4th International Symposium on Independent Component Analysis and Blind Signal Separation (ICA2003), April 2003, Nara, Japan

ADAPTIVE SELECTION FOR MINIMUM β-DIVERGENCE METHOD

Mihoko Minami and Shinto Eguchi

The Institute of Statistical Mathematics and the Graduate University for Advanced Studies
Minato-ku, Tokyo 106-8569, Japan.
mminami@ism.ac.jp, eguchi@ism.ac.jp

ABSTRACT

Many estimators in the literature of blind source separation can be considered as the estimators derived through the framework of the maximum likelihood estimation with various choices of density functions. In other words, they are the minimizers of the Kullback-Leibler divergence between the empirical distribution and a certain form of density function. Unfortunately, this type of estimators is not B-robust (see [8] for B-robustness), that is, it can be easily affected by outliers.

Minami and Eguchi [12] proposed a robust estimator for blind source separation based on the β-divergence. We call it the minimum β-divergence estimator. It was shown that the estimator is locally consistent and B-robust, and the necessary and sufficient condition for asymptotical stability was given.

The tuning parameter β plays a key role on the robust property for the minimum β-divergence estimator. The larger β is, the more robust the corresponding estimator is. However, too large β might produce a less efficient estimator. We propose a selection procedure of the tuning parameter β for the minimum β-divergence estimator.

1. INTRODUCTION

Blind source separation is the problem of recovering the original independent signals when only their linear mixtures are observed. Let $\boldsymbol{s}(t)$ be a vector of m source signals whose components are independent of each other. We cannot observe original signals directly, but observe $\boldsymbol{x}(t)$ linearly transformed by

$$\boldsymbol{x}(t) = A\,\boldsymbol{s}(t), \quad t = 1, 2, ..., n$$

where A is a non-singular unknown matrix of size m. In order to recover independent signals, we estimate recovering matrix W that transforms observed signals $\boldsymbol{x}(t)$ into independent signals $\boldsymbol{y}(t)$:

$$\boldsymbol{y}(t) = W\,\boldsymbol{x}(t), \quad t = 1, 2, ..., n$$

(cf. [10]). If W is properly obtained, $\boldsymbol{y}(t)$ is equal to $\boldsymbol{s}(t)$ except for an arbitrary scaling of each signal component and permutation of indices. Distributions of original signals are unknown although specific density functions might be used to define loss functions or to derive estimating equations.

If components are independent of each other, the joint density of $\boldsymbol{y}$ is expressed as the product of marginal density functions as:

$$q(\boldsymbol{y}) = \prod_{i=1}^{m} q_i(y_i)$$

and the joint density function of $\boldsymbol{x}$ can be expressed as:

$$r(\boldsymbol{x}, W) = |\det(W)| \prod_{i=1}^{m} q_i(\boldsymbol{w}_i \boldsymbol{x}). \tag{1}$$

Thus, minimizing some kind of divergence between the empirical distribution of $\boldsymbol{y}$ and the distribution expressed as (1) is a reasonable strategy to find an estimate.

Most commonly used divergence is the Kullback-Leibler (K-L) divergence. The method based on the K-L divergence is, in other words, an approach through the framework of the maximum likelihood estimation method. Many estimators, including Jutten & Herault's heuristic approach [11], entropy maximization [5], minimizing cross cumulants [6], approximation of mutual information by Gram-Charlier expansion and the natural gradient approach [3], have the same type of estimating functions as those derived from the maximum likelihood approach. Amari & Cardoso [1] showed that this type of estimating functions are unbiased provided means of original signals are zeros. Amari, Chen, & Cichocki [2] gave the necessary and sufficient condition for asymptotic stability. These imply that under certain condition, these methods produce locally consistent estimates for a recovering matrix. However, one problem shared with this type of estimators is that they are not robust to outliers. Their estimating functions are unbounded as functions of signals, and this implies

that a few outliers could change the estimate drastically.

The β-divergence [7] between two density functions with respect to some carrier measure ν is defined as:

$$D_\beta(g,f) = \frac{1}{\beta}\int \left(g^\beta(\boldsymbol{x}) - f^\beta(\boldsymbol{x})\right) g(\boldsymbol{x})\, d\nu(\boldsymbol{x}) \quad (2)$$
$$-\frac{1}{\beta+1}\int \left(g^{\beta+1}(\boldsymbol{x}) - f^{\beta+1}(\boldsymbol{x})\right) d\nu(\boldsymbol{x})$$

for $\beta > 0$. As β goes down to 0, $D_\beta(g,f)$ approaches the K-L divergence, thus, we define the K-L divergence as the β-divergence with $\beta = 0$. The β-divergence is equivalent to the density power divergence $d_\beta(g,\ f)$ [4].

Minami and Eguchi [12] proposed a robust blind source separation method based on the β-divergence. They defined the minimizer of the β-divergence between the joint empirical distribution and the product of marginal distributions of recovered signals as their estimator. We call the method as the minimum β-divergence method and the corresponding estimator as the minimum β-divergence estimator. They showed that its estimating functions are unbiased and gave the necessary and sufficient condition for asymptotic stability. This means that the minimum β-divergence estimator is locally consistent under this condition.

In contrast to the usual estimator based on the K-L divergence, the estimating functions for the minimum β-divergence estimator with $\beta > 0$ can be made bounded by the choice of density functions used to define the object function. Thus, the estimator is robust to outliers unlike the usual method. The larger the value of β is, the more robust the estimator is. However, too large β would produce a less efficient estimator. This paper discusses a method for choosing the tuning parameter β associated with a family of minimum β-divergence methods for blind source separation.

Section 2 reviews the minimum β-divergence method. We discuss the selection method for the tuning parameter in section 3. Section 4 shows numerical examples and section 5 gives concluding remarks.

2. MINIMUM β-DIVERGENCE METHOD

The minimum β-divergence method finds the estimate by minimizing the β-divergence between the empirical distribution $\tilde{r}$ of the observed signal and the product of marginal densities expressed by (1). Instead of unknown density q_i, we use a density function p_i in a specific form, for example, $p_i(z) = c_1 \exp(-c_2 z^4)$ and $p_i(z) = c_2/\cosh(z)$. Moreover, we explicitly include shift parameter $\boldsymbol{\mu} = (\mu_1, \cdots, \mu_m)^T$ into the model. Let

$$r_0(\boldsymbol{x}, W, \boldsymbol{\mu}) = |\det(W)| \prod_{i=1}^{m} p_i(\boldsymbol{w}_i\boldsymbol{x} - \mu_i). \quad (3)$$

We note here that we never assume that the density of recovered signals is p_i when we discuss statistical characteristics of the estimator. In this sense, it would be better to call (3) the pseudo model.

The minimum β-divergence method finds the minimizer of the β-divergence $\widehat{D}_\beta(\tilde{r}, r_0(\cdot, W, \boldsymbol{\mu}))$ between the empirical distribution $\tilde{r}$ of $\boldsymbol{x}$ and $r_0(\boldsymbol{x}, W,\ \boldsymbol{\mu})$ instead of unknown $r(\boldsymbol{x}, W)$. Minimizing $\widehat{D}_\beta(\tilde{r},\ r_0(\cdot, W,\ \boldsymbol{\mu}))$ is equivalent to maximizing the following quasi β-likelihood function:

$$L_\beta(W, \boldsymbol{\mu}) = \frac{1}{n}\sum_{t=1}^{n} l_\beta(\boldsymbol{x}(t); W, \boldsymbol{\mu})$$
$$l_\beta(\boldsymbol{x}; W, \boldsymbol{\mu}) = \frac{1}{\beta}\left(r_0^\beta(\boldsymbol{x}, W, \boldsymbol{\mu}) - 1\right) - (b_\beta(W) - 1)$$

for $\beta > 0$, where

$$b_\beta(W) = \frac{1}{\beta+1}\int r_0^{\beta+1}(\boldsymbol{x}, W, \boldsymbol{\mu}) d\boldsymbol{x}$$
$$= \frac{|\det(W)|^\beta}{\beta+1}\int \prod_{i=1}^{m} p_i^{\beta+1}(z_i) d\boldsymbol{z}.$$

The estimating functions (derivatives of $l_\beta(\boldsymbol{x}; W, \boldsymbol{\mu})$) are given by

$$F_1(\boldsymbol{x}, W, \boldsymbol{\mu}) = r_0^\beta(\boldsymbol{x}, W, \boldsymbol{\mu}) \times \left(I_m - \boldsymbol{h}(W\boldsymbol{x} - \boldsymbol{\mu})\ (W\boldsymbol{x})^T\right) W^{-T} - \beta\ b_\beta(W) W^{-T},$$
$$F_2(\boldsymbol{x}, W, \boldsymbol{\mu}) = r_0^\beta(\boldsymbol{x}, W, \boldsymbol{\mu})\ \boldsymbol{h}(W\boldsymbol{x} - \boldsymbol{\mu})$$

where

$$\boldsymbol{h}(\boldsymbol{y}) = (h_1(y_1), ..., h_m(y_m))^T \quad \text{and}$$
$$h_i(y_i) = -\frac{d \log p_i(y_i)}{dy_i} = -\frac{p_i'(y_i)}{p_i(y_i)}.$$

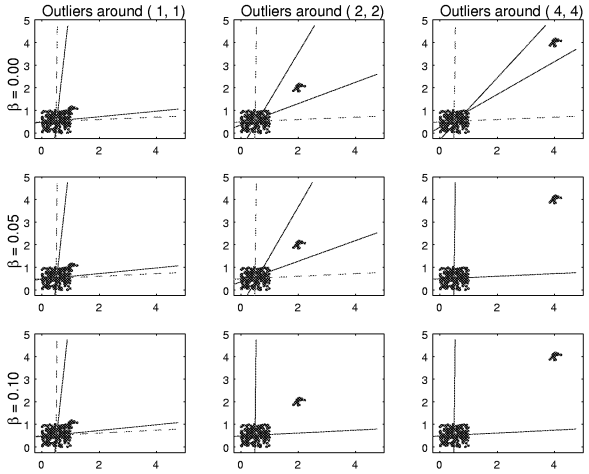
Since the β-divergence with $\beta = 0$ is the K-L divergence, the minimum β-divergence estimator with $\beta = 0$ is equivalent to the estimator derived from the K-L divergence with shift parameters explicitly included.

The minimum β-divergence estimator is locally consistent as the methods derived from the K-L divergence. If observed signals are mixtures of independent signals and W_0 transforms $\boldsymbol{x}$ into $\boldsymbol{y}$ such that components of $\boldsymbol{y}$ are independent of each other,

$$\mathrm{E}_r\left[F_1(\boldsymbol{x}, \Lambda^* W_0, \boldsymbol{\mu}^*)\right] = 0 \quad \text{and}$$
$$\mathrm{E}_r\left[F_2(\boldsymbol{x}, \Lambda^* W_0, \boldsymbol{\mu}^*)\right] = 0$$

with some scaling matrix $\Lambda^* = \mathrm{diag}(\lambda_1^*, .., \lambda_m^*)$ and shift vector $\boldsymbol{\mu}^*$. The necessary and sufficient condition for the negative of the expected Hessian matrix at $\Lambda^* W_0$ and $\boldsymbol{\mu}^*$ to be positive definite was given in [12].

Fig. 1. Effect of outliers



Solid lines express estimates by all data and dotted lines express estimates by data without outliers.

Unlike other properties, the robustness of the estimator does depend on the value of β. An estimator defined as the maximizer of some function is called M-estimator (maximum likelihood type estimator). The robustness of M-estimator is investigated based on its influence function [8]. The influence function measures the asymptotic bias caused by contamination at the point $\boldsymbol{x}$ and an estimator is said to be B-robust if its influence function is bounded. When $\beta > 0$, the minimum β-divergence estimator can be made B-robust by the use of appropriate p_i. On the other hand, when $\beta = 0$, the estimator can never be B-robust no matter what functions are used for p_i.

Estimators with large β tend to be more robust, but too large β might result in a large variance of the estimator. If the underlying distribution of the ith recovered signals were p_i, the minimum β-divergence estimator with $\beta = 0$ (MLE) would be asymptotically most efficient. For blind source separation, the underlying distribution are unknown and the minimum β-divergence estimator with $\beta = 0$ is no more the most efficient estimator. However, it still holds that too large β might result in a poorly efficient estimator. Thus, when two minimum β-divergence estimates with different values of β perform similarly, we might want to use the estimate with the smaller β.

Figure 1 depicts how the estimates with $\beta = 0$ (top), 0.05 (middle) and 0.10 (bottom) are affected by outliers. Data consist of 200 bivariate data whose components follow the uniform distribution on $(0, 1)$ independently and 5 outliers which follow a bivariate Gaussian distribution with covariance matrix $0.1^2\,\mathrm{I}_2$. Mean of outliers are $(1, 1)$ (left), $(2, 2)$ (middle) and $(4, 4)$ (right). Solid lines express estimates by all data and dotted lines express estimates by data without outliers. For this data, ideal estimates are expressed with X-axis ($y = 0$) and Y-axis ($x = 0$). When outliers are at $(1, 1)$, estimates are not much different. When outliers are at $(2, 2)$, estimates with $\beta = 0$ and 0.05 are more affected by outliers than when outliers are at $(1, 1)$ while the estimate with $\beta = 0.10$ is not affected by outliers. When outliers are at $(4, 4)$, only the estimate with $\beta = 0$ is affected by outliers and it is more so than when outliers are at $(2, 2)$.

For the data with outliers centered at $(1, 1)$, three estimates are not much different, thus, a good choice would be $\beta = 0$. For the data with outliers centered at $(2, 2)$, the estimate with $\beta = 0.1$ is good, and for the data with outliers centered at $(4, 4)$, $\beta = 0.05$ would be our choice with consideration of robustness and efficiency.

If we knew ideal estimates as in the above, we could tell which β is good , but it is not the case when we want to estimate a recovering matrix. In the next section, we discuss a selection procedure for β that produces an estimator which is robust enough for the given data and does not lose efficiency too much.

3. SELECTION PROCEDURE FOR β

In order to find an appropriate β for the minimum β-divergence estimator for the given data, we evaluate the estimates by various values of β. There are four aspects for evaluating estimates:

1. Measurement for evaluation
2. Generalization scheme
3. Scaling of estimates for recovering matrix
4. How to decide β

3.1. Measurement for evaluation

We would like to recover independent signals so that:

1. when there are no outliers, independent signals are recovered from all observed signals,
2. when there are possible outliers, possible outliers are ignored and independent signals are recovered only from the main population.

Thus, a measurement for evaluation should give a good evaluation when independent signals are recovered, but should not give too much penalty for the existence of outliers. The K-L divergence between the density of $\boldsymbol{x}$ and the pseud model (3), or equivalently, the pseud log-likelihood does not satisfy this. Our choice is the β-divergence with a fixed and relatively large value of β (say, 0.3 or 0.5) between the distribution of $\boldsymbol{x}$ and

the pseud model (3). The β-divergence with a large value of β satisfies the above condition. We denote the value of β for evaluation by β_0. We define our measure for evaluation of the minimum β-divergence estimator $\widehat{W}_\beta$ as:

$$D_{\beta_0}(\beta) = \mathrm{E}\left[D_{\beta_0}\left(r, r_0(\cdot, \hat{\Lambda}_{\beta,\beta_0}\widehat{W}_\beta, \hat{\boldsymbol{\mu}}_{\beta,\beta_0})\right)\right]$$

where scaling matrix $\hat{\Lambda}_{\beta,\beta_0}$ and shift vector $\hat{\boldsymbol{\mu}}_{\beta,\beta_0}$ will be explained later.

3.2. Generalization scheme

Measurement $D_{\beta_0}(\beta)$ is of the generalization performance of an estimator. The generalization performance relates to its prediction capability on independent test data. If we use the same dataset to evaluate $D_{\beta_0}(\beta)$ as to estimate a recovering matrix, it will underestimate $D_{\beta_0}(\beta)$. If we are in a data-rich situation, the best approach is to divide the dataset into a few parts, and use one set for estimation and another for evaluation. In the other situation, a simple and widely used method by sample re-use is K-fold Cross Validation (CV) method [9]. K-fold CV method uses part of the available data to find the estimate and a different part to test it. For the current problem, we employ K-fold CV method as a generalization scheme.

We split the data into K roughly equal-sized and similarly distributed parts. For the kth part, we find the estimate using the other $K-1$ parts of the data, and calculate β_0-divergence with the kth part of the data. Then we combine calculated β_0-divergence values to obtain the CV estimate.

3.3. Scaling of a recovering matrix

For the blind source separation problem, the scaling and the shifting of recovered signals as well as the scaling of a recovering matrix are arbitrary because scaling and shifting do not affect independence. However, β-divergences are different if scaling is different, that is, for any $\boldsymbol{\mu}_1$ and $\boldsymbol{\mu}_2$

$$D_{\beta_0}(r, r_0(\cdot, W, \boldsymbol{\mu}_1)) \neq D_{\beta_0}(r, r_0(\cdot, \Lambda W, \boldsymbol{\mu}_2))$$

in general, where $\Lambda = \mathrm{diag}(\lambda_1, \cdots, \lambda_m)$ unless $\Lambda = I_m$.

The scaling and shifting conditions for minimum β-divergence method differ by the value of β. In order to evaluate minimum β-divergence estimates properly, we need to rescale and shift estimates with a common condition. For this purpose, we use the scaling and shifting condition for the minimum β-divergence estimator with $\beta = \beta_0$, that is, we rescale the minimum β-divergence estimate $\widehat{W}_\beta$ with β by the diagonal matrix $\hat{\Lambda}_{\beta,\beta_0}$ and use the shift parameter $\hat{\boldsymbol{\mu}}_{\beta,\beta_0}$ for evaluation where $\hat{\Lambda}_{\beta,\beta_0}$ and $\hat{\boldsymbol{\mu}}_{\beta,\beta_0}$ minimize $\widehat{D}_{\beta_0}\left(r, r_0(\cdot, \Lambda\widehat{W}_\beta, \boldsymbol{\mu})\right)$ among diagonal matrix Λ and vector $\boldsymbol{\mu}$.

Table 1. K-fold Cross Validation procedure

Split the data into K parts; $S(1), \cdots, S(K)$.

Let $D^{-k} = \{\boldsymbol{x} | \boldsymbol{x} \notin S(k)\}$.

For $k = 1, \cdots, K$

- Estimate W by maximizing $L_\beta(W, \boldsymbol{\mu})$ using D^{-k},
  $(\widehat{W}_\beta,\ \widehat{\boldsymbol{\mu}}) = \underset{W,\ \boldsymbol{\mu}}{\mathrm{argmax}} \sum_{\boldsymbol{x}\in D^{-k}} l_\beta(\boldsymbol{x}; W, \boldsymbol{\mu})$.
- Estimate Λ_{β,β_0} and $\boldsymbol{\mu}_{\beta,\beta_0}$ by maximizing $L_{\beta_0}(\Lambda\widehat{W}_\beta, \boldsymbol{\mu})$ using D^{-k},
  $(\hat{\Lambda}_{\beta,\beta_0},\ \hat{\boldsymbol{\mu}}_{\beta,\beta_0}) = \underset{\Lambda,\ \boldsymbol{\mu}}{\mathrm{argmax}} \sum_{\boldsymbol{x}\in D^{-k}} l_{\beta_0}(\boldsymbol{x}; \Lambda\widehat{W}_\beta, \boldsymbol{\mu})$.
- Compute $\mathrm{CV}_{(k)}$ using $S(k)$,
  $\mathrm{CV}_{(k)} = -\sum_{\boldsymbol{x}\in S(k)} l_{\beta_0}(\boldsymbol{x},\ \hat{\Lambda}_{\beta,\beta_0}\widehat{W}_\beta,\ \hat{\boldsymbol{\mu}}_{\beta,\beta_0})$

End

Then, $\widehat{D}_{\beta_0}(\beta) = \dfrac{1}{n}\sum_{k=1}^{K} \mathrm{CV}_{(k)}$.

Table 1 summarizes the procedure to find the K-fold CV estimate $\widehat{D}_{\beta_0}(\beta)$.

3.4. How to decide β

As a measure for variation of $\widehat{D}_{\beta_0}(\beta)$, we compute

$$\mathrm{SD}_{\beta_0}(\beta) = \text{the standard error of } \frac{1}{|S(k)|}\mathrm{CV}_{(k)},$$

where $|S(k)|$ denotes the number of elements in the k th part of data $S(k)$.

Plotting $\widehat{D}_{\beta_0}(\beta)$ for β with curves $\widehat{D}_{\beta_0}(\beta) \pm \mathrm{SD}_{\beta_0}(\beta)$ will help us to judge an appropriate β. If the plot shows an elbow shape, the smallest value in the bottom would be a good choice.

Often a "one-standard error" rule is employed with cross-validation, in which we choose the smallest β whose evaluated value $\widehat{D}_{\beta_0}(\beta)$ is no more than one standard error above the smallest $\widehat{D}_{\beta_0}(\beta)$ [9]. However, as long as we know, there is no theoretical justification for this rule.

4. NUMERICAL EXAMPLES

We investigate the performance of the proposed selection procedure using the bivariate uniform data with outliers shown in section 2 and the mixtures of speech signals with/without spike noise shown in [12]

4.1. Bivariate uniform data with outliers

We first computed ten-fold CV estimates (K=10) for the bivariate uniform data with outliers shown in section 2. For β in the range from 0 to 0.3, $\widehat{D}_{\beta_0}(\beta)$ with $\beta_0 = 0.3, 0.4$ and 0.5 were computed with the algorithm given in table 1. Figure 2 depicts the results. In each plot, circles are $\widehat{D}_{\beta_0}(\beta)$ and the smallest value is expressed with an asterisk inside of the circle. Dotted lines are $\widehat{D}_{\beta_0}(\beta) \pm \mathrm{SD}_{\beta_0}(\beta)$.

With all values of $\beta_0 = 0.3, 0.4$ and 0.5, $\widehat{D}_{\beta_0}(\beta)$ show the same characteristics and we make the same decision as follows: For the data with outliers centered at $(1,1)$, $\widehat{D}_{\beta_0}(\beta)$ are not much different, thus, we use $\beta = 0$ for the minimum β-divergence estimator, that is, the method based on the K-L divergence. For the data with outliers centered at $(2,2)$, circles form an elbow shape and the smallest value in the flat bottom part is 0.1, thus, we choose 0.1 for β. For the data with outliers centered at $(4,4)$, circles also form an elbow shape and the smallest value in the flat bottom part is 0.06, thus, we choose 0.06 for β. Figure 1 shows that the above choice of β is appropriate.

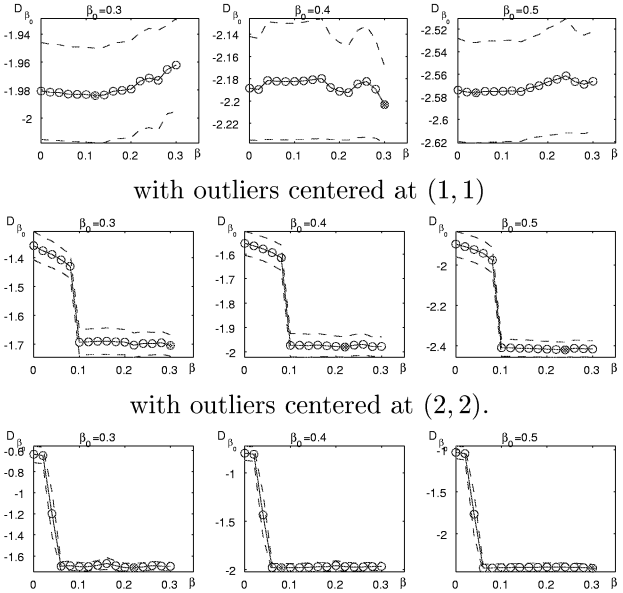
4.2. Mixtures of speech signals

The second data is the mixtures of two speech signals with sample size 1500. Noise is added to the mixed signals with occurrence rate 0.05. The original speech signals and mixed signals with noise were shown and the description of noise was given in [12].

As in the previous example, we computed ten-fold CV estimates $\widehat{D}_{\beta_0}(\beta)$ with $\beta_0 = 0.3, 0.4$ and 0.5 for β in the range from 0 to 0.3 using the algorithm given in table 1. Results are shown in figure 3. In each plot, circles are $\widehat{D}_{\beta_0}(\beta)$ and the smallest value is expressed with an asterisk inside of the circle. Dotted lines are $\widehat{D}_{\beta_0}(\beta) \pm \mathrm{SD}_{\beta_0}(\beta)$.

Again, for both of the noisy speech signals and the noiseless speech signals, $\widehat{D}_{\beta_0}(\beta)$ show the same characteristics with all values of $\beta_0 = 0.3, 0.4$ and 0.5. For the noisy speech signals, circles form an elbow shape, but the angle is not sharp unlike for the bivariate uniform data with outliers centered at $(2,2)$ and $(4,4)$. It looks like a value around $\beta = 0.1$ would be chosen. "one-standard error" rule suggests $\beta = 0.1$ for the measurement with $\beta_0 = 0.3$ and 0.4, and $\beta = 0.08$ for the measurement with $\beta_0 = 0.5$.

Fig. 2. Ten-fold CV for uniform data with outliers



with outliers centered at $(1,1)$

with outliers centered at $(2,2)$.

with outliers centered at $(4,4)$

In each plot, circles are $\widehat{D}_{\beta_0}(\beta)$ and the smallest value is expressed with an asterisk inside of the circle. Dotted lines are $\widehat{D}_{\beta_0}(\beta) \pm \mathrm{SD}_{\beta_0}(\beta)$.

Figure 4 depicts plots of recovered signals(Y-axis) by the minimum β-divergence estimates for recovering matrix with $\beta = 0, 0.04, 0.06, 0.08, 0.10$ and 0.12 versus original independent signals(X-axis). When signals are recovered by the minimum β-divergence estimate for recovering matrix with $\beta = 0.10$, most data points are on a straight line, thus, the recovering matrix estimate with $\beta = 0.10$ recovered signals well. Results are similar with $\beta = 0.08$ and 0.12. With $\beta = 0.06$, most data points are on a straight line, but the line is thicker.

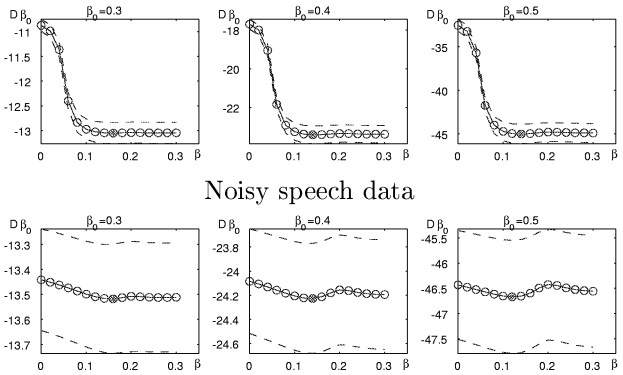
For the noiseless speech data, $\widehat{D}_{\beta_0}(\beta)$ are not much different, thus, we choose $\beta = 0$ for the minimum β-divergence estimator, that is, the method based on the K-L divergence.

We note that the scaling of the recovering matrix estimate by the minimum β-divergence depends on the value of β, so are recovered signals in figure 4.

5. DISCUSSIONS

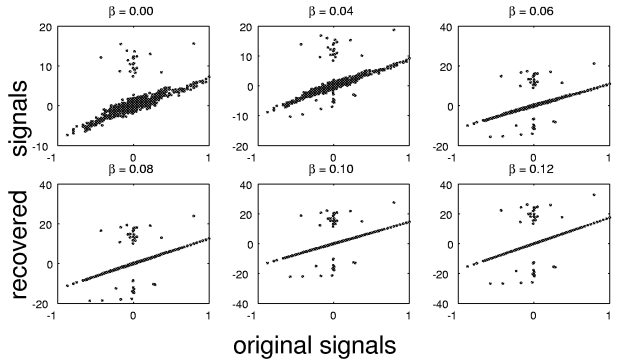
In this paper, we proposed a selection procedure of β for the minimum β-divergence estimator. As a measurement for evaluation of estimators, we use the β-divergence with a fixed and relatively large β_0 (0.3 or 0.5) and we employ K-fold Cross Validation as a generalization scheme. In order to evaluate estimates prop-

Fig. 3. Ten-fold CV for speech data



In each plot, circles are $\widehat{D}_{\beta_0}(\beta)$ and the smallest value is expressed with $*$ inside of the circle. Dotted lines are $\widehat{D}_{\beta_0}(\beta) \pm \mathrm{SD}_{\beta_0}(\beta)$.

Fig. 4. recovered signals vs. original signals



recovered signals by the recovering matrix estimates with $\beta = 0, 0.04, 0.06,\ 0.08, 0.10$ and 0.12 versus original independent signals.

erly, we rescale estimates with the scaling condition for the minimum β-divergence estimator with $\beta = \beta_0$ for evaluation.

As numerical examples, we applied the proposed procedure to the bivariate uniform data with outliers and the mixtures of speech signals with spike noise. In both examples, we could select appropriate β using the proposed procedures.

Two examples for which we showed the proposed selection procedure finds an appropriate β for the minimum β-divergence estimator are typical examples of the dataset with outliers, so called "contamination models" [8]. In [12], it was shown that the minimum β-divergence estimator with proper β finds better estimates than the estimator based on the K-L estimator for the datasets such as the mixtures of multiple ICA models. In our next study, we would like to apply the proposed selection procedure to many types of datasets and brush it up, if necessary, so that it can cope with various situations.

6. REFERENCES

[1] Amari, S. & Cardoso, J. F., "Blind source separation — Semi-parametric statistical approach," IEEE Trans. on Signal Processing, 45, pp. 2692-2700, 1997.

[2] Amari, S., Chen, T. & Cichocki, A., "Stability analysis of learning algorithm for blind source separation," Neural Networks, 10(8), pp. 1345-1351, 1997.

[3] Amari, S., Chichocki, A. & Yang, H.H. (1996). "A new learning algorithm for blind source separation," In Advances in Neural Information Processing 8, pp. 757-763, Cambridge, MA: MIT Press, 1996.

[4] Basu, A., Harris, I.R., Hjort, N.L. & Jones, M.C., (1998). "Robust and efficient estimation by minimizing a density power divergence," Biometrika, 85, pp. 549-559, 1998.

[5] Bell, A. J. & Sejnowski, T. J., "An information-maximization approach to blind separation and blind deconvolution," Neural Computation, 7, pp. 1129-1159, 1995.

[6] Cardoso, J.F. & Souloumiac, A., (1993). "Blind beamforming for non-Gaussian signals," Proc. IEEE, 140,pp. 362-370, 1993.

[7] Eguchi, S. & Kano, Y. (2001). "Robustifying maximum likelihood estimation," Research memorandum 802, Tokyo, Institute of Statistical Mathematics, 2001.

[8] Hampel, F.R., Ronchetti, E.M., Rousseeuw, P.J. & Stahel, W.A., Robust Statistics. New York: Wiley, 1986.

[9] Hastie, T., Tibshirani, R. and Friedman, J The Elements of Statistical Learning, New York: Springer, 2001.

[10] Hyvärinen, A, Karunen, J. & Oja, E., Independent Component Analysis, New York: Wiley, 2001.

[11] Jutten, C. & Herault, J. "Blind separation of sources, Part I: An adaptive algorithm based on neuromimetic architecture," Signal Processing, 24, pp. 1-20, 1991.

[12] Minami, M. & Eguchi, S., "Robust Blind Source Separation by β-Divergence," Neural Computation, 14, pp. 1859-1886, 2002.