^\perp, \end{aligned} \tag{10}$$

which establishes Eq.(9). ■

Lemma 2 *The estimation matrix $\boldsymbol{H}$ is expressed as*

$$\boldsymbol{H} = (\boldsymbol{P}\boldsymbol{\Sigma}^{-1}\boldsymbol{P})^\dagger \boldsymbol{\Sigma}^{-1}. \tag{11}$$

(Proof) Since the null space of $(\boldsymbol{P}\boldsymbol{\Sigma}^{-1}\boldsymbol{P})^\dagger$ is $\mathcal{S}^\perp$, we have

$$(\boldsymbol{P}\boldsymbol{\Sigma}^{-1}\boldsymbol{P})^\dagger\boldsymbol{\Sigma}^{-1} = (\boldsymbol{P}\boldsymbol{\Sigma}^{-1}\boldsymbol{P})^\dagger\boldsymbol{P}\boldsymbol{\Sigma}^{-1} \equiv \overline{\boldsymbol{H}}. \tag{12}$$

Let $\boldsymbol{P}_\perp$ be the orthogonal projection matrix onto $\mathcal{S}^\perp$. Then, for any $\boldsymbol{x} \in \mathbb{R}^d$, we have

$$\overline{\boldsymbol{H}}(\boldsymbol{P}\boldsymbol{x}) = (\boldsymbol{P}\boldsymbol{\Sigma}^{-1}\boldsymbol{P})^\dagger(\boldsymbol{P}\boldsymbol{\Sigma}^{-1}\boldsymbol{P})\boldsymbol{x} = \boldsymbol{P}\boldsymbol{x}, \tag{13}$$
$$\overline{\boldsymbol{H}}(\boldsymbol{\Sigma}\boldsymbol{P}_\perp\boldsymbol{x}) = (\boldsymbol{P}\boldsymbol{\Sigma}^{-1}\boldsymbol{P})^\dagger(\boldsymbol{P}\boldsymbol{P}_\perp)\boldsymbol{x} = \boldsymbol{0}, \tag{14}$$

which are equivalent to Eq.(7). Thus, $\boldsymbol{H} = \overline{\boldsymbol{H}}$. ■

Lemma 2 implies that we can obtain BLUE using the data covariance matrix $\boldsymbol{\Sigma}$, *without* using the noise covariance matrix $\boldsymbol{Q}$. Roughly speaking, the "$\mathcal{T}$-part" of $\boldsymbol{Q}$ should agree with that of $\boldsymbol{\Sigma}$ because the signal $\boldsymbol{s}$ lies only in $\mathcal{S}$. Therefore, it intuitively seems that we can replace $\boldsymbol{Q}$ in $\boldsymbol{H}$ by $\boldsymbol{\Sigma}$ because $\boldsymbol{H}$ only affects the component in $\mathcal{T}$ (see Eq. (7)). The above lemma theoretically supports this intuitive claim.

A practical advantage of the above lemma is that, while estimating the noise covariance matrix $\boldsymbol{Q}$ from the data samples $\{\boldsymbol{x}_i\}_{i=1}^n$ is not a straightforward task in general, $\boldsymbol{\Sigma}$ can be directly estimated in a consistent way as

$$\widehat{\boldsymbol{\Sigma}} = \frac{1}{n}\sum_{i=1}^n \boldsymbol{x}_i\boldsymbol{x}_i^\top. \tag{15}$$

4. ESTIMATING $\mathcal{S}$ BY NON-GAUSSIAN COMPONENT ANALYSIS

Given Eqs.(11) and (15), the remaining issue to discuss is how to estimate the true signal subspace $\mathcal{S}$ (or the orthogonal projection matrix $\boldsymbol{P}$). Here we assume that the signal $\boldsymbol{s}$ follows an unknown *non-Gaussian* distribution, while the

noise $\boldsymbol{n}$ follows a *Gaussian* distribution with mean zero and unknown covariance matrix $\boldsymbol{Q}$. Under this assumption, we can apply our method named the *non-Gaussian component analysis* (NGCA), that allows to identify the signal subspace $\mathcal{S}$. In the following, we will briefly review the main concepts of the NGCA algorithm[1].

Let us now transform the data samples $\{\boldsymbol{x}_i\}_{i=1}^n$ as

$$\boldsymbol{y}_i = \widehat{\boldsymbol{\Sigma}}^{-\frac{1}{2}} \boldsymbol{x}_i \quad \text{for} \quad i = 1, \ldots, n, \tag{16}$$

and let $\widetilde{\mathcal{S}}$ be the corresponding counterpart of $\mathcal{S}$:

$$\widetilde{\mathcal{S}} = \widehat{\boldsymbol{\Sigma}}^{\frac{1}{2}} \mathcal{S}. \tag{17}$$

Then we have the following lemma.

Lemma 3 *For an arbitrary smooth real function $f(t)$ on $\mathbb{R}$ and an arbitrary vector $\boldsymbol{w}$ in $\mathbb{R}^d$, the following vector $\boldsymbol{\beta}$ approximately belongs to $\widetilde{\mathcal{S}}$.*

$$\boldsymbol{\beta} = \tfrac{1}{n} \textstyle\sum_{i=1}^n g(\boldsymbol{w}, \boldsymbol{y}_i), \tag{18}$$

where, for $f'(t)$ being the derivative of $f(t)$,

$$g(\boldsymbol{w}, \boldsymbol{y}) = \boldsymbol{y} f(\boldsymbol{w}^\top \boldsymbol{y}) - \boldsymbol{w} f'(\boldsymbol{w}^\top \boldsymbol{y}). \tag{19}$$

The lemma implies that for a family $\{f_k(\boldsymbol{x})\}_{k=1}^L$ of smooth functions, we can create a family $\{\boldsymbol{\beta}_k\}_{k=1}^L$ of vectors which all approximately belong to $\widetilde{\mathcal{S}}$. Note that the accuracy is theoretically guaranteed in the sense of a *uniform-convergence*—the estimation error *quickly* vanishes as the number of data samples tends to infinity (see the separate paper [2] for full rigorous theoretical analyses).

We then apply principal component analysis (PCA) [4] to $\{\boldsymbol{\beta}_k\}_{k=1}^L$ and extract m leading eigenvectors $\{\boldsymbol{\psi}_i\}_{i=1}^m$ as an estimate of an orthonormal basis in $\widetilde{\mathcal{S}}$. Finally, by pulling back the result into the original space, we obtain the following estimate of $\boldsymbol{P}$:

$$\widehat{\boldsymbol{P}} = \widehat{\boldsymbol{\Sigma}}^{-\frac{1}{2}} \boldsymbol{\Psi} (\widehat{\boldsymbol{\Sigma}}^{-\frac{1}{2}} \boldsymbol{\Psi})^\dagger, \quad \text{where} \quad \boldsymbol{\Psi} = (\boldsymbol{\psi}_1 \cdots \boldsymbol{\psi}_m). \tag{20}$$

Eqs.(18) and (19) imply that the mapping from f to $\boldsymbol{\beta}$ is linear. Therefore, we can arbitrarily change the norm of $\boldsymbol{\beta}$ just by multiplying f by an arbitrary scalar. This can totally corrupt the PCA results since vectors with larger norm have stronger impacts on the PCA solutions. For this reason, the norm of $\{\boldsymbol{\beta}_k\}_{k=1}^L$ should be reasonably normalized. A desirable normalization scheme would be that $\boldsymbol{\beta}$ should have a larger norm if it accurately belongs to $\widetilde{\mathcal{S}}$ (in the sense that the angle between $\boldsymbol{\beta}$ and $\widetilde{\mathcal{S}}$ is small). This can be achieved by normalizing $\{\boldsymbol{\beta}_k\}_{k=1}^L$ by its standard deviation (cf. [2]). A consistent estimator of the variance of $\boldsymbol{\beta}$ can be obtained as

$$N = \tfrac{1}{n} \textstyle\sum_{i=1}^n \|g(\boldsymbol{w}, \boldsymbol{y}_i)\|^2 - \|\boldsymbol{\beta}\|^2. \tag{21}$$

[1] Details of the algorithm including more formal formulation, proofs, extensive theoretical analysis, and additional experimental results are available in the separate paper [2].

For $k = 1, \ldots, L$
  Randomly initialize $\boldsymbol{w}_0 \in \mathbb{R}^d$ such that $\|\boldsymbol{w}_0\| = 1$.
  For $t = 1, \ldots, T$ **% Heuristic update of $\boldsymbol{w}$**
    $\boldsymbol{\beta}_t = \frac{1}{n} \sum_{i=1}^n g_k(\boldsymbol{w}_{t-1}, \boldsymbol{y}_i)$.
    $\boldsymbol{w}_t = \boldsymbol{\beta}_t / \|\boldsymbol{\beta}_t\|$.
  End
  $N_k = \frac{1}{n} \sum_{i=1}^n \|g_k(\boldsymbol{w}_{T-1}, \boldsymbol{y}_i)\|^2 - \|\boldsymbol{\beta}_T\|^2$.
  $\boldsymbol{v}_k = \boldsymbol{\beta}_T / \sqrt{N_k}$. **% Normalization**
End
$\mathcal{V} = \{\boldsymbol{v}_k \mid \|\boldsymbol{v}_k\| \geq \varepsilon\}$. **% Extract informative vectors**
Apply PCA to $\mathcal{V}$ and extract m leading vectors $\{\boldsymbol{\psi}_i\}_{i=1}^m$.
$\widehat{\boldsymbol{P}} = \widehat{\boldsymbol{\Sigma}}^{-\frac{1}{2}} \boldsymbol{\Psi} (\widehat{\boldsymbol{\Sigma}}^{-\frac{1}{2}} \boldsymbol{\Psi})^\dagger$ with $\boldsymbol{\Psi} = (\boldsymbol{\psi}_1 \cdots \boldsymbol{\psi}_m)$.
Output $\widehat{\boldsymbol{H}} = (\widehat{\boldsymbol{P}}^\top \widehat{\boldsymbol{\Sigma}}^{-1} \widehat{\boldsymbol{P}})^\dagger \widehat{\boldsymbol{\Sigma}}^{-1}$.

Fig. 2. Pseudocode of NGCA algorithm.

Given that $\{\boldsymbol{\beta}_k\}_{k=1}^L$ are reasonably normalized, $\|\boldsymbol{\beta}\|$ expresses the amount of "information" on $\widetilde{\mathcal{S}}$ which $\boldsymbol{\beta}$ carries. Namely, the larger $\|\boldsymbol{\beta}\|$ is, the more information on $\widetilde{\mathcal{S}}$ it contains. This implies that the accuracy of the above procedure can be further improved by updating the vector $\boldsymbol{w}$ to increase $\|\boldsymbol{\beta}\|$. Furthermore, if $\|\boldsymbol{\beta}\|$ is still small after the search, such $\boldsymbol{\beta}$ may be ignored since it does not contain enough information on $\widetilde{\mathcal{S}}$. The final NGCA algorithm is sketched in Figure 2.

For identifying the non-Gaussian subspace $\mathcal{S}$, the method of *projection pursuit* (PP) [5, 4] could also be used. PP tries to iteratively find directions with maximum "non-Gaussianity" based on a prefixed *projection index*. A particular implementation of PP actually corresponds to running the NGCA algorithm for a single function $f(t)$ [4]. Therefore, NGCA can be regarded as an extension of PP, which is beneficial in the following sense: It is known that some projection indices are suitable for finding directions with *super-Gaussian* (heavy-tailed) distributions while others are suited for finding *sub-Gaussian* (light-tailed) distributions [4]. Therefore, if the target signal consists of, say, both super- and sub-Gaussian distributions, it is practically difficult to choose the single "right" projection index. On the other hand, in NGCA, we do not need to fix the projection indices and can use many indices at the same time. Thus, suitable indices are automatically found by the algorithm.

5. SIMULATIONS

Here we briefly report exemplary numerical results using the following 4 data sets (see the separate paper [2] for additional experimental analyses). Each data set includes $n = 1000$ samples in $d = 10$ dimension and each sample consists of 8-dimensional independent standard Gaussian. Other $m = 2$ non-Gaussian components are as follows.

(A) Simple Gaussian mixture: 2-dimensional independent Gaussian mixtures with density of each component given by $\phi_{-3,1}(x)/2 + \phi_{3,1}(x)/2$, where $\phi_{\mu,\sigma^2}(x)$ is the Gaussian density with mean μ and variance σ^2.

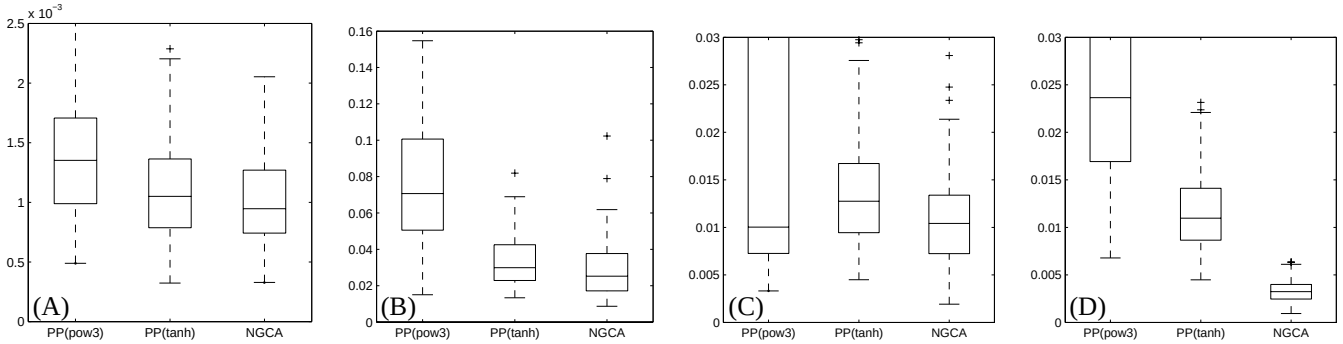

Fig. 3. Boxplots of the estimation error of $\boldsymbol{H}$.

(B) Dependent super-Gaussian: 2-dimensional isotropic distribution with density proportional to $\exp(-\|\boldsymbol{x}\|)$.
(C) Dependent sub-Gaussian: 2-dimensional isotropic uniform with constant positive density for $\|\boldsymbol{x}\| \leq 1$ 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 so that each coordinate has unit variance.

For NGCA, we used the following set of functions:

$$\mathbf{Gauss-pow3}: \ f(t) = t^3 \exp(-t^2/(2\sigma^2)), \quad (22)$$

$$\mathbf{tanh}: \ f(t) = \tanh(bt), \quad (23)$$

$$\mathbf{sin}: \ f(t) = \sin(a_s t), \quad (24)$$

$$\mathbf{cos}: \ f(t) = \cos(a_c t), \quad (25)$$

where $\sigma^2, b, a_s, a_c \in \mathbb{R}^+$ are parameters which enrich the function families. Parameters are set as $a_s, a_c \in [0, 4]$, $b \in [0, 5]$, and $\sigma^2 \in [0.5, 5]$, where each of these ranges was divided into 1000 equispaced values—thus yielding a family of $L = 4000$ functions. We set $\varepsilon = 1.5$ and $T = 10$. We compared NGCA with PP with "pow3" index (corresponding to $f(t) = t^3$) or "tanh" index (corresponding to $f(t) = \tanh(t)$).

Figure 3 shows the boxplots of the estimation error over 100 runs measured by the following criterion:

$$\text{Error}(\widehat{\boldsymbol{H}}, \boldsymbol{H}) = \|\widehat{\boldsymbol{H}} - \boldsymbol{H}\|^2/2m. \quad (26)$$

For the simplest data set (A), NGCA is comparable or slightly better than PPs. 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) [4]. 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 of PP(pow3) is very large for the data set (C) because PP sometimes got trapped in local minima. For the data sets (B) and (C), NGCA appears to be comparable or slightly better than PPs with the "right" index. The superiority of the index integration 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 errors than PPs.

6. CONCLUSIONS

For calculating BLUE, prior knowledge of the noise covariance matrix $\boldsymbol{Q}$ and the signal subspace $\mathcal{S}$ are usually needed. In this paper, we showed that BLUE can be obtained without $\boldsymbol{Q}$, and $\mathcal{S}$ can be successfully identified by the proposed NGCA method. Thus our contribution opens a possibility of obtaining BLUE *without* the prior knowledge of $\boldsymbol{Q}$ and $\mathcal{S}$.

We acknowledge partial financial support by DFG, BMBF, the EU NOE PASCAL (EU # 506778), Alexander von Humboldt foundation, and MEXT (Grant-in-Aid for Young Scientists 17700142).

7. REFERENCES

[1] A. Albert, *Regression and the Moore-Penrose Pseudoinverse*, Academic Press, New York and London, 1972.

[2] G. Blanchard, M. Kawanabe, M. Sugiyama, V. Spokoiny, and K.-R. Müller, "In search of non-Gaussian components of a high-dimensional distribution," *Journal of Machine Learning Research*, 2006, to appear.

[3] H. Ogawa and N.-E. Berrached, "EPBOBs (extended pseudo biorthogonal bases) for signal recovery," *IEICE Transactions on Information and Systems*, vol. E83-D, no. 2, pp. 223–232, 2000.

[4] A. Hyvärinen, J. Karhunen, and E. Oja, *Independent Component Analysis*, John Wiley & Sons, New York, 2001.

[5] J. H. Friedman and J. W. Tukey, "A projection pursuit algorithm for exploratory data analysis," *IEEE Transactions on Computers*, vol. C-23, no. 9, pp. 881–890, 1974.