mathematics

MDPI

Article

Estimating the Major Cluster by Mean-Shift with Updating Kernel

Ye Tian [1,*] and Yasunari Yokota [2]

[1] Graduate School of Engineering, Gifu University, 1-1 Yanagido, Gifu-shi 501-1193, Japan
[2] Department of EECE, Faculty of Engineering, Gifu University, 1-1 Yanagido, Gifu-shi 501-1193, Japan; ykt@edu.gifu-u.ac.jp
* Correspondence: v3814003@edu.gifu-u.ac.jp

Received: 9 July 2019; Accepted: 20 August 2019; Published: 22 August 2019

Abstract: The mean-shift method is a convenient mode-seeking method. Using a principle of the sample mean over an analysis window, or kernel, in a data space where samples are distributed with bias toward the densest direction of sample from the kernel center, the mean-shift method is an attempt to seek the densest point of samples, or the sample mode, iteratively. A smaller kernel leads to convergence to a local mode that appears because of statistical fluctuation. A larger kernel leads to estimation of a biased mode affected by other clusters, abnormal values, or outliers if they exist other than in the major cluster. Therefore, optimal selection of the kernel size, which is designated as the bandwidth in many reports of the literature, represents an important problem. As described herein, assuming that the major cluster follows a Gaussian probability density distribution, and, assuming that the outliers do not affect the sample mode of the major cluster, and, by adopting a Gaussian kernel, we propose a new mean-shift by which both the mean vector and covariance matrix of the major cluster are estimated in each iteration. Subsequently, the kernel size and shape are updated adaptively. Numerical experiments indicate that the mean vector, covariance matrix, and the number of samples of the major cluster can be estimated stably. Because the kernel shape can be adjusted not only to an isotropic shape but also to an anisotropic shape according to the sample distribution, the proposed method has higher estimation precision than the general mean-shift.

Keywords: kernel bandwidth and shape; mean-shift; major cluster; mode estimation; updating kernel

1. Introduction

When measuring a certain physical quantity, a few abnormal values, hereinafter designated as outliers, are included among the normal measured values, thereby exacerbating measurement noise. Frequently in science and engineering, some effort is necessary to estimate the true statistical parameters of the physical quantity from these measured values and the included outliers. Because the measurement noise generally follows a Gaussian distribution with mean zero, all samples from the major cluster are Gaussian-distributed around the true value. The problem described above is summarized to estimate the parameters of the major cluster, such as the mean, covariance matrix, and the number of samples included in the major cluster.

Because the mean equals the mode in a Gaussian distribution, if the outliers do not affect the sample mode of the major cluster, then the problem above can be replaced by a mode-seeking problem of the major cluster. Fukunaga and Hostetler [1] first proposed the mean-shift method, which was subsequently generalized by Cheng [2]. It is therefore known as a convenient iterative method for mode-seeking. The mean-shift was shown to be equivalent to the method that seeks a local maximum by the steepest gradient algorithm for the probability density distribution estimated using the kernel

method [3,4]. Therefore, the bandwidth, which is the size of the used kernel, deeply affects both the estimation accuracy and precision in the mean-shift as well as in kernel density estimation [5].

Usually in kernel density estimation, the bandwidth is determined such that the difference between the true distribution and the estimated distribution is minimized [6–8]. In mean-shift, because the normalized norm affects the convergence speed, a method for determining the bandwidth is proposed for the isotropic kernel [9] and anisotropic kernel [10] such that the norm of the mean-shift vector normalized by the bandwidth is maximized. A method for selecting the most stable bandwidth was also proposed [10,11]. Moreover, mean-shift with bandwidth that varies depending on the coordinate in data space was proposed [9,11]. Nevertheless, these methods entail high calculation costs because they require some provisional estimate of the probability density distribution, which is described as the pilot or initial estimate in some reports of the literature. Other theoretical studies of mean-shift, such as convergence, have been further proven. Li [12] proved its convergence by further imposing some commonly acceptable conditions. Ghassabeh [13] modified the mean-shift to guarantee its convergence. Although the mean-shift has been used widely in many applications [14–16], the use of bandwidth for mean-shift has been largely ignored in studies reported in the literature.

As described herein, we propose a new mean-shift method by which adopting the multi-dimensional Gaussian kernel, the kernel bandwidth and shape are updated to fit the major cluster size and shape in each iteration with no provisional estimation. We first derive a calculation equation for calculating the variance (or covariance matrix) of a major cluster from the sample variance in the kernel (or the sample covariance matrix in the multi-dimensional case) around the mode. Then, as the update progresses in the mean-shift method, the variance (or covariance matrix) of a major cluster is estimated using this calculation equation. In addition, the kernel bandwidth and shape are adjusted adaptively based on this estimated value. Therefore, we propose the mean-shift method with such an updating kernel. The proposed mean-shift requires no predetermination of the kernel bandwidth as necessitated by the general mean-shift method.

This paper is organized as follows. A general mean-shift method is introduced in Section 2. In Section 3, we propose the new mean-shift method in a one-dimensional case. Numerical experiments are presented to evaluate the proposed mean-shift compared to the general mean-shift method in Section 4. An explanation of applications and conclusion are presented respectively in Sections 5 and 6. In the appendices, we describe an extension of the general mean-shift method and the proposed mean-shift to a multi-dimensional case.

2. General Mean-Shift Method

2.1. General Mean-Shift Method

Assuming that the major cluster of N_N points follows a Gaussian distribution with mean μ_N and standard deviation σ_N, we are considering the problem of estimating the mean μ_N of the major cluster when a fewer outliers of N_O points exist in the sample of $N = N_N + N_O$ points. If the mode of the sample is not biased from the mean μ_N under the influence of outliers, then the mean μ_N can be estimated as the mode. The mean-shift is a simple and iterative method to estimate the mode of the major cluster. Letting the sample be $x_n,\ n = 1, \ldots, N$, then the general mean-shift method is realized using the following iterative process:

1. Letting the mean μ_x of sample $x_n,\ n = 1, \ldots, N$ be the initial value of the mean estimator $\hat{\mu}_N$ of major cluster, then
$$\hat{\mu}_N \leftarrow \mu_x. \tag{1}$$
2. Consider a Gaussian distribution $p(x; \mu_W, \sigma_W)$ with the mean μ_W and standard deviation σ_W as the kernel function in the value direction. Here, the mean μ_W of kernel function is found by the mean estimator of major cluster
$$\mu_W \leftarrow \hat{\mu}_N. \tag{2}$$

The standard deviation σ_W is assigned to be an appropriate size as discussed later in Section 2.2.

3. Weight $a_n,\ n = 1, \ldots, N$ for each sample $x_n,\ n = 1, \ldots, N$ weighted by such a Gaussian kernel is

$$a_n = \frac{1}{A} p(x_n; \mu_W, \sigma_W). \tag{3}$$

However, A in Equation (3) above is a normalization coefficient for which the sum of the weight a_n is equal to 1, as

$$A = \sum_{k=1}^{N} p(x_k; \mu_W, \sigma_W). \tag{4}$$

We use this weight a_n to calculate the sample mean μ_x with $x_n,\ n = 1, \ldots, N$ as

$$\mu_x = \sum_{n=1}^{N} a_n x_n. \tag{5}$$

4. The value of mean estimator $\hat{\mu}_N$ of the major cluster is updated by the following equation:

$$\hat{\mu}_N \leftarrow \mu_x. \tag{6}$$

5. If the variation of the value of mean estimator $\hat{\mu}_N$ is equal to or less than the predetermined fixed value, then the update process is terminated. Otherwise, return to 2 and repeat the iteration.

2.2. Shortcomings and Solution of the General Mean-Shift Method

The general mean-shift method estimates the modes of the underlying probability density function. From the definition of a probability density, if the random variable X of N data points $x_i,\ i = 1, 2, 3, \ldots, N$ in one-dimensional space R has density f, then

$$f(x) = \lim_{h \to 0} \frac{1}{2h} P(x - h < X < x + h). \tag{7}$$

For any given h (bin bandwidth or kernel bandwidth), we can estimate $P(x - h < X < x + h)$ by the proportion of the sample falling in the interval $(x - h, x + h)$. Thus, a natural estimator $\hat{f}$ of the density is given by choosing a small h and setting

$$\hat{f}(x) = \frac{1}{2h} \frac{N_x}{N}. \tag{8}$$

Here, N_x denotes the number of samples falling in the interval $(x - h, x + h)$. To express the estimator more transparently, define the weight function $\omega(x; h)$ by

$$\omega(x; h) = \begin{cases} \frac{1}{2h} & |x| < h, \\ 0 & \text{others.} \end{cases} \tag{9}$$

The estimator can be expressed as below [17]:

$$\hat{f}(x) = \frac{1}{N} \sum_{i=1}^{N} \omega(x - x_i; h). \tag{10}$$

Replace the weight function ω by a general kernel function $K(x; \sigma)$ with standard deviation σ, which satisfies the condition

$$\int_{-\infty}^{\infty} K(x) dx = 1, \tag{11}$$

and the kernel estimator for the probability density function $\hat{f}(x)$ at point x can be expressed as

$$\hat{f}(x) = \frac{1}{N}\sum_{i=1}^{N} K(x - x_i; \sigma). \tag{12}$$

The general mean-shift is an attempt to ascertain the local modes of density function $\hat{f}(x)$, which correspond to the zeros of the gradient $\bigtriangledown_x \hat{f}(x) = 0$. Therefore, the type of kernel function $K(x;\sigma)$ and the kernel bandwidth σ both directly affect the performance of general mean-shift method. Fixing the type of kernel function to Gaussian kernel, we specifically examine the influence of the pre-set of the kernel bandwidth in general mean-shift.

To confirm the influence of fixed kernel bandwidth on estimation accuracy in a general mean-shift method, we set various fixed kernel bandwidths in advance. Here, we summarize the numerical and experimentally obtained results for general mean-shift method as discussed in Section 4. Figure 1a presents the bias error between the estimated value in a general mean-shift method and the true value when we select various kernel bandwidths in advance. The horizontal axis shows a selection of different kernel bandwidths. The vertical axes respectively show the bias error between the estimated value for the mean and the true mean value, and the variance of the mean value. While selecting different fixed kernel bandwidths, we estimated the mean of the major cluster, which is distributed as shown in Figure 2 for 1000 trials. Furthermore, we computed the bias errors using the equation described in Section 4.2. Figure 1a shows that, when we enlarge the fixed kernel bandwidth, the mean estimator is more susceptible to outliers. The bias error in general mean-shift method increases. Otherwise, when we decrease the kernel bandwidth, the number of samples involved in the mean estimation decreases. The local mode can easily become the convergence point of the iterative process. In addition, the bias error in general mean-shift method increases. The kernel bandwidth should be set in the range of 0.5–1.5. As shown in Figure 1b, with enlargement of the kernel bandwidth, the estimation variance in general mean-shift method decreases. Therefore, the optimal kernel bandwidth is 1.5. Because the maximum value of these variances is very small and, because it does not exceed 0.06, if we select the kernel bandwidth within this range of 0.5–1.5, we can ensure the unbiasedness and consistency of the mean estimator in general mean-shift method. However, not knowing the true mean of the major cluster beforehand, we cannot calculate the bias error in general mean-shift method. Therefore, we cannot choose the appropriate kernel bandwidth based on the comparison result shown in Figure 1a. Indeed, the proper pre-set of the kernel bandwidth constitutes an important difficulty.

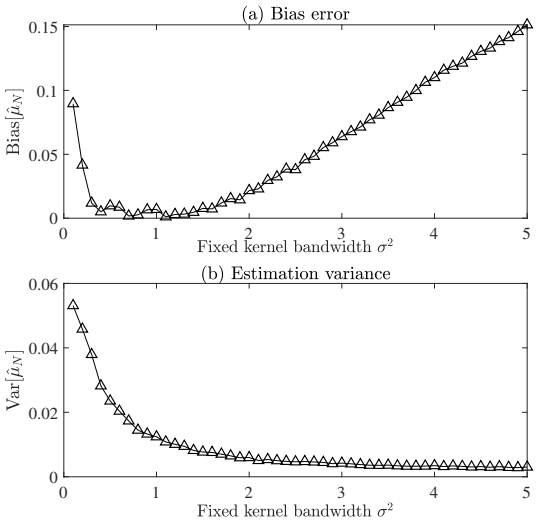

Figure 1. Bias error and estimation variance for various fixed kernel bandwidth σ^2 in a general mean-shift method.

The optimal kernel bandwidth depends on the existence range of outliers, the number of samples belonging to the major cluster and the distribution that the major cluster follows. In the absence of prior knowledge, the kernel bandwidth is often fixed as appropriate to $1/2$ the time of the standard deviation of the whole sample when the whole sample contains the major cluster and the outliers in signal processing [18]. For clustering in image processing or other multiple applications, it is still difficult to preset the kernel bandwidth properly in a general mean-shift method. When the kernel bandwidth is inappropriate, the kernel bandwidth becomes a factor that degrades the estimation accuracy.

As follows, based on the general mean-shift method, we propose a method to change the kernel bandwidth adaptively in accordance with simultaneous estimation of the mean (for a multi-dimensional case, the mean vector) and the standard deviation (for a multi-dimensional case, the covariance matrix) of a major cluster at each iteration. We need not set the kernel bandwidth properly in advance.

3. One-Dimensional Mean-Shift with Updating Kernel

3.1. Derivation of Major Cluster Standard Deviation σ_N from Sample Standard Deviation σ_x

Here, the Gaussian distribution with mean μ and standard deviation σ is represented by $p(x;\mu,\sigma)$. It is abbreviated as $p(x;\sigma)$ especially for $\mu = 0$. We use the two following equations for the two Gaussian distributions:

$$\int_{-\infty}^{\infty} p(x;\sigma_W)p(x;\sigma_N)dx = \frac{1}{\sqrt{2\pi}\sqrt{\sigma_W^2+\sigma_N^2}}, \tag{13}$$

$$\int_{-\infty}^{\infty} x^2 p(x;\sigma_W)p(x;\sigma_N)dx = \frac{\sigma_W^2\sigma_N^2}{\sqrt{2\pi}(\sigma_W^2+\sigma_N^2)^{\frac{3}{2}}}. \tag{14}$$

We assume that the influence of outliers is small such that the sample mode is not biased from the mean μ_N. If the general mean-shift method with the sufficiently small fixed kernel bandwidth decided by the standard deviation of the kernel starts the iteration from an appropriate initial value, then the influence of the outliers on estimation decreases gradually as the estimate converges. Therefore, it is sufficient to consider only the samples from the major cluster $x_n,\ n = 1,\ldots,N_N$ when the estimate converges to their true value. In addition, the mean μ_N of the major cluster and the mean μ_W of the Gaussian kernel coincide near the convergence point. Even if coordinate transformation is performed so that both are 0, generality is not lost. Therefore, we let $\mu_N = \mu_W = 0$ here for analysis. The variance σ_x^2 of the sample $x_n,\ n = 1,\ldots,N_N$ weighted by $a_n,\ n = 1,\ldots,N_N$ is

$$\sigma_x^2 = \sum_{n=1}^{N_N} a_n x_n^2. \tag{15}$$

Weight a_n is a Gaussian kernel given by Equations (3) and (4). In addition, N is replaced by N_N.

The expected value of the sample variance σ_x^2 is calculated after substituting Equation (3) into Equation (15) as

$$E[\sigma_x^2] = E\left[\frac{1}{A}\sum_{n=1}^{N_N} p(x_n;\sigma_W)x_n^2\right]. \tag{16}$$

The variance of $\dfrac{1}{A}$ is sufficiently smaller than the dispersion of other parts. Therefore, it can be approximated to the following equation based on the assumption that the major cluster follows a Gaussian distribution, as

$$E[\sigma_x^2] \simeq \frac{1}{E[A]}E\left[\sum_{n=1}^{N_N} p(x_n;\sigma_W)x_n^2\right]. \tag{17}$$

The approximation is discussed later in Appendix B. Here, we calculate the expected value of A by Equations (4) and (13) as

$$\begin{aligned} E[A] &= E\left[\sum_{k=1}^{N_N} p(x_k;\sigma_W)\right] \\ &= \sum_{k=1}^{N_N} E[p(x_k;\sigma_W)] \\ &= N_N \int_{-\infty}^{\infty} p(x;\sigma_W)p(x;\sigma_N)dx \\ &= \frac{N_N}{\sqrt{2\pi}\sqrt{\sigma_W^2+\sigma_N^2}}. \end{aligned} \tag{18}$$

The expected value of other part becomes

$$\begin{aligned} &E\left[\sum_{n=1}^{N_N} p(x_n;\sigma_W)x_n^2\right] \\ &= \sum_{n=1}^{N_N} E[p(x;\sigma_W)x^2] \\ &= N_N \int_{-\infty}^{\infty} x^2 p(x;\sigma_W)p(x;\sigma_N)dx \\ &= \frac{N_N\sigma_W^2\sigma_N^2}{\sqrt{2\pi}(\sigma_W^2+\sigma_N^2)^{3/2}} \end{aligned} \tag{19}$$

according to Equation (14). In other words, after being weighted by a Gaussian kernel with mean 0 and standard deviation σ_W, the expected value of variance σ_x^2 of the sample which follows a Gaussian distribution with mean 0 and standard deviation σ_N is

$$E[\sigma_x^2] = \frac{\sigma_W^2\sigma_N^2}{\sigma_W^2+\sigma_N^2} \tag{20}$$

according to Equations (18) and (19). Equation (20) above can be transformed to

$$\sigma_N^2 = \frac{\sigma_W^2 E[\sigma_x^2]}{\sigma_W^2 - E[\sigma_x^2]}. \tag{21}$$

This expression shows that standard deviation σ_N can be estimated from the standard deviation σ_x of the sample, which is weighted using a Gaussian kernel with mean 0 and standard deviation σ_W as

$$\hat{\sigma}_N = \sqrt{\frac{\sigma_W^2\sigma_x^2}{\sigma_W^2-\sigma_x^2}}. \tag{22}$$

In addition, using Equation (18), the number N_N of samples belonging to the major cluster can be estimated as

$$\hat{N}_N = A\sqrt{2\pi}\sqrt{\sigma_W^2+\hat{\sigma}_N^2}. \tag{23}$$

Adaptive change of the standard deviation σ_W of the kernel related to the estimated value $\hat{\sigma}_N$ of the standard deviation is sufficient for each update because the mean μ_N of the major cluster and the standard deviation σ_N can also be estimated. Specifically, the standard deviation σ_W of the kernel is assigned to be r times the estimated value $\hat{\sigma}_N$, although it depends on the existence range of outliers. We designate this r as a scale factor. Regarding appropriate r, we will examine this point in a numerical experiment discussed later.

3.2. Mean-Shift with Updating Kernel

Based on the discussion presented in Section 3.1, at each iteration of the general mean-shift method, the standard deviation σ_N is estimated simultaneously in addition to the mean value μ_N. Therefore, we propose a new mean-shift method that adaptively changes the standard deviation σ_W of the kernel. The algorithm is summarized as presented below:

1. Let the mean μ_x of sample $x_n,\ n = 1,\ldots,N$ be the initial value of the mean estimator $\hat{\mu}_N$ of the major cluster and let standard deviation σ_x of this sample be the initial value of the standard deviation estimator σ_N of the major cluster as

$$\hat{\mu}_N \leftarrow \mu_x, \tag{24}$$
$$\hat{\sigma}_N \leftarrow \sigma_x. \tag{25}$$

2. Consider a Gaussian distribution $p(x;\mu_W,\sigma_W)$ with mean μ_W and standard deviation σ_W as the kernel function in the value direction. Here, the mean μ_W and the standard deviation σ_W are given respectively by the estimated value $\hat{\mu}_N$ of the mean and the estimated value $\hat{\sigma}_N$ of the standard deviation of the major cluster:

$$\mu_W \leftarrow \hat{\mu}_N, \tag{26}$$
$$\sigma_W \leftarrow r\hat{\sigma}_N. \tag{27}$$

   Here, mean μ_W and variance σ_W of the Gaussian kernel are not estimators, although they change when the kernel updates.
3. Weight $a_n,\ n = 1,\ldots,N$ for each sample $x_n,\ n = 1,\ldots,N$ weighted by such a Gaussian kernel $p(x;\mu_W,\sigma_W)$ is calculated using Equations (3) and (4). We use this weight a_n to calculate the sample mean μ_x and standard deviation σ_x with $x_n,\ n = 1,\ldots,N$ as shown below:

$$\mu_x = \sum_{n=1}^{N} a_n x_n, \tag{28}$$
$$\sigma_x = \sqrt{\sum_{n=1}^{N} a_n (x_n - \mu_x)^2}. \tag{29}$$

4. The values of mean estimator $\hat{\mu}_N$, standard deviation estimator $\hat{\sigma}_N$, and number of samples estimator $\hat{N}_N$ of the sample are updated, respectively, by the following equations:

$$\hat{\mu}_N \leftarrow \mu_x, \tag{30}$$
$$\hat{\sigma}_N \leftarrow \sqrt{\frac{\sigma_W^2 \sigma_x^2}{\sigma_W^2 - \sigma_x^2}}, \tag{31}$$
$$\hat{N}_N \leftarrow A\sqrt{2\pi}\sqrt{\sigma_W^2 + \hat{\sigma}_N^2}. \tag{32}$$

5. If the variations of the values of these estimators are equal to or less than the predetermined fixed value, then the update process is terminated. Otherwise, return to 2 and repeat the iteration.

4. Numerical Experiment

4.1. Update Process of Mean-Shift with an Updatable Kernel

For the proposed method, we use iteration to confirm the process by which the estimated values of the mean vector, the covariance matrix, and the number of samples converge to true values of the major cluster. Although no restriction is made of the dimension of data to which the proposed method is applicable, to illustrate and explain the distribution of data and update process, two-dimensional

data are targeted for analysis. Herein, we obtain the major cluster with $N_N = 3000$ points generated in two-dimensional normal distribution with the mean vector $\boldsymbol{\mu}_N = (0,0)^T$ and variance covariance matrix as

$$\boldsymbol{C}_N = \begin{pmatrix} 3 & 2 \\ 2 & 3 \end{pmatrix}.$$

The outliers with $N_O = 200$ points are distributed uniformly within the range of $x_1 \in [-2,-1]$, $x_2 \in [3,4]$. Figure 2 shows an example of the generated sample in (x_1, x_2) space. Symbol • in the figure represents the coordinates of each point. The points spreading in the central elliptical shape belong to the major cluster. Other points distributed in a square shape on the upper left are outliers. In the figure, the solid ellipse represents a contour line where 99% of the M-dimensional normal distribution defined by the mean vector $\boldsymbol{\mu}_N$ and the covariance matrix $\boldsymbol{C}_N$ fall within it. Later, we present the mean vector $\boldsymbol{\mu}_N$ and covariance matrix $\boldsymbol{C}_N$, or their estimates.

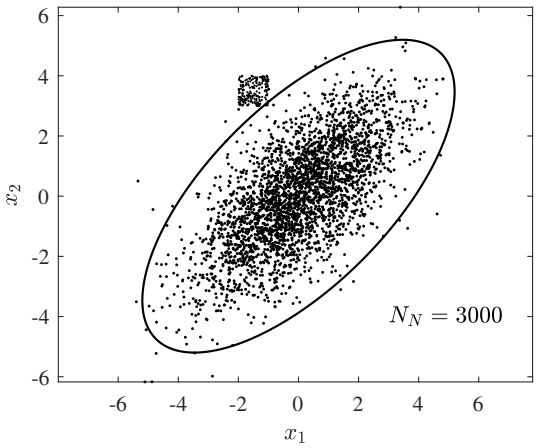

Figure 2. Example of a sample set for numerical experiments.

In general, as discussed in Appendix C.2, the initial estimated value of the mean $\hat{\boldsymbol{\mu}}_N$ and covariance matrix $\hat{\boldsymbol{C}}_N$ for major cluster can be assigned respectively to the mean and covariance matrix of all samples. However, we set the initial kernel having mean vector $\hat{\boldsymbol{\mu}}_N = (-2,3)^T$ and covariance matrix

$$\hat{\boldsymbol{C}}_N = \begin{pmatrix} 1.25 & -0.75 \\ -0.75 & 1.25 \end{pmatrix}$$

intentionally to be located and shaped sufficiently apart from the major cluster. To demonstrate how the estimated value converges to the true value with updating, the scale factor is $r = 1.0$. The update ends when it satisfies all conditions for which the sum of squares of the change amount $\hat{\boldsymbol{\mu}}_N$ is 0.01 or fewer, the sum of squares of the change amount of $\hat{\boldsymbol{C}}_N$ is 0.01 or fewer, and the square of the change amount of $\hat{N}_N$ is 30 or less.

As described earlier, the solid ellipse shown in Figure 3 represents the estimated value of mean vector $\hat{\boldsymbol{\mu}}_N$, covariance matrix $\hat{\boldsymbol{C}}_N$, and number $\hat{N}_N$ of samples for each update in the proposed method. In Figure 3, the estimated values $\hat{\boldsymbol{\mu}}_N, \hat{\boldsymbol{C}}_N, \hat{N}_N$ are shown to converge to the true values $\boldsymbol{\mu}_N, \boldsymbol{C}_N, N_N$ corresponding to Figure 2 as the update progresses, although they start from more or less bad initial values. Here, for the estimated value $\hat{N}_N$, we have accuracy to one decimal place.

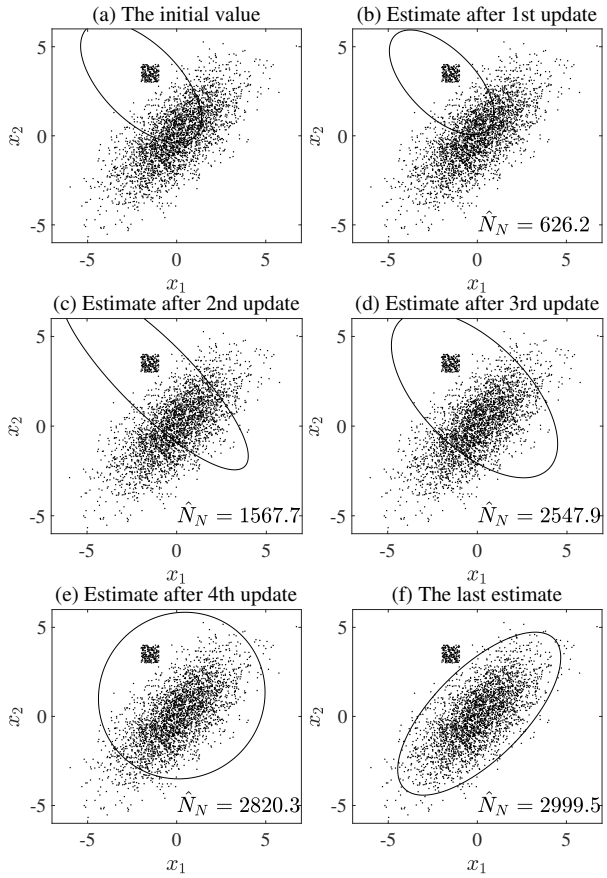

Figure 3. Updates of the estimated major cluster.

4.2. Influence of Kernel Bandwidth on Estimation Accuracy (Unbiasedness)

An exceedingly important property required for estimators is unbiasedness: a property by which the expected value of the estimated value coincides with the true value. If no statistical bias in the estimated value exists, then it represents that the estimation is accurate. Assuming that the parameter is θ, we investigate the unbiasedness of the estimator $\hat{\theta}$. If parameter θ is a scalar, then the bias error is the difference $E[\hat{\theta}] - \theta$ between the expected value and the true value θ of the estimator. Otherwise, if parameter θ is a vector or matrix, then the bias error is the square root $\sqrt{\|E[\hat{\boldsymbol{\theta}}] - \boldsymbol{\theta}\|^2}$ of the sum of squares over all the elements. It can be evaluated whether the bias error is zero. As explained below, it demonstrates that the initial value of the kernel bandwidth has less influence on the unbiasedness of the estimated value in the proposed method discussed in Appendix C than in the general mean-shift method introduced in Appendix A.

The distributions that major cluster and outliers follow, the numbers of samples N_N, N_O, scale factor r, and update ending condition are the same as those described in Section 4.1. The initial estimated value of mean vector $\hat{\boldsymbol{\mu}}_N$ is the mean vector of all samples. The initial estimated value of covariance matrix is assigned to $\hat{\boldsymbol{C}}_N = \sigma^2 \boldsymbol{I}$. In the general mean-shift method, the covariance matrix of the kernel is $\boldsymbol{C}_W = \sigma^2 \boldsymbol{I}$. Under the conditions presented above, the mean vector $\boldsymbol{\mu}_N$, the covariance matrix $\boldsymbol{C}_N$, and the number N_N of samples are estimated using the general mean-shift method and the proposed method. In addition, because it is impossible to obtain the expected value in numerical experiments, the expected value is replaced by the average value of the estimated values for 1000 trials that change the random number.

In the proposed method, σ^2 is the initial value of the kernel bandwidth. It corresponds to the pre-set value of the kernel bandwidth in a general mean-shift method. When this σ^2 is changed to various values, the bias errors of the estimated value of the mean vector $\boldsymbol{\mu}_N$, covariance matrix $\boldsymbol{C}_N$, and number N_N of samples are calculated. Results are presented respectively in Figure 4a–c.

The horizontal axis shows the selection of different kernel bandwidth. The vertical axes respectively show the bias errors for estimators $\boldsymbol{\mu}_N$, $\boldsymbol{C}_N$, and N_N. In this figure, symbol ◯ corresponds to the proposed method. The symbol △ represents the bias errors in a general mean-shift method. However, because the covariance matrix and number of samples cannot be estimated in a general mean-shift method, only the results obtained using the proposed method are shown in Figure 4b,c. The scale on the vertical axis of the figures is fixed to represent 10% of errors at full scale. In the following figures, the same scale applied to these figures will be used unless specified otherwise.

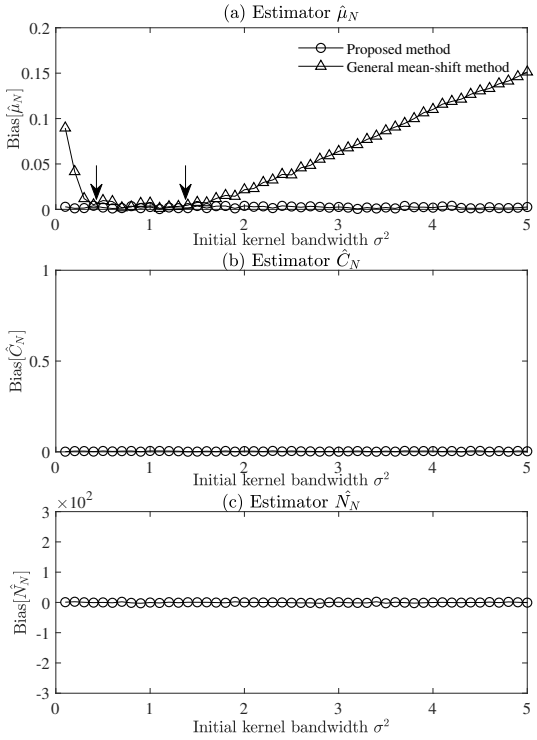

Figure 4. Bias errors for various initial kernel bandwidths σ^2 in the proposed method and the general mean-shift method.

Figure 4a shows that the bias error increases linearly and that the unbiasedness is lost when the kernel bandwidth σ^2 approximately exceeds the range of 0.5–1.5 represented by symbol ↓ in a general mean-shift method because, as the kernel becomes larger, the outliers fall within the range of the kernel, which greatly affects the mean estimation of the major cluster. For this reason, the proper set of the kernel bandwidth is an important difficulty in a general mean-shift method. However, the kernel bandwidth is adjusted according to the estimated value of covariance matrix of a major cluster at each iteration in the proposed method. Therefore, it is less susceptible to the influence of initial value σ^2. Furthermore, in Figure 4b,c, it is the same situation in the estimations of covariance matrix $\boldsymbol{C}_N$ and number N_N of samples.

While maintaining the ratio of the number N_N of samples of major cluster and the number N_O of samples of the outliers to 3000:200 and changing the number $N = N_N + N_O$ of samples from 1000 to 90,000, the variance of each estimate value of the mean vector $\boldsymbol{\mu}_N$, covariance matrix $\boldsymbol{C}_N$, and number N_N of samples are obtained using our proposed method, as shown in Figure 5. The horizontal axis shows the selection of different numbers of samples corresponding to the whole samples. The vertical axes respectively represent the bias errors for estimators $\boldsymbol{\mu}_N$, $\boldsymbol{C}_N$, and N_N. Because the proposed method is independent of the initial value σ^2, the initial value σ^2 is fixed to 1.5. Figure 5 shows that these estimators are unbiased for a finite number of samples.

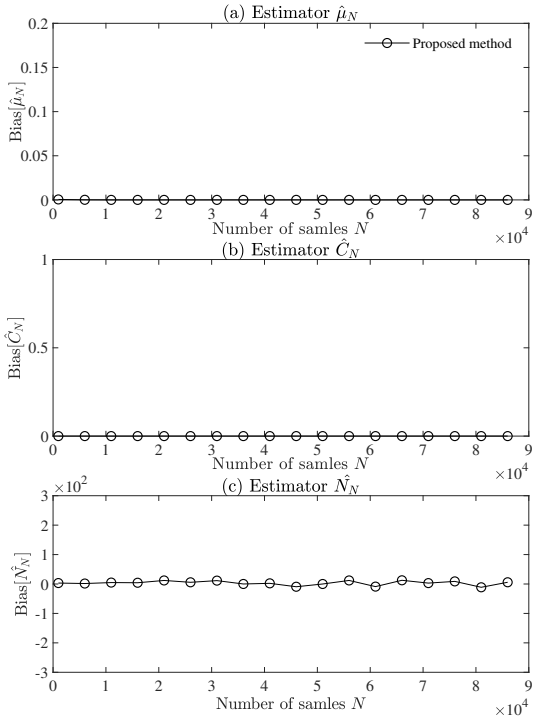

Figure 5. Bias errors for various numbers N of samples in the proposed method.

4.3. Influence of the Scale Factor r Value on Estimation Accuracy

In the proposed method, we need not select the initial value of kernel bandwidth in advance because the kernel bandwidth is changed adaptively. The pre-set of the initial value shows some difficulty in influencing the estimation accuracy. Instead, the problem of optimal setting of the scale factor r occurred. Scale factor r represents the ratio of the kernel bandwidth (standard deviation) to the major cluster width (standard deviation). Therefore, the smaller the scale factor, the smaller the kernel bandwidth (standard deviation) is set with respect to the major cluster width (standard deviation). From the viewpoint of estimation accuracy, the kernel bandwidth (standard deviation) should be sufficiently large but not cover the outliers. In other words, if the outliers exist at the distance from the mode of major clusters more than three times the standard deviation of the major cluster, according to three-sigma rule of thumb, the kernel bandwidth should be the same as the standard deviation of major cluster, which means $r = 1$. Otherwise, if there are a certain number of outliers within a standard deviation away from the mode of the major cluster, the kernel bandwidth is expected to be 1/3 of the standard deviation of the major cluster, which means $r = 1/3$. If the distribution of the major cluster and the outliers is specified completely, then it is possible to derive the theoretical formula of the optimal scale factor r as a parameter. However, because the purpose is to estimate the distribution of the major cluster and the outliers, then, even if a theoretical formula for scale factor r is derived, it cannot be used for estimation. Derivation of the theoretical formula for scale factor r has no great value. Therefore, as described below, we investigate the influence of the selected value of this scale factor on the estimation accuracy.

The distributions that major cluster and outliers follow, number N_N, N_O of samples, and update ending condition are the same as those in Section 4.1. As shown in Section 4.2, the initial values of the estimated value of mean vector and covariance matrix $\hat{\boldsymbol{\mu}}_N, \hat{\boldsymbol{C}}_N$ are given, respectively, by the mean vector and covariance matrix of the whole samples. We select scale factor r to be various values and estimate the mean vector $\boldsymbol{\mu}_N$, covariance matrix $\boldsymbol{C}_N$, and number N_N of samples using the proposed method. The bias errors of each estimated value is presented in Figure 6a,c. The horizontal axis represents the selection of various scale factors r. The vertical axes respectively represent the bias errors for estimators $\boldsymbol{\mu}_N$, $\boldsymbol{C}_N$, and N_N.

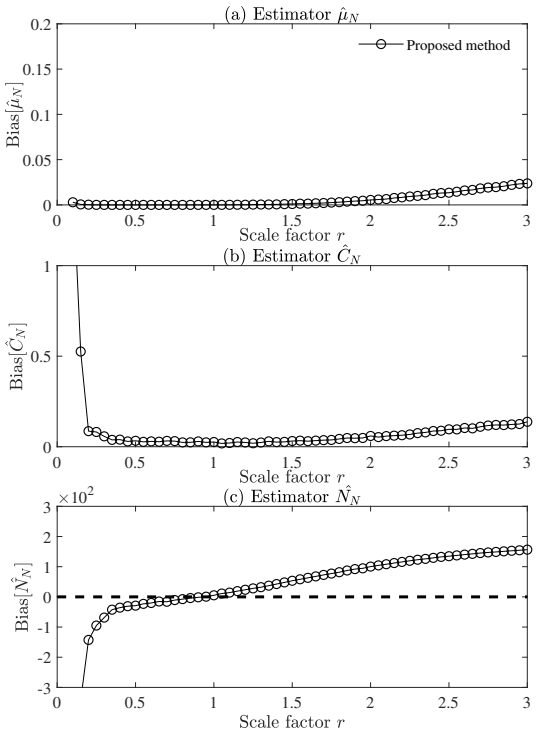

Figure 6. Bias errors for various scale factors r in the proposed method.

Figure 6 shows that the bias errors of any estimate increases and that the unbiasedness is lost when scale factor r is selected as a value larger than a certain value because, when the kernel bandwidth increases, it becomes more susceptible to outliers, as with the general mean-shift method shown in Figure 6. However, when scale factor r is selected as a small value, the bias error is increased extremely. The unbiasedness is lost relative to the covariance matrix $\boldsymbol{C}_N$ and number N_N of samples, although it is not readily apparent on mean vector $\boldsymbol{\mu}_N$. The reason for this is explainable as presented below.

If we select scale factor r as a value smaller than one, the kernel bandwidth becomes small because of a lack of the practical number of samples that contribute to the estimation. For that reason, the estimation precisions of mean vector $\boldsymbol{\mu}_x$ and standard deviation $\sigma_{x,m}$ are deteriorated. The deterioration of this estimation accuracy results from the small number of samples. Consequently, the estimated error has normality, but does not include bias error. As shown in Equation (A32), the estimated value $\hat{\boldsymbol{\mu}}_N$ of the mean vector is the sample mean vector $\boldsymbol{\mu}_x$. The estimation equation of the standard deviation $\hat{\sigma}_N$ and number $\hat{N}_N$ of samples is a nonlinear function of the sample standard deviation $\sigma_{x,m}$, as shown in Equations (A19) and (A20). In general, normality is lost by a nonlinear transformation. Therefore, the estimation errors of both the standard deviation $\hat{\sigma}_N$ and the number $\hat{N}_N$ of samples are converted to the bias errors by the nonlinear transformations, even if the estimation error of the sample standard deviation $\sigma_{x,m}$ had normality.

Figure 6 shows that the appropriate value of the scale factor r is in the range of $0.5 \leq r \leq 1.5$, but it depends on the characteristic of the target data. For example, the lower limit increases when the number of samples is small. The upper limit decreases when the outliers approach a major cluster. Comparing the bias error with the general mean-shift indicates that the selection of scale factor r need not be the same as the situation of kernel bandwidth as shown in Figure 4 because the range in which the bias error can be kept low is wide.

4.4. Verification of Consistency

The goodness of the estimator is evaluated by accuracy and precision. Accuracy is evaluated as the bias error, as discussed in Section 4.2, whereas the precision is evaluated by the variance of estimated values. Before comparing the estimation precision of a general mean-shift method with the proposed

method, one must confirm the consistency of the estimated values in both methods. Consistency is an important property required for the estimator. It indicates the characteristics by which the variance of the estimated values approaches 0 as the number of samples used for estimation increases.

The distributions that major cluster and outliers follow, in addition to the update ending conditions, are the same as those described in Section 4.1. As shown in Appendix C.2, the initial values $\hat{\boldsymbol{\mu}}_N, \hat{\boldsymbol{C}}_N$ of the estimate values of the mean vector and covariance matrix are given respectively by mean vector $\boldsymbol{\mu}_x$ and covariance matrix $\boldsymbol{C}_x$ of the whole samples. To ensure that the estimator is unbiased, we select the scale factor as $r = 1.0$ based on the discussion of the proposed method in Section 4.3, and the kernel as $\boldsymbol{C}_W = \sigma^2 \boldsymbol{I}, \sigma^2 = 1.5$ based on the discussion for a general mean-shift method in Section 4.2.

While maintaining the ratio of the number N_N of samples of major cluster and the number N_O of samples of the outliers to 3000 : 200 and changing the number $N = N_N + N_O$ of samples from 1000 to 90,000, the variance of each estimate value of the mean vector $\boldsymbol{\mu}_N$, covariance matrix $\boldsymbol{C}_N$, and number N_N of samples is obtained using both methods. The estimation variance is replaced by the sample variance of each estimate for 1000 trials as the sample number changes. The estimation variances $\mathrm{Var}[\hat{\boldsymbol{\mu}}_N]$, $\mathrm{Var}[\hat{\boldsymbol{C}}_N]$, $\mathrm{Var}[\hat{N}_N]$ are shown in Figure 7a–c. The horizontal axis shows the logarithm of various numbers of samples $\hat{N}_N$. The vertical axes respectively show logarithms for estimation variances $\mathrm{Var}[\hat{\boldsymbol{\mu}}_N]$, $\mathrm{Var}[\hat{\boldsymbol{C}}_N]$, $\mathrm{Var}[\hat{N}_N]$. In this figure, symbol ◯ corresponds to the proposed method. Symbol △ represents the general mean-shift method. Because the covariance matrix and number of samples can not be estimated in the general mean-shift method, only the results obtained using the proposed method are presented in Figure 7b,c.

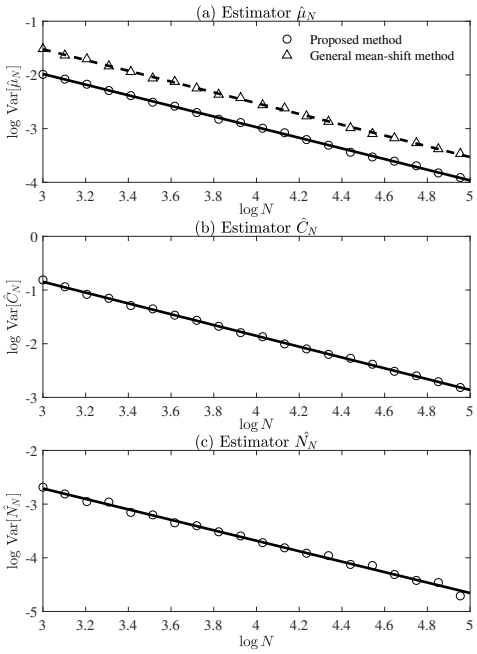

Figure 7. Variance of the estimates $\hat{\boldsymbol{\mu}}_N, \hat{\boldsymbol{C}}_N, \hat{N}$ for various numbers N of samples in the proposed method and the general mean-shift method.

From Figure 7a–c, it is readily apparent that the variance $\mathrm{Var}[\cdot] \to 0$ for sample number $N \to \infty$. Therefore, estimators $\hat{\boldsymbol{\mu}}_N, \hat{\boldsymbol{C}}_N, \hat{N}_N$ have consistency. Figure 7a–c are drawn as a logarithmic graph; the slope should be −1 in fact. Therefore, the relation between the sample number of samples and the estimation variance is approximated using a linear polynomial with the slope fixed at −1. The straight line represented by the approximate linear polynomial is superimposed by a solid line in these figures. These results demonstrate the validity of the approximation. Here, we simply define the estimation variance as the 0-order coefficient of the approximate linear polynomial or the virtual estimation

variance corresponding to sample number $N = 1$. Regarding to the estimation variance, we compare the estimation precision of proposed method with the general mean-shift method.

4.5. Estimation Precisions of the Proposed and General Mean-Shift Methods

The distributions that major cluster and outliers follow, the number N_N, N_O of samples, and the update ending condition are the same as those described in Section 4.1. As shown in Appendix C.2, the initial values of the estimated value of mean vector and covariance matrix $\hat{\boldsymbol{\mu}}_N, \hat{\boldsymbol{C}}_N$ are given, respectively, by the mean vector $\boldsymbol{\mu}_x$ and covariance matrix $\boldsymbol{C}_x$ of the whole samples.

We select scale factor r to be various values and use the proposed method to estimate the mean vector $\boldsymbol{\mu}_N$, covariance matrix $\boldsymbol{C}_N$, and number N_N of samples. Figure 8a–c respectively present the estimation variances corresponding to the estimated values of the mean vector $\boldsymbol{\mu}_N$, covariance matrix $\boldsymbol{C}_N$, and number N_N of samples. The horizontal axis shows the logarithm of various scale factor r. The vertical axes respectively show the estimation variances $\mathrm{Var}[\hat{\boldsymbol{\mu}}_N]$, $\mathrm{Var}[\hat{\boldsymbol{C}}_N]$, $\mathrm{Var}[\hat{N}_N]$. Similarly, letting the covariance matrix of kernel be $\boldsymbol{C}_W = \sigma^2 \boldsymbol{I}$, we estimate the mean vector $\boldsymbol{\mu}_N$ using the general mean-shift method while the kernel bandwidth σ^2 is changed to various values. The estimation variance of estimated value $\hat{\boldsymbol{\mu}}_N$ is presented in Figure 9.

From Figure 8a–c, the estimation variance of each estimated value of the mean vector $\boldsymbol{\mu}_N$, covariance matrix $\boldsymbol{C}_N$, and number N_N of samples decreases with respect to r, monotonically. If r is small, then the kernel bandwidth decreases. The number of substantial points involved in the estimation decreases. Therefore, the estimation precision deteriorates. On one hand, if r is large, then the estimation precision decreases. Because bias error occurs as shown in Figure 6, it is not desirable as an estimator. However, the estimation variance related to general mean-shift method decreases monotonically with respect to kernel bandwidth σ^2, as shown in Figure 9. The reason is exactly the same as in the case of the proposed method.

Finally, the estimation precision of a general mean-shift method and that of the proposed method are compared. Regarding the general mean-shift method, the estimation is unbiased if $\sigma^2 \leq 1.5$, as shown in Figure 4. However, the estimation precision increases as σ^2 becomes larger, as shown in Figure 9. In the general mean-shift method, the optimal selected value of the kernel bandwidth is $\sigma^2 = 1.5$. The estimation variance at kernel bandwidth $\sigma^2 = 1.5$ is read from Figure 9: its value is shown by a horizontal dotted line in Figure 8a. In the proposed method, $0.5 \leq r \leq 1.5$ is the suitable range of the scale factor r. In this range, the estimation variance of the proposed method is half or less than half of that of the general mean-shift method. The proposed method has higher estimation precision than the general mean-shift method that has the optimal kernel bandwidth for the following reason. In the general mean-shift method, the kernel shape is expressed as an isotopic shape because the covariance matrix of the kernel is represented as a diagonal matrix in which all diagonal elements are equal. Otherwise, in the proposed method, the kernel shape can take an arbitrary anisotropic shape because the covariance matrix of the kernel can take an arbitrary matrix that satisfies the condition as a covariance matrix. The practical number of samples that contribute to the estimation can be maximized by adjusting the kernel shape to the distribution of samples.

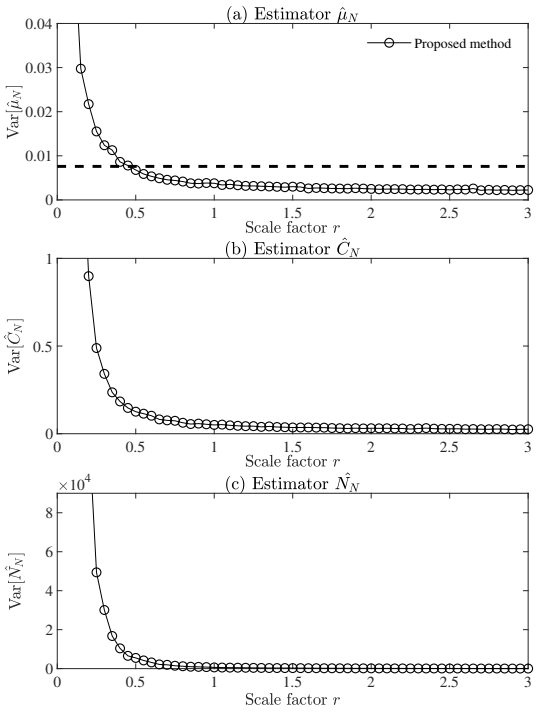

Figure 8. Estimating the variance of the estimates $\hat{\boldsymbol{\mu}}_N, \hat{\boldsymbol{C}}_N, \hat{N}$ for various scale factors r of the proposed method.

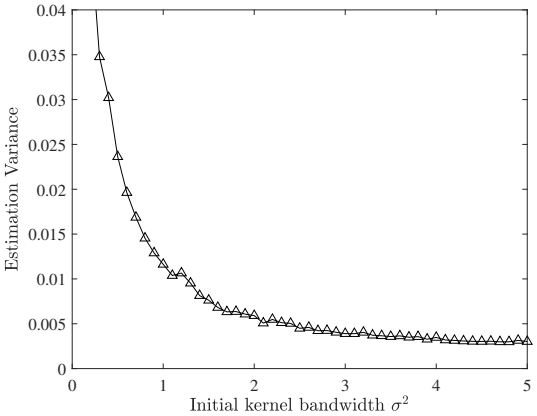

Figure 9. Estimating the variance of the estimate $\hat{\boldsymbol{\mu}}_N$ for various kernel bandwidths σ^2 in the general mean-shift method.

4.6. Discussion

Numerical experiments in two-dimensions described in Sections 4.2, 4.3 and 4.4 yield results for the major cluster and outlier model shown in Figure 2. The purpose of our numerical experiment is to confirm whether the estimators (mean, covariance, number of sample of the major cluster) in the proposed method are unbiased and consistent without a proper pre-set of kernel bandwidth. If these estimators are consistent unbiased estimators, then the proposed method can achieve accurate estimation of the mean, covariance, and the number of samples of the major cluster. We chose the two-dimensional numerical experiment to observe the dynamic changes of the kernel more intuitively during the iteration. The iteration process is shown in Figure 3. In the numerical experiments described herein, the major cluster follows the Gaussian distribution. If the proposed method performs well on other distributions, then the scope of application of the proposed method can be expected to expand. We discuss the scope of application of the proposed method in two aspects as presented below.

For a one-dimensional signal processing field, the assumption of normality is not regarded as being such a severe strong assumption. Yokota and Ye [19] proposed the radical root, or r-th root, transform

of the power spectrum series that follows the chi-square distribution, such that the transformed series follows a quasi-Gaussian distribution. Lotter and Vary [20] proposed a spectral amplitude estimator with a parametric super-Gaussian speech model for approximating the probability density distribution of the real speech spectral amplitudes. In fact, the parametric super-Gaussian distribution can approximate the Rayleigh–Laplace–Gamma distribution or other distributions exactly. Ye and Yokota [21] applied the radical root transformation to the super-Gaussian distributions. Thereby, they confirmed that the super-Gaussian distribution after r-th radical root transformation can be quasi-Gaussian distributed. By radical root transformation [21], the proposed method is applicable for major clusters that follow different distributions other than a Gaussian distribution. However, for clustering in image processing or other multiple dimensional applications, the major cluster following a Gaussian distribution is truly a strong assumption.

In addition to the problem addressed in this paper, many methods exist to solve this problem other than the mean-shift method. They have been discussed as described below. Under the normality assumption, Grubbs' test [22–24] and Thompson Tau test [25] are known as methods for testing whether the sample farthest from the sample mean is an outlier. These tests are applied sequentially from the samples that are outermost from the sample mean, but the number of outliers is only valid at most to several. Moreover, applying the tests to multi-dimensional data are not easy. If the outliers follow a Gaussian distribution and if the number of clusters in which the outliers are distributed is known, then, by applying a Gaussian mixture model [26–28], the mean and covariance matrix of major cluster can be estimated easily using the Expectation-maximization(EM) algorithm [29,30]. However, such an assumption cannot be applied generally to the outliers.

In fact, selection of the kernel bandwidth is an important issue that strongly affects the result of the general mean-shift algorithm compared to setting of the kernel type. Therefore, we only used the Gaussian kernel to make a presumption here. However, there are many commonly used kernel functions in addition to the Gaussian kernel, such as the Epanechnikov Kernel, the Uniform Kernel, the Quartic Kernel, and the Triweight Kernel. Application of it to other kernel functions according to the derivation of this article will undoubtedly make this research more comprehensive and general. Such application is expected to be an important part of our future research.

5. Application

Considering a stochastic process $x(t)$, the short-time Fourier spectrum centering on time t with a suitable window length is denoted as $X(t,f)$. Here, f represents the frequency. Let $X_f(t) \equiv X(t,f)$ be denoted as the spectrum series if frequency f is fixed. By applying the non-steady-state analysis of the stochastic process, the spectrogram $P(t,f) = |X(t,f)|^2$ denotes the power of the short-time Fourier spectrum $X(t,f)$. Because the frequency f is fixed, $P_f(t)$ will be designated as the power spectrum series.

Yokota and Ye [19] proposed a power spectrum estimation method robust for sudden noise. The method uses the radical root transformation to quasi-Gaussian distribution. The following concludes the process of the noise estimation algorithm proposed by Yokota and Ye [19]:

(1) Obtainpower spectrogram $P(t,f)$ from the noisy signal. We chose a pulse code modulation(PCM) recording of a noisy signal that contains a certain amount of sudden noise for analysis and computes the spectrogram with a Hamming window length of 10 ms achieving a 50% overlap between adjacent frames by short-time Fourier transformation. Figure 10a presents an example of a noisy signal for analysis and the corresponding spectrogram.

(2) Perform the following process for each frequency f:

(2-1) Use the radical root transformation in the power spectrum series $P_f(t)$ with the transformation parameter $r^* = 3.314$. Thereby, obtain the new quasi-Gaussian distributed power spectrum series $P_f^{1/r^*}(t)$. Figure 10b portrays a histogram of the power spectrum series at $f = 512$ Hz before the transformation.

(2-2) Compute the mode value of transformed power spectrum series $P_f^{1/r^*}(t)$ by kernel density estimation [5]. Then, put the mode value as the corresponding time average value $P_{noise}(f)$ of the noise power spectrum series. Figure 10c depicts a histogram of the transformed power spectrum series at f = 512 Hz, the kernel density estimation [5] with proper kernel bandwidth and the major cluster estimation using our proposed method.

(2-3) Compute the time average value $P(f)$ of the noise power spectrum series from the time average value $P_{noise}(f)$ as

$$P(f) = \left(\frac{P_{noise}(f)}{\Gamma\left(\frac{r^*+1}{r^*}\right)} \right)^{r^*} . \tag{33}$$

(3) Obtain $P(f)$ as an estimation of the noise power spectral density.

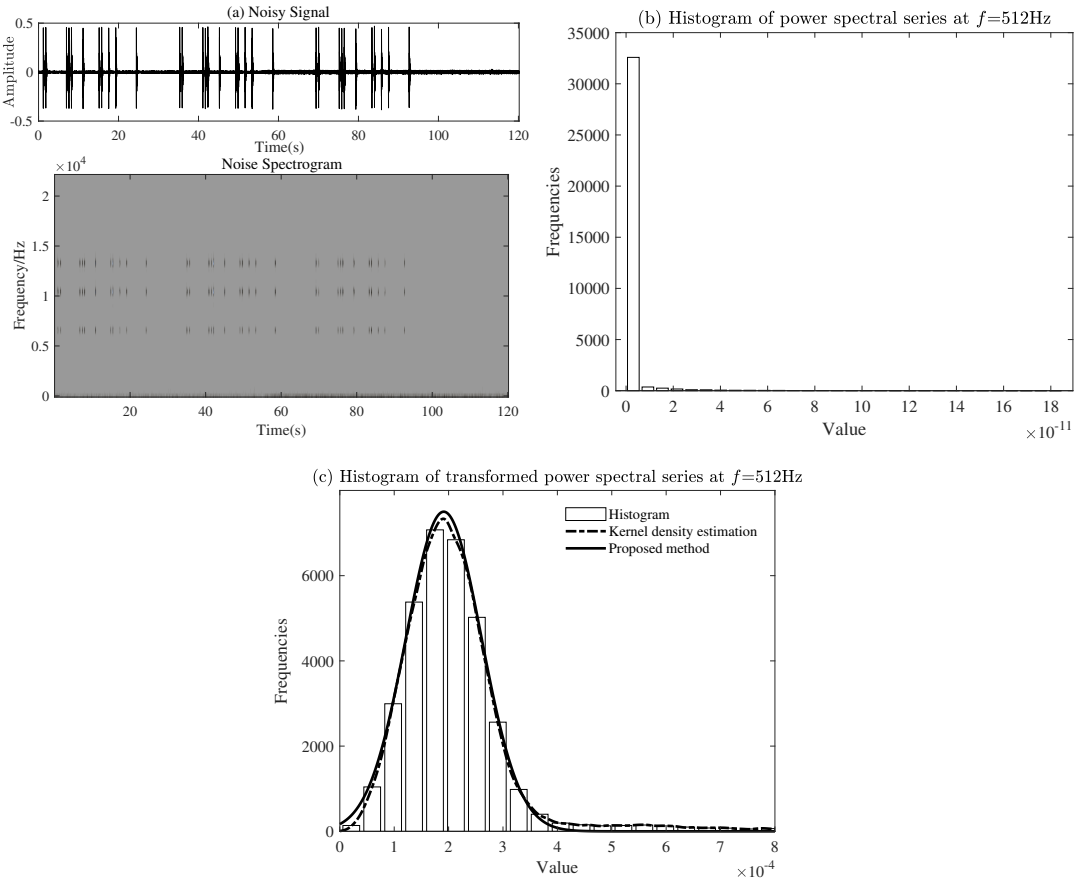

Figure 10. (**a**) example of a noisy signal for analysis and the corresponding spectrogram; (**b**) histogram of the power spectrum series of the noisy signal at f = 512 Hz; (**c**) histogram of the transformed power spectrum series of the noisy signal at f = 512 Hz.

In the noise estimation algorithm [19], the mode estimation accuracy directly affects the noise estimation result. As Figure 10c shows, kernel density estimation [5] can be replaced by our proposed method for comparison. The proper pre-setting of kernel bandwidth is also important in kernel density estimation [5]. It exhibits a strong influence on the resulting estimate similarly to the general mean-shift method. To illustrate its effects, we obtained the noise power spectrum series from the PCM recording, which is shown in Figure 10a, for analysis. Figure 11 portrays the relation between the kernel bandwidth and kernel density estimation. The histogram shows the true density. The broken

curve is under-smoothed because it includes too many spurious data artifacts arising from use of 0.000001 bandwidth, which is too small. The dotted curve is over-smoothed because using 0.0001 bandwidth obscures much of the underlying structure. The solid curve with 0.00003 bandwidth is regarded as optimally smoothed because its density estimate is close to the true density.

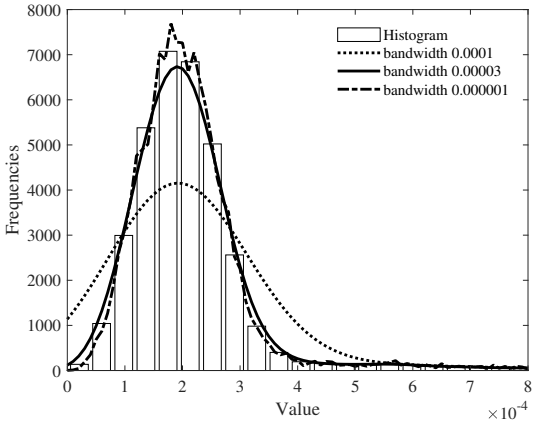

Figure 11. Relation between kernel bandwidth and kernel density estimation.

To assess the performance of the proposed method, general mean-shift method, and kernel density estimation for a noise estimation algorithm [19], this study uses PCM recordings of air-conditioning noise with some sudden noise, as shown in Figure 10a and without sudden noise, respectively, as test data and the true value. Noisy signal data in PCM recordings are not compressed. They have no power consumption. Figure 12 presents comparison results for noise estimation using the proposed method and kernel estimation. Here, we preset the kernel bandwidth as 0.0001. As Figure 12 shows, in the case in which an inappropriate kernel bandwidth is set in advance, noise estimation using our proposed method closely approximates the true noise, but the estimation accuracy using the kernel estimation is not high.

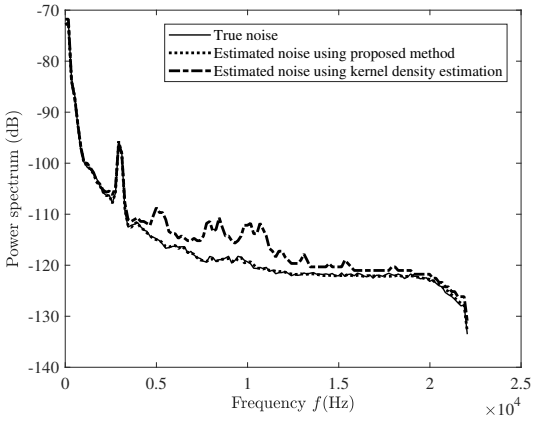

Figure 12. Comparison of the proposed method to kernel estimation for noise estimation.

6. Conclusions

The study described in this paper has addressed the problem of proper pre-setting for the fixed search kernel in a general mean-shift method. To improve the estimation accuracy, a new mean-shift method was proposed in which the mean vector and covariance matrix of the major cluster are estimated at each iteration. Then, the kernel bandwidth and shape are adjusted corresponding to the estimates. In numerical experiments, we compared the estimation accuracy and precision of the proposed method and of the general mean-shift method. The experimentally obtained results demonstrate that the estimation accuracy and precision of the proposed mean-shift are higher than

those of a general mean-shift method. Moreover, the proposed mean-shift can estimate the covariance matrix and the number of samples of major clusters effectively and correctly. Neither can be estimated using the general mean-shift method. These results were confirmed through formal experimentation, the results of which indicated the superior performance of our method compared to that of the general mean-shift method.

Author Contributions: Y.T. and Y.Y. conceived of and designed the methodology and experiments. Y.T. wrote the manuscript. Y.Y. reviewed and edited the manuscript. All authors read and approved the final manuscript.

Funding: This research received no external funding.

Acknowledgments: This work was supported by the Human Resource Development and R&D project on Manufacturing Technology for the Aerospace Industry: subsidy from Gifu Prefecture, Japan.

Conflicts of Interest: The authors declare that they have no conflict of interest related to this report or to the study it describes.

Appendix A. General Mean-Shift for a Multi-Dimensional Situation

Even when the target for data are multi-dimensional, it is fundamentally the same as the one-dimensional data. Sample $\boldsymbol{x}_n,\ n = 1,\ldots,N$ of the M-dimensional column vector includes the major cluster of N_N points and a few outliers. The major cluster follows an M-dimensional Gaussian distribution with mean vector $\boldsymbol{\mu}_N$ and covariance matrix $\boldsymbol{C}_N$. Here, the mode of the major cluster is not biased from the mean vector $\boldsymbol{\mu}_N$ under the influence of N_O point outliers. The iteration process in the multi-dimensional mean-shift method is the following:

1. Let the mean vector $\boldsymbol{\mu}_x$ of sample $\boldsymbol{x}_n,\ n = 1,\ldots,N$ be the initial value of the mean estimator $\hat{\boldsymbol{\mu}}_N$ of the major cluster
$$\hat{\boldsymbol{\mu}}_N \leftarrow \boldsymbol{\mu}_x. \tag{A1}$$
2. Consider a M-dimensional Gaussian distribution $p(\boldsymbol{x};\boldsymbol{\mu}_W,\boldsymbol{C}_W)$ with mean vector $\boldsymbol{\mu}_W$ and covariance matrix $\boldsymbol{C}_N$ as the kernel function in value direction. Here, the mean mean vector $\boldsymbol{\mu}_W$ of kernel function is ascertained by the mean estimator of major cluster
$$\boldsymbol{\mu}_W \leftarrow \hat{\boldsymbol{\mu}}_N. \tag{A2}$$
In addition, covariance matrix $\boldsymbol{C}_N$ is assigned to be an appropriate size as discussed in Section 2.2.
3. The weight $a_n,\ n = 1,\ldots,N$ for each sample $\boldsymbol{x}_n,\ n = 1,\ldots,N$ weighted by such a Gaussian kernel is
$$a_n = \frac{1}{A} p(\boldsymbol{x}_n;\boldsymbol{\mu}_W,\boldsymbol{C}_W). \tag{A3}$$
However,
$$A = \sum_{k=1}^{N} p(\boldsymbol{x}_k;\boldsymbol{\mu}_W,\boldsymbol{C}_W). \tag{A4}$$
We use this weight a_n to calculate the sample mean vector $\boldsymbol{\mu}_x$ with $\boldsymbol{x}_n,\ n = 1,\ldots,N$ as
$$\boldsymbol{\mu}_x = \sum_{n=1}^{N} a_n \boldsymbol{x}_n. \tag{A5}$$
4. The value of mean vector estimator $\hat{\boldsymbol{\mu}}_N$ for the major cluster is updated using the following equation:
$$\hat{\boldsymbol{\mu}}_N \leftarrow \boldsymbol{\mu}_x. \tag{A6}$$
5. If the value variation of mean vector estimator $\hat{\boldsymbol{\mu}}_N$ is equal to or less than the predetermined fixed value, the update process is terminated. Otherwise, return to 2 and repeat the iteration.

Appendix B. Proof of Equation (17)

Equation (16) can be rewritten as

$$E[C_x] = E\left[\frac{\frac{1}{N_N}\sum_{n=1}^{N_N} p(x_n;C_W)x_n^2}{\frac{1}{N_N}\sum_{k=1}^{N_N} p(x_k;C_W)}\right]. \tag{A7}$$

The denominator $\frac{1}{N_N}\sum_{k=1}^{N_N} p(x_k;C_W)$ and numerator $\frac{1}{N_N}\sum_{n=1}^{N_N} p(x_n;C_W)x_n^2$ are both random variables. Obviously, if the standard deviation of the denominator is sufficiently small compared to the expected value of the denominator, Equation (16) can be approximated as shown below because the denominator can be regarded as a simple variable rather than a random variable

$$E[\boldsymbol{C}_x] \simeq \frac{E\left[\frac{1}{N_N}\sum_{n=1}^{N_N} p(x_n;C_W)\boldsymbol{x}_n^2\right]}{E\left[\frac{1}{N_N}\sum_{k=1}^{N_N} p(x_k;C_W)\right]}, \tag{A8}$$

as shown in Equation (17). Hereafter, it is proved that the standard deviation can be as small as possible with respect to the expected value of the denominator when the number of samples $N_N \to \infty$.

Proof. The denominator on the right side of Equation (A7) has the form of

$$y = \frac{1}{N_N}\sum_{n=1}^{N_N} x_n. \tag{A9}$$

□

The expected value $E(y)$ and the standard deviation $\sigma(y)$ are

$$E(y) = E(x_n) > 0, \tag{A10}$$

$$\sigma(y) = \frac{1}{\sqrt{N_N}}\sigma(x_n). \tag{A11}$$

If the number N_N of samples is sufficiently large, which means $N_N \to \infty$, $\sigma(y)$ for $E(y)$ converges to 0. $p(x_n;C_W)$ is non-negative because it is a probability density distribution. That is, since the random variable x_n follows the probability density distribution $f(x)$ defined by $x \geq 0$, the expected value $E[x_n]$ of x_n is always positive. Regardless of the number of samples N_N, it becomes $E[y] = E[x_n]$, so that, with the number of samples $N_N \to \infty$, the denominator can reduce the standard deviation as much as possible relative to the expected value.

The expected values and standard deviations for various probability density distributions $f(x)$ defined by $x \geq 0$ are presented in Table A1. The table shows that, for all probability density distributions shown in this table, the standard deviation $\sigma(x_n)$ does not become larger than the expected value $E(x_n)$ beyond the order. The same is probably true for other probability density distributions not listed in this table. Therefore, corresponding to the number of samples $N_N = 100$, the standard deviation $\sigma(y)$ can be about one-tenth of the expected value $E(y)$. Practically speaking, Equation (A8), i.e., the approximation of Equation (A7), holds.

Table A1. Expected value and standard deviation of probability density distribution $f(x)$ defined by $x \geq 0$.

$f(x)$	Expectation	S.D.
gamma	$k\theta$	$\sqrt{k}\theta$
χ^2	k	$\sqrt{2k}$
exponential	$1/\lambda$	$1/\lambda$
Erlang	$k\mu$	$\sqrt{k}\mu$
Rayleigh	$\sigma\sqrt{\pi/2}$	$\sigma\sqrt{2-\pi/2}$
log-normal	$e^{\mu+\sigma^2/2}$	$e^{\mu+\sigma^2/2}\sqrt{e^{\sigma^2}-1}$
Pareto	$\frac{ab}{a-1}$	$\frac{\sqrt{a}b}{(a-1)\sqrt{a-2}}$

Appendix C. Multi-Dimensional Mean-Shift with Updating Kernel

Appendix C.1. Derivation of Standard Deviation of a Major Cluster from the Sample

Here, we extend derivation of the estimated value for standard deviation σ_N in the one-dimensional derived in Section 3.1 to multi-dimensional. The major cluster is assumed to follow a multi-dimensional (M-dimensional) normal distribution. Although the covariance matrix generally does not become a diagonal matrix, it is possible to re-coordinate the coordinate axes so that the covariance matrix becomes a diagonal matrix by appropriate orthogonal transformation. Furthermore, the coordinate axes are shifted such that the mean vector becomes a zero vector. In this section, we consider the variable $(x_1, \ldots, x_M)$ in such a transformed coordinate system. We let the variables be $\boldsymbol{x} = (x_1, \ldots, x_M)^T$ and denote the standard deviation of each variable by $\boldsymbol{\sigma}_N = (\sigma_{N,1}, \ldots, \sigma_{N,M})^T$. On the newly revised coordinate axes, because the covariance is zero, a M-dimensional normal distribution is represented as a direct product of the one-dimensional normal distribution of each variable as

$$p(\boldsymbol{x}; \boldsymbol{\sigma}_N) = \prod_{m=1}^{M} p(x_m; \sigma_{N,m}). \tag{A12}$$

The kernel function in the value direction is also assumed to be a Gaussian distribution with a mean zero vector and a diagonal covariance matrix. Because the standard deviation of each variable is $\boldsymbol{\sigma}_W = (\sigma_{W,1}, \ldots, \sigma_{W,M})^T$, the Gaussian distribution of kernel function is

$$p(\boldsymbol{x}; \boldsymbol{\sigma}_W) = \prod_{m=1}^{M} p(x_m; \sigma_{W,m}). \tag{A13}$$

Using this Gaussian kernel, the weight $a_n, n = 1, \ldots, N_N$ for the sample $\boldsymbol{x}_n = (x_{1,n}, \ldots, x_{M,n})^T$, $n = 1, \ldots, N_N$ can be denoted as

$$a_n = \frac{1}{A} p(\boldsymbol{x}_n; \boldsymbol{\sigma}_W). \tag{A14}$$

However, A in the above equation is

$$A = \sum_{k=1}^{N_N} p(\boldsymbol{x}_k; \boldsymbol{\sigma}_W). \tag{A15}$$

The sample variance $\sigma_{x,m}^2$ weighted by a_n is

$$\sigma_{x,m}^2 = \sum_{n=1}^{N_N} a_n x_{m,n}^2, \quad m = 1, \ldots, M. \tag{A16}$$

For the same reason, under the one-dimensional case, by substituting Equation (A14) into Equation (A16), the expected value of the sample variance $\sigma_{x,m}^2$ can be approximated as

$$E[\sigma_{x,m}^2] \simeq \frac{1}{E[A]} E\left[\sum_{n=1}^{N_N} p(\boldsymbol{x}_n; \sigma_W) x_{m,n}^2\right]. \tag{A17}$$

Applying Equation (A12) to Equation (A15) and using Equation (13), the expected value of A is found as

$$\begin{aligned} E[A] &= E\left[\sum_{k=1}^{N_N} p(\boldsymbol{x}_k; \sigma_W)\right] \\ &= \sum_{k=1}^{N_N} E[p(\boldsymbol{x}_k; \sigma_W)] \\ &= N_N \prod_{j=1}^{M} \int_{-\infty}^{\infty} p(x_j; \sigma_{W,j}) p(x_j; \sigma_{N,j}) dx_j \\ &= \frac{N_N}{(2\pi)^{M/2}} \prod_{j=1}^{M} (\sigma_{W,j}^2 + \sigma_{N,j}^2)^{-1/2}, \end{aligned} \tag{A18}$$

while using Equation (14), the remainder of Equation (A17) is

$$\begin{aligned} & E\left[\sum_{n=1}^{N_N} p(\boldsymbol{x}_n; \sigma_W) x_{m,n}^2\right] \\ &= \sum_{n=1}^{N_N} E[p(\boldsymbol{x}_n; \sigma_W) x_{m,n}^2] \\ &= N_N \prod_{j=1}^{M} \int_{-\infty}^{\infty} x_m^2 p(x_j; \sigma_{W,j}) p(x_j; \sigma_{N,j}) dx_j \\ &= \frac{\sigma_{W,m}^2 \sigma_{N,m}^2}{\sigma_{W,m}^2 + \sigma_{N,m}^2} \frac{N_N}{(2\pi)^{M/2}} \prod_{j=1}^{M} (\sigma_{W,j}^2 + \sigma_{N,j}^2)^{-1/2}. \end{aligned} \tag{A19}$$

That is, according to Equation (A18) and Equation (A19), Equation (A17) becomes

$$E[\sigma_{x,m}^2] = \frac{\sigma_{W,m}^2 \sigma_{N,m}^2}{\sigma_{W,m}^2 + \sigma_{N,m}^2}. \tag{A20}$$

The equation above can be transformed to

$$\sigma_{N,m}^2 = \frac{\sigma_{W,m}^2 E[\sigma_{x,m}^2]}{\sigma_{W,m}^2 - E[\sigma_{x,m}^2]}. \tag{A21}$$

The standard deviation $\sigma_{N,m}$ of a major cluster can be estimated as

$$\hat{\sigma}_{N,m} = \sqrt{\frac{\sigma_{W,m}^2 \sigma_{x,m}^2}{\sigma_{W,m}^2 - \sigma_{x,m}^2}}, \tag{A22}$$

when using the standard deviation $\sigma_{x,m}$ of the sample weighted with a Gaussian kernel with standard deviation σ_W. Furthermore, using Equation (A18), we can estimate the number N_N of samples belonging to a major cluster as

$$\hat{N}_N = A(2\pi)^{M/2} \prod_{j=1}^{M} (\sigma_{W,j}^2 + \hat{\sigma}_{N,j}^2)^{1/2}. \tag{A23}$$

The standard deviation σ_W of the Gaussian kernel is assigned adaptively as r times the estimated value $\hat{\sigma}_N$ of the standard deviation at each iteration. The appropriate value of the scale factor r is discussed later in relation to a numerical experiment.

Appendix C.2. Mean-Shift Method with Updating Kernel

1. The mean vector $\boldsymbol{\mu}_x$ and the covariance matrix $\boldsymbol{C}_x$ of the whole samples are determined using the following equations:

$$\boldsymbol{\mu}_x = \frac{1}{N} \sum_{n=1}^{N} \boldsymbol{x}_n, \tag{A24}$$

$$\boldsymbol{C}_x = \frac{1}{N} \sum_{n=1}^{N} (\boldsymbol{x}_n - \boldsymbol{\mu}_x)(\boldsymbol{x}_n - \boldsymbol{\mu}_x)^T. \tag{A25}$$

The initial values of the mean vector $\hat{\boldsymbol{\mu}}_N$ and the covariance matrix $\hat{\boldsymbol{C}}_N$ of the major cluster are assigned as

$$\hat{\boldsymbol{\mu}}_N \leftarrow \boldsymbol{\mu}_x, \tag{A26}$$

$$\hat{\boldsymbol{C}}_N \leftarrow \boldsymbol{C}_x. \tag{A27}$$

2. One can consider a multi-dimensional Gaussian distribution $p(\boldsymbol{x}; \boldsymbol{\mu}_W, \boldsymbol{C}_W)$ with mean vector $\boldsymbol{\mu}_W$ and covariance matrix $\boldsymbol{C}_W$ as the kernel function in the value direction. Here, the mean vector $\boldsymbol{\mu}_W$ and covariance matrix $\boldsymbol{C}_W$ of the kernel function are determined as

$$\boldsymbol{\mu}_W \leftarrow \hat{\boldsymbol{\mu}}_N, \tag{A28}$$

$$\boldsymbol{C}_W \leftarrow r^2 \hat{\boldsymbol{C}}_N. \tag{A29}$$

Actually, r^2 in the above equation is derived from the fact that the covariance matrix has the squared order of the standard deviation.

3. Weight a_n for each sample $\boldsymbol{x}_n$ weighted by such a Gaussian kernel is calculated using Equations (A3) and (A4). The mean vector $\boldsymbol{\mu}_x$ and the covariance matrix $\boldsymbol{C}_x$ are determined using the following equations:

$$\boldsymbol{\mu}_x = \sum_{n=1}^{N} \boldsymbol{a}_n \boldsymbol{x}_n, \tag{A30}$$

$$\boldsymbol{C}_x = \sum_{n=1}^{N} a_n (\boldsymbol{x}_n - \boldsymbol{\mu}_x)(\boldsymbol{x}_n - \boldsymbol{\mu}_x)^T. \tag{A31}$$

4. The value of mean vector estimator $\hat{\boldsymbol{\mu}}_N$ is updated using the following equation:

$$\hat{\boldsymbol{\mu}}_N \leftarrow \boldsymbol{\mu}_x. \tag{A32}$$

Let

$$\boldsymbol{C}_W = \boldsymbol{V}_W \boldsymbol{\Lambda}_W \boldsymbol{V}_W^T \tag{A33}$$

be an eigenvalue decomposition of the covariance matrix $\boldsymbol{C}_W$, which can be represented as a symmetric matrix of the kernel. The diagonal elements of the diagonalized matrix $\boldsymbol{\Lambda}_W$ are eigenvalues of $\boldsymbol{C}_W$; they represent the variances $\sigma^2_{W,1}, \ldots, \sigma^2_{W,M}$ along the directions represented by each of the column vectors of orthogonal matrix $\boldsymbol{V}_W$. In addition, the diagonal element of

$$\boldsymbol{\Lambda}_x = \boldsymbol{V}_W^T \boldsymbol{C}_x \boldsymbol{V}_W \tag{A34}$$

is the variance $\sigma^2_{x,1}, \ldots, \sigma^2_{x,M}$ of $\boldsymbol{V}_W$ in the column vector direction in the sample covariance matrix $\boldsymbol{C}_x$. According to Equation (A20), we can estimate the number N_N of samples belonging to the major cluster by the standard deviation $\sigma_{N,1}, \ldots, \sigma_{N,M}$, which is obtained by $\sigma^2_{W,1}, \ldots, \sigma^2_{W,M}$ and $\sigma^2_{x,1}, \ldots, \sigma^2_{x,M}$ in Equation (A19). Let $\hat{\boldsymbol{\Lambda}}_N$ be the diagonal matrix that has the estimated $\hat{\sigma}_{N,1}, \ldots, \hat{\sigma}_{N,M}$ as the diagonal elements. Using $\hat{\boldsymbol{\Lambda}}_N$, the covariance matrix $\hat{\boldsymbol{C}}_N$ is updated with the following equation:

$$\hat{\boldsymbol{C}}_N \leftarrow \boldsymbol{V}_W \hat{\boldsymbol{\Lambda}}_N \boldsymbol{V}_W^T. \tag{A35}$$

The estimated value $\hat{N}_N$ of the number of samples belonging to a major cluster is updated using the following equation:

$$\hat{N}_N \leftarrow A(2\pi)^{M/2} \prod_{j=1}^{M} (\sigma^2_{W,j} + \hat{\sigma}^2_{N,j})^{1/2}. \tag{A36}$$

5. If the value variations of $\hat{\boldsymbol{\mu}}_N, \hat{\boldsymbol{C}}_N, \hat{N}_N$ are equal to or less than the predetermined fixed value, then the update process is terminated. Otherwise, return to 2 and repeat the iteration.

References

1. Fukunaga, K.; Hostetler, L. The estimation of the gradient of a density function, with applications in pattern recognition. *IEEE Trans. Inf. Theory* **1975**, *21*, 32–40. [CrossRef]
2. Cheng, Y. Mean shift, mode seeking, and clustering. *IEEE Trans. Pattern Anal. Mach. Intell.* **1995**, *17*, 790–799. [CrossRef]
3. Comaniciu, D.; Meer, P. Mean shift analysis and applications. In Proceedings of the Seventh IEEE International Conference on Computer Vision, Kerkyra, Greece, 20–27 September 1999; pp. 1197–1203.
4. Comaniciu, D.; Meer, P. Mean shift: A robust approach toward feature space analysis. *IEEE Trans. Pattern Anal. Mach. Intell.* **2002**, *24*, 603–619. [CrossRef]
5. Wand, M.P.; Jones, M.C. *Kernel Smoothing*; Springer: London, UK, 1995.
6. Sheather, S.J.; Jones, M.C. A reliable data-based bandwidth selection method for kernel density estimation. *J. R. Stat. Soc. Ser. B* **1991**, *53*, 683–690. [CrossRef]
7. Chen, S. Optimal bandwidth selection for kernel density functionals estimation. *J. Probab. Stat.* **2015**, *70*, 1–21. [CrossRef]
8. Slaoui, Y. Data-Driven Bandwidth Selection for Recursive Kernel Density Estimators Under Double Truncation. *Sankhya B* **2018**, *80*, 341–368. [CrossRef]
9. Comaniciu, D.; Ramesh, V.; Meer, P. The variable bandwidth mean shift and data-driven scale selection. In Proceedings of the 1991 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Maui, HI, USA, 3–6 June 1991; pp. 438–445.
10. Okada, K.; Comaniciu, D.; Krishnan, A. Scale selection for anisotropic scale-space. In Proceedings of the 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Washington, DC, USA, 27 June–2 July 2004; pp. 594–601.
11. Comaniciu, D. An algorithm for data-driven bandwidth selection. *IEEE Trans. Pattern Anal. Mach. Intell.* **2003**, *25*, 281–288. [CrossRef]
12. Li, X.; Hu, Z.; Wu, F. A note on the convergence of the mean shift. *Pattern Recognit.* **2007**, *40*, 1756–1762. [CrossRef]
13. Ghassabeh, Y.A.; Rudzicz, F. Modified mean shift algorithm. *IET Image Process.* **2018**, *12*, 2171–2177.
14. Birchfield, S.T.; Rangarajan, S. Mean shift blob tracking with kernel histogram filtering and hypothesis testing. *Pattern Recognit. Lett.* **2005**, *26*, 605–614.

15. Leichter, I. Mean shift trackers with cross-bin metrics. *IEEE Trans. Pattern Anal. Mach. Intell.* **2011**, *34*, 695–706. [CrossRef] [PubMed]
16. Vojir, T.; Noskova, J.; Matasa, J. Robust scale-adaptive mean-shift for tracking. *Pattern Recognit. Lett.* **2014**, *49*, 250–258. [CrossRef]
17. Silverman, B.W. *Density Estimation for Statistics and Data Analysis*; Chapman & Hall: London, UK, 1986.
18. Kawamura, Y.; Yokota, Y.; Matsumaru, N.; Shirai, K. 24 Hours monitoring system of heart rate variability to predict septic shock. *IEICE Tech. Rep.* **2012**, *112*, 29–34.
19. Yokota, Y.; Ye, T. Quasi-Gaussian distributed power spectrum series by radical root transform and application to robust power spectrum density estimation against for sudden noise. *IEICE Trans. Fundam. (Jpn. Ed.)* **2016**, *3*, 149–158.
20. Lotter, T.; Vary, P. Noise reduction by maximum a posteriori spectral amplitude estimation with super-Gaussian speech modelling. In Proceedings of the International Workshop on Acoustic Echo and Noise Control, Kyoto, Japan, 8 September 2003; pp. 83–88.
21. Ye, T.; Yokota, Y. Noise estimation for speech enhancement based on quasi-Gaussian distributed power spectrum series by radical root transformation. *IEICE Trans. Fundam. Electron. Commun. Comput. Sci.* **2017**, *6*, 1306–1314. [CrossRef]
22. Grubbs, F.E. Sample criteria for testing outlying observations. *Ann. Math. Stat.* **1950**, *21*, 27–58. [CrossRef]
23. Grubbs, F.E.; Beck, G. Extension of sample sizes and percentage points for significance tests of outlying observations. *Technometrics* **1972**, *14*, 847–854. [CrossRef]
24. Zeller, C.B.; Lachos, V.H.; Labra, F.V. Influence diagnostics for Grubbs's model with asymmetric heavy-tailed distributions. *Stat. Pap.* **2014**, *55*, 671–690. [CrossRef]
25. Thompson, R. A note on restricted maximum likelihood estimation with an alternative outlier model. *J. R. Stat. Soc. Ser. B (Methodol.)* **1985**, *47*, 53–55. [CrossRef]
26. Rasmussen, C.E. The infinite Gaussian mixture model. In *Advances in Information Processing Systems 12*; MIT Press: Cambridge, MA, USA, 2000; pp. 554–560.
27. Blomer, J.; Bujna, K. Adaptive seeding for Gaussian mixture models. In Proceedings of the 20th Pacific-Asia Conference, PAKDD 2016, Auckland, New Zealand, 19–22 April 2016; Volume 9652, pp. 296–308.
28. Viroli, C.; McLachlan, G.J. Deep Gaussian mixture models. *Stat. Comput.* **2017**, *29*, 43–51. [CrossRef]
29. Dempster, A.P.; Laird, N.M.; Rubin, D.B. Maximum likelihood from incomplete data via the EM algorithm. *J. R. Stat. Soc. Ser. B (Stat. Methodol.)* **1977**, *39*, 1–38. [CrossRef]
30. Melnykov, V.; Melnykov, I. Initializing the EM algorithm in Gaussian mixture models with an unknown number of components. *J. Comput. Stat. Data Anal.* **2012**, *56*, 1381–1395. [CrossRef]