ANOMALY DETECTION FOR ENVIRONMENTAL NOISE MONITORING

BY

DUC H. PHAN

THESIS

Submitted in partial fulfillment of the requirements
for the degree of Master of Science in Electrical and Computer Engineering
in the Graduate College of the
University of Illinois at Urbana-Champaign, 2018

Urbana, Illinois

Advisor:

Professor Douglas L. Jones

Abstract

Octave-band sound pressure level is the preferred measure for continuous environmental noise monitoring over raw audio because accepted standards and devices exist, these data do not compromise voice privacy, and thus an octave-band sound meter can legally collect data in public. By setting up an experiment that continuously monitors octave-band sound pressure level in a residential street, we show daily noise-level patterns correlated to human activities. Directly applying well-known anomaly detection algorithms including one-class support vector machine, replicator neural network, and principal component analysis based anomaly detection shows low performance in the collected data because these standard algorithms are unable to exploit the daily patterns. Therefore, principal component analysis anomaly detection with time-varying mean and the covariance matrix over each hour, is proposed in order to detect abnormal acoustic events in the octave band measurements of the residential-noise-monitoring application. The proposed method performs at 0.83 in recall, 0.88 in precision and 0.85 in F-measure on the evaluation data set.

Acknowledgments

This research was sponsored by the TerraSwarm Research center, one of six centers administered by the STARnet phase of the Focus Center Research Program (FCRP), a Semiconductor Research Corporation program sponsored by MARCO and DARPA; research funds from the College of Engineering, University of Illinois at Urbana-Champaign (UIUC); and the Coordinated Science Laboratory, UIUC. I have been greatly supported and guided by my advisor, Professor Douglas L. Jones. My labmates Erik Johnson, Long Le, and Alex Asilador, and Jamie Norton have been a great source of help and discussion.

Contents

1. Introduction

Environmental noise influences quality of life [1], [2]. Cars, planes, and commercial bustle can disturb sleep, create stress or hearing problems, and impact cognitive development in children. Another aspect of environmental noise is that it is directly related to human activities and lifestyles. As shown in Figure 1.1, a daily noise level pattern monitored near a road reveals the association between environmental noise and human activity. From midnight to early morning corresponding to the sleeping time of many families, the average noise level is low compared to the rest of the day. After that the average noise level increases due to the need to commute between house and school or workplaces. Variations in the noise level are significant between day time and night time. Therefore, monitoring and analyzing the environmental noise has been an active research area for decades [3]-[6].

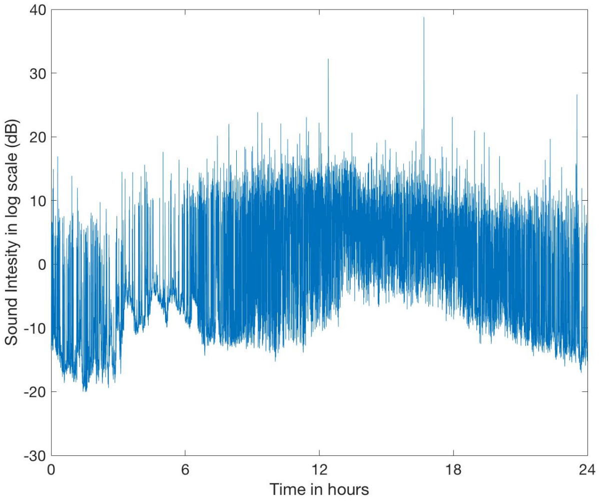

Figure 1.1 A daily noise level pattern monitored near a road. Both short-term fluctuations and long-term variation across the day are apparent.

One important tool for environmental noise analysis is anomaly or outlier detection, in which a generative process for the daily noise pattern is defined to provide the typical noise characteristics of a particular area, and anomalies are points, or events, that are unlikely generated by the generative model. Since long-term monitoring always yields a massive amount of data, researchers and analysts cannot investigate every single data point. Hence, anomalies mark potentially interesting events and instances that merit further investigation. As examples, anomalies in residential streets might include police or ambulance sirens, car crashes, or people shouting and screaming. If surveillance cameras are also installed, anomalies in environmental noise could selectively trigger raw audio and camera recording and transmission, reducing network traffic and data storage.

Anomaly detection in acoustic environmental noise faces several challenges. First, in order to protect speech privacy, the algorithm should only be applied on coarse measures such as noise intensity, octave band, or one-third-octave band measurements [7] over intervals considerably larger than phoneme duration, but not directly to the raw audio. Secondly, the normal and anomaly definitions are time-dependent at a given monitored area, because noise-level patterns change over time as shown in Figure 1.1. Thirdly, if octave bands or one-third-octave bands are used, the algorithm has to work on high-dimensional data.

Fortunately, the variation of acoustic noise strongly relates to human activities; therefore, it is reasonable to assume 24-hour periodicity on generative models in an urban setting. The periodicity suggests means for reducing the complexity of the anomaly detection algorithm.

This thesis applies well-known approaches in anomaly detection including one-class support vector machine (SVM), replicator neural network (RNN), and principal-component-analysis based anomaly detection to a continuous octave-band noise intensity monitor in a residential area, before proposing a time-varying principal-component-analysis-based anomaly detection which improves the performance significantly. The proposed method treats measurements at each hour independently. At each hour, typical measurements are approximately generated by a multivariate Gaussian distribution, and anomalies are input samples which are unlikely to be present under the corresponding normative distribution.

The rest of this thesis is organized as follows. Chapter 2 surveys related works in environmental noise monitoring and analysis. Chapter 3 provides background about octave band measurement; Gaussian, log-normal, and chi-squared distributions; principal component analysis (PCA); anomaly detection algorithms including PCA based anomaly detection, one-class SVM, and RNN; and an evaluation framework. Chapter 4 discusses the data collected in the study. Chapter 5 presents the proposed method in detail. Chapter 6 evaluates the performance of the suggested algorithm. Lastly, Chapter 7 draws conclusions and discusses about directions for further work.

2. Related Works

Many works in environmental noise monitoring recently have focused on detection and classification of acoustic events. Salamon and Bello [8] designed a convolutional deep neural network which has “3 convolutional layers interleaved with 2 pooling operations, followed by 2 fully connected (dense) layers”. In the preprocessing steps, Salamon and Bello [9] transform raw audio data into log-scaled mel-spectrogram representation before extracting time-frequency patches as input for the networks. Scream and gunshot detection are the topic of study of Valenzise et al. [10]. The authors converted 23 ms audio frames into feature vectors including zero-crossing rate (ZCR), mel-frequency cepstral coefficients (MFCC), and some other spectral and distribution- based measurements before using two independent Gaussian mixture models to discriminate gunshots and screams from environmental noise respectively. Matrix factorizations such as non-negative matrix factorization (NMF), principal component analysis (PCA) and their variants seeking good representations of environmental acoustic scenes are examined in by Bisot et al. [11].

In addition, anomaly detection of acoustic events have been studied. Ntalampiras et al. [12] use a Gaussian Mixture Model (GMM) to form statistical representations of normal events, thresholding the likelihood of the incoming data based on selected anomalies returned by the GMM. Chakrabarty and Elhilali [13] apply a Restricted Boltzmann machine (RBM) and conditional RBM as the generative model of the acoustic environment, and anomalies are identified based on their likelihood.

In all of the aforementioned works, transient acoustic events are subjects for classification and detection. In other words, the acoustic scene must always be represented with sufficient information that can reconstruct the corresponding raw audio. Therefore, these approaches may not be easily deployed in public environments due to privacy concerns about human speech.

Fortunately, environmental noise monitoring by means of sound pressure level, octave band, one-third-octave-band measurements can be conducted in public areas. Hardware and infrastructure is available for up to city-scale noise monitoring with reasonable cost. Mydlarz et al. [14], [15] implement a low-cost microelectromechanical systems (MEMS) microphone array (less than $100 USD per sensor node) that complies with the standard IEC 61672-1[16] for electroacoustic sound level meters. In addition, Mydlarz et al. [15] provide a sensor network designed for city-scale deployment. Hence, we believe that in many cases, classification and detection applied for environmental monitoring should start with the assumption that only sound-pressure levels over at least one second intervals are available as inputs.

Given this survey of related works, our contributions can be summarized as follows: First, the study is conducted under the assumption that only octave-band sound pressure level is available as raw data. Secondly, the nonstationary nature of sound levels in residential areas and their daily patterns are illustrated. Thirdly, different anomaly detection algorithms applied to a continuously monitoring sound pressure level data set are reported. Lastly, the extension of robust PCA-anomaly detection [17] by introducing piecewise constant mean and covariance parameters in multivariate Gaussian distribution of the generative model produces an anomaly detection that can adapt to the dynamics of residential noise level.

3. Background

This chapter provides the theoretical framework for the discussion and explanations in the following chapters. Octave-band measures are introduced, followed by discussion of statistical properties including the normal, log-normal, and chi-square distributions before principal component analysis (PCA) is considered. Next, anomaly detection techniques comprising PCA-based anomaly detections, one-class support vector machine (SVM), and replicator neural network (RNN) are studied. Lastly, performance metrics of anomaly detection algorithms consisting of recall, precision, and F-measure are introduced.

3.1 Octave Band

As stated in ANSI S1.43-1997 (r 2007) [7], a sound-pressure-level meter can split the frequency sprectrum into octave bands and provide measurements of the average energy level of each band over intervals of one second. Therefore, it is approved by US law to collect octave-band mesurements in public environments, since there is no known method to reconstruct the speech information from these mesurements.

In addition, intensive scientific study of human speech has determined that the information as to both the words spoken and the speaker's identity are carried in the fine spectro-temporal (time-frequency) structure of the signal. Therefore, if we sufficiently undersample in this domain (or in another domain from which this information is provably irrecoverable), the speech/speaker are fundamentally irrecoverable.

Furthermore, the noise level at octave bands provides richer information than a single noise intensity at the same sampling rate, and it saves processing resources since it can be implemented on hardware.

In this thesis, the raw audio is collected in order to validate the results of analysis; therefore, we need to transform audio data into octave band features as shown in Figure 3.1. Raw audio data are split into overlapped frames. At each frame, signal energy is presented in the form of octave band components. Applying Parseval's theorem and signal processing theory [18], an octave band component can be calculated by summing the squares of the fast Fourier transform (FFT) magnitudes corresponding to the cutoff frequencies of that octave band and then dividing by the length of the FFT block; the cutoff frequencies for each octave band are approximately equal to the standard specification in [7]. In the final step, the average of these octave-band components over data frames equivalent to T seconds of recording provides the octave-band vector of interest. Note that there are extra calibrations and scaling steps [7], [19] in order to generate the actual sound pressure level; however, without loss of generality, the analysis in this thesis can skip these steps because all the data is measured from a single microphone with a stationary setting.

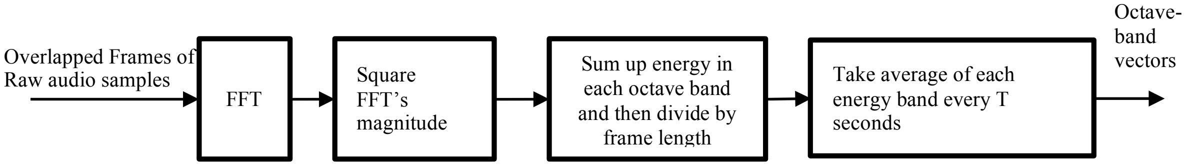

Figure 3.1 Transformation of raw audio data into octave-band features. First, raw audio data are split into overlapped frames. At each frame, signal energy is present in the form of octave-band components before going to the last stage. In the final step, the average of these octave-band components over data frames equivalent to T seconds of recording provides the octave-band vector of interest.

3.2 The Gaussian Distribution

A widely used distribution of continuous radom variables which plays an important role in this thesis is the Gaussian distribution, also known as the normal distribution [20], [21]. In one-dimensional random variables X, the probability density function of a Gaussian random variable is defined as

$$p_X(x|\mu,\sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(x-\mu)^2} \tag{3.1}$$

where μ is the mean, σ^2 is the variance and x is a realization of X. If a normal distribution has the mean μ of zero and the variance σ^2 of one, then the distribution is called standard normal.

When an N-dimensional random variable $\boldsymbol{X}$ is studied, the probability density function becomes

$$p_{\boldsymbol{X}}(\boldsymbol{x}|\boldsymbol{\mu},\boldsymbol{\Sigma}) = \frac{1}{(2\pi)^{\frac{N}{2}}|\boldsymbol{\Sigma}|^{\frac{1}{2}}} e^{-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})} \tag{3.2}$$

where $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ are an N-dimensional mean vector and NxN covariance matrix respectively. $|\boldsymbol{\Sigma}|$ represents the determinant of $\boldsymbol{\Sigma}$**,** and $\boldsymbol{x}$ is a realization of random vector $\boldsymbol{X}$**.** The Gaussian random vector parameterized by $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ is often denoted as $N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. In the special case, where $\boldsymbol{\Sigma}$ is a diagonal matrix having $\sigma_1^2, \dots, \sigma_N^2$ as diagonal elements, elements $X_1 \dots X_N$ of vector X become independent random variables. Therefore the probablity density function can reduce to the form of Equation (3.3) [20].

$$p_{\boldsymbol{X}}(\boldsymbol{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \prod_{n=1}^{N} p(x_i|\mu_i, \sigma_i^2) \qquad (3.3)$$

where

$$\boldsymbol{x} = [x_1 \dots x_N]^T \qquad (3.4)$$

$$\boldsymbol{\mu} = E[\boldsymbol{X}] = [\mu_1 \dots \mu_N]^T \qquad (3.5)$$

$$\boldsymbol{\Sigma} = E[(\boldsymbol{X} - \boldsymbol{\mu})(\boldsymbol{X} - \boldsymbol{\mu})^T] = \begin{bmatrix} \sigma_1^2 & 0 & \cdots & 0 \\ 0 & \sigma_2^2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sigma_N^2 \end{bmatrix} \qquad (3.6)$$

Note that $E[\boldsymbol{x}]$ denotes the expectation of random vector $\boldsymbol{X}$. Definition and properties of the expection can be found in [20] and [21].

In the general case, $\boldsymbol{\Sigma}$ is a positive-definite matrix, so it can be factorized into the form of Equation (3.7) [22]:

$$\boldsymbol{\Sigma} = \boldsymbol{U\Lambda U^T} \qquad (3.7)$$

where $\boldsymbol{\Lambda}$ is a diagonal matrix and $\boldsymbol{U}$ is an *N*x*N* dimenisional matrix which contains an orthonormal basis for an N-dimensional real vector space; $\boldsymbol{U^T U} = \boldsymbol{I}$, where $\boldsymbol{I}$ is the identity matrix.

Let us consider vector $\boldsymbol{Z}$ in *N*-dimensional space taking the form of Equation (3.8). It can be shown that $\boldsymbol{Z}$ has a Gaussian distribution because the sum of linearly scaled Gaussian random variables produces another Gaussian random variable [20].

$$\boldsymbol{Z} = \begin{bmatrix} Z_1 \\ \vdots \\ Z_N \end{bmatrix} = \boldsymbol{U^T X} \tag{3.8}$$

$$E[\boldsymbol{Z}] = E[\boldsymbol{U^T X}] = \boldsymbol{U^T} E[\boldsymbol{X}] = \boldsymbol{U^T \mu} \tag{3.9}$$

$$\begin{aligned} E[(\boldsymbol{Z} - E[\boldsymbol{Z}])(\boldsymbol{Z} - E[\boldsymbol{Z}])^T] &= E[\boldsymbol{U}^T(\boldsymbol{X} - \boldsymbol{\mu})(\boldsymbol{X} - \boldsymbol{\mu})^{\boldsymbol{T}}\boldsymbol{U}] \\ &= \boldsymbol{U^T} E[(\boldsymbol{X} - \boldsymbol{\mu})(\boldsymbol{X} - \boldsymbol{\mu})^T]\boldsymbol{U} \\ &= \boldsymbol{U^T \Sigma U} \\ &= \boldsymbol{U^T U \Lambda U^T U} \\ &= \boldsymbol{\Lambda} \end{aligned} \tag{3.10}$$

Furthermore, the components $Z_1 \dots Z_N$ of vector $\boldsymbol{Z}$ are independent random variables; the covariance matrix of random vector $\boldsymbol{Z}$ is the diagonal matrix $\boldsymbol{\Lambda}$. In other words, $\boldsymbol{Z}$ is the Gaussian random vector $N(\boldsymbol{U^T \mu}, \boldsymbol{\Lambda})$. The linear transfomation from vector $\boldsymbol{X}$ to vector $\boldsymbol{Z}$ is also known as the decorrelation of the Gaussian random variable; this transformation also provides a probablistic interpretation for PCA disscussed in Section 3.5.1.

3.3 The Log-normal Distribution

The sample space of a Gaussian random variable is the set of real numbers R; therefore, the Gaussian distribution is sometimes not directly suitable for a generative model of a non-negative data set. In such cases, the log-normal distribution could be a reasonable choice because the realization of a log-normal random variable is a positive number. A similar argument holds when modeling a high-dimensional data set.

Formally, the log-normal random variable Y is the random variable whose lograrithm has a Gaussian distribution [23]. In other words, if Y is a log-normal random variable, then $X = \ln Y$ is a Gaussian random variable.The probability density function of the log-normal distribution is given by

$$p_Y(y|\mu, \sigma^2) = \frac{1}{y\sigma\sqrt{2\pi}} e^{-\frac{(\ln y - \mu)^2}{2\sigma^2}} \quad , x > 0 \tag{3.11}$$

where μ and σ^2 are the mean and variance of the Gaussian variable $X = \ln Y$.

Figure 3.2 shows examples of log-normal distributions with different parameter values. From these examples, it can be clearly seen that the log-normal distribution is a very plausible model for positive heavy-tailed data sets with appropriate choice of parameters.

In a similar manner, if $\boldsymbol{X} = [X_1 \dots X_N]^T$ is an N-dimensional Gaussian random vector, then $Y = e^X = \begin{bmatrix} e^{X_1} \\ \vdots \\ e^{X_N} \end{bmatrix}$ is a multivariate log-normal random vector [24].

The log-normal distribution has been successfully applied in different fields such as modeling the time to repair a maintainable system in reliability analysis [25], and modeling the firing rate across a population of neurons [26], [27]. As shown in Chapters 4 and 5, this thesis presents another application of the log-normal distribution.

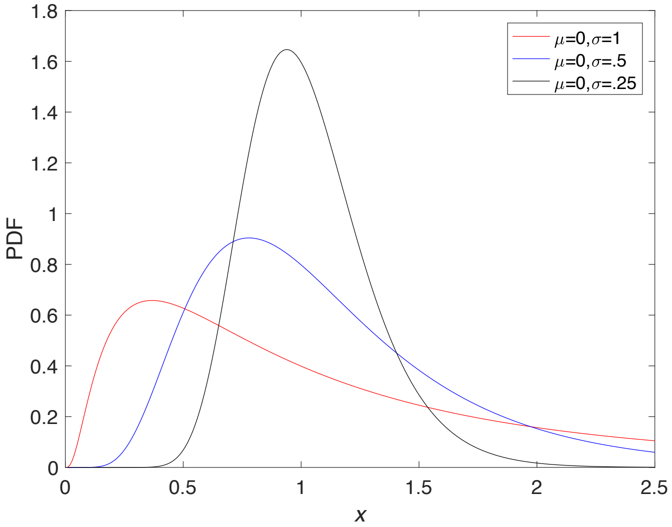

Figure 3.2 Examples of log-normal distributions with different parameters.

3.4 The Chi-Squared Distribution

Chi-squared or χ^2 distribution is another well-known distribution related to the normal distribution. The chi-squared distribution, which plays an important role in statistics and hypothesis testing [28], is formally defined as follows.

If $X_1 \dots X_k$ are independent, standard normally distributed, then random variable Q shown in Equation (3.12) has chi-squared distribution and is denoted as $Q \sim \chi^2(k)$, where k is called the number of degrees of freedom.The chi-squared probability density function is defined in Equation (3.13), and examples of chi-squared denistiy functions with different degrees of freedoom are shown in Figure 3.3.

$$Q = \sum_{i=1}^{k} X_i^2 \tag{3.12}$$

$$p_Q(q|k) = \frac{q^{\frac{k}{2}-1} e^{-\frac{q}{2}}}{2^{\frac{k}{2}}\,\Gamma\left(\frac{k}{2}\right)}\;, for\;\; q > 0 \tag{3.13}$$

where

$$\Gamma(\alpha) = \int_{0}^{+\infty} x^{\alpha-1} e^{x} dx \tag{3.14}$$

is called the gamma function.

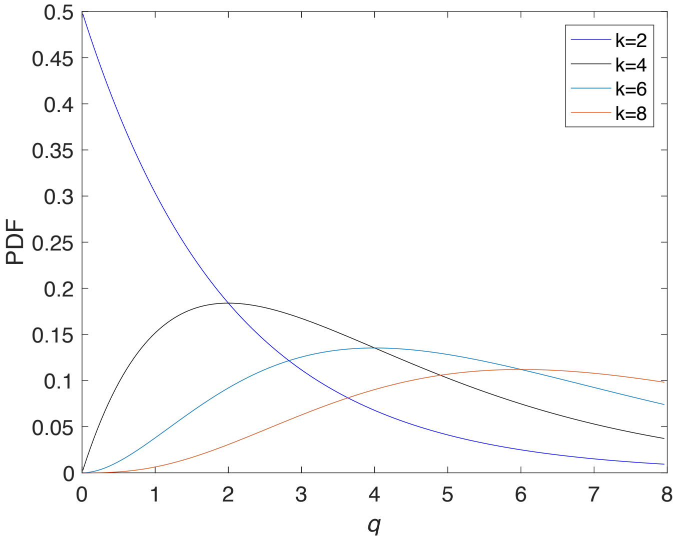

Figure 3.3 Chi-squared probability density functions with various degrees of freedoms.

In addition, the cumulative distribution function (CDF) of the chi-squared distribution is given in Equation (3.15):

$$F_Q(q|k) = P_Q(Q \leq q) = \frac{\gamma\left(\frac{k}{2}, \frac{q}{2}\right)}{\Gamma\left(\frac{k}{2}\right)}\text{, for } q > 0 \tag{3.15}$$

where

$$\gamma(s, x) = \int_0^x t^{s-1} e^{-t} dt \tag{3.16}$$

Lastly, given k degrees of freedom, the chi-squared value of p-value, $\chi_k^2(p)$, is defined as

$$\chi_k^2(p) = q \text{ such that } P_Q(Q > q) = 1 - F_Q(q|k) = p, p \in [0,1] \tag{3.17}$$

3.5 Principal Component Analysis

The principal component analysis (PCA) is a linear dimension-reduction method also known as the Karhunen–Loève transform in stochastic settings [20], [29], [30]. Namely, given an *N*-dimensional random vector $\boldsymbol{X}$ with mean vector $\boldsymbol{\mu} \in \boldsymbol{R^N}$and a multivariate distribution $\boldsymbol{P}$, and letting $\boldsymbol{Y} = \boldsymbol{A}(\boldsymbol{X} - \boldsymbol{\mu})$ and $\widetilde{\boldsymbol{X}} = \boldsymbol{BY} + \boldsymbol{\mu}$ where A is an R×N matrix and $\boldsymbol{B}$ is an N×R matrix (*R* ≤ *N*), the PCA task is finding matrices $\boldsymbol{A}$ and $\boldsymbol{B}$ such that the cost function in Equation (3.18) is minimized.

$$J(\boldsymbol{A}, \boldsymbol{B}) = E_{\boldsymbol{P}}\left[\left\|\boldsymbol{X} - \widetilde{\boldsymbol{X}}\right\|^2\right] = E\left[\left(\boldsymbol{X} - \widetilde{\boldsymbol{X}}\right)^{\boldsymbol{T}}\left(\boldsymbol{X} - \widetilde{\boldsymbol{X}}\right)\right] \tag{3.18}$$

where $E_{\boldsymbol{P}}[\cdot]$ denotes the expectation operator over the multivariate distribution $\boldsymbol{P}$.

Letting $\boldsymbol{\Sigma} = E_{\boldsymbol{P}}[(\boldsymbol{X} - \boldsymbol{\mu})(\boldsymbol{X} - \boldsymbol{\mu})^T]$ be the covariance matrix of random vector $\boldsymbol{X}$**,** from Equation (3.7), the covariance matrix can be rewritten as

$$\boldsymbol{\Sigma} = \boldsymbol{U\Lambda U^T} \tag{3.19}$$

where $\boldsymbol{U^T U} = \boldsymbol{I}$, and $\boldsymbol{\Lambda}$ is a diagonal matrix with a nonnegative diagonal element. Without loss of generality, $\boldsymbol{\Lambda}$ can be written as

$$\boldsymbol{\Lambda} = \begin{bmatrix} \lambda_1 & 0 & \cdots & 0 \\ 0 & \lambda_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \lambda_N \end{bmatrix}, \qquad \text{with } \ \lambda_1 \geq \lambda_2 \ldots \geq \lambda_N \geq 0 \tag{3.20}$$

Let $\boldsymbol{U_1} \ldots \boldsymbol{U_N}$ be the column vectors forming matrix $\boldsymbol{U}$, then $\boldsymbol{U_i}$ is an eigenvector corresponding to eigenvalue λ_i of the covariance matrix. Namely, $\boldsymbol{\Sigma U_i} = \lambda_i \boldsymbol{U_i}$ [22].

It can be shown that $J(\boldsymbol{A}, \boldsymbol{B})$ is minimized when matrix $\boldsymbol{A}$ equals the transpose of matrix $\boldsymbol{B}$, $\boldsymbol{A} = \boldsymbol{B^T}$and matrix $\boldsymbol{B}$ is formed by the first R ($R \leq N$) eigenvectors, $\boldsymbol{B} = [\boldsymbol{U_1} \ldots \boldsymbol{U_R}]$ [20]. Therefore, $\boldsymbol{U_1}, \ldots, \boldsymbol{U_N}$ are also called principal components. In addition, while they are in general uncorrelated random variables, the elements of $\boldsymbol{Y} = \boldsymbol{A}(\boldsymbol{X} - \boldsymbol{\mu})$ are independent random variables if $\boldsymbol{P}$ is a Gaussian distribution.

Intuitively, PCA can be easily understood by first looking at Figure 3.4, in which sample data is translated to the origin and then rotated in order to align with the axes by $\boldsymbol{Z} = \boldsymbol{U^T}(\boldsymbol{X} - \boldsymbol{\mu})$. After that, components which have larger variances are selected. For example, if the dimension of the sample data in Figure 3.4 is reduced to one, then $\boldsymbol{Y} = [Z_1]$ and $A^T = B = [\boldsymbol{U_1}]$. Note that the PCA discussion has so far assumed knowledge of the $\boldsymbol{U}$ orthonormal basis; however, in reality, this basis can be efficiently solved by singular value decomposition (SVD) given the covariance matrix [22].

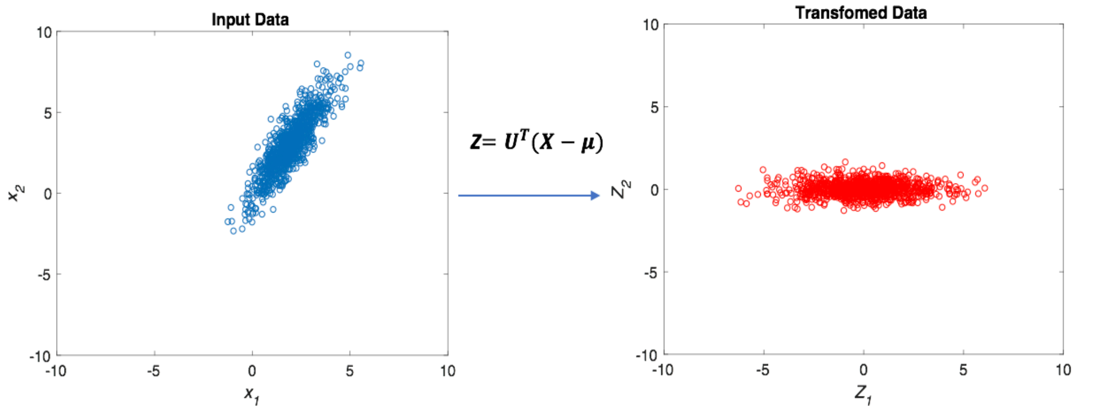

Figure 3.4 An Intuitive explanation of principal component analysis. The left plot shows the sample data generated by a two-dimensional Gaussian random vector, and the right plot shows the translated and rotated sample data in PCA.

Finally, let us consider Figure 3.5, where the sample data after removing the mean are weighted and projected onto principal components as shown in Equation (3.21). As a result, the elements of random vector $\boldsymbol{Q}$ are uncorrelated and have unit variance. In other words, the covariance matrix of random vector $\boldsymbol{Q}$ is the identity matrix. The transformation from random vector $\boldsymbol{X}$ to random vector $\boldsymbol{Q}$ provides a foundation for explaining PCA-based anomaly detection in Section 3.6.

$$\boldsymbol{Q} = (\,\Sigma^{-\frac{1}{2}})\boldsymbol{U}^{\boldsymbol{T}}(\boldsymbol{X} - \boldsymbol{\mu}) \tag{3.21}$$

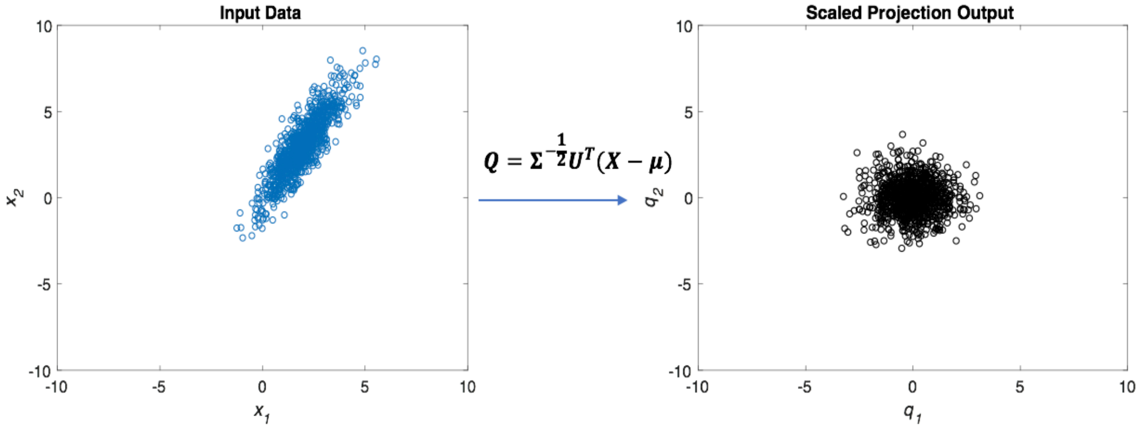

Figure 3.5 The left plot shows the sample data generated for a two-dimensional Gaussian random vector, and the right plot shows the scaled projections of sample data on the principal components.

3.6 Anomalies and Anomaly Detection

Anomalies are patterns of data which do not follow a well-defined notion of typical behaviors. Figure 3.6 demonstrates a simple example of anomalies in a two-dimensional data set. The data show two normal regions as blue dots, while anomalies are red dots which are far away from normal regions.

Anomaly detection techinques often try to find a mapping from data instances into scores or ranking numbers. An analyst can either declare a few instances with top scores or choose a threshold to select anomalies. Classification-based, neural-network-based, statistical and spectral approaches are common solutions for anomaly detection problems operating in high-dimensional unlabeled data (unsupervised learning settings) [17].

In classification-based techniques, algorithms try to learn the boundary of the typical points; a testing instance is anomalous if it lies outside the normal boundary. Classification-based works best if the training data do not contain anomalies. One-class SVM and its variants are the mainstream techniques because with the kernel trick SVM can learn a complex non-linear boundary. In addition, one-class SVM has been successfully applied to several anomaly detection applications [31-33]. In unsupervised settings, the training data may contain anomalies, so the anomaly score can be assigned as a signed distance from the decision boundary; the points lying within the boundary have positive scores and points lying outside the boundary have negative scores [34].

In neural-network-based techniques, anomalies are assumed to have a small fraction compared to typical data in a given set of data; therefore the weights in the hidden layer of the neural net are influenced mainly by the typical behaviors of the generating process. During the learning phase, the data are compressed into hidden layers of the neural network. In the testing phase data are reconstructed from the trained network. The errors between the input data and the reconstructed data are used as scores of anomaly detections [17]. Even though deep convolutional neural network based techniques have been developed [35], we believe that a large training data set is required for applying deep neural networks. Therefore, given the size of experimental data in this study, we select a replicator neural network (RNN), a simple but powerful and widely-used neural network [36-37], to apply to our observed data.

In statistical anomaly detection techniques, a stochastic generative model of the observed data is estimated. Anomalies are data instances occurring in the low probability of region of the stochastic model. If a multivariate Gaussian distribution closely approximates the distribution of a given training data set, then robust principal component analysis (PCA) based anomaly detection is an appropriate choice. In addition, if typical data instances and anomalies appear to be different in a lower dimensional subspace, the PCA-based technique also provides a tool to find the lower dimension subspace of interest. In other words, PCA-based anomaly detection algorithm is also a key technique in spectral based anomaly detection [17].

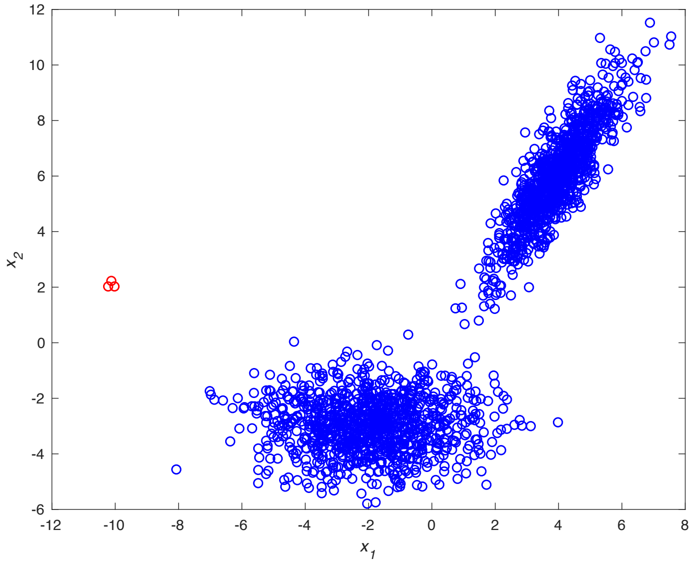

Figure 3.6 An example of anomalies in a two-dimensional data set.

Next, the detailed background knowledge for PCA-based anomaly detection, one-class SVM, and RNN are presented.

3.6.1 PCA-based anomaly detection

In PCA-based anomaly detection, data can be either generated by a multivariate Gaussian distribution or embedded into a lower-dimensional subspace in which anomalies appear significantly different from the normal instances.

The general procedure of PCA-based anomaly detection starts by estimating principal components of the covariance matrix of the training data, or matrix $\boldsymbol{U}$ in Equation (3.19). In the testing phase, each point is assigned an anomaly score based on the point distance from the principal components. Specifically, if column vectors $U_1 \dots U_N$ are principal components corresponding to eigenvalues $\lambda_1 \leq \cdots \leq \lambda_n$ of training data covariance matrix $\boldsymbol{\Sigma}$, and $\boldsymbol{\mu}$ is the mean vector of the training data, the anomaly score of a point $\boldsymbol{x} = [x_1\ x_2 \dots x_N]^{\boldsymbol{T}}$ is given by Equation (3.22):

$$S(\boldsymbol{x}) = \sum_{\boldsymbol{i=1}}^{q} \frac{((U_{\boldsymbol{i}})^{\boldsymbol{T}}(\boldsymbol{x} - \boldsymbol{\mu}))^2}{\lambda_i}, q \leq N \tag{3.22}$$

If the assumption that typical data $\boldsymbol{X}$ is a Gaussian random vector, it can be verified that $\frac{U_i^T(X-\mu)}{\sqrt{\lambda_i}}$ is a standard normal distribution. Thus, *S(**X**)* has the chi-squared distribution of q degrees of freedom, $\chi^2(q)$, by definition in Equation (3.12). As a result, by applying Equation (3.17), a point $\boldsymbol{x}$ is anomalous if

$$S(x) \geq \chi_q^2(p) \tag{3.23}$$

In practice, the mean vector and covariance matrix of the training data with *M* samples are estimated by a maximum-likelihood estimator as shown in Equations (3.24) and (3.25) [20].

$$\boldsymbol{\mu} = \frac{1}{M} \sum_{k=1}^{M} x_k \tag{3.24}$$

$$\boldsymbol{\Sigma} = \frac{1}{M} \sum_{k=1}^{M} (\boldsymbol{x_k} - \boldsymbol{\mu})((\boldsymbol{x_k} - \boldsymbol{\mu})^T \tag{3.25}$$

3.6.2 One-class SVM

Given a data set $D = \{\boldsymbol{x}_1 \dots \boldsymbol{x}_m\}, \boldsymbol{x}_i \in \boldsymbol{R^N}$ having *m* data points in *N*- dimensional space, and a transformation $\phi: R^N \to F$ which project the data into a feature space *F*, one-class SVM learns a decision boundary by maximizing the separation between the data points and the origin in the transformed space [34]. More precisely, the decision boundary has the form given by

$$g(x) = \boldsymbol{w}^T \phi(x) - p \tag{3.26}$$

where $\boldsymbol{w} \in \boldsymbol{R^N}$is weight vector and p is a bias term; the primary objective of one-class SVM is given by Equation (3.27)

$$\min_{\boldsymbol{w},\xi,p} \frac{\left|\left|\boldsymbol{w}\right|\right|^2}{2} - p + \frac{1}{vm}\sum_{i=1}^{m} \xi_i \tag{3.27}$$

$$\text{subject to}: \boldsymbol{w}^T \phi(\boldsymbol{x}_i) \geq p - \xi_i, \xi_i \geq 0$$

where ξ_i is a slack variable for point x_i; the slack variable provides a relaxation for a point which can lie outside of the decision boundary. p is the distance from the decision boundary to the origin in the feature space, and v is the upper bound of the fraction of the anomalies in the data set [34]. The optimization in Equation (3.27) is transformed into its dual form as given by [33]

$$\min_{\lambda} \frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{m} \lambda_i k(\boldsymbol{x_i}, \boldsymbol{x_j})\lambda_j \tag{3.28}$$

$$\text{subject to: } 0 \leq \lambda_i \leq \frac{1}{mv}, \sum_{i=1}^{m} \lambda_i = 1$$

where $k(\boldsymbol{x_i}, \boldsymbol{x_j}) = \phi(\boldsymbol{x_i}) \cdot \phi(\boldsymbol{x_2})$ is the dot product of points x_i, x_j in the feature space, and decision boundary becomes [33]

$$g(\boldsymbol{x}) = \sum_{i=1}^{m} \lambda_i k(\boldsymbol{x_i}, \boldsymbol{x}) - p \tag{3.29}$$

Note that from Equations (3.28) and (3.29), only the kernel $k(\boldsymbol{x_i}, \boldsymbol{x_j})$ needs to be defined without knowing the transformation $\phi(\boldsymbol{x})$ explicitly. In addition, the Gaussian kernel as defined in Equation (3.30) is used to guarantee the existence of the decision boundary *g(**x**)* because data transformed with Gaussian kernel lies in same quadrant; given any two points in the data set, its Gaussian kernel output is non-negative.

$$k(\boldsymbol{x_i}, \boldsymbol{x}) = \exp\left(-\frac{\left|\left|\boldsymbol{x_i} - \boldsymbol{x}\right|\right|^2}{2\sigma^2}\right) \tag{3.30}$$

where σ is the parameter of the kernel function. Small values of σ could lead to overtraining (memorizing training set ***D)*** and many support vectors which are points $\boldsymbol{x}$ such that g($\boldsymbol{x}$) equals to one, while large values of σ can ignore particular characteristics of the data set. For implementation, we will start with $\sigma = 0.01$ and increase it gradually until the number of support vectors does not decline significantly [34].

In one-class SVM *g(**x**)* can be used as the anomaly score; if a point ***x*** has its g(x) value is close to zero or negative, it is potentially an anomaly.

Finally, MATLAB machine learning and statistical toolbox [38] is used for solving the Equation (3.28) with the Gaussian kernel in order to implement one-class SVM.

3.6.3 Replicator neural network

The replicator neural network was first introduced by Hawkins et al. [36]. The network structure includes three hidden layers between the input and output layers. In addition, the output layer has the same size as the input layer. However, Ciesielski and Ha [39] later discovered that using one hidden layer can provide equivalent performance. Therefore, we will explore the replicator neural network with one hidden layer as shown in Figure 3.7.

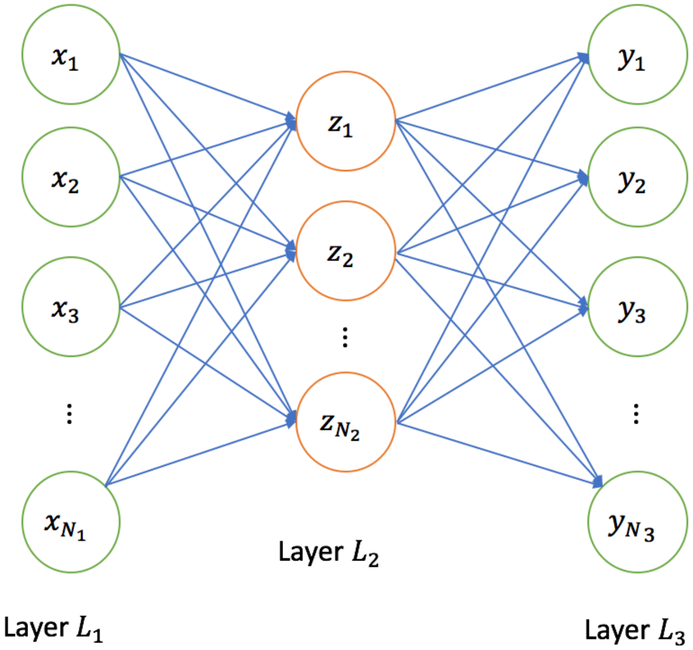

Figure 3.7 A fully connected replicator network with one hidden layer

Let $w_{ij}^{(l)}$ be the weight that joins the input node *i* of layer *l* and output node *j* of layer *l*+*1*, and let N_l be number of nodes in the layer *l.* The values of the hidden layer and output layer are defined in Equation (3.31) and (3.32) respectively as follows:

$$z_j = a\left(\sum_{i=1}^{N_1} w_{ij}^{(1)} x_i + b_j^{(1)}\right) \tag{3.31}$$

$$y_j = a\left(\sum_{i=1}^{N_2} w_{ij}^{(2)} z_i + b_j^{(2)}\right) \tag{3.32}$$

where $b_j^{(l)}$ denotes the bias term for node j in layer l, and $a\,(\cdot)$ denotes an activation function applied to the hidden layer. In RNN, a sigmoid function as shown in equation (3.33) is used as the activation function [37].

$$a(x) = \frac{1}{1 + e^{-x}} \tag{3.33}$$

During the training phase, given a training set $D = \{\boldsymbol{x}^{(1)} \ldots\ \boldsymbol{x}^{(N)}\}$, $\boldsymbol{x}^{(j)} \in \boldsymbol{R}_{\boldsymbol{N_1}}$, the network parameters are derived by minimizing overall the least-squared error as shown in Equation (3.34):

$$\min_{w_{ij}^{(1)}, w_{ij}^{2}, b_j^{(1)}, b_j^{(2)}} \sum_{j=1}^{N} \sum_{i=1}^{N_1} (x_i^{(j)} - y_i^{(j)})^2 \tag{3.34}$$

Note that the objective function in Equation (3.34) is non-convex because of the sigmoid activation function; therefore, numerical methods for optimizing this objective function only guarantee a local minimum, but not the global minima.

In the testing phrase, if the error, in Equation (3.35), between the input $\boldsymbol{x} = [x_1 \ldots x_{N_1}]^{\boldsymbol{T}}$ and its reconstruction $\boldsymbol{y} = [y_1 \ldots y_{N_1}]$ through the RNN is larger than some threshold, the point $\boldsymbol{x}$ is declared as an anomaly.

$$e(\boldsymbol{x}) = \sum_{i=1}^{N_1} (x_i - y_i)^2 \tag{3.35}$$

For implementation of the RNN algorithm, we select the MATLAB neural network toolbox [40] to solve Equation (3.24). Furthermore, the number of nodes in the hidden layer in the RNN is selected empirically for best performance given a specific application or training set.

3.7 Performance Measures

Precision, recall and F-measure are common methods for preformance evaluation of an anomaly detection algorithm [41]. Given a data set D, which has M instances of true anomalies, an anomaly detection algorithm can declare K instances of D as anomalies, but only T instances $(T \leq K, T \leq M)$ belong to the true anomaly group. As a result, the precision, recall and F-measure of the alogrithm are defined respectively as

$$Precision = \frac{T}{K} \tag{3.36}$$

$$Recall = \frac{T}{M} \tag{3.37}$$

$$F_{measure} = 2 \cdot \frac{Recall * Precision}{Recall + \ Precision} \tag{3.38}$$

Intuitively, high precision means the algorithm detects more relevant than irrelevant intances, while high recall refers to the alogrithm's ability to return most of the relevant instances. On the other hand, the F-measure is an attempt to combine precsion and recall into a single number, where a larger number represents the better performance.

4. The Collected Data

In the experiment, we are interested in abnormal acoustic scenes instead of transient events. In the context of street noise, acoustic scenes are selected as 10 seconds of audio data. Therefore, the observed data set $\boldsymbol{D} = \{\boldsymbol{x_1}, \dots \boldsymbol{x_N}\}, \boldsymbol{x_i} \in \boldsymbol{R^8}$ is a sequence of non-overlapped 10-second-averages of octave-band noise levels. The first octave band cuts off at 62.5 Hz and 125 Hz while the last one starts at 8 kHz and ends at 16 kHz. The data were collected by setting up an omnidirectional microphone facing a section of a one-way street in a residential area. In this experiment, the raw data are kept for performance evaluation in Chapter 6. In addition, in the context of this thesis, given a measurement $\boldsymbol{x_i}$, its noise intensity or noise level is referred to as

$$y_i = \left|\left|\boldsymbol{x_i}\right|\right|^2 = \sqrt{x_{1i}^2 + \cdots + x_{8i}^2},\ \boldsymbol{x_i} = [x_{1i} \dots x_{8i}]^T \qquad (4.1)$$

Given that the collected data started at 7pm, by inspecting the noise-level patterns plotted in Figure 4.1 and Figure 4.2, one can observe the repetitive patterns which strongly relate to human activities; the noise level is generally low during nighttime with fewer variations than during daytime when the level goes up and varies more. Furthermore, in this situation, the noise level is mainly influenced by the amount of vehicle traffic at the area of recording.

Analysis of the collected data suggests that the generative model of the data measurements is nonstationary. Figure 4.3 presents the histogram of the noise intensities, while Figure 4.4 depicts the histograms of noise intensities grouped by hours. It can be clearly seen that the distributions in Figure 4.4 are very different from the overall distribution in Figure 4.3. Therefore, a time-varying process is required for modeling data generation.

Furthermore, the distribution of values in each band is also time-varying. The clues can be seen in Figures 4.5 and 4.6 which show histograms of noise level in the frequency band from 250 to 500 Hz and the frequency band from 500 to 1000 Hz, respectively. The means and variances of these distributions tend to decrease during nighttime and early morning while they increase from early morning to midday.

Lastly, the acquired measurements strongly relate to human activities in the surroundings. These activities are generally subject to schedules such as kids going to school at 8-9 am, people going to work from 9am-10am, etc. Hence, the data collected at a given hour can be assumed to be independent from that collected at other hours, thereby leading to the proposed method of modeling typical data patterns presented in Chapter 5.

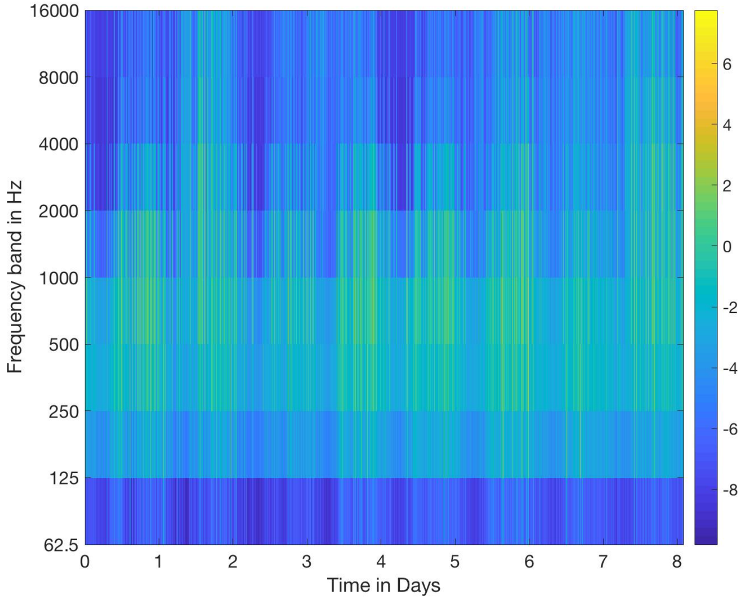

Figure 4.1 The collected octave-band log-normed vector sequence for eight consecutive days. Note that the value of each element of the octave band vector is in log scale; the starting time of the collected data is 7pm.

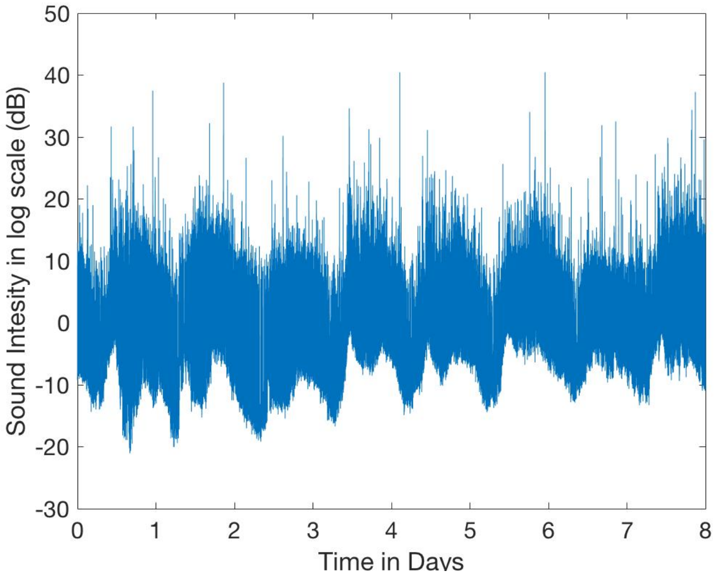

Figure 4.2 The noise intensity corresponding to collected octave-band vectors; the starting time of the collected data is 7pm.

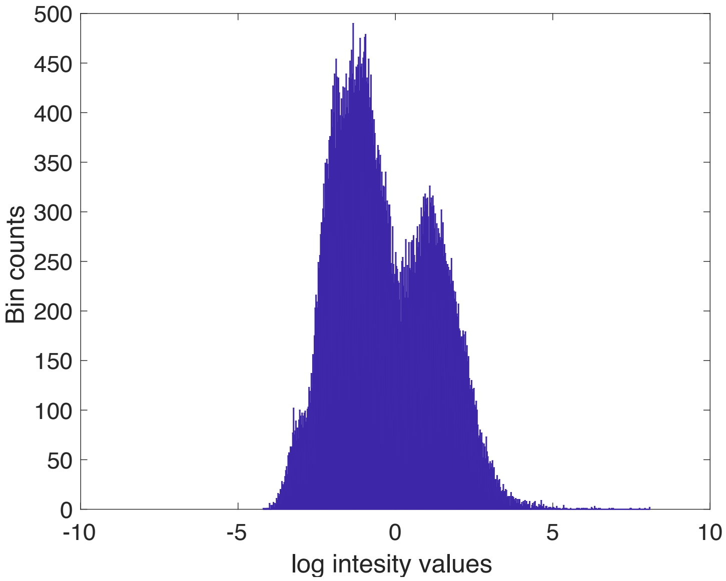

Figure 4.3 The histogram of log noise level intensity values of the collected data. These data appear to be generated by a tri-modal distribution.

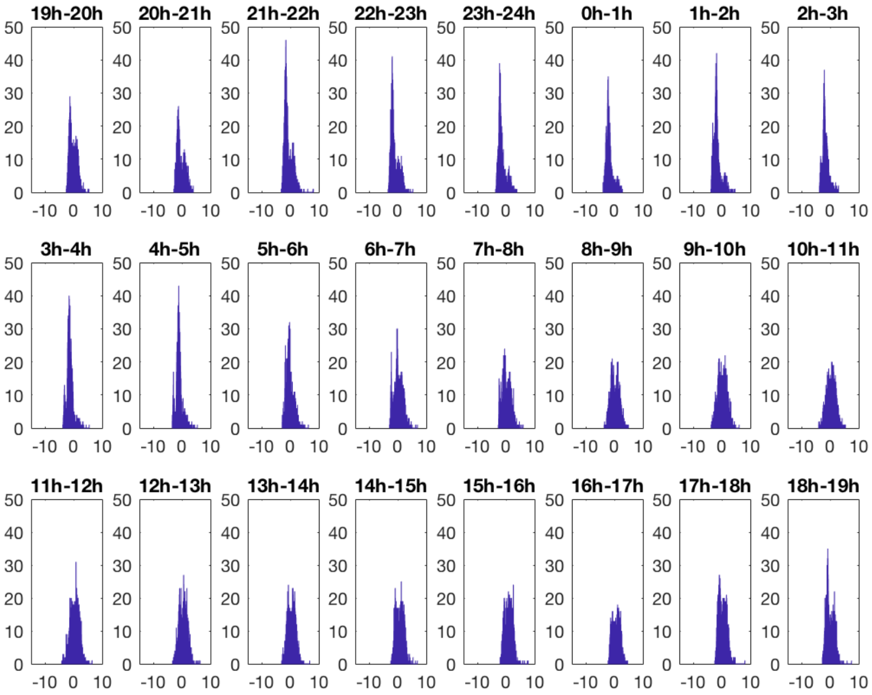

Figure 4.4 The histogram of log noise level values of the collected data grouped by hours; the first block is from 19h to 20h in the top left plot. These data can be closely approximated by a single log-normal distribution for each hour.

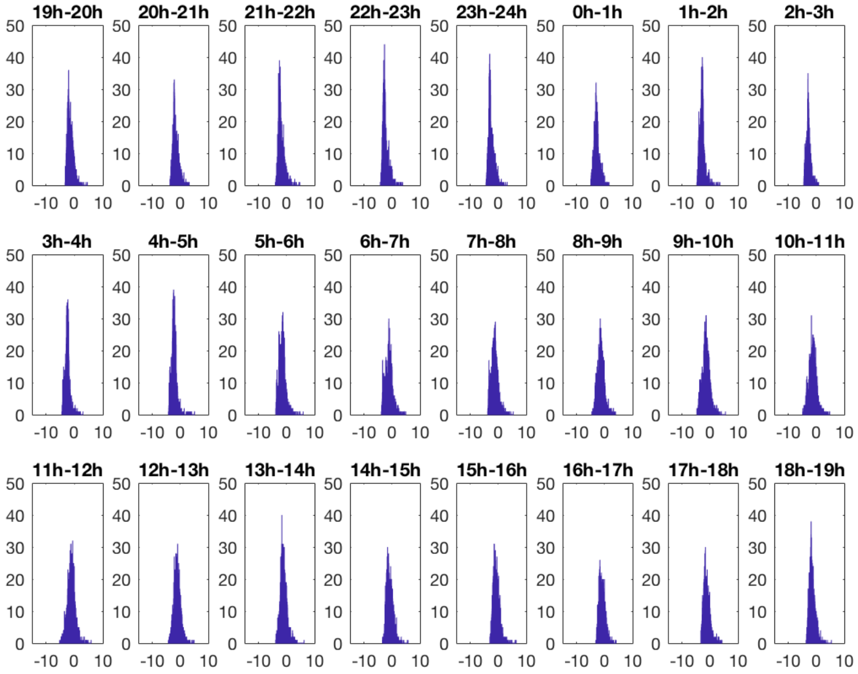

Figure 4.5 The histograms of log noise-level at the frequency band from 250 to 500 Hz grouped by hours; the first block is from 19h to 20h in the top left plot.

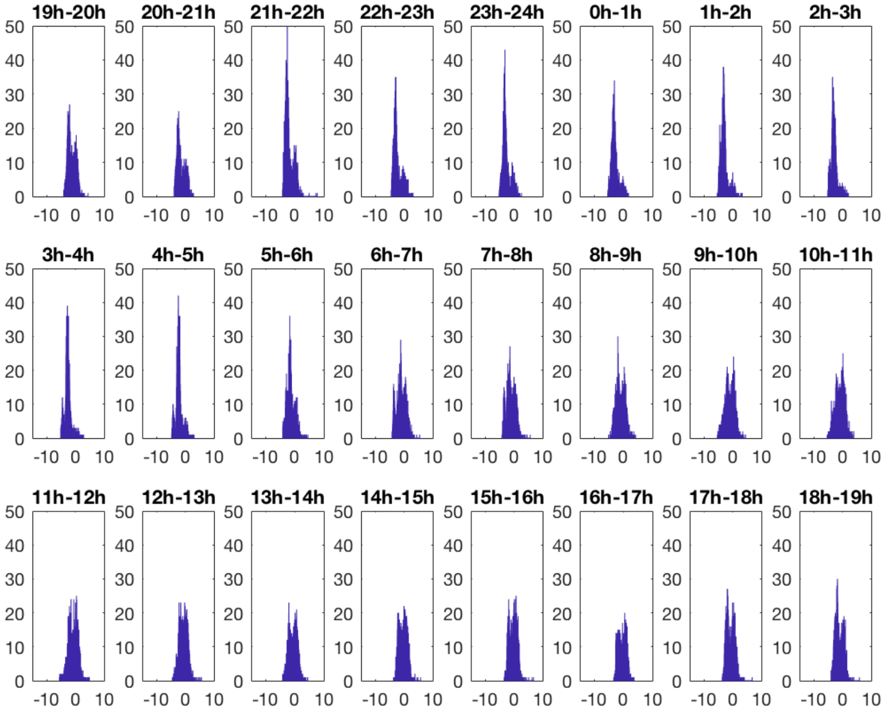

Figure 4.6 The histograms of log noise-level at the frequency band from 500 to 1000 Hz grouped by hours; the first block is from 19h to 20h in the top left plot.

5. Proposed Method

As shown later in Chapter 6, given the collected measurements, in unsupervised learning setting, one-class SVM does not perform as well as RNN or PCA. In addition, the boundary decision of the one-class SVM is sensitive to anomalies in the evaluated data set thereby leading to lower performance than the other techniques. On the other hand, PCA-based anomaly detection and RNN perform similarly in this case; however, applying these two techniques directly does not exploit the daily pattern and the changes in distribution over a day as pointed out in the previous chapter. Therefore, one could suggest that the collected data is split into hours before applying either PCA based anomaly detection or RNN independently for each hour.

Noting that many histogram instances, especially from morning to late afternoon hours (from 7am to 6 pm), show signatures of Gaussian distributions for both noise intensities in Figure 4.4 and octave-band measurements in Figures 4.5 and 4.6. As a consequence, approximating the generative model of observed data at each hour of a day by a Gaussian distribution is very natural. Furthermore, RNN with its modeling flexibility is not guaranteed to capture full characteristics of a Gaussian distributed data set, because numerical solvers only return a local minima solution for its objective function; Section 3.6.3 notes that the objective function are non-convex. Thus, extension of PCA-based anomaly detection with its probabilistic interpretation provides a more robust and stable solution. Indeed, the evaluations in Chapter 6 also show that applying PCA-based techniques independently for each hour of the collected data provides better performance than using RNN in a similar setting.

For the above mentioned reasons, we proposed the following method. At a given time n, $\boldsymbol{X}(n) = [\boldsymbol{X}_{\boldsymbol{1}}(n) \dots \boldsymbol{X}_{\boldsymbol{8}}(n)]^{\boldsymbol{T}}$ is drawn from a log-normal distribution such that

$$\boldsymbol{Z}(\boldsymbol{n}) = \log(\boldsymbol{X}(n)) \sim N(\boldsymbol{\mu}(n), \boldsymbol{\Sigma}(n)) \tag{5.1}$$

where $N(\boldsymbol{\mu}(n), \boldsymbol{\Sigma}(n))$ is a multivariate Gaussian with time-varying mean vector, $\boldsymbol{\mu}(n)$**,** and covariance matrix $\boldsymbol{\Sigma}(n)$. In this model, $\boldsymbol{\mu}(n)$ and $\boldsymbol{\Sigma}(n)$ are approximated as piecewise constant and periodic over 24 hours; their values only change over each hour. There are three reasons which lead to this approximation. First, human activities are subject to schedule, which probably causes the daily patterns as shown in Figure 4.2. Second, if one takes a closer look at a particular day, one sees a slow variation of the underlying statistic, even nearly stationary for some hours as shown in Figure 5.1 for noise-level intensities or Figure 5.2 for intensities at a particular band in octave-band measurements. Last but foremost, by letting $\boldsymbol{\mu}(n)$ and $\boldsymbol{\Sigma}(n)$ be piecewise constant, the complexity of the model is reduced significantly because within an hour, $\boldsymbol{\mu}(n)$ and $\boldsymbol{\Sigma}(n)$ become constant and can be independently estimated by Equations (3.24) and (3.25) given the data samples which belong to that hour. In addition, a piecewise constant function provides some degrees of freedom to model a time-varying function. The quicker a piecewise constant function can change values, the closer it approximates a target function; therefore, if a larger set of data could be collected, our model parameters, $\boldsymbol{\mu}(n)$ and $\boldsymbol{\Sigma}(n)$, can be varied much faster, such as every 10 minutes, while they are reliably estimated.

After assuming the generative model as a stationary Gaussian distribution within an hour block, the anomaly detection technique described in Section 3.4 can be applied independently for each hour. In addition, to avoid small noise intensities or quiet points being marked as anomalies, the noise intensity of an anomaly, $\boldsymbol{x}$(n), has to satisfy Equation (5.2) as well as Equation (3.23)

$$\boldsymbol{x}^{\boldsymbol{T}}(n) \ \cdot \boldsymbol{x}(n) \geq E[\boldsymbol{x}^{\boldsymbol{T}}(n) \ \cdot \boldsymbol{x}(n)] \tag{5.2}$$

where

$$E[\boldsymbol{x}^{\boldsymbol{T}}(n) \ \cdot \boldsymbol{x}(n)] = \frac{1}{M} \sum_{samples\ in\ hour\ j} \boldsymbol{x}^{\boldsymbol{T}}(k) \cdot \boldsymbol{x}(k)$$

if *n* belongs to hour *j* and *M* is the number of samples collected in hour *j*.

In summary, if column vectors $U_1(n)$ ... $U_N(n)$ are principal components corresponding to eigenvalues $\lambda_1(n) \leq \cdots \leq \lambda_n$(n) of covariance matrix $\boldsymbol{\Sigma}$**(**n**)**, the anomaly score of a point $\boldsymbol{x}(n) = [x_1(n)\ x_2(n) \ldots x_8(n)]^{\boldsymbol{T}}$ is given by Equation (5.3).

$$S(\boldsymbol{x}(n)) = \sum_{\boldsymbol{i=1}}^{q} \frac{\left[\left(U_i(\boldsymbol{n})\right)^{\boldsymbol{T}} \left(\log\left(\boldsymbol{x}(\boldsymbol{n})\right) - \boldsymbol{\mu}(\boldsymbol{n})\right)\right]^{\boldsymbol{2}}}{\lambda_i(n)}, q \leq N \tag{5.3}$$

$\boldsymbol{x}$(n) is an anomaly if the following condition in Equation (5.4) is valid:

$$\mathbf{1}\{\boldsymbol{x}^{\boldsymbol{T}}(n)\ ) \cdot \boldsymbol{x}(\boldsymbol{n})) \geq E[\boldsymbol{x}^{\boldsymbol{T}}(n) \ \cdot \boldsymbol{x}(n)]\} \cdot S(x) \geq \chi_q^2(p) \tag{5.4}$$

and given *n* within hour *j* across training data, $\boldsymbol{\mu}(n)$ and $\boldsymbol{\Sigma}(n)$ are estimated as

$$\boldsymbol{\mu}(\boldsymbol{n}) = \frac{1}{M} \sum_{samples\ in\ hour\ j} log(\boldsymbol{x_k}) \tag{5.5}$$

$$\boldsymbol{\Sigma}(n) = \frac{1}{M} \sum_{samples\ in\ hour\ j} (log(\boldsymbol{x}_k) - \boldsymbol{\mu}(\boldsymbol{n}))\big((log(\boldsymbol{x}_k) - \boldsymbol{\mu}(\boldsymbol{n})\big)^T \tag{5.6}$$

where M is the number of samples belonging to hour *j*. Note that $U_1(n) \ldots U_N(n)$ and $\lambda_1(n) \ldots \lambda_2(n)$ are also piecewise constant.

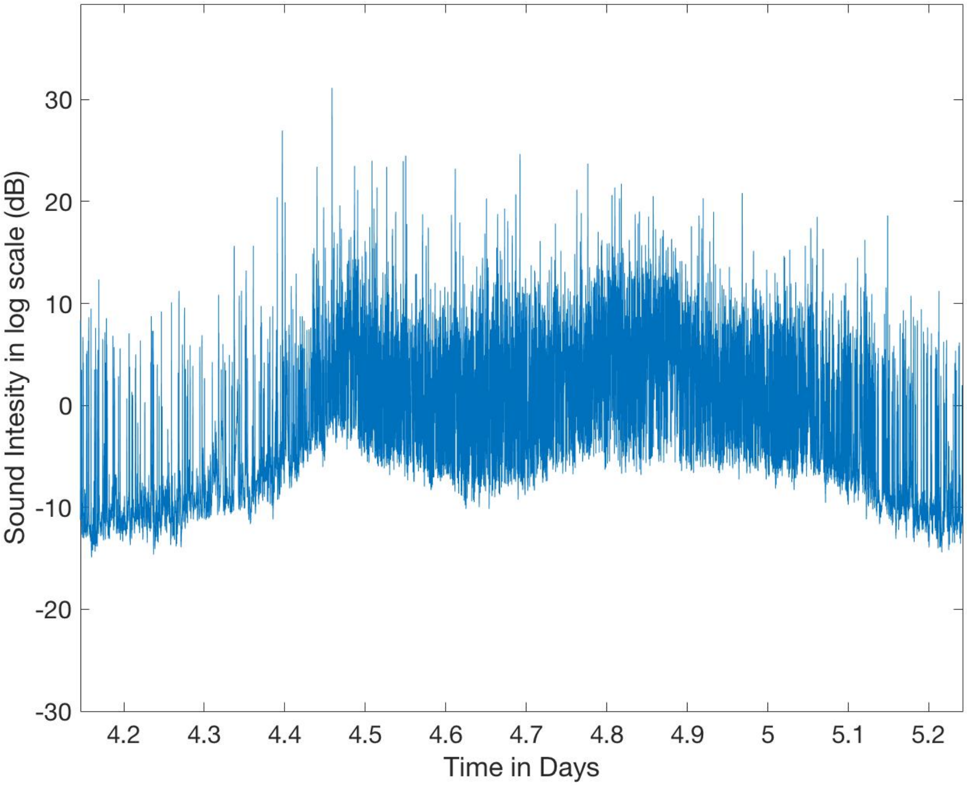

Figure 5.1 Sound intensity measures over one day. From 4.5 to 4.8, data samples seem to be generated by a stationary statistic.

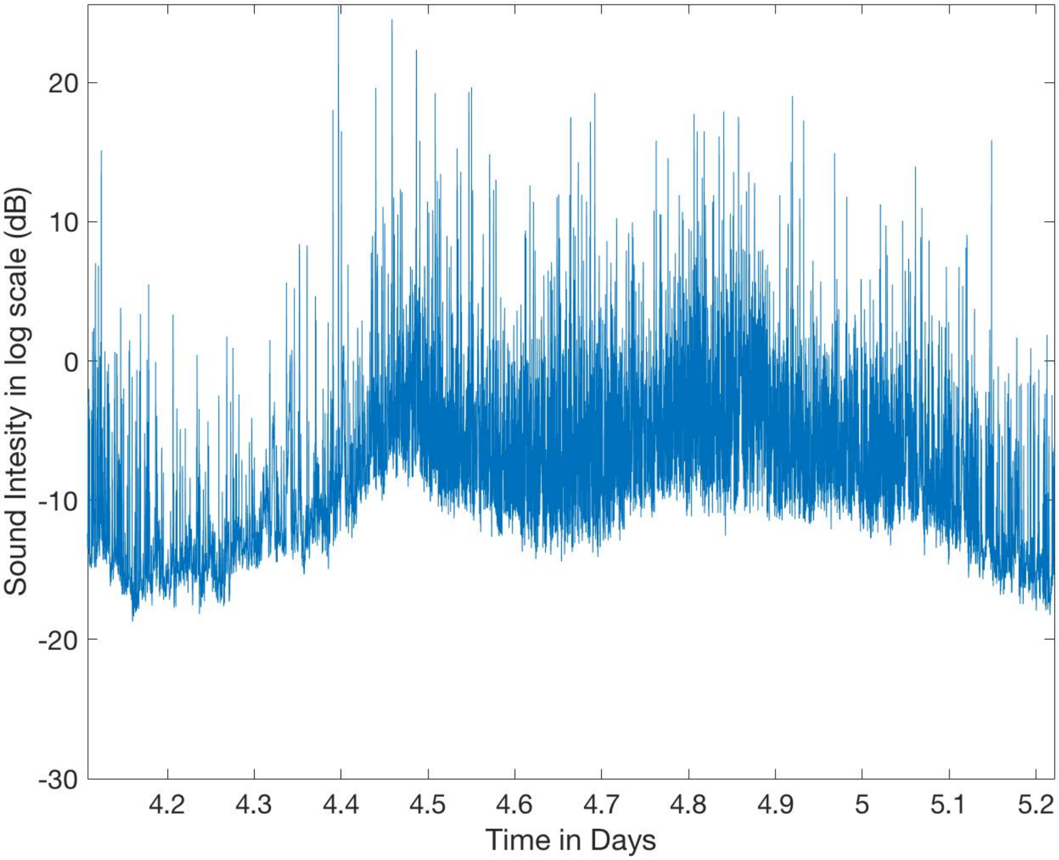

Figure 5.2 Intensities at the band from 250 to 500 Hz in octave measurements. From 4.5 to 4.8, data points seem to be generated by a stationary statistic.

6. Evaluation

The performance of the proposed method is evaluated by using the eight-day continuously recorded data which contains labelled anomalies including ambulances, police cars, loud vehicles and other unexpected sounds. Extra police and ambulance vehicle sounds are also injected from the SONYC data set [42], [43]. We believe residents will regard acoustic scenes having these types of sound as abnormal and unpleasant moments; police and ambulance sirens, especially, can strongly correlate to accidents or crimes happening in the residential area.

Each sound clip from the SONYC data set is sampled at 16 kHz, so its corresponding octave-band values cannot include the measurement from 8 kHz to 16 kHz. Therefore, the injected sound needs to go through an up-sampling-by-two operation [18] before being added at random temporal positions to the recorded data. In order to simulate the settings of the recorded data, the injected anomalies are also scaled by the ratio of the average energy of ambulance and police siren sounds in the recordings to that of the injected data.

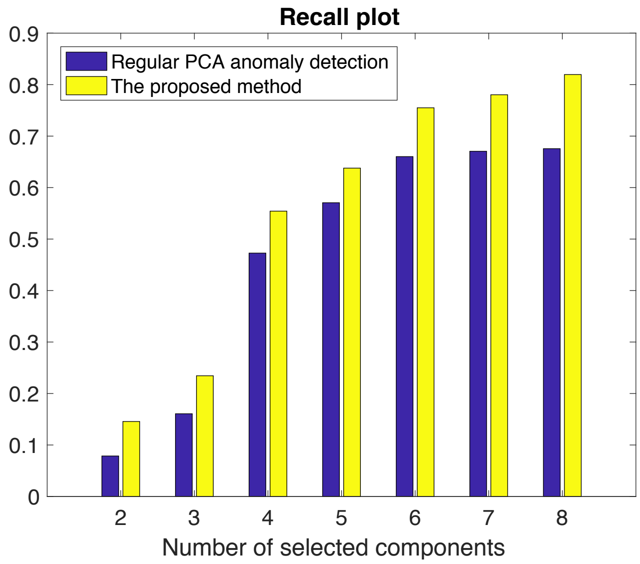

Figure 6.1 Recall measures for regular PCA anomaly detection and the proposed method for various numbers of principal components selected.

Detection performance is presented by precision, recall and F-measure for different numbers of principal components. For comparison, the performance of the proposed algorithm is compared to the traditional PCA-based anomaly detection presented in Section 3.6.1. Note that the p-value in Equation (3.23) is set to 0.982 for evaluations.

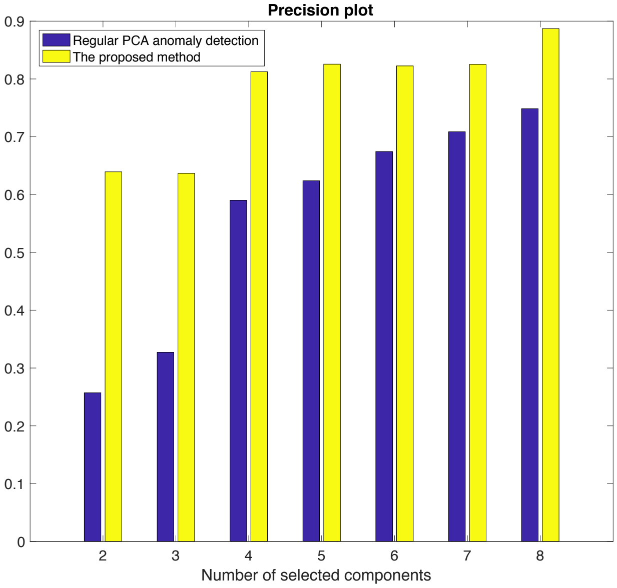

Figure 6.2 Precision measures for regular PCA anomaly detection and the proposed method for various numbers of principal components selected.

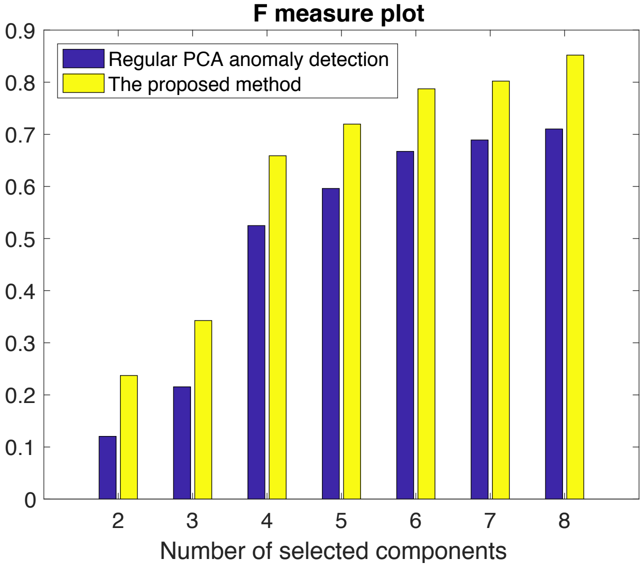

Figure 6.3 F-measures for regular PCA anomaly detection and the proposed method for various numbers of principal components selected.

Figures 6.1, 6.2, and 6.3 show evaluation results. First, the more principal components selected, the better the performances of all methods, as shown in Figure 6.3. Recall that the performance of the proposed method increases gradually as the number of selected components increases, while the corresponding precision measures slightly fluctuate around 0.82 beyond five principal components selected. The proposed method outperforms the regular PCA-based anomaly detection by at least 10% on all measures for a given number of principal components, and the best performance occurs when all principal components are selected.

In comparison with other anomaly detection techniques presented in Section 3.6, one-class SVM is calibrated for the given data set with the parameter for Gaussian kernel in Equation (3.30) $\sigma = 1.242$ and the model parameter in Equation (3.27) $v = 0.01$. Replicator neural network (RNN) with five hidden nodes is applied to the whole training set while RNN with six hidden nodes is applied to each hour of the training set; the size of the hidden layer in the RNN are selected by adding hidden nodes until the performance does not increase significantly. Note that the criterion in Equation (5.5) for avoiding marking measurements with low energy as anomalies is applied to all implemented algorithms for fair comparison.

Table 6.1. The Performance comparison of the proposed method against one-class SVM, RNN, and regular PCA-based anomaly detection algorithms

Method	Precision	Recall	F-measure
One-class support vector machine (SVM)	0.59	0.53	0.56
Replicator neural network (RNN)	0.75	0.66	0.70
Regular PCA-based anomaly detection	0.75	0.67	0.71
RNN trained for each hour	0.79	0.82	0.80
The proposed Method	0.89	0.81	0.84

Table 6.1 shows the best performance result we could achieve for each algorithm. One-class SVM has the lowest performance on the given data set because when anomalies are present in the training data, the learned discriminative boundary covers them and causes performance to degrade. RNN and regular PCA-based anomaly detection have similar performances and their performances are significantly better than the one-class SVM technique; however, they still fall behind our proposed techniques by more than 10% in all measures. Furthermore, when RNN is applied independently for each hour, its performance improves significantly and is only 4% lower than the performance of the proposed method in F-measure. The difference between the performance of RNN trained for each hour and the proposed method can be explained by the fact that when data is split into hours for modeling the generative process, the underlying distribution closely approximates a Gaussian distribution as shown in Chapter 3, and PCA-based anomaly

detection is the optimal method if the training data are normally distributed. Furthermore, RNN can model flexible distributions, but only local minima of its RNN can be returned by numerical solvers.

For completeness, receiver operating characteristic (ROC) curves [44] of the algorithms in Table 6.1 are plotted in Figure 6.4; the ROC curve of the proposed technique is above all the ROC curves of the other algorithms when the probability of false alarm is less than 0.25, while the ROC curve of one-class SVM is the lowest. The ROC curves of regular PCA-based anomaly detection and RNN applied to the whole data set are very similar. The ROC curves also show that RNN performance improves significantly when it is applied independently into each hour of the evaluated data set. Note that in Figure 6.4, the probability of false alarm is equal to subtracting the precision measure in Equation (3.26) from one, while the probability of detection is equivalent to the recall measure defined in Equation (3.27).

As a closing remark, given that 0.6% of the evaluation data set of the continuous eight-day record are anomalies, the precision of 0.89 means only 0.066% of the data instances are false alarms or triggered incorrectly as anomalies as shown in Equation 6.1. Thus, the proposed algorithm has a false-alarm rate of roughly four times per hour and the capability of discovering 81% of anomalies presented in the evaluation set.

$$0.006 \cdot (1 - precision) \ \cdot 100\% = 0.066\% \tag{6.1}$$

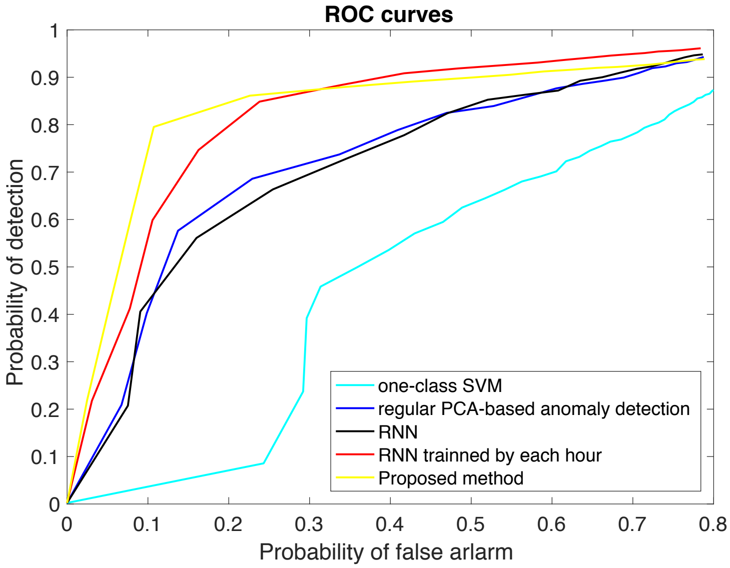

Figure 6.4 Receiver operating characteristic curves for the algorithm in Table 6.1

7. Conclusion and Further Work

Directly applying traditional anomaly detection algorithms such as the PCA-based technique, one-class SVM, and RNN do not provide the best solution for SPL type measurements in the problem of environmental noise monitoring in a residential area as presented in Table 6.1 and Figure 6.4. When modifying the original models in order to exploit the daily patterns and the non-stationarity of the experimental data, the performances increase significantly. For example, RNN trained by each hour outperforms its original model by 10 % percent in F-measures, and the proposed extension of regular PCA anomaly detection boosts the performance up by 14% in F-measures.

In fact, when data is normally distributed, PCA-based anomaly detection provides the optimal solution for anomaly detection, and the histogram of the observed data at each hour can be closely approximated by Gaussian distributions thereby leading to the introduction of time varying models for mean and variances over one-hour intervals in the proposed method. In other words, the presented technique treats the collected data in log scale at each hour independently and models its generation by a multivariate Gaussian and detects anomalies accordingly. Despite the simplicity of the proposed method, it can reach 0.85 F-measure with 0.83 recall and 0.89 precision without dimension reduction, and outperform regular PCA-based anomaly detection, RNN and one-class SVM. In addition, the technique is suitable for high-dimensional data because it only adds extra parameters linearly with increasing dimension of the data, thereby reducing the number of samples required for parameter estimation.

Besides an anomaly detection method, this thesis also introduces a practical setup for real-world environmental noise monitoring. By collecting 10-second average octave band noise level, the required bandwidth and memory for data storage and communication are reduced significantly, while the low resolution of the collected data strengthens privacy protection. Therefore, the system can potentially be deployed and scaled easily in residential areas without legal restrictions by consuming battery and solar power. Furthermore, this thesis shows that octave-band measurements provide sufficient information for detecting unknown anomalies which may include interesting sound events such as police and ambulance sirens, and large-vehicle engines.

In future work, data from various residential areas need to be collected for testing with the proposed techniques. In addition, a question of interest is how the proposed technique can work or be extended if multiple microphones could be deployed in a given area.

References

[1] Seidman, Michael D., and Robert T. Standring. "Noise and quality of life." *International Journal of Environmental Research and Public Health* 7.10 (2010): 3730-3738.

[2] Schreckenberg, Dirk, et al. "Aircraft noise and quality of life around Frankfurt Airport." *International Journal of Environmental Research and Public Health* 7.9 (2010): 3382-3405.

[3] Couvreur, Christophe, et al. "Automatic classification of environmental noise events by hidden Markov models." *Applied Acoustics* 54.3 (1998): 187-206.

[4] Tsai, Kang-Ting, Min-Der Lin, and Yen-Hua Chen. "Noise mapping in urban environments: A Taiwan study." *Applied Acoustics* 70.7 (2009): 964-972.

[5] Mydlarz, Charlie, Justin Salamon, and Juan Pablo Bello. "The implementation of low-cost urban acoustic monitoring devices." *Applied Acoustics* 117 (2017): 207-218.

[6] Salamon, Justin, and Juan Pablo Bello. "Feature learning with deep scattering for urban sound analysis." *Signal Processing Conference (EUSIPCO), 2015 23rd European*. IEEE, 2015.

[7] American National Standards Institute. ANSI S1.43-1997 (r 2007), Specifications for Integrating-Averaging Sound Level Meters. New York: American National Standards Institute; 2007

[8] Salamon, Justin, and Juan Pablo Bello. "Deep convolutional neural networks and data augmentation for environmental sound classification." *IEEE Signal Processing Letters* 24.3 (2017): 279-283.

[9] Salamon, Justin, and Juan Pablo Bello. "Unsupervised feature learning for urban sound classification." *Acoustics, Speech and Signal Processing (ICASSP), 2015 IEEE International Conference on*. IEEE, 2015.

[10] Valenzise, Giuseppe, et al. "Scream and gunshot detection and localization for audio-surveillance systems." *Advanced Video and Signal Based Surveillance, 2007. AVSS 2007. IEEE Conference on*. IEEE, 2007.

[11] Bisot, Victor, et al. "Acoustic scene classification with matrix factorization for unsupervised feature learning." *Acoustics, Speech and Signal Processing (ICASSP), 2016 IEEE International Conference on*. IEEE, 2016.

[12] Ntalampiras, Stavros, Ilyas Potamitis, and Nikos Fakotakis. "Probabilistic novelty detection for acoustic surveillance under real-world conditions." *IEEE Transactions on Multimedia* 13.4 (2011): 713-719.

[13] Chakrabarty, Debmalya, and Mounya Elhilali. "Abnormal sound event detection using temporal trajectories mixtures." *Acoustics, Speech and Signal Processing (ICASSP), 2016 IEEE International Conference on*. IEEE, 2016.

[14] Mydlarz, Charlie, et al. "The design and calibration of low cost urban acoustic sensing devices." Euronoise, 2015.

[15] Mydlarz, Charlie, Justin Salamon, and Juan Pablo Bello. "The implementation of low-cost urban acoustic monitoring devices." *Applied Acoustics* 117 (2017): 207-218.

[16] International Electrotechnical Commission. "Electroacoustics-Sound level meters-Part 1: Specifications (IEC 61672-1: 2002)." (2003): 61672-1.

[17] Chandola, Varun, Arindam Banerjee, and Vipin Kumar. "Anomaly detection: A survey." *ACM Computing Surveys (CSUR)* 41.3 (2009): 15.

[18] Oppenheim, Alan V., and W. Schafer Ronald. *Discrete-time Signal Processing.* New Jersey, Prentice Hall Inc., 1989.

[19] Veggeberg, Kurt. "Octave analysis explored: a tutorial." *EE-Evaluation Engineering* 47.8 (2008): 40-44.

[20] Hajek, Bruce. *Random Processes for Engineers*. Cambridge University Press, 2015

[21] Bishop, Christopher M. *Pattern Recognition and Machine Learning*. Springer, 2006.

[22] Strang, Gilbert, and Kai Borre. *Linear Algebra, Geodesy, and GPS*. Siam, 1997.

[23] Johnson, Norman L., Samuel Kotz, and N. Balakrishnan. *Continuous Univariate Distributions (Vol. 1).* Wiley & Sons, 1995

[24] Tarmast, Ghasem. "Multivariate log-normal distribution." *Proceedings of 53rd Session of International Statistical Institute* (2001).

[25] O'Connor, Patrick, and Andre Kleyner. *Practical Reliability Engineering*. John Wiley & Sons, 2012.

[26] Mizuseki, Kenji, and György Buzsáki. "Preconfigured, skewed distribution of firing rates in the hippocampus and entorhinal cortex." *Cell reports* 4.5 (2013): 1010-1021.

[27] Buzsáki, György, and Kenji Mizuseki. "The log-dynamic brain: how skewed distributions affect network operations." *Nature Reviews. Neuroscience* 15.4 (2014): 264.

[28] Sematech and N. I. S. T. Engineering Statistics Handbook. *NIST SEMATECH* (2006).

[29] Hotelling, Harold. "Analysis of a complex of statistical variables into principal components." *Journal of Educational Psychology* 24.6 (1933): 417-441.

[30] Smith, Lindsay I. "A tutorial on principal components analysis." *Cornell University, USA* 51.52 (2002): 65.

[31] Perdisci, Roberto, Guofei Gu, and Wenke Lee. "Using an ensemble of one-class svm classifiers to harden payload-based anomaly detection systems." *Data Mining, 2006. ICDM'06. Sixth International Conference on*. IEEE, 2006.

[32] Manevitz, Larry M., and Malik Yousef. "One-class SVMs for document classification." *Journal of Machine Learning Research* 2.Dec (2001): 139-154.

[33] Dreiseitl, Stephan, et al. "Outlier detection with one-class SVMs: an application to melanoma prognosis." *AMIA Annual Symposium Proceedings*. Vol. 2010. American Medical Informatics Association, 2010.

[34] Amer, Mennatallah, Markus Goldstein, and Slim Abdennadher. "Enhancing one-class support vector machines for unsupervised anomaly detection." *Proceedings of the ACM SIGKDD Workshop on Outlier Detection and Description*. ACM, 2013.

[35] Sabokrou, Mohammad, et al. "Deep-cascade: cascading 3d deep neural networks for fast anomaly detection and localization in crowded scenes." *IEEE Transactions on Image Processing* 26.4 (2017): 1992-2004.

[36] Hawkins, Simon, et al. "Outlier detection using replicator neural networks." *International Conference on Data Warehousing and Knowledge Discovery*. Springer, Berlin, Heidelberg, 2002.

[37] Tóth, László, and Gábor Gosztolya. "Replicator neural networks for outlier modeling in segmental speech recognition." *International Symposium on Neural Networks*. Springer, Berlin, Heidelberg, 2004.

[38] Statistics and Machine Learning Toolbox Documentation, web page. Available at https://www.mathworks.com/help/stats/index.html. Accessed January 2018

[39] Ciesielski, Vic, and Vinh Phuong Ha. "Texture detection using neural networks trained on examples of one class." *Australasian Joint Conference on Artificial Intelligence*. Springer, Berlin, Heidelberg, 2009.

[40] Neural Network Toolbox Documentation, web page. Available at https://www.mathworks.com/help/nnet/. Accessed January 2018

[41] Powers, David Martin. "Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation." *Journal of Machine Learning Technologies*, 2011

[42] Salamon, Justin, Christopher Jacoby, and Juan Pablo Bello. "A dataset and taxonomy for urban sound research." *Proceedings of the 22nd ACM International Conference on Multimedia*. ACM, 2014.

[43] SONYC: Sound of New York City, web page. Available at https://wp.nyu.edu/sonyc/. Accessed September 2017

[44] Fawcett, Tom. "An introduction to ROC analysis." *Pattern Recognition Letters* 27.8 (2006): 861-874.

.