DOCTORAL DISSERTATION

Research on noise reduction based on mode estimation utilizing Gaussian property

December, 2019

Electronics and Information Systems Engineering Division
Graduate School of Engineering
Gifu University
Japan

Tian Ye

Research on noise reduction based on mode estimation utilizing Gaussian property

by

Tian Ye

Submitted in partial satisfaction of the requirements for
the degree of Doctor of Philosophy
in Engineering

Electronics and Information Systems Engineering Division
Graduate School of Engineering
Gifu University
Japan

December, 2019

Research on noise reduction based on mode estimation utilizing Gaussian property

Tian Ye

Submitted in partial satisfaction of the requirements for
the degree of Doctor of Philosophy
in Engineering

December, 2019

Abstract

How to suppress the influence of noise on the estimation result has always been an important issue in the field of signal processing and data processing which are concerned in this thesis. The following are two important issues related to noise reduction for speech enhancement and data analysis, respectively.

The first one is for speech enhancement. Speech is a fundamental method of human communication. With advances in technology of digital signal processing, speech processing equipment, likes cellular phones and professional mobile radio, are an integral part of everyday life. Environmental noise is one of negative factors which widely exist in speech processing equipment for signal processing, such as traffic noise, train noise, office noise etc. Suppression of the acoustic background noise is a relevant and challenging problem. Apart from reducing listener fatigue and improving the quality and intelligibility of speech, noise reduction which can be called speech enhancement, is crucial to obtain good performance of the speech signal processing. For most speech enhancement algorithms, an estimate for the noise spectrum is assumed to be available. Such an estimate is crucially important for speech-enhancement performance. The noise estimate strongly affects the enhanced signal quality. Annoying residual noise will be audible if the noise estimate is too low. Alternatively, if the noise estimate is too high, then speech will be destroyed possibly resulting in loss of intelligibility.

The other one is related to data analysis. As measuring a certain physical quantity, the relatively small number of abnormal values or outliers, hereinafter referred to as outliers, are included in the normal measured values that contain measure-

ment noise. It is one of the frequently occurred works in science and engineering to estimate the true statistical parameters of the physical quantity from these measured values including outliers. Because the measurement noise follows a Gaussian distribution with mean zero in general, all the samples form the major cluster are Gaussian-distributed around the true value. The problem mentioned 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.

In order to solve the problems above, the research focuses on the noise estimation in speech enhancement for speech signal processing and the estimation of major cluster for data analysis. And the construction of this thesis is summarized as follows.

Chapter 1 is the introduction which describes developing and the main problems both in speech enhancement and major cluster estimation. The motivation and organization of the thesis is followed.

Chapter 2 contents the proposed noise estimation method for speech enhancement based on quasi-Gaussian distributed power spectrum series by radical root transformation. This contribution presents and analyzes the statistical regularity related to the noise power spectrum series and the speech spectrum series. It also undertakes a thorough inquiry of the quasi-Gaussian distributed power spectrum series obtained using the radical root transformation. Consequently, a noise-estimation algorithm is proposed for speech enhancement. This method is effective for separating the noise power spectrum from the noisy speech power spectrum. In contrast to standard noise-estimation algorithms, the proposed method requires no speech activity detector. It was confirmed to be conceptually simple and well suited to real-time implementations. Practical experiment tests indicated that our method is preferred over previous methods.

Chapter 3 proposes a new estimation method for the major cluster by the mean-shift with updating kernel. The mean-shift method which is known as a convenient mode-seeking method. Utilizing a principle that the sample mean over an analysis window, hereafter referred to the kernel according to custom, in the data space where the samples are distributed is biased toward the densest direction of samples from the center of the kernel, the mean-shift method tries to seek the densest point of samples, or the sample mode, iteratively. A smaller kernel causes convergence to a local mode appeared due to statistical fluctuation; on the other hand, a larger kernel causes estimating a biased mode affected by other clusters, abnormal values, or outliers if there are existing other than the major cluster. Therefore, the optimal

selection of the kernel size, which is referred to the bandwidth in many references, is an important problem. In this paper, under the assumptions that the major cluster follows a Gaussian probability density distribution and the outliers do not affect the sample mode of the major cluster, adopting Gaussian kernel, we proposed a new mean-shift in which both the mean vector and covariance matrix of the major cluster are estimated in each iteration, then the kernel size and shape are updated adaptively. Numerical experiments indicate that the mean vector, covariance matrix and the number of samples belonging to the major cluster can be stably estimated. Because the kernel shape can be adjusted not only to an isotopic shape but also to an anisotropic shape according to the sample distribution, the proposed method is shown to have higher estimation precision compared to the general mean-shift.

Chapter 4 will draw the conclusion of the research. At the end of the thesis, prospective ideas of future works will be explored.

Keywords: Spectrogram, power spectrum series, quasi-Gaussian distribution, speech activity detector, power spectral density, radical root transformation, mean-shift, mode estimation, kernel bandwidth and shape, outliers, updating kernel

Acknowledgements

This dissertation could not have been fulfilled if it were not for the understanding and support of the following people:

The author is deeply grateful for a lot of things beyond words alone can express. Above all, the author is most thankful to his family for their great support.

The author's deepest gratitude goes first and foremost to his main supervisor Prof. Dr. Yasunari Yokota for the participation and meaningful ideas during his weekly lab seminar, the tirelessness effort and professional discussion with all Lab students which is indirectly encourage the author and as his motivator during his six-years study in Gifu University.

The author would like to express his gratitude to Prof. Dr. Satoru Hayamizu and Associate Prof. Dr. Motoki Shiga for their valuable discussion through the dissertation defense.

The author thanks all the members of Yokota laboratory for the daily discussion on the related matter.

Finally, the author would like to express his deep gratitude to those who have contributed in one way or another to the completion of this work. Last but not least, the author would like to thank you, for your interest in his dissertation.

Contents

Abstract **v**

Acknowledgements **ix**

1 Introduction **1**

1.1 Noise Estimation in Speech-Enhancement 2

1.2 Estimating the Major Cluster by Mean-Shift with Updating Kernel . 3

2 Noise Estimation for Speech Enhancement Based on Quasi-Gaussian Distributed Power Spectrum Series by Radical Root Transformation **7**

2.1 Quasi-Gaussian distributed power spectrum series of noise and speech 7

2.1.1 Power spectrum series 7

2.1.2 Probability density distribution of the power spectrum series . 8

2.1.3 Box–Cox transformation [24] and radical root transformation [7] 12

2.1.4 Probability density distribution after radical root transformation [7] 13

2.1.5 Evaluation of the normality about the power spectrum series after radical root transformation [7] 14

2.2 Proposed noise estimation algorithm based on the quasi-Gaussian distributed power spectrum series 18

2.3 Experimental results 20

2.3.1 Typical conventional noise estimation: Martin's minimum trackingalgorithm [2] 20

2.3.2 Characteristic of the Martin's minimum tracking algorithm . . 21

2.3.3 Comparison of results obtained using the proposed method and Martin's minimum tracking algorithm 22

2.4 Discussion 23

2.5 Conclusion 25

3 Estimating the Major Cluster by Mean-Shift with Updating Kernel 27

3.1 General Mean-Shift Method 27
3.1.1 General Mean-Shift Method 27
3.1.2 Shortcomings and Solution of the General Mean-Shift Method 28
3.2 One-Dimensional Mean-Shift with Updating Kernel 31
3.2.1 Derivation of Major Cluster Standard Deviation σ_N from Sample Standard Deviation σ_x 31
3.2.2 Mean-Shift with Updating Kernel 33
3.3 Numerical Experiment 35
3.3.1 Update Process of Mean-Shift with an Updatable Kernel . . . 35
3.3.2 Influence of Kernel Bandwidth on Estimation Accuracy (Unbiasedness) 36
3.3.3 Influence of the Scale Factor r Value on Estimation Accuracy 39
3.3.4 Verification of Consistency 42
3.3.5 Estimation Precisions of the Proposed and General Mean-Shift Methods 44
3.3.6 Discussion 46
3.4 Application 47
3.5 Conclusions 50

4 Conclusions and Future Works 53

4.1 Conclusions 53
4.2 Future Works 53

List of publications 59

A General Mean-Shift for a Multi-Dimensional Situation 61

B Proof of Equation (3.17) 63

C Multi-Dimensional Mean-Shift with Updating Kernel 65

C.1 Derivation of Standard Deviation of a Major Cluster from the Sample 65
C.2 Mean-Shift Method with Updating Kernel 68

List of Figures

2.1 Scattergram of spectrum series $X_f(t)$ at $f = 512$Hz on the complex plane and result fitted with two-dimensional Gaussian distribution: (a) Scattergram related to to air-conditioning noise, (b) Scattergram related to vacuum cleaner noise, (c) Histogram for (a), (d) Histogram for (b), (e) Two-dimensional distribution compared to (c), and (f) Two-dimensional distribution compared to (d). 9

2.2 Histogram of speech spectral amplitudes and fitted approximation by super-Gaussian distribution: (a) Chinese speech signal, (b) Japanese speech signal, (c) Japanese voiced part, and (d) Japanese unvoiced part. 11

2.3 Original super-Gaussian distribution and comparison between the speech amplitude series after radical root transformation with optimal parameter r and Gaussian distribution according to a different parameter set (β, v): top panel, r; bottom panel, $r = 1.656$. 17

2.4 (a) A noisy speech signal for analysis and its corresponding spectrogram. (b) Histogram of the power spectrum series of the noisy speech signal at f=512 Hz. (c) Histogram of the transformed power spectrum series of the noisy speech signal at f=512 Hz and the Gaussian mixture model. 19

2.5 Top panel: Plot of noisy speech power spectrum and noise estimate for noisy speech at f=500 Hz. Bottom panel: Plot of true and estimated noise power spectrum for the same noisy speech at f=500 Hz. 22

2.6 Noisy speech signal under the condition noise, the corresponding spectrogram and the comparison between the proposed method and the Martin's method [2] for the noisy speech signal under condition noise: (a) air-conditioning noise, (b) vacuum cleaner noise (low gear), (c) vacuum cleaner noise (high gear). 24

3.1 Bias error and estimation variance for various fixed kernel bandwidth σ^2 in a general mean-shift method. 30
3.2 Example of a sample set for numerical experiments. 35
3.3 Updates of the estimated major cluster. 37
3.4 Bias errors for various initial kernel bandwidths σ^2 in the proposed method and the general mean-shift method. 39
3.5 Bias errors for various numbers N of samples in the proposed method. 40
3.6 Bias errors for various scale factors r in the proposed method. 41
3.7 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. . . . 43
3.8 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. 45
3.9 Estimating the variance of the estimate $\hat{\boldsymbol{\mu}}_N$ for various kernel bandwidths σ^2 in the general mean-shift method. 46
3.10 (**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. 49
3.11 Relation between kernel bandwidth and kernel density estimation. . . 50
3.12 Comparison of the proposed method to kernel estimation for noise estimation. 51

List of Tables

2.1 Comparison of the KL divergence between the true noise spectrum and noise estimation. 23
2.2 Recording condition 23

B.1 Expected value and standard deviation of probability density distribution $f(x)$ defined by $x \geq 0$. 64

Chapter 1

Introduction

Noise refers to unnecessary information or data other than the information of the processing object. Measurement data may cause individual data to be unrealistic or lost due to environmental interference or human factors during its acquisition and transmission. These data are often referred to as noise or outliers. In order to restore the objective authenticity of the data in order to get better analysis results, it is necessary to perform noise reduction analysis on the original data or to eliminate outliers. How to suppress the influence of noise or outliers on the estimation result has always been an important issue in the field of signal processing and data processing which are concerned in this thesis. The following are two important issues related to noise reduction for speech enhancement and data analysis, respectively.

The first one is for speech enhancement. Speech is a fundamental method of human communication. With advances in technology of digital signal processing, speech processing equipment, likes cellular phones and professional mobile radio, are an integral part of everyday life. Environmental noise is one of negative factors which widely exist in speech processing equipment for signal processing, such as traffic noise, train noise, office noise etc. Suppression of the acoustic background noise is a relevant and challenging problem. Apart from reducing listener fatigue and improving the quality and intelligibility of speech, noise reduction which can be called speech enhancement, is crucial to obtain good performance of the speech signal processing. For most speech enhancement algorithms, an estimate for the noise spectrum is assumed to be available. Such an estimate is crucially important for speech-enhancement performance. The noise estimate strongly affects the enhanced signal quality. Annoying residual noise will be audible if the noise estimate is too low. Alternatively, if the noise estimate is too high, then speech will be destroyed possibly resulting in loss of intelligibility.

The other one is related to data analysis. As measuring a certain physical quan-

tity, the relatively small number of abnormal values or outliers, hereinafter referred to as outliers, are included in the normal measured values that contain measurement noise. It is one of the frequently occurred works in science and engineering to estimate the true statistical parameters of the physical quantity from these measured values including outliers. Because the measurement noise follows a Gaussian distribution with mean zero in general, all the samples form the major cluster are Gaussian-distributed around the true value. The problem mentioned 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.

In order to solve the problems above, the research focuses on the noise estimation in speech enhancement for speech signal processing and the estimation of major cluster for data analysis.

1.1 Noise Estimation in Speech-Enhancement

For most speech-enhancement algorithms, an estimate for the noise spectrum is assumed to be available. Such an estimate is crucially important for speech-enhancement performance. The noise estimate strongly affects the enhanced signal quality. Annoying residual noise will be audible if the noise estimate is too low. Alternatively, if the noise estimate is too high, then speech will be destroyed possibly resulting in loss of intelligibility. The simplest approach estimates and updates the noise spectrum during the silent segments (e.g., during pauses) of the signal using a voice-activity detection (VAD) algorithm [1].

Without using a speech activity detector [1], several noise-estimation algorithms have been proposed for speech enhancement applications. Martin [2] proposed a method for estimating the noise power spectral density based on tracking the minimum of recursively smoothed periodograms over a finite window from the noisy speech. Doblinger [3] updated the noise estimate by tracking the minimum of the noisy speech continuously in each frequency bin. Hirsch and Ehrlicher [4] improved the noise estimate by comparing the noisy speech power spectrum to a prior noise estimate. Cohen [5] proposed a minima-controlled recursive algorithm (MCRA), which updates the noise estimate by tracking the noise-only regions of the noisy speech spectrum. In the improved MCRA approach (Cohen [6]), a different method was proposed to track the noise-only regions of the spectrum based on the estimated speech-presence probability. This probability is controlled by the minima.

In brief, the previously described noise-estimation algorithms [2]–[6] developed

for speech enhancement algorithms are all based on the Minimum Statistics quoted in Martin's method [2]. Although the smoothing factor and the window length strongly influence the noise estimation performance using Martin's method [2], no good criteria exist to ascertain the optimal value of the parameters. Therefore, the estimation accuracy of these methods [2]–[6] is poor.

Yokota and Ye [7] proposed the radical root, or r-th root, transform of the power spectrum series such that the transformed series follow a quasi-Gaussian distribution. Furthermore, a power spectrum estimation method robust for sudden noise was proposed. Considering using the speech and the background noise instead of the abrupt noise and the stationary Gaussian stochastic process, respectively, we can estimate the background noise power spectrum for speech enhancement by applying Yokota's method in principle. However, in Yokota's method, the proportion of the time of the abrupt noise to the whole signal is small, whereas in the case of speech enhancement, the speech segment is usually much longer. When we estimate the noise power spectrum, it is strongly affected by the speech power spectrum. Therefore, it is impossible to apply Yokota's method under this condition directly.

In this paper, we approximate the probability density distribution of the speech power spectrum series with the super-Gaussian distribution and calculate the range of the necessary optimal parameter r of the radical root transformation to make the probability density distribution normal distributed. By using the optimal parameter r, we present that both the power spectrum series of the speech and the background noise can be quasi-normalized. Therefore, after applying the radical root transformation to the mixed power spectrum series consisting of the speech and the background noise, the mixed power spectrum series follow a mixed Gaussian distribution. It is possible to estimate the parameters of each distribution by using the EM algorithm. This is proposed as a noise power spectrum estimation method aiming at speech enhancement. Practical experiments conducted with different noise types confirm the validity of the proposed method.

1.2 Estimating the Major Cluster by Mean-Shift with Updating Kernel

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 [8] first proposed the mean-shift method, which was subsequently generalized by Cheng [9]. 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 [10, 11]. 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 [12].

Usually in kernel density estimation, the bandwidth is determined such that the difference between the true distribution and the estimated distribution is minimized [13–15]. In mean-shift, because the normalized norm affects the convergence speed, a method for determining the bandwidth is proposed for the isotropic kernel [16] and anisotropic kernel [17] 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 [17, 18]. Moreover, mean-shift with bandwidth that varies depending on the coordinate in data space was proposed [16, 18]. 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 [19] proved its convergence by further imposing some commonly acceptable conditions. Ghassabeh [20] modified the mean-shift to guarantee its convergence. Although the mean-shift has been used widely in many applications [21–23], 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.

Chapter 2
Noise Estimation for Speech Enhancement Based on Quasi-Gaussian Distributed Power Spectrum Series by Radical Root Transformation

2.1 Quasi-Gaussian distributed power spectrum series of noise and speech

2.1.1 Power spectrum series

Considering a stochastic process $x(t)$, the short-time Fourier spectrum centering on the 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. 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.

2.1.2 Probability density distribution of the power spectrum series

Power spectrum series of the noise

The spectrum series $X_f(t)$ is a complex stochastic series. If $x(t)$ is a Gaussian stochastic process, then the real part $\mathrm{Re}[X_f(t)]$ and the imaginary part $\mathrm{Im}[X_f(t)]$ will both follow a Gaussian distribution with a zero mean and equal variance. In other words, $X_f(t)$ follows a two-dimensional Gaussian distribution centered at $0+0i$ with a covariance matrix $\sigma^2 I$ on the complex plane. Here I denotes the identity matrix. Also, σ^2 denotes the variance of the real part $\mathrm{Re}[X_f(t)]$; the imaginary part is $\mathrm{Im}[X_f(t)]$.

To confirm that the noise power spectrum series $X_f(t)$ of real environment noise actually follows two-dimensional Gaussian distribution in the complex plane, we make pulse code modulation (PCM) recordings for the air-conditioning noise and the vacuum cleaner noise. We then calculate $X_f(t)$ with a Hamming window length of 10 ms, achieving a 50% overlap between adjacent frames using short-time Fourier transformation.

Here, the mean vector of $\mathrm{Re}[X_f(t)]$ and $\mathrm{Im}[X_f(t)]$ is defined as $\begin{pmatrix} mean(\mathrm{Re}[X_f(t)]) \\ mean(\mathrm{Im}[X_f(t)]) \end{pmatrix}$. Furthermore, the covariance matrix related to $\mathrm{Re}[X_f(t)]$ and $\mathrm{Im}[X_f(t)]$ is defined as $\begin{pmatrix} cov(\mathrm{Re}[X_f(t)], \mathrm{Re}[X_f(t)]) & cov(\mathrm{Re}[X_f(t)], \mathrm{Im}[X_f(t)]) \\ cov(\mathrm{Im}[X_f(t)], \mathrm{Re}[X_f(t)]) & cov(\mathrm{Im}[X_f(t)], \mathrm{Im}[X_f(t)]) \end{pmatrix}$. The mean vectors for air-conditioning noise and vacuum cleaner noise are, respectively, $\begin{pmatrix} 0.0001 \\ 0.0000 \end{pmatrix}$ and $\begin{pmatrix} 0.0026 \\ 0.0034 \end{pmatrix}$. All mean vectors are close to the zero vector, which implies that the mean of $X_f(t)$ on the complex plane is approximately $0 + 0i$. The covariance matrices of $\mathrm{Re}[X_f(t)]$ and $\mathrm{Im}[X_f(t)]$ for air-conditioning noise and vacuum cleaner noise are, respectively, $\begin{pmatrix} 0.0047 & 0.0000 \\ 0.0000 & 0.0047 \end{pmatrix}$ and $\begin{pmatrix} 12.3382 & -0.0004 \\ -0.0004 & 12.3566 \end{pmatrix}$. The covariances between $\mathrm{Re}[X_f(t)]$ and $\mathrm{Im}[X_f(t)]$ are clearly close to zero. Therefore, real part $\mathrm{Re}[X_f(t)]$ and imaginary part $\mathrm{Im}[X_f(t)]$ are non-correlated. We plot the two-dimensional Gaussian distribution consisting of the mean vectors and the covariance matrices compared to the actual histogram of spectrum series $X_f(t)$. Scattergrams of spectrum series $X_f(t)$ for the noises of two types are depicted in Fig. 2.1(a) and Fig. 2.1(b). Fig. 2.1(c) and Fig. 2.1(d) present histograms of spectrum series $X_f(t)$.

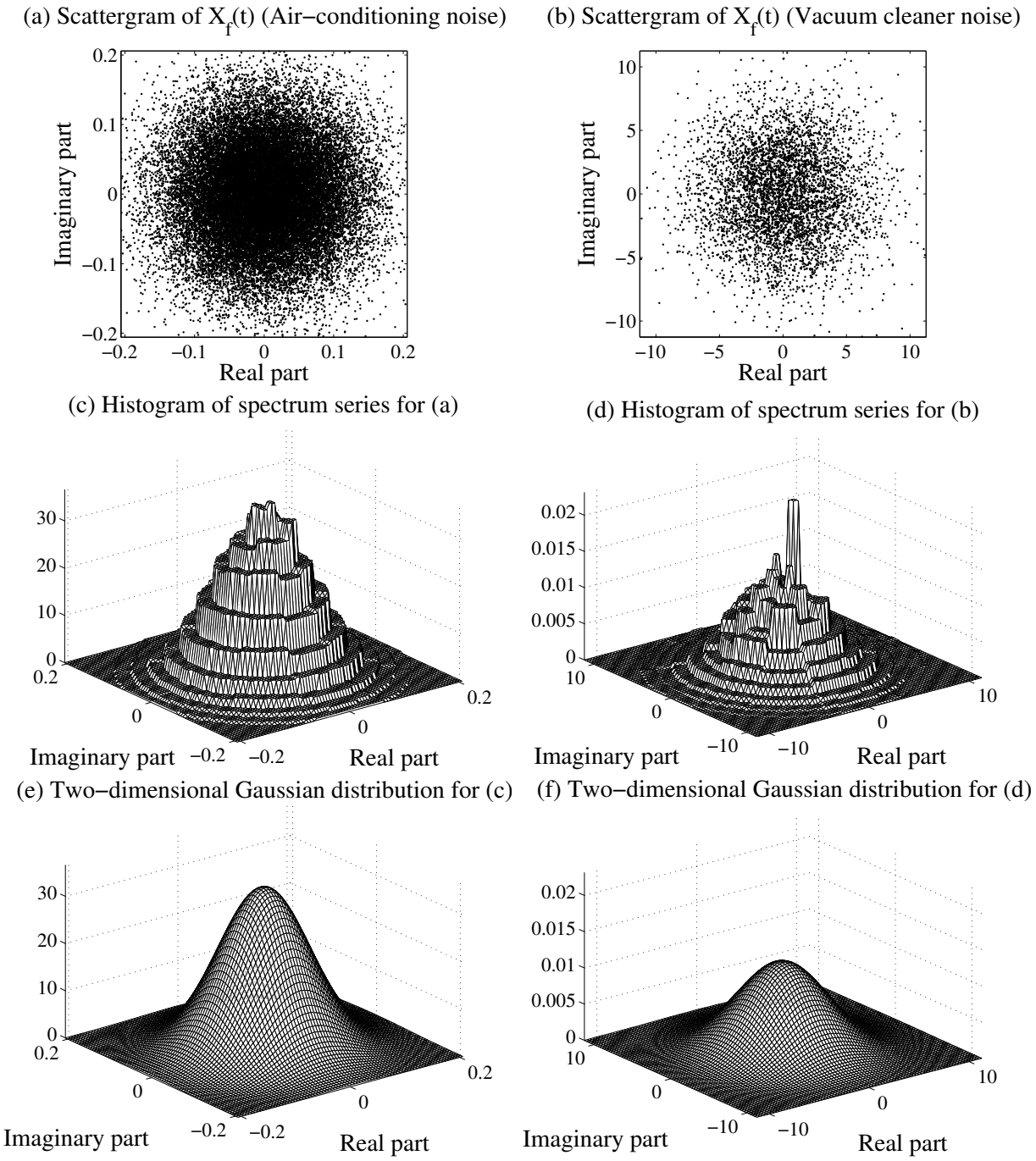


Figure 2.1: Scattergram of spectrum series $X_f(t)$ at $f = 512$Hz on the complex plane and result fitted with two-dimensional Gaussian distribution: (a) Scattergram related to to air-conditioning noise, (b) Scattergram related to vacuum cleaner noise, (c) Histogram for (a), (d) Histogram for (b), (e) Two-dimensional distribution compared to (c), and (f) Two-dimensional distribution compared to (d).

Using the two-dimensional distribution to fit Fig. 2.1(c) and Fig. 2.1(d), the fitting results are presented in Fig. 2.1(e) and Fig. 2.1(f). Comparisons between Figs. 2.1(c) and 2.1(d) and Figs. 2.1(e) and 2.1(f), respectively reveal that the actual noise power spectrum series follows the two-dimensional Gaussian distribution.

Normalizing the variance σ^2 to one, the power spectrum series $P_f(t) = |X_f(t)|^2 = \mathrm{Re}[X_f(t)]^2 + \mathrm{Im}[X_f(t)]^2$ follows a χ^2 distribution with the degree of freedom $k = 2$. In general, the probability density distribution of the χ^2 distribution with the degree of freedom k is expressed as

$$p(x;k) = \frac{(1/2)^{k/2}}{\Gamma(k/2)} x^{k/2-1} e^{-x/2}, \tag{2.1}$$

where $\Gamma(\cdot)$ is the gamma function, and the expectation of this distribution is equal to the degrees of freedom k. Then the probability density distribution of the power spectrum series $P_f(t)$ is

$$p(x;2) = \frac{1}{2} e^{-\frac{1}{2}x}, \tag{2.2}$$

as the degree of freedom $k = 2$.

Power spectrum series of the speech

The speech signal is a non-stationary signal that the spectrum is changing markedly over time. Lotter and Vary [25] proposed a spectral amplitude estimator with a parametric super-Gaussian speech model for approximating the probability density distribution of the real speech spectral amplitudes $A_f(t) = |X_f(t)|$. The power spectrum series $P_f(t)$ will be determined as $P_f(t) = A_f^2(t)$. The Probability Density Function (PDF) $p(a)$ of the speech spectral amplitudes $A_f(t)$ can be approximated by the following parametric function in super-Gaussian speech model as

$$p\,(a) = \frac{\mu^v}{\Gamma(v)} \frac{a^{v-1}}{\sigma_s^v} \exp\left(-\mu \frac{a}{\sigma_s}\right), \tag{2.3}$$

where σ_S^2 denotes the variance of the PDF. Parameters μ and v determine the PDF shape. μ gives the slope of the decay to higher values, whereas v strongly influences the value of the PDF at small values. For brevity, let $\beta = \frac{u}{\sigma_s}$. Thereby, Eq. (2.3) can be simplified as

$$p\,(a) = \frac{\beta^v}{\Gamma(v)} a^{v-1} \exp(-\beta a)\,. \tag{2.4}$$

To confirm that it is always possible to approximate the real PDF of the speech spectral amplitudes $A_f(t)$ with Eq. (2.4), the following experiments were conducted.

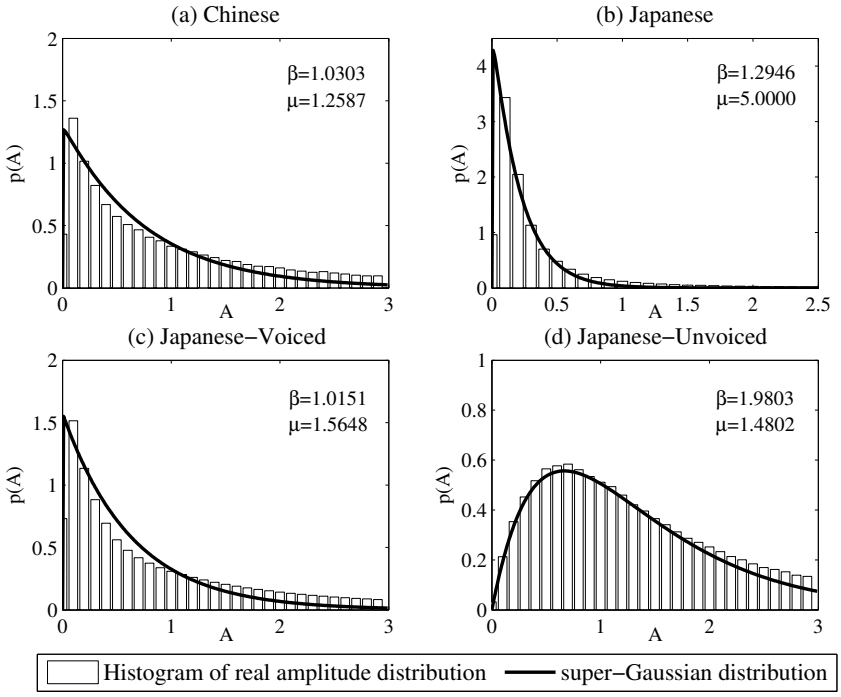


Figure 2.2: Histogram of speech spectral amplitudes and fitted approximation by super-Gaussian distribution: (a) Chinese speech signal, (b) Japanese speech signal, (c) Japanese voiced part, and (d) Japanese unvoiced part.

For these experiments, we used the voice files of the standard language with about one hour speech related to both Chinese and Japanese including in the database from textbooks with CDs [26]–[27]. To estimate the PDF of the speech spectral amplitudes $A_f(t)$, we calculate $A_f(t)$ with a Hamming window length of 10 ms, achieving a 50% overlap between adjacent frames by short-time Fourier transformation. Then we approximate the real PDF of the speech spectral amplitudes $A_f(t)$ with Eq. (2.4).

Fig. 2.2 shows the goodness of the approximation to the real PDF of speech spectral amplitudes. Fig. 2.2(a) portrays a histogram of Chinese speech spectral amplitudes and the best fitted approximation by Eq. (2.4) with the optimal parameter set $(\beta, \mu) = (1.0303, 1.2587)$. Speech is divisible into numerous voiced and unvoiced regions. The classification of speech signals as voiced–unvoiced provides a preliminary acoustic segmentation for speech processing applications such as speech synthesis, speech enhancement, and speech recognition. The Japanese voiced–unvoiced part is chosen here to verify the approximation performance for special speech signals. Similarly, Figs. 2.2(b), 2.2(c), and 2.2(d) present corresponding results for the Japanese speech signal, Japanese voiced part, and Japanese unvoiced part. Fig. 2.2 implies that the super-Gaussian distribution represented by Eq. (2.4) always provides an extremely good approximation for the real speech spectrum amplitude series.

2.1.3 Box–Cox transformation [24] and radical root transformation [7]

Noise power spectral series and speech spectrum series are not normally distributed. Therefore, it is difficult to distinguish noise power spectral series from speech spectrum series. To make the two distributions more normal distribution-like, we use radical root transformation [7], which is based on the framework of the Box–Cox transformation [24].

The one-parameter Box–Cox transformation [24] is defined as

$$f(x) = \begin{cases} \dfrac{x^{\lambda}-1}{\lambda} & \lambda \neq 0, \\ \log(x) & \lambda = 0. \end{cases} \tag{2.5}$$

This transformation holds for $x > 0$. Under the situation of $\lambda \neq 0$, when λ tends to zero, the limit of the formula above in Eq. (2.5) is equal to the following in Eq. (2.5). For this reason, such a form of formula exists in Box–Cox transformation

under the situation of $\lambda \neq 0$. However, the purpose of this study is to make χ^2 distribution and super-Gaussian distribution normally distributed. Therefore, it is unnecessary to consider the situation of $\lambda = 0$. For such a purpose, the Box–Cox transformation is fundamentally equivalent to $f(x) = x^{\lambda}$, although the average and the variance differ. In this formula, using a substitution $r \equiv \frac{1}{\lambda}$, one can obtain the radical root transformation [7] as

$$f(x) = x^{\frac{1}{r}}, \qquad 0 < r < \infty. \tag{2.6}$$

The radical root transformation [7] can be regarded as the brief modified form of the Box–Cox transformation [24].

2.1.4 Probability density distribution after radical root transformation [7]

Generally, by applying the probability density distribution $p(x)$ of a random variable x, the probability density distribution $q(y)$ after transformation $y = f(x)$ is

$$q(y) = p(g(y))\frac{dx}{dy}, \tag{2.7}$$

based on the transform relation $q(y)dy = p(x)dx$. In Eq. (2.7), $x = g(y)$ is the inverse function of $y = f(x)$.

Power spectrum series of the noise

Considering the noise power spectrum series, the probability density distribution $p(x)$ is represented by Eq. (2.2). After substituting $f(x) = x^{1/r}$ in Eq. (2.7), the transformed probability density distribution $q(y)$ is given as

$$q(y;r) = \frac{1}{2}e^{-\frac{1}{2}y^r} r y^{r-1}. \tag{2.8}$$

The expectation $m'_y(r)$ and variance $\sigma'^2_y(r)$ of the distribution $q(y;r)$ are, respectively, [7],

$$m'_y(r) = 2^{\frac{1}{r}}\Gamma\left(\frac{r+1}{r}\right), \tag{2.9}$$

$$\sigma'^2_y(r) = -2^{\frac{r+2}{r}}\Gamma\left(\frac{r+1}{r}\right)^2 + 4^{\frac{1}{r}}\Gamma\left(\frac{r+1}{r}\right)^2 + 4^{\frac{1}{r}}\Gamma\left(\frac{r+2}{r}\right). \tag{2.10}$$

Eq. (2.9) and Eq. (2.10) are derived under circumstances in which the expectation of the χ^2 distribution matches $k = 2$ degrees of freedom. When the expectation of the χ^2 distribution is m_x, the expectation $m_y(r)$ and variance $\sigma_y^2(r)$ of the transformed distribution $q(y;r)$ can also be derived [7]. Furthermore, the relation between the expectations before and after radical root transformation are derived [7] as

$$m_x = \left(\frac{m_y(r)}{\Gamma\left(\frac{r+1}{r}\right)} \right)^r . \tag{2.11}$$

Power spectrum series of the speech

Because the speech spectral amplitude series follow a super-Gaussian distribution approximately, the probability distribution $p(a)$ is represented by Eq. (2.4). After substituting $f(x) = x^{1/r}$ in Eq. (2.7), the transformed probability density distribution $q(y)$ is

$$q(y;r) = \frac{r\beta^v}{\Gamma(v)} y^{vr-1} \exp(-\beta y^r) . \tag{2.12}$$

The expectation $m'_y(r)$ and variance $\sigma'^2_y(r)$ of the transformed distribution $q(y;r)$ are

$$\begin{aligned} m'_y(r) &= \int_0^\infty y q(y;r) dy \\ &= \frac{1}{\beta^{\frac{1}{r}}} \frac{\Gamma\left(v+\frac{1}{r}\right)}{\Gamma(v)}, \end{aligned} \tag{2.13}$$

$$\begin{aligned} \sigma'^2_y(r) &= \int_0^\infty (y - m_y(r))^2 q(y;r) dy \\ &= \int_0^\infty y^2 q(y;r) dy - (\int_0^\infty y q(y;r) dy)^2 \\ &= \frac{1}{\beta^{\frac{2}{r}}} \frac{\Gamma(v+\frac{2}{r})}{\Gamma(v)} - (\frac{1}{\beta^{\frac{1}{r}}} \frac{\Gamma(v+\frac{1}{r})}{\Gamma(v)})^2. \end{aligned} \tag{2.14}$$

2.1.5 Evaluation of the normality about the power spectrum series after radical root transformation [7]

Power spectrum series of the noise

The expectation is equal to the mode in normal distribution. Under conditions where the mode and expectation of the converted distribution are equal, the optimal value of parameter r in radical root transformation can be obtained approximately using

numerical analysis [7]. The first derivative of the probability density distribution $q(y;r)$ represented by Eq. (2.8) is

$$q'(y;r) = -\frac{r}{4}e^{-\frac{1}{2}y^r}(ry^{2r-2} - 2ry^{r-2} + 2y^{r-2}). \quad (2.15)$$

When $q'(y;r) = 0$, the mode of the distribution is inferred as

$$Mode_y(r) = \left(\frac{2r-2}{r}\right)^{\frac{1}{r}}. \quad (2.16)$$

If the mode is equal to the expectation $Mode_y(r) = m'_y(r)$, then the optimal value of parameter r^* is determined [7] as

$$r^* \simeq 3.312. \quad (2.17)$$

Power spectrum series of speech

In probability theory and information theory, the Kullback–Leibler divergence is a non-symmetric measure of the difference between two probability distributions. The KL divergence between the probability density distribution $q(y;r)$ shown in Eq. (2.12). The Gaussian distribution $p_g(y)$ is

$$D_{KL(p_g||q)} = \int_{-\infty}^{\infty} p_g(y) \log \frac{p_g(y)}{q(y;r)} dy, \quad (2.18)$$

where the mean of the Gaussian distribution $p_g(y)$ is $m'_y(r)$ represented by Eq. (2.13), and the variance is $\sigma'^2_y(r)$ represented by Eq. (2.14). One can ascertain the optimal value of parameter r under the condition that the KL divergence between the two distributions reaches the minimum.

In Section 2.2.2, we have used a super-Gaussian distribution to approximate the real PDF of speech spectral amplitudes obtained from the Chinese speech signal, Japanese speech signal, Japanese voiced part, and Japanese unvoiced part. The fitted parameter sets are $(\beta, v) = (1.0303, 1.2587)$, $(\beta, v) = (1.2946, 5.0000)$, $(\beta, v) = (1.0151, 1.5648)$, and $(\beta, v) = (1.9803, 1.4802)$. Because of the test of the Chinese speech signal, Japanese speech signal, Japanese voiced part, and Japanese unvoiced part, the range of the choice of the parameter sets (β, v) can be ascertained. This study only examines the situation under the condition $(0 < \beta < 5, 0 < v < 5)$. We can obtain the numerical solution by setting the condition. Each parameter set (β, v) gets one optimal value of parameter r.

Using r-th radical root transformation, the range of KL divergences between the transformed probability density distribution and the corresponding Gaussian distribution is $[5.1306 * 10^{-5}, 0.0406]$. The KL divergences between them are extremely small, which implies that the probability density distribution transformed from super-Gaussian distribution obeys the Gaussian distribution.

To present the performance of r-th radical root transformation, we apply radical root transformation to the super-Gaussian distributions corresponding to 16 parameter sets (β, v). Then we plot the transformed probability distribution and the corresponding Gaussian distribution. The top panel in Fig. 2.3 presents the relation between the transformed probability density distribution corresponding the optimal transformation parameter r and the Gaussian distribution. That relation implies that the super-Gaussian distribution after r-th radical root transformation can be quasi-Gaussian distributed.

The optimal value of transformation parameter $r^* = 3.312$ related to the noise power spectrum series is discussed in Section 2.5.1. Because of $P_f(t) = A_f^2(t)$, applying radical root transformation to the speech spectral amplitude series

$$P_f^{\frac{1}{r^*}}(t) \quad = \quad A_f^{\frac{1}{r^*/2}}(t), \tag{2.19}$$

the transformation parameter is $r = r^*/2 = 1.656$. Using radical root transformation related to parameter $r = 1.656$, the range of KL divergences between the transformed probability density distribution and the corresponding Gaussian distribution is $[0.0039, 0.1496]$. The KL divergences are also small. Therefore, for all the parameter sets under the condition $(0 < \beta < 5, 0 < v < 5)$, using the common transformation parameter $r = 1.656$, the transformed probability density distribution can also be converted approximately into the corresponding Gaussian distribution.

To confirm that the super-Gaussian distributions after the radical root transformation related to common transformation parameter $r = 1.656$ are quasi-Gaussian distributed, we plot the super-Gaussian distributions after radical root transformation related to the same 16 parameter sets (β, v) used before, as well as the corresponding Gaussian distribution. The bottom panel of Fig. 2.3 presents the relation between the transformed probability density distribution corresponding the common transformation parameter $r = 1.656$ and the Gaussian distribution. Compared to the parameter sets (β, v) obtained from different speech signals, it indicates that the speech power spectrum series after radical root transformation can be quasi-Gaussian distributed approximately.

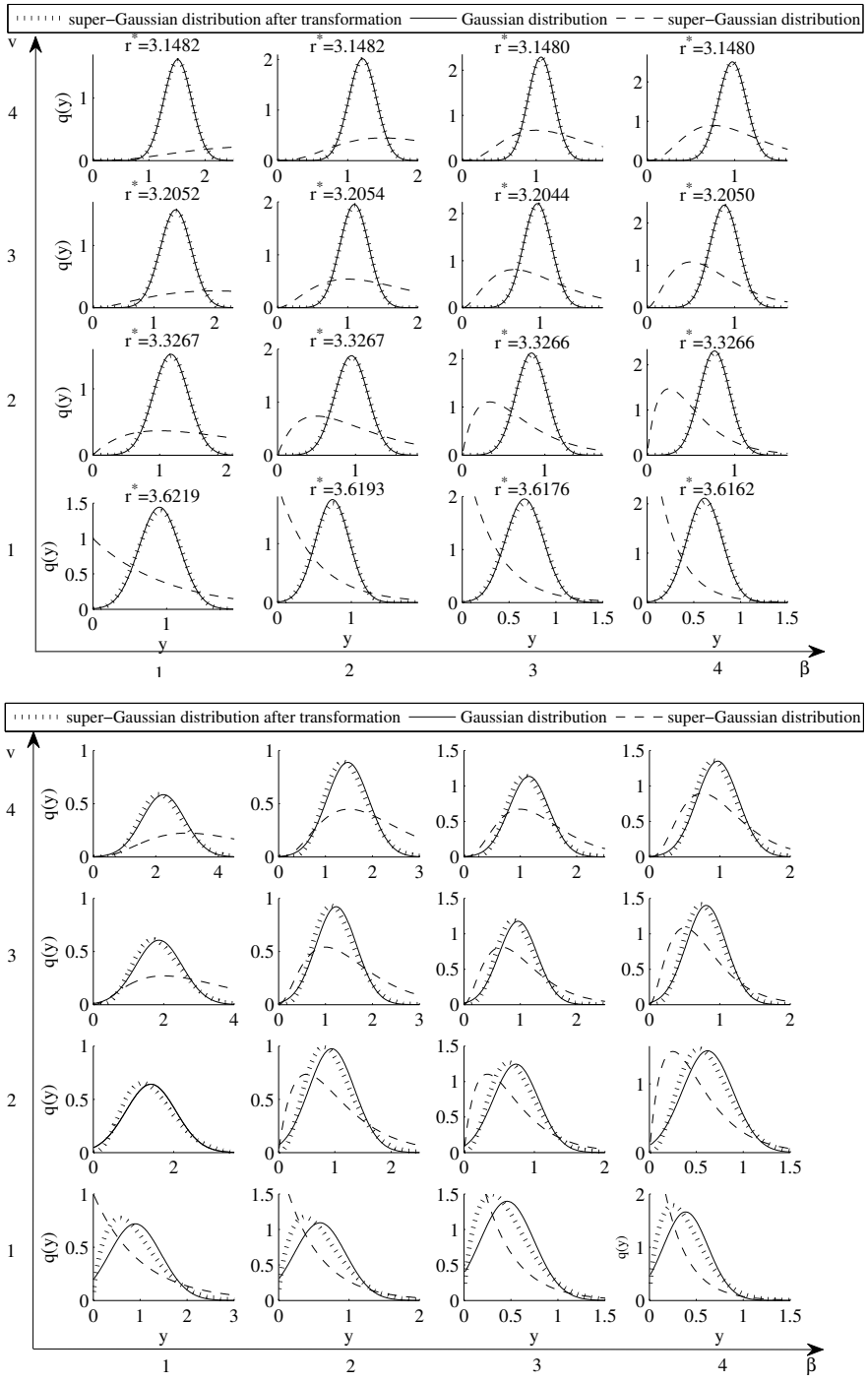


Figure 2.3: Original super-Gaussian distribution and comparison between the speech amplitude series after radical root transformation with optimal parameter r and Gaussian distribution according to a different parameter set (β, v): top panel, r; bottom panel, $r = 1.656$.

2.2 Proposed noise estimation algorithm based on the quasi-Gaussian distributed power spectrum series

As discussed earlier, using the radical root transformation with the same parameter $r^* = 3.312$, the transformed noise power spectrum series and the transformed speech power spectrum series can be quasi-Gaussian distributed. The proposed method relies on one basic assumption: The noise power spectrum series and the speech power spectrum series are independent in a noisy speech signal. That is,

$$P_f(t) = |X_f(t)|^2 + |D_f(t)|^2, \tag{2.20}$$

where $P_f(t)$, $|X_f(t)|^2$ and $|D_f(t)|^2$ respectively denote the power spectrum of noisy speech, clean speech, and noise. In addition, f and t respectively stand for the frequency index and time index. Therefore, the mixed power spectrum series after the radical root transformation follows a two-dimensional Gaussian mixture distribution. The parameters of the Gaussian mixture model can be computed easily using the EM algorithm. Therefore, we can separate the transformed noise power spectrum series from the total transformed power series. The following concludes the process of the proposed noise estimation algorithm:

(1) (2-1) (1) Obtain the power spectrogram $P(t, f)$ from the noisy speech signal. We choose a PCM recording of noisy speech signal for analysis and compute the spectrogram with a Hamming window length of 10 ms, achieving a 50% overlap between adjacent frames by short-time Fourier transformation. Fig. 2.4(a) presents an example of a noisy speech 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 $r^* = 3.312$, and obtain the new quasi-Gaussian distributed power spectrum series $P_f^{1/r^*}(t)$. Fig. 2.4(b) depicts a histogram of the power spectrum series at $f = 512$Hz before the transformation.

(2–2) Compute the Gaussian mixture model parameters using the EM algorithm. Consequently, the weights, the means, and the variances of two Gaussian distribution in the Gaussian mixture model are obtained. Here, we assume that the noise is stationary and that the sound is intermittent. Based on this assumption, the two Gaussian distributions appear because of the existence of noise-only segment and noise-speech mixture segment (not speech-only segment). The power of the noise-

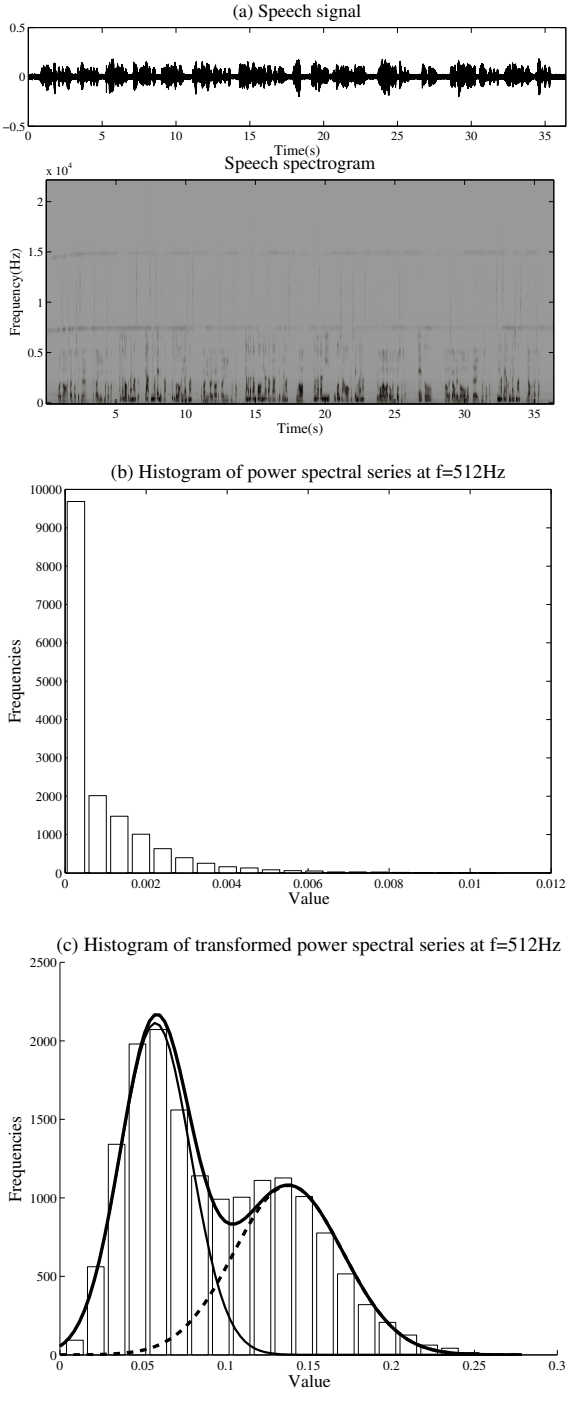


Figure 2.4: (a) A noisy speech signal for analysis and its corresponding spectrogram. (b) Histogram of the power spectrum series of the noisy speech signal at f=512 Hz. (c) Histogram of the transformed power spectrum series of the noisy speech signal at f=512 Hz and the Gaussian mixture model.

only segment is markedly smaller than the power of noise-speech mixture segment. Therefore, the Gaussian distribution with smaller mean in the Gaussian mixture model corresponds to noise. Then put the smaller mean as the corresponding time average value $P_{noise}(f)$ of the noise power spectrum series. Fig. 2.4(c) plots a histogram of the transformed power spectrum series at $f = 512$Hz and the related Gaussian mixture model.

(2–3) According to Eq. (2.11), compute the time average value $P(f)$ of the noise power spectrum series from the time average value $P_{noise}(f)$ as

$$P(f) = \left(\frac{P_{noise}(f)}{\Gamma\left(\frac{r^*+1}{r^*}\right)} \right)^{r^*} . \tag{2.21}$$

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

2.3 Experimental results

2.3.1 Typical conventional noise estimation: Martin's minimum trackingalgorithm [2]

Martin's method [2] is based on minimum statistics and smoothing of the noisy speech power spectral density. This method relies on two major observations. The first is statistical independence of speech and noise represented by Eq. (2.20). The second is that the noisy speech power spectrum often becomes equal to the noise power spectrum. This happens during speech pauses and also between words and syllables. Therefore, the estimate of noise power spectral density is obtained by tracking the minimum of the noisy speech in each frequency separately. For searching the minimum, a first-order recursive version of the noisy speech power spectrum series is used as

$$\bar{P}_f(t) = \alpha \bar{P}_f(t-1) + (1-\alpha) P_f(t), \tag{2.22}$$

where α is the smoothing constant. Later, the minimum is tracked for each window as

$$\begin{aligned} \bar{P}_{min}(f,t) = \min[&\bar{P}_f(t), \bar{P}_f(t+1), \\ &..., \bar{P}_f(t+L-1)], \end{aligned} \tag{2.23}$$

where L denotes the window length. This minimum is always smaller than (or in trivial cases equal to) the average value of noise power. Therefore, a bias correction is necessary. Finally, the minimum value is multiplied using a bias correction factor $B_f(t)$ which depended mainly on the variance of the noisy signal. It is given as

$$\sigma_N^2(f,t) = B_f(t)\bar{P}_{min}(f,t). \tag{2.24}$$

When the frequency f is fixed, the estimated noise power spectrum is

$$\hat{P}_f(t) = \frac{1}{T}\sum_{t=0}^{T}\sigma_N^2(f,t), \tag{2.25}$$

where T is the signal length.

2.3.2 Characteristic of the Martin's minimum tracking algorithm

The choice of the window length L is based on the notion that it would encompass at least one silence period of the noisy speech. It can be expected to track at least one frame of the noise-only region. However, no method exists to adjust L based on the speech peak width. Actually, L is chosen as sufficiently large to encompass the broadest peak possible in any speech waveform generally. In contrast, the value of the smoothing constant α strongly affects the noise power spectrum estimation results. Ideally, for better noise tracking, α should be close to zero when speech is present. To date, no appropriate criteria exist to select the optimal parameters for window length L and smoothing constant α.

Although the noise power estimator $\sigma_N^2(f,t)$ is amended by the bias correction factor $B_f(t)$, it is still not an unbiased estimator. Noise power estimation is still smaller than the actual mean noise power. To demonstrate the performance of Martin's method [2], we make PCM recordings in practical for an ambient noise and a noisy speech signal under this noise. We then use Martin's method [2] to obtain the noise power spectrum. Fig. 2.5 presents the power spectrum of noisy speech and the estimated noise power spectrum of noisy speech with the parameter set as ($L = 100,\ \alpha = 0.9$). The estimated noise spectrum is compared with the true noise spectrum for the same example. Fig. 2.5(a) portrays the power spectrum and noise estimation $\bar{P}_{min}(f,t)$ for a noisy speech signal. Fig. 2.5(b) presents a comparison between the true noise power σ_f^2 and the estimated noise power $\hat{P}(f)$ for the same noisy speech signal, which implies that Martin's method [2] does not perform well

in practice. The true noise power spectrum is estimated directly from the noise-only segment of the experimental recordings. However, in proposed method and Martin's method, noise power spectrum is estimated from the overall experimental recordings without distinguishing between the noise-only segment and the noise-speech mixture segment. Therefore, neither method requires a voice-activity detection (VAD) algorithm [1].

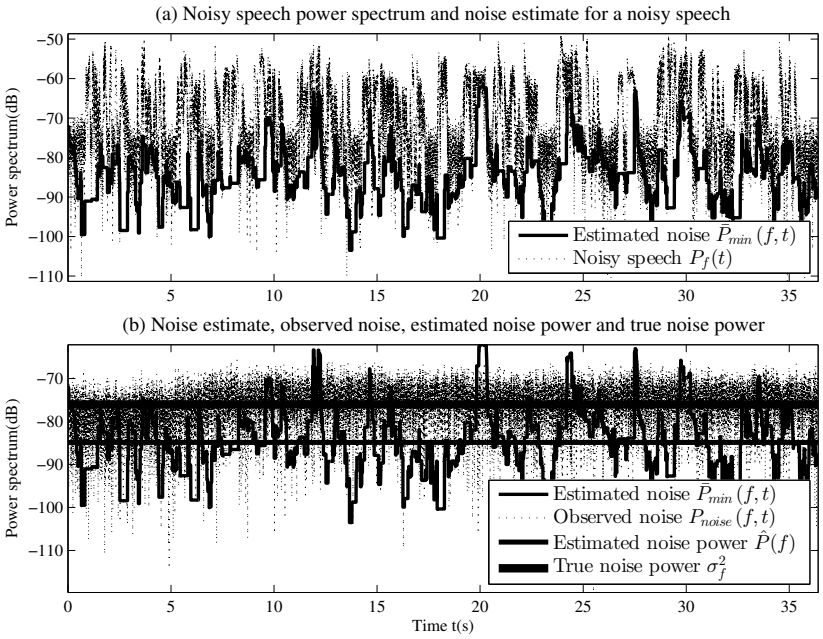


Figure 2.5: Top panel: Plot of noisy speech power spectrum and noise estimate for noisy speech at f=500 Hz. Bottom panel: Plot of true and estimated noise power spectrum for the same noisy speech at f=500 Hz.

2.3.3 Comparison of results obtained using the proposed method and Martin's minimum tracking algorithm

To compare the performance of the proposed method and Martin's method [2], this study uses PCM recordings of dialogues between two people under three noise conditions. Speech signal data in PCM recordings is not compressed, and has no power consumption. Table 2 presents the recording condition. Three noise conditions are air-conditioning noise, vacuum cleaner noise (low gear), and vacuum cleaner noise (high gear). We obtain noise power spectrum density of three noisy speech signals using two methods. The top panel in Fig. 2.6(a) portrays a noisy speech signal for recordings under the air-conditioning noise and the corresponding spectrogram.

Table 2.1: Comparison of the KL divergence between the true noise spectrum and noise estimation.

Method \ Noise condition	Air conditioner	Vacuum cleaner (low gear)	Vacuum cleaner (high gear)
Proposed method	0.1324	0.1105	0.1280
Martin's method	2.8492	2.4134	1.0693

Table 2.2: Recording condition

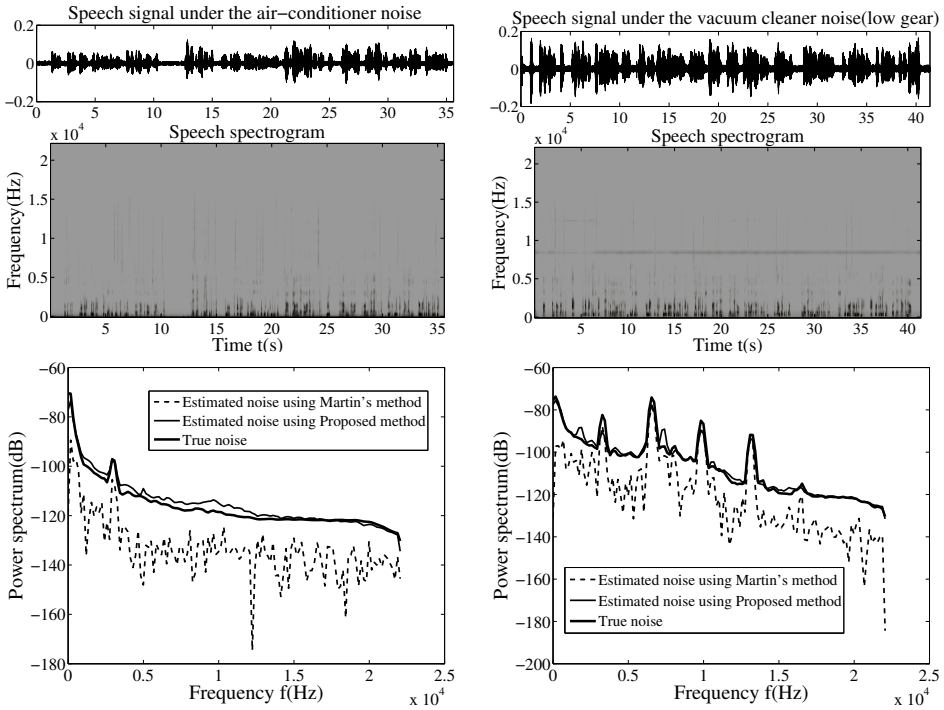
Sampling frequency	44.1 (kHz)
Quantization accuracy	16(bit)

The bottom panel in Fig. 2.6(a) presents a comparison of noise estimation between the proposed method and Martin's method [2] for noisy speech signal under air-conditioning noise. Fig. 2.6(b) presents results related to the noisy speech signal under vacuum cleaner noise (low gear). Fig. 2.6(c) presents corresponding results about for the noisy speech signal under vacuum cleaner noise (high gear).

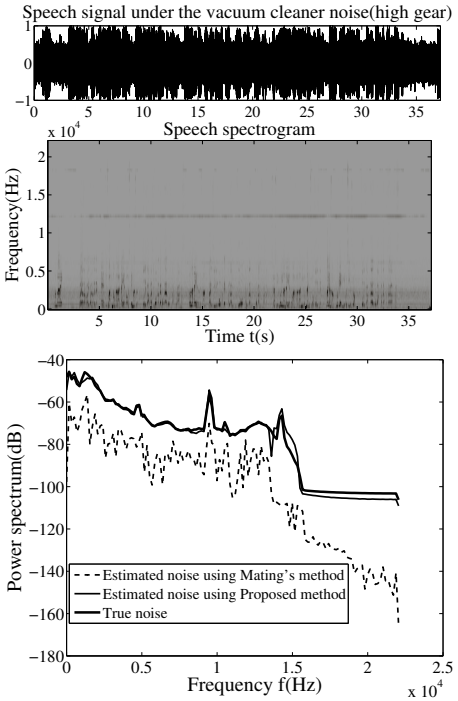
From comparison of the noise power spectrum estimation between proposed method and Martin's method [2], one can find that the proposed method has higher accuracy than Martin's method [2]. After using the proposed method and Martin's method [2], the KL divergences between true noise power spectrum and noise estimation can be computed. Table 1 presents KL divergences between the true noise power spectrum and noise estimation using two methods. Under three noise conditions, the KL divergences between the true noise power spectrum and noise estimation with proposed method are 0.1324, 0.1105, and 0.1280. The KL divergences between the true noise power spectrum and noise estimation with Martin's method [2] are 2.8492, 2.4134 and 1.0693. Therefore, the KL divergences between the true noise power spectrum and noise estimation obtained using the proposed method are much smaller. Therefore, the proposed method has consistent performance compared to Martin's method [2].

2.4 Discussion

The experimentally obtained results demonstrate that the estimation accuracy of proposed method is higher than that of Martin's method under a single noise con-



(a) (b)



(c)

Figure 2.6: Noisy speech signal under the condition noise, the corresponding spectrogram and the comparison between the proposed method and the Martin's method [2] for the noisy speech signal under condition noise: (a) air-conditioning noise, (b) vacuum cleaner noise (low gear), (c) vacuum cleaner noise (high gear).

dition. Moreover, our method can track the fluctuation of power spectrum peak more effectively. Here, we approach the application feasibility of our method under a multiple mixed noise condition. For the proposed method it is fundamentally important, that noise is assumed to be stationary; speech is assumed to be is intermittent. Even if two or more noises exist, if they are all stationary, then their sum can be regarded as a single noise. In this case, because two mixed distributions are observed, one can apply the proposed method directly.

Our method relies strongly on the premise that noise is stationary. Nevertheless, noise might change in real environments. When applying the proposed method to real environments, it seems that the algorithms such as updating the estimate of noise power spectrum by repeating noise estimation at regular intervals is necessary. Because the noise estimation accuracy influences these algorithms strongly, it is difficult to address this point in the present study. However, many speech enhancement systems and noise estimation systems can be achieved as an online system with small delay such as one frame delay. Our method is based on the statistical characteristics of the noise power spectrum and the speech power spectrum in the frequency domain. The temporal variation result cannot be obtained in the noise estimate power spectrum. It requires several seconds for the noisy speech signal. Consequently, it is difficult to use our method as an online system with small delay, although Martin's method is useful as an online system with one frame delay. Therefore, this restriction in the proposed method is notably severe for commercial uses such as mobile phones. In experimental results, the PCM recordings are about 40 seconds. It takes about 6 seconds (Intel(R) Core(TM) i7-3537U CPU @2.00GHz 2.50GHz) to estimate the background noise power spectrum. Therefore, it involves a delay of at least 46 seconds for speech enhancement. However, under the condition that the background noise power spectrum does not change frequently, we can apply our method to speech enhancement after 46 seconds immediately from the beginning until the background noise changes. If we allow the performance degradation due to applying the background noise power spectrum before change, we believe that the proposed method can be applied online as well.

2.5 Conclusion

The study described in this paper has addressed noise estimation for speech enhancement of noisy speech. Based on our previous work [7], we extended the statistical characteristics analysis for a noise spectrum and a speech spectrum. The

noise estimate was obtained in each frequency bin of the noisy speech spectrum. Our experiments assessing various types demonstrate that our method improves the noise estimation accuracy compared to Martin's method. However, unlike other methods, our method is not dependent on fixed parameters such as window length and smoothing constant. The algorithm turns out to be fairly generic. In experiments using different noise types, we did not observe a need to return the algorithm parameters to a single-noise condition. Under multiple noise conditions, we also examined the application feasibility of our method. These results were confirmed by formal experimentation, which indicated superior performance of our proposed method compared to Martin's method.

Chapter 3

Estimating the Major Cluster by Mean-Shift with Updating Kernel

3.1 General Mean-Shift Method

3.1.1 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

$$\hat{\mu}_N \ \leftarrow \ \mu_x. \tag{3.1}$$

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

$$\mu_W \ \leftarrow \ \hat{\mu}_N. \tag{3.2}$$

   The standard deviation σ_W is assigned to be an appropriate size as discussed later in Section 3.1.2.

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

$$a_n = \frac{1}{A} p(x_n; \mu_W, \sigma_W). \tag{3.3}$$

However, A in Equation (3.3) above is a normalization coefficient for which the sum of the weight a_n is equal to 1, as

$$A = \sum_{k=1}^{N} p(x_k; \mu_W, \sigma_W). \tag{3.4}$$

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

$$\mu_x = \sum_{n=1}^{N} a_n x_n. \tag{3.5}$$

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

$$\hat{\mu}_N \leftarrow \mu_x. \tag{3.6}$$

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.

3.1.2 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

$$f(x) = \lim_{h \to 0} \frac{1}{2h} P(x - h < X < x + h). \tag{3.7}$$

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

$$\hat{f}(x) = \frac{1}{2h} \frac{N_x}{N}. \tag{3.8}$$

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

$$\omega(x;h) = \begin{cases} \frac{1}{2h} & |x| < h, \\ 0 & \text{others.} \end{cases} \tag{3.9}$$

The estimator can be expressed as below [28]:

$$\hat{f}(x) = \frac{1}{N}\sum_{i=1}^{N}\omega(x-x_i;h). \tag{3.10}$$

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

$$\int_{-\infty}^{\infty} K(x)dx = 1, \tag{3.11}$$

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

$$\hat{f}(x) = \frac{1}{N}\sum_{i=1}^{N}K(x-x_i;\sigma). \tag{3.12}$$

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 3.3. Figure 3.1a 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 3.2 for 1000 trials. Furthermore, we computed the bias errors using the equation described in

Section 3.3.2. Figure 3.1a 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 3.1b, 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 3.1a. Indeed, the proper pre-set of the kernel bandwidth constitutes an important difficulty.

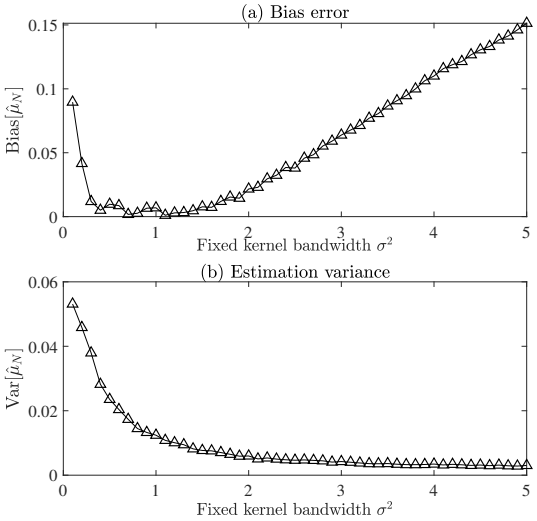


Figure 3.1: Bias error and estimation variance for various fixed kernel bandwidth σ^2 in a general mean-shift method.

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 [29]. 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.

3.2 One-Dimensional Mean-Shift with Updating Kernel

3.2.1 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:

$$\int_{-\infty}^{\infty} p(x;\sigma_W)p(x;\sigma_N)dx = \frac{1}{\sqrt{2\pi}\sqrt{\sigma_W^2+\sigma_N^2}}, \tag{3.13}$$

$$\int_{-\infty}^{\infty} x^2 p(x;\sigma_W)p(x;\sigma_N)dx = \frac{\sigma_W^2\sigma_N^2}{\sqrt{2\pi}(\sigma_W^2+\sigma_N^2)^{\frac{3}{2}}}. \tag{3.14}$$

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

$$\sigma_x^2 = \sum_{n=1}^{N_N} a_n x_n^2. \tag{3.15}$$

Weight a_n is a Gaussian kernel given by Equations (3.3) and (3.4). In addition, N is replaced by N_N.

The expected value of the sample variance σ_x^2 is calculated after substituting Equation (3.3) into Equation (3.15) as

$$E[\sigma_x^2] = E\left[\frac{1}{A}\sum_{n=1}^{N_N} p(x_n;\sigma_W)x_n^2\right]. \tag{3.16}$$

The variance of $\frac{1}{A}$ 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

$$E[\sigma_x^2] \simeq \frac{1}{E[A]}E\left[\sum_{n=1}^{N_N} p(x_n;\sigma_W)x_n^2\right]. \tag{3.17}$$

The approximation is discussed later in Appendix B. Here, we calculate the expected value of A by Equations (3.4) and (3.13) as

$$\begin{aligned} E[A] &= E\left[\sum_{k=1}^{N_N} p(x_k;\sigma_W)\right] \\ &= \sum_{k=1}^{N_N} E[p(x_k;\sigma_W)] \\ &= N_N\int_{-\infty}^{\infty} p(x;\sigma_W)p(x;\sigma_N)dx \\ &= \frac{N_N}{\sqrt{2\pi}\sqrt{\sigma_W^2+\sigma_N^2}}. \end{aligned} \tag{3.18}$$

The expected value of other part becomes

$$\begin{aligned} &E\left[\sum_{n=1}^{N_N} p(x_n;\sigma_W)x_n^2\right] \\ &\quad= \sum_{n=1}^{N_N} E[p(x;\sigma_W)x^2] \\ &\quad= N_N\int_{-\infty}^{\infty} x^2 p(x;\sigma_W)p(x;\sigma_N)dx \\ &\quad= \frac{N_N\sigma_W^2\sigma_N^2}{\sqrt{2\pi}(\sigma_W^2+\sigma_N^2)^{3/2}} \end{aligned} \tag{3.19}$$

according to Equation (3.14). 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

$$E[\sigma_x^2] = \frac{\sigma_W^2 \sigma_N^2}{\sigma_W^2 + \sigma_N^2} \tag{3.20}$$

according to Equations (3.18) and (3.19). Equation (3.20) above can be transformed to

$$\sigma_N^2 = \frac{\sigma_W^2 E[\sigma_x^2]}{\sigma_W^2 - E[\sigma_x^2]}. \tag{3.21}$$

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

$$\hat{\sigma}_N = \sqrt{\frac{\sigma_W^2 \sigma_x^2}{\sigma_W^2 - \sigma_x^2}}. \tag{3.22}$$

In addition, using Equation (3.18), the number N_N of samples belonging to the major cluster can be estimated as

$$\hat{N}_N = A\sqrt{2\pi}\sqrt{\sigma_W^2 + \hat{\sigma}_N^2}. \tag{3.23}$$

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.

3.2.2 Mean-Shift with Updating Kernel

Based on the discussion presented in Section 3.2.1, 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

$$\hat{\mu}_N \leftarrow \mu_x, \tag{3.24}$$

$$\hat{\sigma}_N \leftarrow \sigma_x. \tag{3.25}$$

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:

$$\mu_W \leftarrow \hat{\mu}_N, \tag{3.26}$$

$$\sigma_W \leftarrow r\hat{\sigma}_N. \tag{3.27}$$

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 (3.3) and (3.4). 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:

$$\mu_x = \sum_{n=1}^{N} a_n x_n, \tag{3.28}$$

$$\sigma_x = \sqrt{\sum_{n=1}^{N} a_n (x_n - \mu_x)^2}. \tag{3.29}$$

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:

$$\hat{\mu}_N \leftarrow \mu_x, \tag{3.30}$$

$$\hat{\sigma}_N \leftarrow \sqrt{\frac{\sigma_W^2 \sigma_x^2}{\sigma_W^2 - \sigma_x^2}}, \tag{3.31}$$

$$\hat{N}_N \leftarrow A\sqrt{2\pi}\sqrt{\sigma_W^2 + \hat{\sigma}_N^2}. \tag{3.32}$$

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.

3.3 Numerical Experiment

3.3.1 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 3.2 shows an example of the generated sample in (x_1, x_2) space. Symbol $\bullet$ 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.

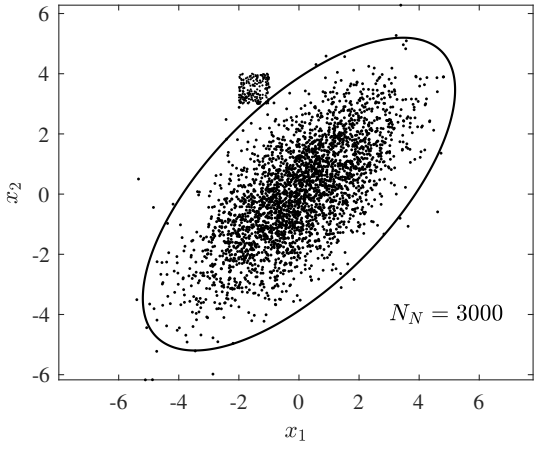


Figure 3.2: Example of a sample set for numerical experiments.

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 3.3 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 3.3, 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 3.2 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.

3.3.2 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

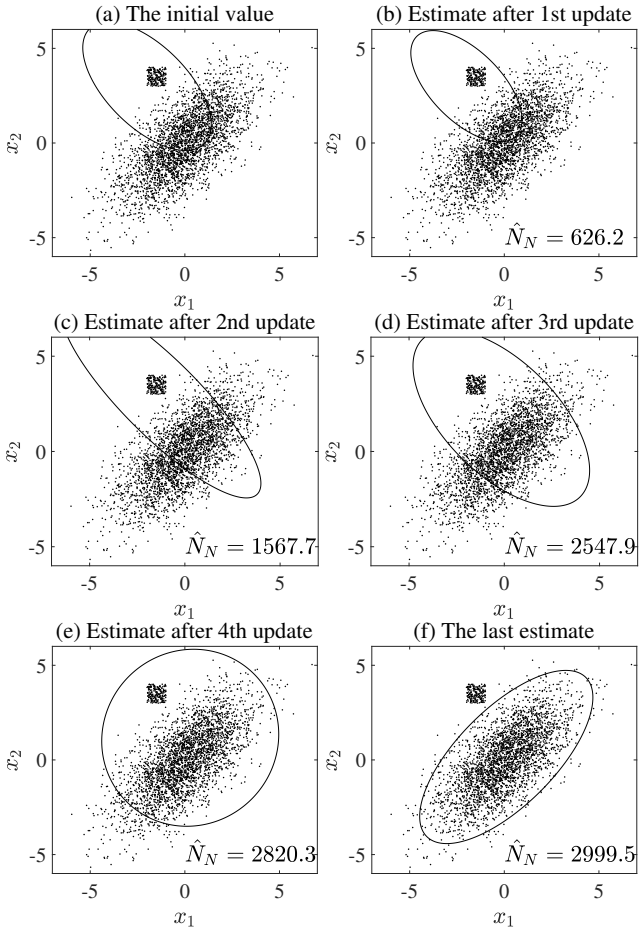


Figure 3.3: Updates of the estimated major cluster.

N_N, N_O, scale factor r, and update ending condition are the same as those described in Section 3.3.1. 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 3.4a–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 $\bigcirc$ corresponds to the proposed method. The symbol $\triangle$ 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 3.4b,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 3.4a 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 $\downarrow$ 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 3.4b,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

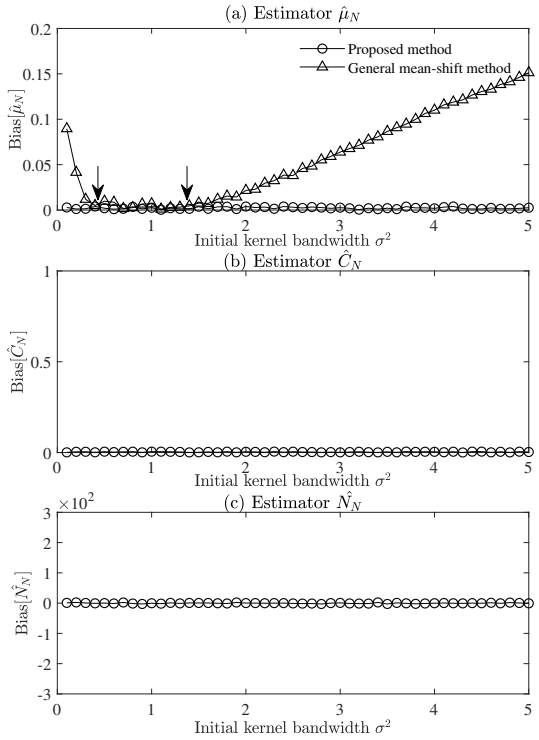


Figure 3.4: Bias errors for various initial kernel bandwidths σ^2 in the proposed method and the general mean-shift method.

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 3.5. 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 3.5 shows that these estimators are unbiased for a finite number of samples.

3.3.3 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

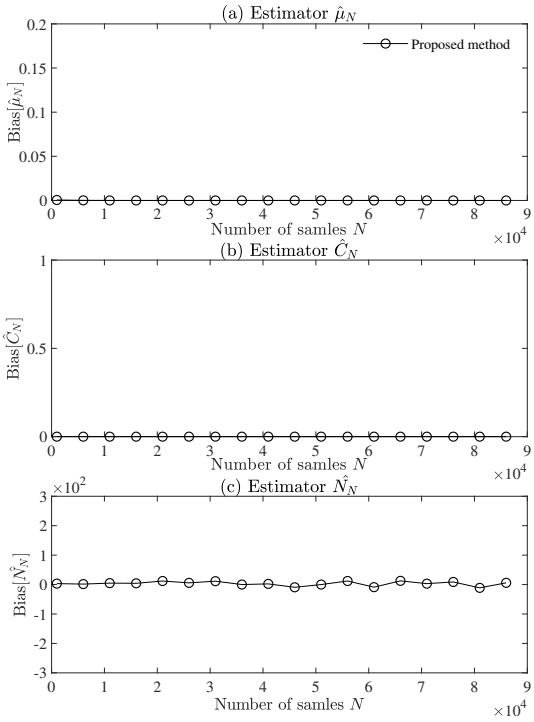


Figure 3.5: Bias errors for various numbers N of samples in the proposed method.

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 3.3.1. As shown in Section 3.3.2, 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 3.6a,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.

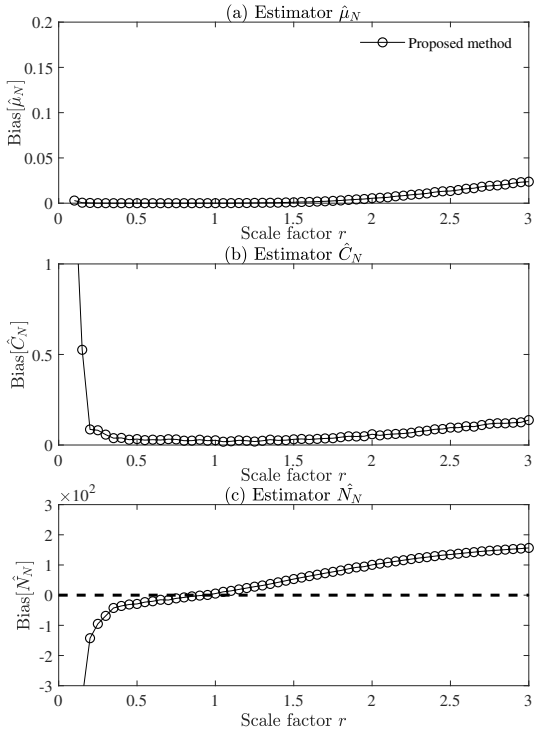


Figure 3.6: Bias errors for various scale factors r in the proposed method.

Figure 3.6 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 3.6. 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 3.6 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 3.4 because the range in which the bias error can be kept low is wide.

3.3.4 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 3.3.2, 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 3.3.1. 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 3.3.3, 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 3.3.2.

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 3.7a–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 3.7b,c.

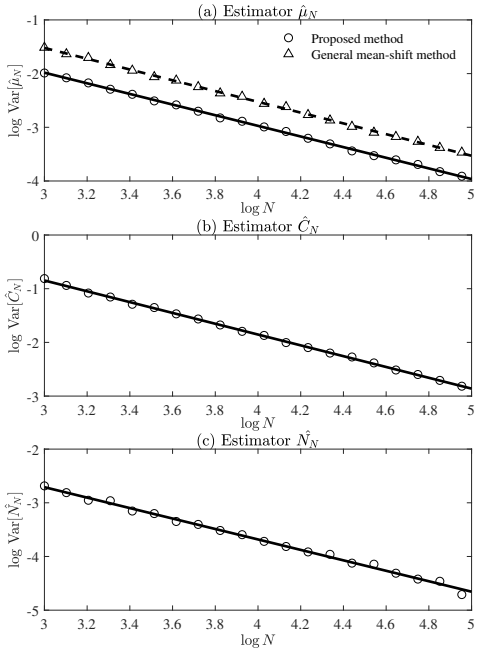


Figure 3.7: Variance of the estimates $\hat{\boldsymbol{\mu}}_N, \hat{\boldsymbol{C}}_N, \hat{N}$ for various numbers N of samples in the proposed method and the general mean-shift method.

From Figure 3.7a–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 3.7a–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.

3.3.5 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 3.3.1. 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 3.8a–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 3.9.

From Figure 3.8a–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 3.6, 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 3.9. 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 3.4. However, the estimation precision increases as σ^2 becomes larger, as shown in Figure 3.9. 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 3.9: its value is shown by a horizontal dotted line in Figure 3.8a. 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.

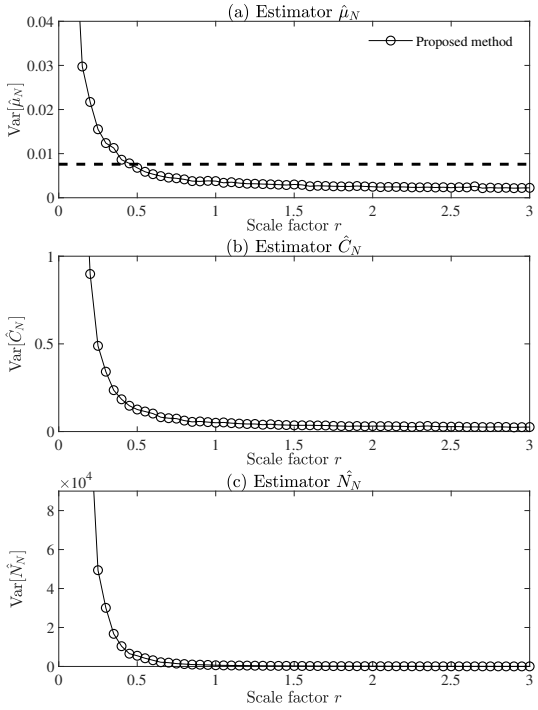


Figure 3.8: Estimating the variance of the estimates $\hat{\boldsymbol{\mu}}_N, \hat{\boldsymbol{C}}_N, \hat{N}$ for various scale factors r of the proposed method.

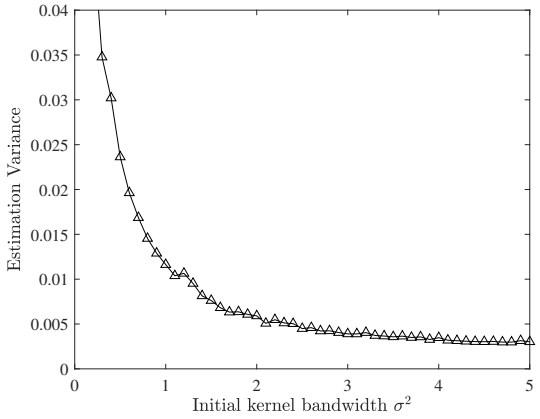


Figure 3.9: Estimating the variance of the estimate $\hat{\boldsymbol{\mu}}_N$ for various kernel bandwidths σ^2 in the general mean-shift method.

3.3.6 Discussion

Numerical experiments in two-dimensions described in Sections 3.3.2, 3.3.3 and 3.3.4 yield results for the major cluster and outlier model shown in Figure 3.2. 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 3.3. 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 [7] 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 [25] 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 distri-

bution or other distributions exactly. Ye and Yokota [31] 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 [7], 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 [31–33] and Thompson Tau test [34] 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 [35–37], the mean and covariance matrix of major cluster can be estimated easily using the Expectation-maximization(EM) algorithm [38, 39]. 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.

3.4 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 [7] 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 [7]:

(1) Obtain power 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 3.10a 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 3.10b 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 [12]. Then, put the mode value as the corresponding time average value $P_{noise}(f)$ of the noise power spectrum series. Figure 3.10c depicts a histogram of the transformed power spectrum series at $f = 512$ Hz, the kernel density estimation [12] 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

$$P(f) = \left(\frac{P_{noise}(f)}{\Gamma\left(\frac{r^*+1}{r^*}\right)} \right)^{r^*} . \tag{3.33}$$

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

In the noise estimation algorithm [7], the mode estimation accuracy directly affects the noise estimation result. As Figure 3.10c shows, kernel density estimation [12] can be replaced by our proposed method for comparison. The proper pre-setting of kernel bandwidth is also important in kernel density estimation [12]. It exhibits a strong influence on the resulting estimate similarly to the general mean-shift method.

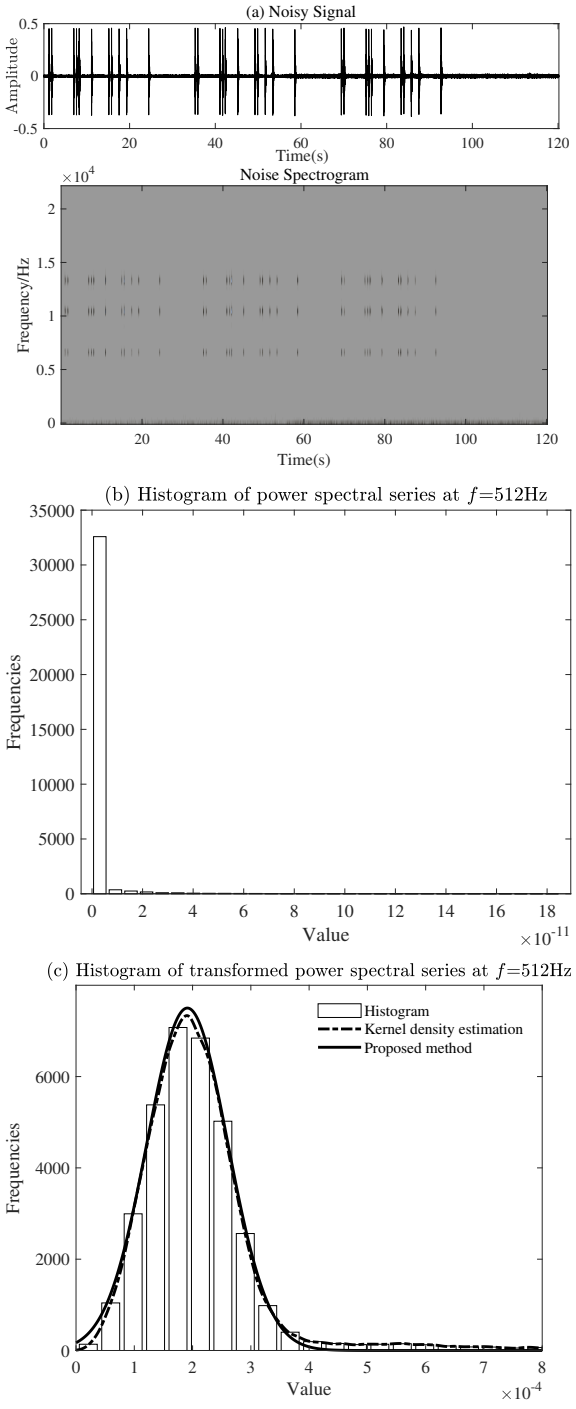


Figure 3.10: (**a**) example of a noisy signal for analysis and the corresponding spectrogram; (**b**) histogram of the power spectrum series of the noisy signal at $f = 512$ Hz; (**c**) histogram of the transformed power spectrum series of the noisy signal at $f = 512$ Hz.

To illustrate its effects, we obtained the noise power spectrum series from the PCM recording, which is shown in Figure 3.10a, for analysis. Figure 3.11 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.

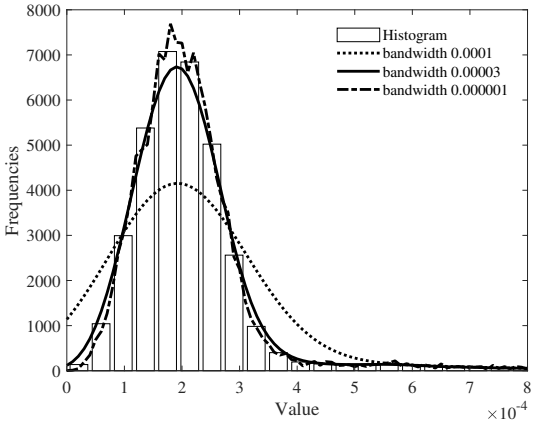


Figure 3.11: Relation between kernel bandwidth and kernel density estimation.

To assess the performance of the proposed method, general mean-shift method, and kernel density estimation for a noise estimation algorithm [7], this study uses PCM recordings of air-conditioning noise with some sudden noise, as shown in Figure 3.10a 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 3.12 presents comparison results for noise estimation using the proposed method and kernel estimation. Here, we preset the kernel bandwidth as 0.0001. As Figure 3.12 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.

3.5 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 estima-

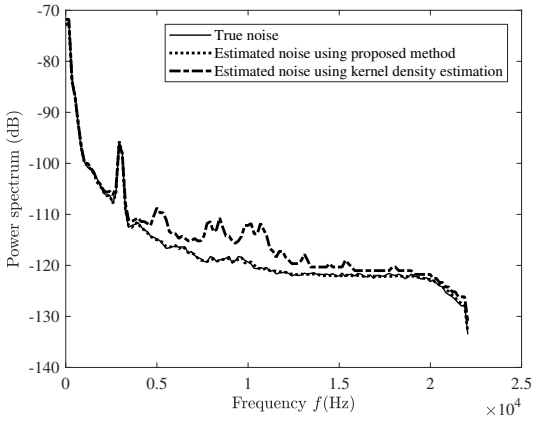


Figure 3.12: Comparison of the proposed method to kernel estimation for noise estimation.

tion 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.

Chapter 4

Conclusions and Future Works

4.1 Conclusions

In this dissertation, for speech enhancement, based on Gaussian analysis of noise power spectrum and speech power spectrum, we have proposed a simple noise estimation algorithm to accurately estimate the noise power spectrum. In addition, we have also propose a mean-shift algorithm with updating kernel to accurately estimate the mean, standard deviation and number of samples of the major cluster. Both methods can effectively suppress the influence of outliers.

4.2 Future Works

The research topic on speech enhancement has been ended. The author's plans for the estimation of the major cluster by mean-shift with updating kernel are as following:

- To comparison of accuracy and calculation cost of proposed method with other mainstream algorithms which are mentioned in this dissertation for details.
- To confirm the validity of proposed method by applying other target data.

Bibliography

[1] J. Sohn and N. Kim, "A statistical model-based voice activity detection," IEEE Signal Process. Lett. vol. 6, no. 1, pp. 1–3, 1999.

[2] R. Martin, "Spectral subtraction based on minimum statistics," Proc. of EUSIPCO, pp. 1182–1185, 1994.

[3] G. Doblinger, "Computationally efficient speech enhancement by spectral minima tracking in subbands," Proc. of Eurospeech, pp. 1513–1516, 1995.

[4] H. Hirsch and C. Ehrlicher, "Noise estimation techniques for robust speech recognition," Proc. of IEEE Int. Conf. Acoust., Speech, Signal Processing, pp. 153–156, 1995.

[5] I. Cohen, "Noise estimation by minima controlled recursive averaging for robust speech enhancement," IEEE Signal Processing Letters, vol. 9, no. 1, pp. 12–15, 2002.

[6] I. Cohen, "Noise spectrum estimation in adverse environments: Improved minima controlled recursive averaging," IEEE Trans. on Speech and Audio Processing, vol. 11, no. 5, pp. 466–475, 2003.

[7] Y. Yokota; T. Ye, "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.), vol. J99-A, no. 3, pp. 149–158, 2016.

[8] K. Fukunaga and L. Hostetler, "The estimation of the gradient of a density function, with applications in pattern recognition," IEEE Trans. Inf. Theory, vol. 21, no. 1, pp. 32–40, 1975.

[9] Y. Cheng, "Mean shift, mode seeking, and clustering," IEEE Trans. Pattern Anal. Mach. Intell., vol. 17, no. 8, pp. 790–799, 1995.

[10] D. Comaniciu and P. Meer, "Mean shift analysis and applications," In Proceedings of the Seventh IEEE International Conference on Computer Vision, Kerkyra, Greece, vol. 2, pp. 1197–1203, 1999.

[11] D. Comaniciu and P. Meer, "Mean shift: A robust approach toward feature space analysis," IEEE Trans. Pattern Anal. Mach. Intell., vol. 24, no. 5, pp. 603–619, 2002.

[12] M.P. Wand and M.C. Jones, Kernel Smoothing; Springer: London, UK, 1995.

[13] S.J. Sheather and M.C. Jones, "A reliable data-based bandwidth selection method for kernel density estimation," J. R. Stat. Soc. Ser. B, vol. 53, no. 3, pp. 683–690, 1991.

[14] S. Chen, "Optimal bandwidth selection for kernel density functionals estimation," J. Probab. Stat., vol. 2015, Article ID 242683, pp. 1–21, 2015.

[15] Y. Slaoui, "Data-driven bandwidth selection for recursive kernel density estimators under double truncation," Sankhya B, vol. 80, no. 2, pp. 341–368, 2018.

[16] D. Comaniciu; V. Ramesh; P. Meer, "The variable bandwidth mean shift and data-driven scale selection," In Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Maui, HI, USA, vol. 1, pp. 438–445, 1991.

[17] K. Okada; D. Comaniciu; A. Krishnan, "Scale selection for anisotropic scale-space: application to volumetric tumor characterization," In Proceedings of the 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Washington, DC, USA, pp.594–601, 2004.

[18] D. Comaniciu, "An algorithm for data-driven bandwidth selection," IEEE Trans. Pattern Anal. Mach. Intell., vol. 25, no. 2, pp. 281–288, 2003.

[19] X. Li; Z. Hu; F. Wu, "A note on the convergence of the mean shift," Pattern Recognit., vol. 40, no. 6, pp. 1756–1762, 2007.

[20] Y.A. Ghassabeh and F. Rudzicz, "Modified mean shift algorithm," IET Image Process, vol. 12, no. 12, pp. 2172–2177, 2018.

[21] S.T. Birchfield and S. Rangarajan, "Mean shift blob tracking with kernel histogram filtering and hypothesis testing," Pattern Recognit. Lett., vol. 26, no. 5, pp. 605–614, 2005.

[22] I. Leichter, "Mean shift trackers with cross-bin metrics," IEEE Trans. Pattern Anal. Mach. Intell., vol. 34, no. 4, pp. 695–706, 2011.

[23] T. Vojir; J. Noskova; J. Matasa, "Robust scale-adaptive mean-shift for tracking," Pattern Recognit. Lett., vol. 49, pp.250–258, 2014.

[24] G.E.P. Box and D.R. Cox, "An analysis of transformations," J. Royal Statistical Society, vol. 26, no. 2, pp. 211–252, 1964.

[25] T. Lotter and P. Vary, "Noise reduction by maximum a posteriori spectral amplitude estimation with super-Gaussian speech modelling," in Proc. International Workshop on Acoustic Echo and Noise Control, Kyoto, Japan, pp. 83–86, Sep. 2003.

[26] Mandarin Training Research Center, "Specific materials for National Proficiency Test of Mandarin (Fifth Edition)," China Peace Publishing House, Beijing (China), ISBN978-7-5137-0948-4, 2015.

[27] Three Agent Network, "Minna no Nihongo Shokyuu I (Elementary Japanese)," 3A Corporation, Tokyo (Japan), ISBN978-4-88319-603-6, 2012.

[28] B.W. Silverman, Density Estimation for Statistics and Data Analysis; Chapman & Hall: London, UK, 1986.

[29] Y. Kawamura; Y. Yokota; N. Matsumaru; K. Shirai, "4 Hours monitoring system of heart rate variability to predict septic shock," IEICE Tech. Rep., vol. 112, no. 63, pp. 29–34, 2012.

[30] T. Ye; Y. Yokota, "Noise estimation for speech enhancement based on quasi-Gaussian distributed power spectrum series by radical root transformation," IEICE Trans. Fundam. Electron. Commun. Comput. Sci., vol.E100-A, no. 6, pp. 1306–1314, 2017.

[31] F.E. Grubbs, "Sample criteria for testing outlying observations," Ann. Math. Stat., vol. 21, no. 1, pp. 27–58, 1950.

[32] F.E. Grubbs; G. Beck, "Extension of sample sizes and percentage points for significance tests of outlying observations," Technometrics, vol. 14, no. 4, pp. 847–854, 1972.

[33] C.B. Zeller; V.H. Lachos; F.V. Labra, "Influence diagnostics for Grubbs's model with asymmetric heavy-tailed distributions," Stat. Pap., vol. 55, no. 3, pp. 671–690, 2014.

[34] R. Thompson, "A note on restricted maximum likelihood estimation with an alternative outlier model," J. R. Stat. Soc. Ser. B (Methodol.), vol. 47, no. 1, pp. 53–55, 1985.

[35] C.E. Rasmussen, "The infinite Gaussian mixture model," In Advances in Information Processing Systems 12; MIT Press: Cambridge, MA, USA, pp. 554–560, 2000.

[36] J. Blomer and K. Bujna, "Adaptive seeding for Gaussian mixture models," In Proceedings of the 20th Pacific-Asia Conference, PAKDD 2016, Auckland, New Zealand, pp. 296–308, 2016.

[37] C. Viroli and G.J. McLachlan, "Deep Gaussian mixture models," Stat. Comput., vol. 29, no. 1, pp. 43–51, 2019.

[38] A.P. Dempster; N.M. Laird; D.B. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," J. R. Stat. Soc. Ser. B (Stat. Methodol.), vol. 39, no. 1, pp. 1–38, 1977.

[39] V. Melnykov and I. Melnykov, "Initializing the EM algorithm in Gaussian mixture models with an unknown number of components," J. Comput. Stat. Data Anal., vol. 56, no. 6, pp. 1381–1395, 2012.

Chapter
List of publications

Publications related to this thesis

Journal Papers:

1. Yasunari Yokota, Tian Ye, "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.), vol. J99-A, no. 3, pp. 149–158, 2016.

2. Tian Ye, Yasunari Yokota, "Noise estimation for speech enhancement based on quasi-Gaussian distributed power spectrum series by radical root transformation," IEICE Trans. Fundam. Electron. Commun. Comput. Sci., vol.E100-A, no. 6, pp. 1306–1314, 2017.

3. Tian Ye, Yasunari Yokota, "Estimating the Major Cluster by Mean-Shift with Updating Kernel," Mathematics, vol. 7, iss. 9, paper ID. 771, pp. 1–25, 2019.

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

$$\hat{\boldsymbol{\mu}}_N \ \leftarrow \ \boldsymbol{\mu}_x. \tag{A.1}$$

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

$$\boldsymbol{\mu}_W \ \leftarrow \ \hat{\boldsymbol{\mu}}_N. \tag{A.2}$$

   In addition, covariance matrix $\boldsymbol{C}_N$ is assigned to be an appropriate size as discussed in Section 3.1.2.

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

$$a_n \ = \ \frac{1}{A} p(\boldsymbol{x}_n; \boldsymbol{\mu}_W, \boldsymbol{C}_W). \tag{A.3}$$

However,

$$A = \sum_{k=1}^{N} p(\boldsymbol{x}_k; \boldsymbol{\mu}_W, \boldsymbol{C}_W). \tag{A.4}$$

We use this weight a_n to calculate the sample mean vector $\boldsymbol{\mu}_x$ with $\boldsymbol{x}_n$, $n = 1, \ldots, N$ as

$$\boldsymbol{\mu}_x = \sum_{n=1}^{N} a_n \boldsymbol{x}_n. \tag{A.5}$$

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

$$\hat{\boldsymbol{\mu}}_N \leftarrow \boldsymbol{\mu}_x. \tag{A.6}$$

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 (3.17)

Equation (3.16) can be rewritten as

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

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

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

as shown in Equation (3.17). 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 (B1) has the form of

$$y = \frac{1}{N_N}\sum_{n=1}^{N_N} x_n. \tag{B.3}$$

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

$$E(y) = E(x_n) > 0, \tag{B.4}$$

$$\sigma(y) = \frac{1}{\sqrt{N_N}}\sigma(x_n). \tag{B.5}$$

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 B.1. 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 (B2), i.e., the approximation of Equation (B1), holds.

Table B.1: Expected value and standard deviation of probability density distribution $f(x)$ defined by $x \geq 0$.

$\boldsymbol{f(x)}$	**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	$\mathrm{e}^{\mu+\sigma^2/2}$	$\mathrm{e}^{\mu+\sigma^2/2}\sqrt{\mathrm{e}^{\sigma^2}-1}$
Pareto	$\dfrac{ab}{a-1}$	$\dfrac{\sqrt{a}b}{(a-1)\sqrt{a-2}}$

Appendix C
Multi-Dimensional Mean-Shift with Updating Kernel

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 3.2.1 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

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

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

$$p(\boldsymbol{x}; \boldsymbol{\sigma}_W) = \prod_{m=1}^{M} p(x_m; \sigma_{W,m}). \tag{C.2}$$

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

$$a_n = \frac{1}{A} p(\boldsymbol{x}_n; \boldsymbol{\sigma}_W). \tag{C.3}$$

However, A in the above equation is

$$A = \sum_{k=1}^{N_N} p(\boldsymbol{x}_k; \boldsymbol{\sigma}_W). \tag{C.4}$$

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

$$\sigma_{x,m}^2 = \sum_{n=1}^{N_N} a_n x_{m,n}^2, \quad m = 1, \ldots, M. \tag{C.5}$$

For the same reason, under the one-dimensional case, by substituting Equation (C3) into Equation (C5), the expected value of the sample variance $\sigma_{x,m}^2$ can be approximated as

$$E[\sigma_{x,m}^2] \simeq \frac{1}{E[A]} E\left[\sum_{n=1}^{N_N} p(\boldsymbol{x}_n; \boldsymbol{\sigma}_W) x_{m,n}^2\right]. \tag{C.6}$$

Applying Equation (C1) to Equation (C4) and using Equation (13), the expected value of A is found as

$$\begin{aligned} E[A] &= E\left[\sum_{k=1}^{N_N} p(\boldsymbol{x}_k; \boldsymbol{\sigma}_W)\right] \\ &= \sum_{k=1}^{N_N} E[p(\boldsymbol{x}_k; \boldsymbol{\sigma}_W)] \\ &= N_N \prod_{j=1}^{M} \int_{-\infty}^{\infty} p(x_j; \sigma_{W,j}) p(x_j; \sigma_{N,j}) dx_j \\ &= \frac{N_N}{(2\pi)^{M/2}} \prod_{j=1}^{M} (\sigma_{W,j}^2 + \sigma_{N,j}^2)^{-1/2}, \end{aligned} \tag{C.7}$$

while using Equation (14), the remainder of Equation (C6) is

$$\begin{aligned} E\left[\sum_{n=1}^{N_N} p(\boldsymbol{x}_n;\boldsymbol{\sigma}_W)x_{m,n}^2\right] \\ &= \sum_{n=1}^{N_N} E[p(\boldsymbol{x}_n;\boldsymbol{\sigma}_W)x_{m,n}^2] \\ &= N_N \prod_{j=1}^{M} \int_{-\infty}^{\infty} x_m^2 p(x_j;\sigma_{W,j})p(x_j;\sigma_{N,j})dx_j \\ &= \frac{\sigma_{W,m}^2\sigma_{N,m}^2}{\sigma_{W,m}^2+\sigma_{N,m}^2}\frac{N_N}{(2\pi)^{M/2}}\prod_{j=1}^{M}(\sigma_{W,j}^2+\sigma_{N,j}^2)^{-1/2}. \end{aligned} \tag{C.8}$$

That is, according to Equation (C7) and Equation (C8), Equation (C6) becomes

$$E[\sigma_{x,m}^2] = \frac{\sigma_{W,m}^2\sigma_{N,m}^2}{\sigma_{W,m}^2+\sigma_{N,m}^2}. \tag{C.9}$$

The equation above can be transformed to

$$\sigma_{N,m}^2 = \frac{\sigma_{W,m}^2 E[\sigma_{x,m}^2]}{\sigma_{W,m}^2 - E[\sigma_{x,m}^2]}. \tag{C.10}$$

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

$$\hat{\sigma}_{N,m} = \sqrt{\frac{\sigma_{W,m}^2\sigma_{x,m}^2}{\sigma_{W,m}^2-\sigma_{x,m}^2}}, \tag{C.11}$$

when using the standard deviation $\sigma_{x,m}$ of the sample weighted with a Gaussian kernel with standard deviation $\boldsymbol{\sigma}_W$. Furthermore, using Equation (C7), we can estimate the number N_N of samples belonging to a major cluster as

$$\hat{N}_N = A(2\pi)^{M/2}\prod_{j=1}^{M}(\sigma_{W,j}^2+\hat{\sigma}_{N,j}^2)^{1/2}. \tag{C.12}$$

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

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:

$$\boldsymbol{\mu}_x = \frac{1}{N}\sum_{n=1}^{N}\boldsymbol{x}_n, \tag{C.13}$$

$$\boldsymbol{C}_x = \frac{1}{N}\sum_{n=1}^{N}(\boldsymbol{x}_n-\boldsymbol{\mu}_x)(\boldsymbol{x}_n-\boldsymbol{\mu}_x)^T. \tag{C.14}$$

   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

$$\hat{\boldsymbol{\mu}}_N \leftarrow \boldsymbol{\mu}_x, \tag{C.15}$$

$$\hat{\boldsymbol{C}}_N \leftarrow \boldsymbol{C}_x. \tag{C.16}$$

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

$$\boldsymbol{\mu}_W \leftarrow \hat{\boldsymbol{\mu}}_N, \tag{C.17}$$

$$\boldsymbol{C}_W \leftarrow r^2\hat{\boldsymbol{C}}_N. \tag{C.18}$$

   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:

$$\boldsymbol{\mu}_x = \sum_{n=1}^{N}\boldsymbol{a}_n\boldsymbol{x}_n, \tag{C.19}$$

$$\boldsymbol{C}_x = \sum_{n=1}^{N}a_n(\boldsymbol{x}_n-\boldsymbol{\mu}_x)(\boldsymbol{x}_n-\boldsymbol{\mu}_x)^T. \tag{C.20}$$

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

$$\hat{\boldsymbol{\mu}}_N \leftarrow \boldsymbol{\mu}_x. \tag{C.21}$$

Let

$$\boldsymbol{C}_W = \boldsymbol{V}_W \boldsymbol{\Lambda}_W \boldsymbol{V}_W^T \tag{C.22}$$

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 $\sigma_{W,1}^2, \ldots, \sigma_{W,M}^2$ along the directions represented by each of the column vectors of orthogonal matrix $\boldsymbol{V}_W$. In addition, the diagonal element of

$$\boldsymbol{\Lambda}_x = \boldsymbol{V}_W^T \boldsymbol{C}_x \boldsymbol{V}_W \tag{C.23}$$

is the variance $\sigma_{x,1}^2, \ldots, \sigma_{x,M}^2$ of $\boldsymbol{V}_W$ in the column vector direction in the sample covariance matrix $\boldsymbol{C}_x$. According to Equation (C9), 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 $\sigma_{W,1}^2, \ldots, \sigma_{W,M}^2$ and $\sigma_{x,1}^2, \ldots, \sigma_{x,M}^2$ in Equation (C8). 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:

$$\hat{\boldsymbol{C}}_N \leftarrow \boldsymbol{V}_W \hat{\boldsymbol{\Lambda}}_N \boldsymbol{V}_W^T. \tag{C.24}$$

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

$$\hat{N}_N \leftarrow A(2\pi)^{M/2} \prod_{j=1}^{M} (\sigma_{W,j}^2 + \hat{\sigma}_{N,j}^2)^{1/2}. \tag{C.25}$$

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.