Regionally optimised time-frequency distributions using finite mixture models

M.J. Coates and W.J. Fitzgerald
Signal Processing Laboratory, Department of Engineering,
University of Cambridge, Cambridge, UK

March 19, 1999

Abstract

A method is presented for the improvement of the resolution and clarity of bilinear time frequency distributions generated from signals consisting of a number of approximately time-frequency disjoint components. The method involves the determination of the parameters of a finite mixture of Gaussians, which is used to model an initial time-frequency distribution. The expectation-maximisation algorithm and the functional merging technique are used to derive the parameter set, including the number of Gaussians in the mixture. The mixture model indicates the number of (linear) components in the signal, and the regions they occupy in the time-frequency plane. This information is used to isolate the components, and smoothing kernels are designed using the properties of each isolated component. During the generation of the smoothing kernels, a set of basis functions is derived for each component, which describes the time-frequency region it occupies. This basis can be used for time-frequency filtering, enabling operations such as signal decomposition and noise reduction to be performed.

1 Introduction

Time-frequency distributions (TFDs) are important in the analysis of signals generated in a wide variety of environments. This paper describes a method to improve the resolution of time-frequency distributions for signals consisting of a number of approximately time-frequency disjoint components. The method involves the construction of a finite mixture model to represent the time-frequency energetic structure of a signal; this model indicates the number of time-frequency components in the signal and their locations in the time-frequency plane. Such information is useful both for generating an enhanced TFD, and also for applications such as signal decomposition, noise suppression and coherence measurements.

The TFDs discussed in this paper are those generated by applying smoothing filters to the Wigner distribution (WD). These distributions are defined as:

$$P_x(t,f) = \int_{t'} \int_{f'} \Phi_{P,t,f}(t',f') W_x(t-t', f-f')\, dt'\, df' \tag{1}$$

where W_x is the Wigner distribution,

$$W_x(t,f) = \int_{-\infty}^{\infty} x(t+\frac{\tau}{2}) x^*(t-\frac{\tau}{2}) e^{-j2\pi f\tau}\, d\tau. \tag{2}$$

$\Phi_{P,t,f}$ is a two-dimensional low-pass smoothing filter (kernel), which specifies the shape and nature of the local region used to determine the energy at each location in the time-frequency (TF) plane. The subscript indicates the possible time and frequency dependence of the filter; if there is no dependence, the TFD is a member of Cohen's bilinear class [6]. These distributions can be expressed as the two-dimensional Fourier transform of the symmetric ambiguity function, weighted by the smoothing kernel:

$$P_x(t,f) = \frac{1}{2\pi} \int \int A_x(\tau,\nu) \phi_{P,t,f}(\tau,\nu) e^{j2\pi(\nu t - \tau f)}\, d\tau\, dv. \tag{3}$$

Here, $\phi_{P,t,f}(\tau,\nu)$ is the two-dimensional Fourier transform of Φ. The symmetric ambiguity function is defined as:

$$A_x(\tau,\nu) \stackrel{\text{def}}{=} \int_t x\left(t+\frac{\tau}{2}\right) x^*\left(t-\frac{\tau}{2}\right) e^{-j2\pi\nu t}\, dt. \tag{4}$$

The choice of the kernel $\phi_{P,t,f}(\tau,\nu)$ is critical to the appearance and quality of a TFD. Numerous kernels have been developed, each displaying advantageous properties for particular classes of signals. A fixed kernel cannot achieve a good representation, as defined by minimal smearing of auto-components and strong suppression of cross-component interference, for every type of signal encountered. To achieve satisfactory performance for a wide variety of signals, the kernel must be dependent on the analysed signal.

There has been much work on the design of algorithms to develop signal-adaptive kernels. The kernels fall into three main groups. The first group consists of kernels which are adapted using the entire signal, and remain of constant shape over the entire time-frequency plane; examples can be found in [4, 3, 5, 24]. Such kernels are unsatisfactory when the time-frequency behaviour of the signal changes with time. The second group consists of kernels which vary with time, and are designed using time-localised properties; prominent examples are those described in [10, 11]. TFDs generated using kernels in this group degrade when the time-supports of signal components overlap or are closely spaced. The third group of kernels vary over both time and frequency, and are devised using time-frequency localised properties of the signal [1, 12, 13].

There are a number of other techniques aimed at improving the appearance of TFDs, notably hybrid linear/bilinear schemes [18], the reassignment methods [2], and image-processing based algorithms [20, 15]. Although all methods aim to reduce the interference between separate components during the generation of time-frequency distributions, none of the techniques directly address the issue of identifying and isolating the components prior to kernel design. Such an isolation allows the kernels which are applied within a TF region occuped by a particular component to be designed according to the properties of that component alone.

The kernel design method proposed in this paper involves a complete isolation and extraction of separate time-frequency components. Distinct kernels are then designed for each extracted component. The method revolves around the construction of a model of the time-frequency energy distribution; the model indicates the number of components, and the regions they occupy in the time-frequency plane. The structure of the model is a finite mixture of Gaussian densities, or equivalently a normalised sum of weighted radial basis functions. Each Gaussian density represents a single (linear) time-frequency component.

The following section of the paper provides an overview of the method for generating a time-frequency distribution. The third section discusses the techniques for determining the parameters of the model: the number of components, and the time-frequency regions they occupy. The fourth section details how kernels are designed for the separate components once they have been isolated, and discusses the computational cost involved in the algorithm. The fifth section illustrates the application of the kernel design process using a synthetic and an experimentally-obtained signal. Conclusions follow in the final section.

2 Overview of the TFD generation process

The generation of the regionally optimised TFD consists of the following seven stages:

1. An initial TFD is generated. A TFD which suppresses the majority of interference terms, even at the cost of substantial smearing, is preferable. The component concentration is recovered in the final distribution when regionally optimised kernels are

applied. The spectrogram, smoothed pseudo-Wigner distribution, or adaptive optimal kernel (AOK) TFD [11] are all satisfactory.

2. The TFD is modelled using a finite mixture model (see Section 3) of N Gaussians, where N is a substantial over-estimate of the number of signal components. The model is initialised to provide good coverage of the initial TFD.

3. The parameters of the N-Gaussian finite mixture model are determined. The parameter determination is an optimisation problem, the aim being to obtain the best representation of the energy distribution of the signal. This is similar to the problem of probability density estimation, and methods used to perform that task can be applied. In particular, a variant of the *expectation-maximisation* (EM) algorithm [7] (see Appendix A) is used to optimise the model parameters.

4. At this stage, the model provides a reasonable approximation to the energy distribution, but the number of time-frequency components is substantially over-estimated. The *functional merging* technique [21] (see Section 3.2) is applied to estimate the true number of components in the distribution.

5. Using the developed time-frequency energy model, a Bayesian classification approach [8, 22] is used to define the time-frequency region occupied by each component, effectively segmenting the time-frequency plane.

6. Based on the defined regions, time-varying filters [14, 9] are used to extract the identified components, and determine *regional ambiguity functions* (see Section 4.1).

7. The regional ambiguity functions are used in conjunction with the time-varying kernel optimisation methodology of Baraniuk and Jones [11] to develop distinct regionally localised kernels (see Section 4.2). These smoothing kernels are applied solely within their corresponding time-frequency regions.

3 The Time-Frequency Energy Model

3.1 The Model Structure and Parameter Estimation

A key constituent of the regional optimisation procedure is the modelling of the energy distribution. The aim is to determine the number of modes in the distribution and their shape, under the assumption that each mode approximately corresponds to a component. The energy distribution is normalised over the time-frequency plane, allowing the application of techniques for modelling probability distributions. The modelling is performed by adapting a Gaussian mixture model to approximate the number of modes in the underlying multi-modal

distribution. The two-dimensional Gaussian mixture model is defined as a linear combination of Gaussian densities:

$$F_N(\mathbf{x}) = \sum_{i=1}^{N} \lambda_i f_i(\mathbf{x}|\boldsymbol{\theta}_i), \tag{5}$$

where $\mathbf{x}$ are time-frequency location vectors of the form $\mathbf{x} = [t, f]$, N is the number of normal distributions $f_i(\mathbf{x}|\boldsymbol{\theta}_i)$, and λ_i are the mixing weights, constrained to be positive and of unity sum. The parameter vector of each normal distribution in the mixture $\boldsymbol{\theta}_i$ is comprised of the mean $\boldsymbol{\mu}_i$ and the full covariance matrix Σ_i. Each normal distribution within the mixture model can thus be expressed as:

$$f_i(\mathbf{x}|\boldsymbol{\theta}_i) = \frac{1}{2\pi|\Sigma_i|^{1/2}} \exp\left(-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_i)^T\Sigma_i^{-1}(\mathbf{x}-\boldsymbol{\mu}_i)\right) \tag{6}$$

The adaptation of the parameters is based on maximising the likelihood of the parameters of the model for the given initial distribution. When the task is viewed in this manner, the initial distribution takes the guise of sampled "data" in a probability density estimation problem. The aim of the adaptation is to maximise the likelihood that, by taking sufficient samples from the model distribution, the initial distribution is realised. Denoting the initial distribution as $P_{\text{init}}(t, f) = P_{\text{init}}(\mathbf{x})$, the negative log-likelihood of the initial distribution is:

$$\begin{aligned} E = -\ln \mathcal{L} &= -\sum_{\mathbf{x}} P_{\text{init}}(\mathbf{x}) \ln F_N(\mathbf{x}) \\ &= -\sum_{\mathbf{x}} P_{\text{init}}(\mathbf{x}) \ln \left\{ \sum_{i=1}^{N} \lambda_i f_i(\mathbf{x}|\boldsymbol{\theta}_i) \right\}. \end{aligned} \tag{7}$$

In this expression, $F_N(\mathbf{x})$ is the probability of the model generating energy at $\mathbf{x} = [t, f]$. The quantity E in (7) can be regarded as the error function; mimimising E is equivalent to maximising the likelihood of the initial distribution.

A variant of the expectation-maximisation (EM) algorithm [7] can be applied to estimate the parameters which minimise the error function E in (7). The variant of the EM algorithm is reviewed in Appendix A. After its application, the TFD is well represented by the finite mixture model, but the number of modes in the distribution is over-approximated. As the primary aim is to identify the number of modes (assuming each mode corresponds to a component), and only approximate the time-frequency regions they occupy, it is important to improve the approximation of the number of the modes. Many of the techniques designed to adapt the number of Gaussians comprising a mixture model are based on *pruning* [19] or *growing* [17, 16] the model, but both approaches generate models wherein the number of Gaussians does not truly reflect the complexity of the underlying distribution.

3.2 Determining the model order: Functional merging

An alternative technique for adapting the number of Gaussians is *functional merging* [21]. This procedure provides a much better indication of the true complexity, or number of modes, within a distribution. Functional merging is based on the principle that Gaussians modelling the same mode should be closer together than those modelling separate modes. Critical to the operation of functional merging is the choice of a distance measure to gauge the distance between two distributions, in this case Gaussians. The *Kullback-Leibler*, *Bhattacharyya* [23, 22] and *arc-cosine* distances [16] are all related to maximum likelihood, and reduce to the *Mahalanobis* distance when used to measure the distance between two Gaussian distributions with equal covariance matrices. The advantage of the arc-cosine distance is that it has a closed-form expression when used to measure the distance between mixtures of Gaussians, rendering its numerical evaluation straightforward.

The arc-cosine distance, Ω, is defined in a Hilbert space, $\mathcal{H}$, as the angle between two distributions, $f(\cdot)$ and $g(\cdot)$:

$$\Omega = \arccos\left(\frac{\langle f, g\rangle}{\|f\|_2 \cdot \|g\|_2}\right), \tag{8}$$

where $\langle\cdot,\cdot\rangle$ is the inner product, $\langle f, g\rangle = \int_{\mathcal{H}} f(\mathbf{x})g(\mathbf{x})d\mathbf{x}$. When used to measure the distance between a mixture of Gaussians, $F_N(\mathbf{x})$, and a single Gaussian distribution, $f_j(\mathbf{x}|\boldsymbol{\theta}_j)$, the inner product becomes a weighted sum of inner products between Gaussians:

$$\langle F_N, f_j\rangle = \sum_{i=1}^{N} \lambda_i \langle f_i(\mathbf{x}|\boldsymbol{\theta}_i), f_j(\mathbf{x}|\boldsymbol{\theta}_j)\rangle \tag{9}$$

where

$$\langle f_i(\mathbf{x}|\boldsymbol{\theta}_i), f_j(\mathbf{x}|\boldsymbol{\theta}_j)\rangle = \frac{|\Sigma|^{1/2}}{2\pi|\Sigma_i|^{1/2}|\Sigma_j|^{1/2}} \exp\left(-\frac{1}{2}(\boldsymbol{\mu}_i^T\Sigma_i^{-1}\boldsymbol{\mu}_i + \boldsymbol{\mu}_j^T\Sigma_j^{-1}\boldsymbol{\mu}_j - \boldsymbol{\mu}^T\Sigma^{-1}\boldsymbol{\mu})\right)$$

for

$$\begin{aligned}\Sigma &= (\Sigma_i^{-1} + \Sigma_j^{-1})^{-1}\\ \boldsymbol{\mu} &= \Sigma(\Sigma_i^{-1}\boldsymbol{\mu}_i + \Sigma_j^{-1}\boldsymbol{\mu}_j).\end{aligned}$$

The functional merging technique starts by determining the arc-cosine distances between every pair of Gaussians in the mixture model. The two closest Gaussians in the mixture model, $f_i(\mathbf{x}|\boldsymbol{\theta}_i)$ and $f_j(\mathbf{x}|\boldsymbol{\theta}_j)$, are selected for examination. The arc-cosine distance between the mixture of these two Gaussians

$$F_n(\mathbf{x}) = \lambda_i f_i(\mathbf{x}|\boldsymbol{\theta}_i) + \lambda_j f_j(\mathbf{x}|\boldsymbol{\theta}_j),$$

and a single proposed replacement Gaussian $f_{\text{new}}(\mathbf{x}|\boldsymbol{\theta}_{\text{new}})$ is determined. The parameters of the new Gaussian, λ_{new} and $\boldsymbol{\theta}_{\text{new}} = \{\boldsymbol{\mu}_{\text{new}}, \Sigma_{\text{new}}\}$, are chosen to minimise this distance. The minimisation is equivalent to the maximisation of the likelihood of $F_n(\mathbf{x})$ given the new Gaussian, which involves matching the zeroth, first and second order moments of the new Gaussian to those of the mixture of two Gaussians it is replacing. Thus:

$$\lambda_{\text{new}} = \int_{\mathcal{H}} F_n(\mathbf{x})\, d\mathbf{x} = \lambda_i + \lambda_j$$
$$\boldsymbol{\mu}_{\text{new}} = \frac{1}{\lambda_{\text{new}}} \int_{\mathcal{H}} F_n(\mathbf{x})\mathbf{x}\, d\mathbf{x} = \frac{\lambda_i}{\lambda_{\text{new}}}\boldsymbol{\mu}_i + \frac{\lambda_j}{\lambda_{\text{new}}}\boldsymbol{\mu}_j$$

and from:

$$\lambda_{\text{new}} \int_{\mathcal{H}} f_{\text{new}}(\mathbf{x}|\boldsymbol{\theta}_{\text{new}})\mathbf{x}\mathbf{x}^T\, d\mathbf{x} = \int_{\mathcal{H}} F_n(\mathbf{x})\mathbf{x}\mathbf{x}^T\, d\mathbf{x}$$

it follows that:

$$\Sigma_{\text{new}} = \frac{\lambda_i}{\lambda_{\text{new}}}(\Sigma_i + (\boldsymbol{\mu}_i - \boldsymbol{\mu}_{\text{new}})(\boldsymbol{\mu}_i - \boldsymbol{\mu}_{\text{new}})^T) + \frac{\lambda_j}{\lambda_{\text{new}}}\left(\Sigma_j + \left(\boldsymbol{\mu}_j - \boldsymbol{\mu}_{\text{new}}\right)\left(\boldsymbol{\mu}_j - \boldsymbol{\mu}_{\text{new}}\right)^T\right).$$

The value of the arc-cosine distance is stored, and a new mixture model constructed, replacing the two Gaussians with the single new Gaussian. The arc-cosine distances between the new Gaussian and all other Gaussians in the model are determined. The process of selecting the two closest Gaussians in the mixture and merging them is repeated until the arc-cosine distance between the mixture of the closest two Gaussians and the single proposed replacement exceeds a maximum distance $\tau_{\max}$. The maximum distance is set to a value beyond which merging should not occur; a substantial range of $\tau_{\max}$ values generate the same model. If a threshold distance τ is chosen, and the closest Gaussians successively merged until the distance between the new single Gaussian and the former mixture of two Gaussians exceeds τ, a mixture model consisting of M Gaussians results. The value of M is plotted against the threshold distance τ for the range of thresholds from zero to the maximum distance $\tau_{\max}$. This results in the construction of a graph similar to that of Figure 1(a), which was produced for the noise-free synthetic signal examined in Section 5.1.

[Figure 1 here]

The value of M should be consistent over the largest range of thresholds when the majority of Gaussians representing the same mode have been merged, and no Gaussians representing separate modes have been merged. Let Υ_i denote the continuous range of thresholds for which the mixture model consists of M_i Gaussians. A model probability function $P(M_i|\Upsilon_i)$, which

is based on the principle just stated, can be defined as:

$$P(M_i|\Upsilon_i) = \frac{\text{length}(\Upsilon_i)}{\tau_{\max}} \tag{10}$$

The model generated by the functional merging process which contains M_j components, where M_j maximises $P(M_i|\Upsilon_i)$, is selected to approximate the underlying distribution. This model indicates the number of time-frequency components detected in the distribution, and the regions they occupy. An example of the model probability function is shown in Figure 1(b); this demonstrates that the most probable model is the mixture consisting of two Gaussians.

4 Regional kernel optimisation

4.1 Regional ambiguity function

The kernels designed in this paper vary according to the time-frequency region over which they are defined. This regional localisation is accomplished by defining a regional ambiguity function (RAF). Each RAF is localised to a time-frequency region R_k, and can be expressed as:

$$A(R_k;\tau,v) \equiv \int_{t\in R_k} x_k\left(t+\frac{\tau}{2}\right) x_k^*\left(t-\frac{\tau}{2}\right) e^{-j2\pi vt}\,dt \tag{11}$$

where

$$x_k(t) = (\mathbf{S_k}x)(t) = \int_{t'} S_k(t,t')x(t')\,dt' = \sum_{v=1}^{N_k}\langle x, s_v\rangle s_v(t). \tag{12}$$

Here $x_k(t)$ is the orthogonal projection of $x(t) \in \mathcal{L}_2(R)$ on $\mathcal{S}_k$, a linear signal space covering the region R_k energetically and with little energy outside the region. $S_k(t,t')$ is the kernel of the orthogonal projection operator $\mathbf{S}_k$, and $\{s_v(t)\}_{v=1}^{N_k}$ is an orthonormal basis of $\mathcal{S}_k$, where N_k is the dimension of the space. This projection onto a linear signal space is the linear time-frequency filtering procedure proposed by Hlawatsch [9]. The design of the signal space is described in [9]; the advantages of such linear time-frequency filters are discussed in [14]. The time-frequency filtered signal, $x_k(t)$, is the portion of $x(t)$ whose time-frequency energy resides approximately exclusively within the region R_k.

Time-dependence of the kernel within each region requires the definition of a regional short-time ambiguity function (RSTAF):

$$\begin{aligned} &A(t,R_k;\tau,v) \\ &\qquad = \int_{u\in R_k} x_k^*\left(u-\frac{\tau}{2}\right) w^*\left(u-t-\frac{\tau}{2}\right) x_k\left(u+\frac{\tau}{2}\right) w\left(u-t+\frac{\tau}{2}\right) e^{-j2\pi vu}\,du \end{aligned} \tag{13}$$

with $w(u)$ a symmetrical windowing function. This is similar in nature to the short-time ambiguity function (STAF) defined in [11]. The introduction of the RSTAF enables time-varying optimised kernels to be designed for specific regions in the time-frequency plane, corresponding to the areas occupied by separable components. Through the use of the window, only the portion of the signal in the interval $[t-T, t+T]$ is incorporated in the calculation of the STAF at any particular time-instant t.

4.2 Kernel design

The kernels designed in this paper have Gaussian radial cross-section as in [3], chosen because of the combination of flexibility and computational ease. The radially-Gaussian kernels can be expressed in the ambiguity plane (using polar coordinates) as:

$$\Phi(r,\psi) = \exp\left(-\frac{r^2}{2\sigma^2(\psi)}\right)$$

where $r = \sqrt{(2\pi v)^2 + \tau^2}$. The spread function, σ, controls the spread of the Gaussian at radial angle $\psi = \arctan(\tau/2\pi v)$.

The kernels are designed by solving an optimisation problem similar to that formulated in [3]. The optimisation problem has constraints and a performance index designed to suppress cross-components whilst passing auto-components with as little distortion as possible. The problem is formulated in the ambiguity plane, exploiting its property of separation of auto- and cross-components. For the kernel corresponding to region R_k (component k) at time $t \in R_k$, the optimisation problem can be stated as:

$$\Phi_{\text{opt}}(t, R_k) = \arg\max_{\Phi} \int_0^{2\pi} \int_0^{\infty} \left|A(t, R_k; r, \psi)\Phi(r,\psi)\right|^2 r\, dr\, d\psi \tag{14}$$

subject to:

$$\Phi(r,\psi) = \exp\left(-\frac{r^2}{2\sigma^2(\psi)}\right) \tag{15}$$

and

$$\frac{1}{4\pi^2}\int_0^{2\pi} \int_0^{\infty} \left|\Phi(r,\psi)\right|^2 r\, dr\, d\psi = \frac{1}{4\pi^2}\int_0^{2\pi} \sigma^2(\psi)\, d\psi \leq \alpha \tag{16}$$

where $A(t, R_k; r, \psi)$ is the RSTAF in polar coordinate form and α is a parameter controlling the volume under the kernel. Baraniuk and Jones [11] have described a method for the solution of an analogous optimisation problem, with the STAF, $A(t; r, \psi)$, replacing $A(t, R_k; r, \psi)$; this method can be applied directly to the problem presented here.

4.3 Computational Complexity

The computational cost of generating the regionally optimised is difficult to quantitatively assess, because of the high dependence on the nature of the signal. The complexities of the stages of the algorithm are now approximated. A block of length K samples is considered, with M frequency points being generated. N is used to denote the number of components in the initial model, and N' the number of components after functional merging.

1. Stage 1 (Generation of the initial distribution): $O(KM \log M)$ for a fixed-kernel decomposition and $O(KM^2)$ for an adaptive optimal kernel distribution.

2. Stages 2-4 (Construction of mixture model): The complexity of the modified EM algorithm is approximately $O(INR)$, where I is the number of iterations, and R is the number of time-frequency locations used to update the model. For the majority of signals, R is in the range $\left[\frac{KM}{4} \ldots \frac{3KM}{4}\right]$, i.e., between 1/4 and 3/4 of the time-frequency locations have sufficient energy to affect the development of the model.

3. Stages 5-6 (Classification and filter development): If the projection filter is used, the complexity is approximately $O(N'M^3)$. The Weyl filter [14] yields similar performance, and can reduce the complexity to approximately $O(N'M \log M)$.

4. Stage 7 (Regional kernel development): If the length of each component is denoted K_i, the complexity of this stage is approximately $O(\sum_{i=1}^{N'} K_i M^2)$.

In general, the adapted EM algorithm is the most computationally demanding section of the procedure. A significant saving can be obtained by constructing the finite mixture model at a lower resolution (either in time, frequency, or both) than is desired for the final distribution. The complexity of the EM algorithm revolves around the number of components in the initial model, and the number of TF locations which contain sufficient energy in the initial distribution to demand inclusion in the model adaptation. The use of a lower resolution enables a reduction in both of these factors. The regional classification and subsequent localised kernel design and application can be performed at the higher resolution.

Despite these savings, the regional optimisation technique is computationally expensive in comparison to time-varying adaptive kernel approaches such as the adaptive optimal kernel (AOK) [11]. The signals analysed in the following section provide an indication of the increased computational cost. For the 128 sample signal in Section 5.1, in which 2 components are identified, the regional approach requires approximately 2.5 times the computation required by the AOK approach. For the 440 sample signal in Section 5.2, in which 11 linear components are identified, the regional approach requires approximately 8 times the computation of the AOK approach.

5 Results of regional optimisation

This section demonstrates the application of the regionally optimised kernels to synthetic and natural signals. This provides an opportunity to illustrate the operation of key parts of the design algorithm, in particular component detection and region assignment.

5.1 Synthetic data - example 1

The set of data used in this section was constructed to demonstrate the necessity of utilising a well-chosen smoothing kernel in the generation of an interpretable time-frequency distribution. The signal was designed so that the two components comprising the signal, a chirp (modulated Gaussian) and a Gaussian, were separable in the time-frequency plane, but overlapped in time, and also in frequency. Such an arrangement highlights the advantages of optimising kernels for specific time-frequency regions.

The signal can be expressed as:

$$x(t) = \text{rect}(10, 40) \times e^{-j0.03(t-16)^2} e^{j\frac{\pi}{3.8}(t-16)} + 2e^{-0.025(t-34)^2} e^{j\frac{\pi}{4.2}(t-34)} \tag{17}$$

for $\text{rect}(t_1, t_2) = u(t - t_1) - u(t - t_2)$, where $u(t)$ is the Heaviside step function. The real part of the noiseless signal is displayed in Figure 2(a), the Wigner distribution of the signal in Figure 2(e), and the spectrogram in Figure 2(f). The cross-components of the Wigner distribution make identification of the time-frequency nature of the signal difficult; the smearing of the spectrogram has the same effect.

[Figure 2 here]

The algorithm for generating the regionally optimised distribution was applied to the synthetic signal, using the AOK technique of [11] to generate the initial TFD. The initial distribution is displayed in Figure 3(a), together with the component region allocation, which was determined from the mixture model. The functionally merged model is shown in Figure 3(b). The enhanced resolution of the regionally optimised distribution for the noiseless signal is apparent in Figure 3(c).

[Figure 3 here]

A noisy version of the signal (see Figure 2(a)) was generated by adding white complex Gaussian noise, such that the SNR (ratio of total signal power to total noise power) was 4 dB. An initial AOK distribution was generated, and used to determine the component regions and the regionally optimised distributions (Figures 3(d) and 3(e)). The isolated components were reconstructed based on the time-varying filters, and are compared to the original components in Figures 2(b) and 2(d). The signal was de-noised by summing the recovered components. Figure 2(c) displays the real part of the de-noised signal, and compares the absolute noise before and after the de-noising. For the noisy signal in Figure 2(a), the procedure improves the SNR from 4 dB to 8.5 dB.

[Figure 4 here]

5.2 Vibration signal

The application of the algorithm to an experimentally-obtained set of data is examined in this section. The data-set illustrates the limitations of the Gaussian mixture model when the major components in the signal have non-linear time-frequency behaviour. The signal is a measurement of the impulse response of a beam. A 7.2 m long mild-steel beam of rectangular cross-section 32.1 $\times$ 6.3 mm was suspended horizontally on light cords. One end of the beam was lightly tapped with a soft-tipped hammer designed to generate only low-frequency vibrations (below 1kHz). The impact response was measured using an accelerometer attached to the beam close to the point of impact, and the data was captured using a data logger at a sampling frequency of 4096 Hz.

For analysis purposes, the data was down-sampled to 2048 Hz, and the analytic signal generated by applying the Hilbert transform. TFDs were generated for a section of the signal consisting of 440 sample points, as displayed in Figure 5(a). The signal represents the passage of bending waves of different frequencies as they travel along the beam, and are reflected at each end. Since the group velocity of the bending waves is dependent on frequency, the high-frequency groups travel faster than the low-frequency groups, as can be observed in Figure 5(b), which is the initial TFD of the signal, generated using the AOK method.

[Figure 5 here]

The region allocation during the generation of a regionally optimised distribution is displayed in Figure 5(c). The functional merging determined that the most probable model contained eleven Gaussians. The first three components in the distribution display marked non-linear time-frequency behaviour, and the limitations of the Gaussian mixture model resulted in the first two components being represented by three Gaussians, and the third by two. Despite the significant overestimate of components, the mixture model is still useful for the redesign of kernels, as each region contains sufficient local information about the time-frequency nature of the signal. The noise and interference terms visible in the initial distribution have been reduced in the regionally optimised distribution, displayed in Figure 5(d). Some small artefacts are introduced where the division of the components occurs.

6 Conclusion

Bilinear time-frequency distributions provide useful information about the nature of non-stationary signals. Whenever signals consist of more than a solitary time-frequency component, the application of smoothing kernels in the generation of such distributions is necessary to allow interpretation of the distributions. Furthermore, satisfactory distributions can only be constructed for a wide range of signals if the kernels are signal-dependent.

It has been proposed in this paper that for signals consisting of multiple time-frequency components, in addition to being dependent on the signal, the smoothing kernel should vary for each time-frequency component. A technique for the adaptation of finite mixtures to model time-frequency distributions has been presented. The technique indicates the number of (linear) components in a distribution, and their locations in the TF plane. The use of the model to improve time-frequency distributions through the construction of regionally optimised kernels has been discussed. The importance of such a procedure has been indicated by applying the technique to a signal consisting of components which are closely spaced in the time-frequency plane, and overlap in both time and frequency. In addition to improving the clarity and resolution of the time-frequency distribution, the regional optimisation procedure produces sets of basis functions describing the approximate time-frequency region occupied by each component. These sets can be utilised in a variety of procedures including signal decomposition and noise reduction.

One limitation of the finite mixture model is that it is comprised of Gaussians, resulting in every TF distribution being modelled as a set of components restricted to approximately linear time-frequency behaviour. A possible solution to this problem is to associate a Gaussian mixture model (a subset of the global model) with each component.

References

[1] J.C. Andrieux, M.R. Fein, G. Mourgues, P. Bertrand, B. Izrar, and V.T. Nguyen. Optimum smoothing of the Wigner-Ville distribution. *IEEE Trans. Acoust., Speech, Signal Proc.*, 35(6):764–769, June 1987.

[2] F. Auger and P. Flandrin. Improving the readability of time-frequency and time-scale representations by the reassignment method. *IEEE Trans. Signal Processing*, 43(5):1068–1089, May 1995.

[3] R.G. Baraniuk and D.L. Jones. Signal-dependent time-frequency analysis using a radially Gaussian kernel. *Signal Processing*, 32:263–284, 1993.

[4] R.G. Baraniuk and D.L. Jones. A signal-dependent time-frequency representation: optimal kernel design. *IEEE Trans. Signal Processing*, 41(4):1589–1601, Apr. 1993.

[5] R.G Baraniuk and L.F. Wisur-Olsen. Optimal phase kernels for time-frequency analysis. Technical Report ECE 9611, Rice University, 1996.

[6] L. Cohen. Time-frequency distributions - a review. *Proc. IEEE*, 77(7):941–981, July 1989.

[7] A.P. Dempster, N.M. Laird, and D.B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *J. Royal Statistical Soc. B*, 39(1):1–38, 1977.

[8] R.O. Duda and P.E. Hart. *Pattern classification and scene analysis.* Wiley, New York, 1974.

[9] F. Hlawatsch and W. Kozek. Time-frequency projection filters and time-frequency signal expansions. *IEEE Trans. Signal Processing*, 42(12):3321–3334, Dec. 1994.

[10] D.L. Jones and R.G. Baraniuk. A simple scheme for adapting time-frequency representations. *IEEE Trans. Signal Processing*, 42(12):3530–3535, Dec. 1994.

[11] D.L. Jones and R.G. Baraniuk. An adaptive optimal-kernel time-frequency representation. *IEEE Trans. Signal Processing*, 43(10):2361–2371, Oct. 1995.

[12] D.L. Jones and T.W. Parks. A high-resolution data-adaptive time-frequency representation. *IEEE Trans. Acoust., Speech, Signal Proc.*, 38(12):2127–2135, Dec. 1990.

[13] G. Jones and B. Boashash. Generalized instantaneous parameters and window matching in the time-frequency plane. *IEEE Trans. Signal Processing*, 45(5):1264–1275, May 1997.

[14] W. Kozek and F. Hlawatsch. A comparative study of linear and nonlinear time-frequency filters. In *Proc. 1992 IEEE-SP Int. Symp. Time-Frequency and Time-Scale Analysis*, pages 163–166, 1992.

[15] S. Krishnamachari and W.J. Williams. Adaptive kernel design in the generalized marginals domain for time-frequency analysis. In *Proc. 1994 IEEE Int. Conf. Acoust. Speech. Signal Proc.*, pages 341–344, 1994.

[16] C. Molina and M. Niranjan. Pruning with replacement on limited resource allocating networks by F-projections. *Neural Computation*, 8:345–356, 1996.

[17] J.C. Platt. A resource allocating network for function interpolation. *Neural Computation*, 3:213–225, 1991.

[18] S. Qian and D. Chen. Decomposition of the Wigner distribution and time-frequency distribution series. *IEEE Trans. Signal Processing*, 42(10):2836–2842, Oct. 1994.

[19] R. Reed. Pruning algorithms— a survey. *IEEE Trans. Neural Networks*, 4:740–747, 1993.

[20] B. Ristic and B. Boashash. Kernel design for time-frequency signal analysis using the radon transform. *IEEE Trans. Signal Processing*, 41(5):1996–2008, 1993.

[21] C.M. Stow, A.C.T. Kennington, C. Molina, and W.J. Fitzgerald. Experimental issues of functional merging on probability density estimation. In *Proc. 1997 IEE Int. Conf. Artificial Neural Networks*, volume 1, pages 3–8, 1997.

[22] C.W. Therrien. *Decision, estimation and classification: an introduction to pattern recognition and related topics.* Wiley, New York, 1989.

[23] D.M. Titterington, A.F.M. Smith, and U.E. Makov. *Statistical analysis of finite mixture distributions*. John Wiley and Sons, Chichester, 1985.

[24] G.T. Venkatesan and M.G. Amin. Time-frequency distribution kernels using FIR filter design techniques. *IEEE Trans. Signal Processing*, 45(6):1645–1650, June 1997.

A The EM algorithm

The expectation-maximisation (EM) algorithm was introduced by Dempster *et al.* [7] as a means of solving non-linear optimisation problems with missing data. The algorithm can also be applied to the problem of estimating a probability density, when a data-set of samples from the probability distribution is available, by using it to adjust the parameters of a finite mixture model used to approximate the underlying distribution. The optimisation of the parameters of the model would be straightforward if it were known which component of the model was responsible for generating each data point. A hypothetical complete data set is considered, in which a labelling of each data point is provided, and this is the context in which the EM algorithm is applied; the "missing" data is the labelling.

The task is to adapt the parameters of a normal mixture model, $F_N(\boldsymbol{x})$, to best approximate an underlying distribution $\tilde{F}(\boldsymbol{x})$. The EM algorithm is iterative in nature, updating parameter estimates during each iteration. Initial parameter estimates must be provided, and the algorithm halts once convergence has been achieved to within acceptable limits.

Firstly, the *unsupervised* EM algorithm is reviewed. This is utilised in the situation where the dataset consists solely of locations $\boldsymbol{x}_i$, $i = 1, \ldots, m$, and no knowledge is available about the underlying probabilities of the samples, $y_i = p(\boldsymbol{x}_i|\tilde{F}(\boldsymbol{x}_i))$. For this case, the EM equations to update the parameters of a Gaussian mixture with full covariance matrices are:

$$\begin{aligned}
\boldsymbol{\mu}_j &= \frac{\sum_{i=1}^{m} B\boldsymbol{x}_i}{\sum_{i=1}^{m} B} \\
\boldsymbol{\Sigma}_j &= \frac{\sum_{i=1}^{m} B(\boldsymbol{x}_i - \boldsymbol{\mu}_j)(\boldsymbol{x}_i - \boldsymbol{\mu}_j)^T}{\sum_{i=1}^{m} B} \\
\lambda_j &= \frac{1}{m}\sum_{i=1}^{m} B
\end{aligned}$$

where B is the posterior probability of mixture component j being responsible for data point $\boldsymbol{x}_i$. This is obtained from Bayes' theorem as:

$$B = P(j|\boldsymbol{x}_i) = \frac{\lambda_j f(\boldsymbol{x}_i|\boldsymbol{\theta}_j)}{F_N(\boldsymbol{x}_i)}.$$

The *supervised* EM algorithm arises when the dataset consists of a uniformally distributed set of samples x_i, $i = 1, \ldots, m$, with known associated probabilities, $y_i = p(\boldsymbol{x}_i|\tilde{F}(\boldsymbol{x}_i))$. For this case, the B term in the update equations becomes:

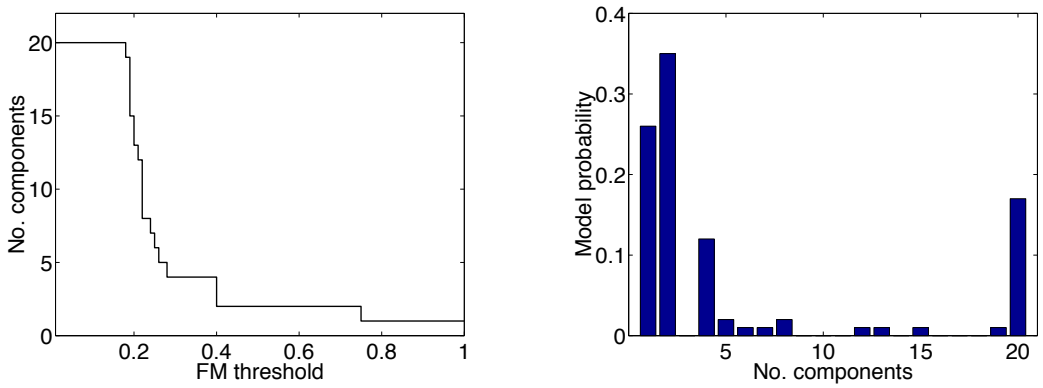
$$B = y_i P(j|\boldsymbol{x}_i).$$

The value y_i represents an explicit target value for the mixture model.

In this paper, the modelling of the initial time-frequency distribution is addressed by

casting the problem as one of probability density estimation, with the $\mathbf{x}_i$ vectors taking the form of time-frequency locations, $[t, f]$. The initial distribution acts as a target distribution, and the *supervised* version of the EM algorithm is applied to determine the parameters of the mixture model. A good initialisation of the parameters is important to aid convergence and avoid sub-optimal solutions. The covariance matrices are set to the identity matrix (for suitably scaled time and frequency vectors), and the weights are initially the same for all components. The means are assigned to the most energetic locations in the initial distribution, subject to the constraint that all means must be separated by a suitable distance.

In experiments conducted, it was observed that satisfactory convergence of the EM algorithm generally occurred within 10-15 iterations. The initial number of components required depends on the complexity of the signal; experiments have shown that for the majority of signals the inclusion of 20-40 Gaussians in the initial model provides sufficient coverage, and results in a good approximation to the initial distribution.



(a) (b)

Figure 1: Functional merging algorithm for synthetic signal - example 2. (a) Number of components in the model as the merging threshold is varied. (b) Model probability.

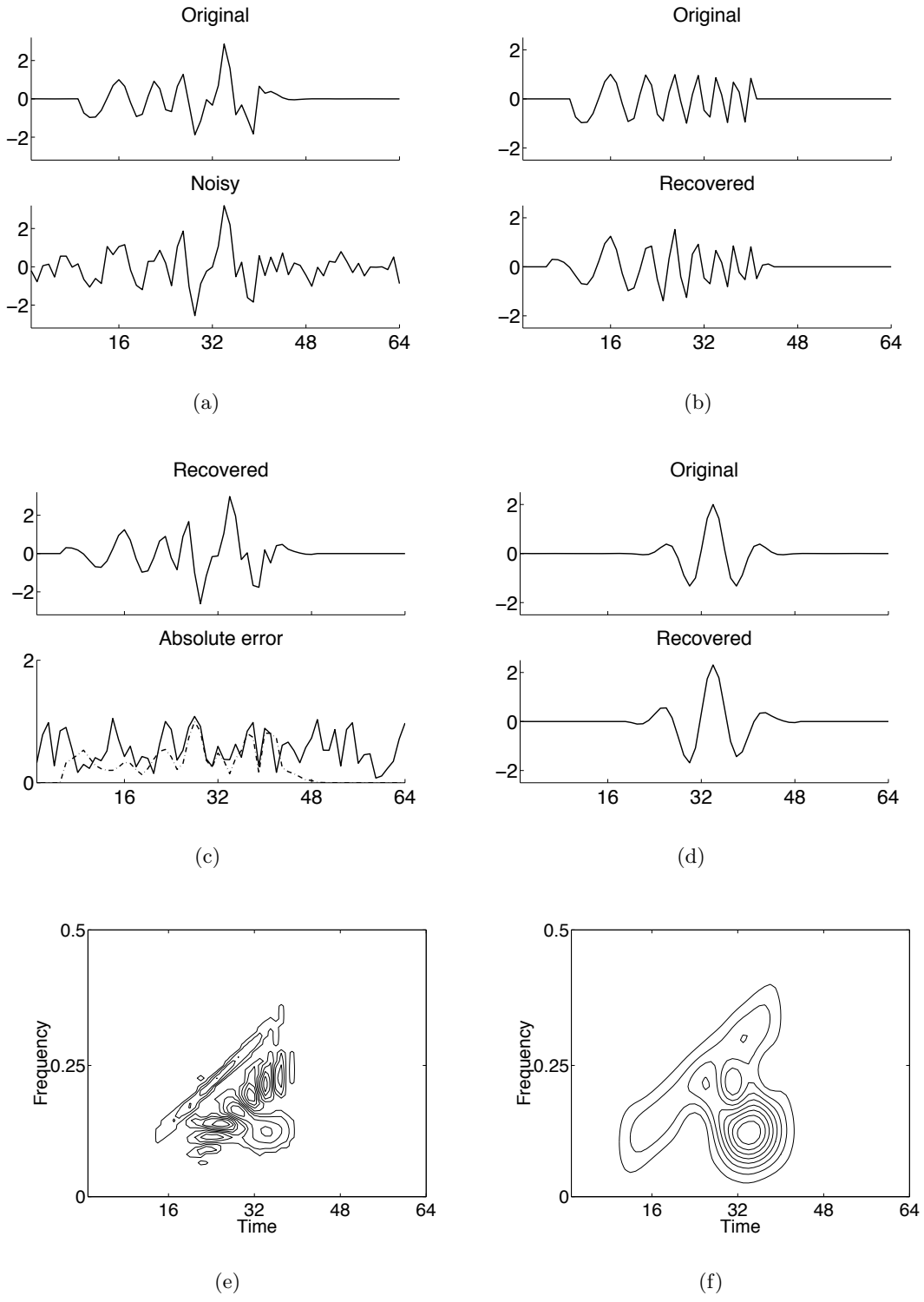


Figure 2: Synthetic signal analysis. (a) Real parts of original signal and noisy version (SNR = 4 dB). (b) The chirp component and the component recovered from the noisy signal. (c) The recovered signal and a comparison of absolute errors (solid - before de-noising; dashed - after). (d) The Gaussian component and the recovered component. (e) Wigner distribution of original signal. (f) Spectrogram of original signal.

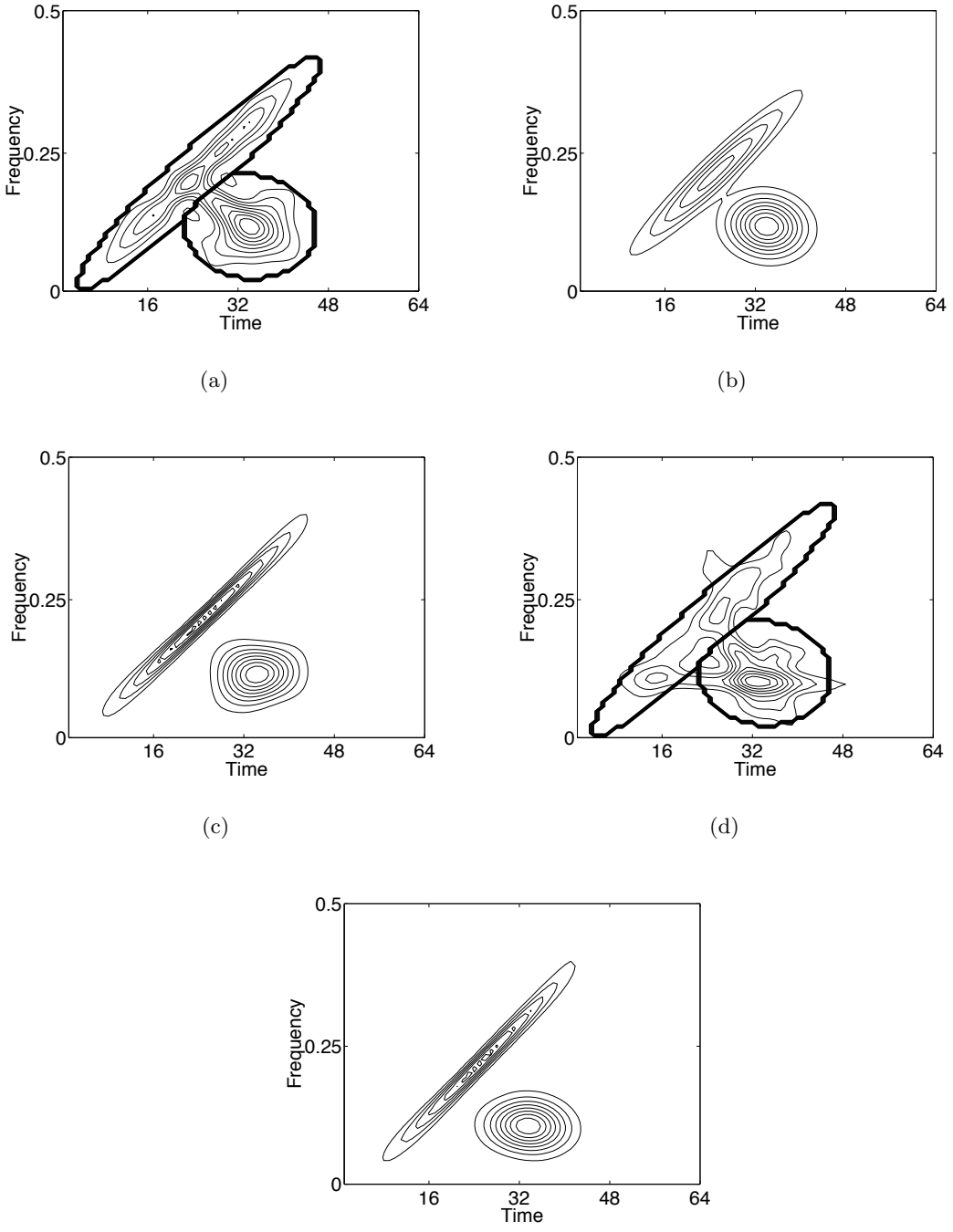


Figure 3: Synthetic signal analysis. (a) Initial TFD (AOK method) of original signal and the developed regions. (b) TFD approximation after functional merging. (c) Regionally optimised TFD of original signal. (d) Initial TFD (AOK method) of noisy signal and developed regions. (e) Regionally optimised TFD of noisy signal.

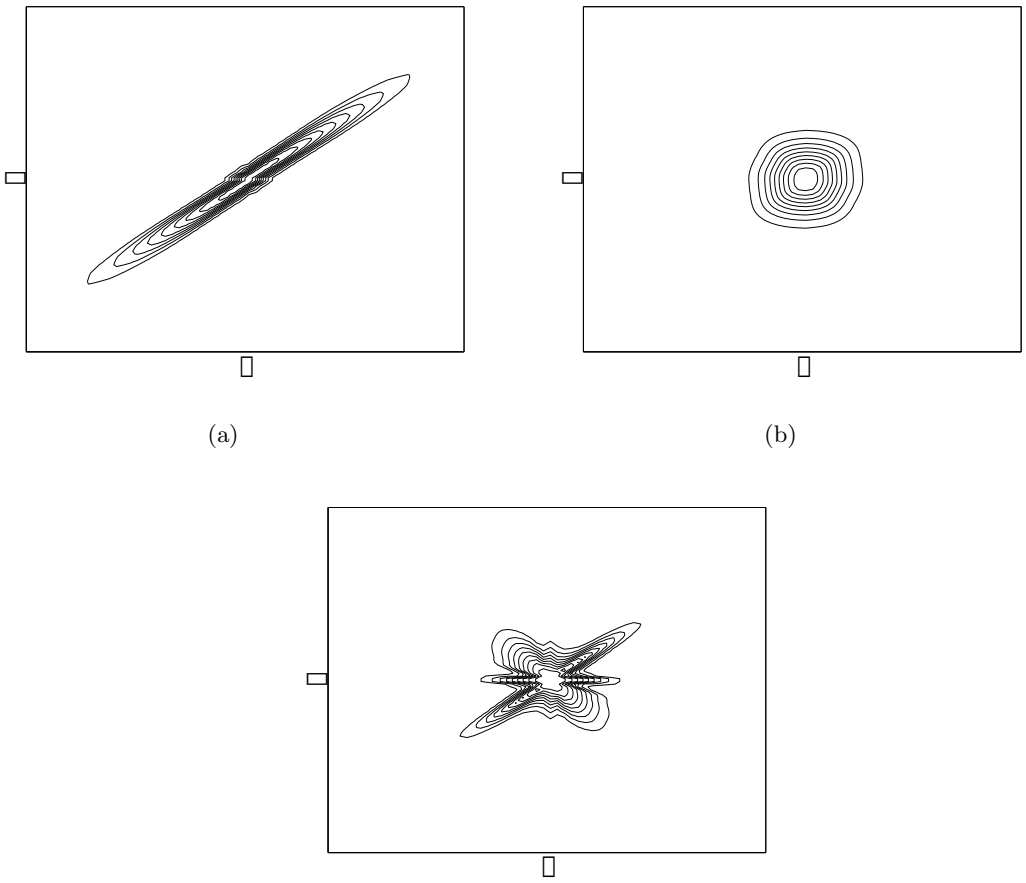


Figure 4: Kernel shapes for the synthetic signal. (a) Regionally optimised kernel for chirp component. (b) Regionally optimised kernel for Gaussian component. (c) Adaptive optimal kernel shape at time t=32.

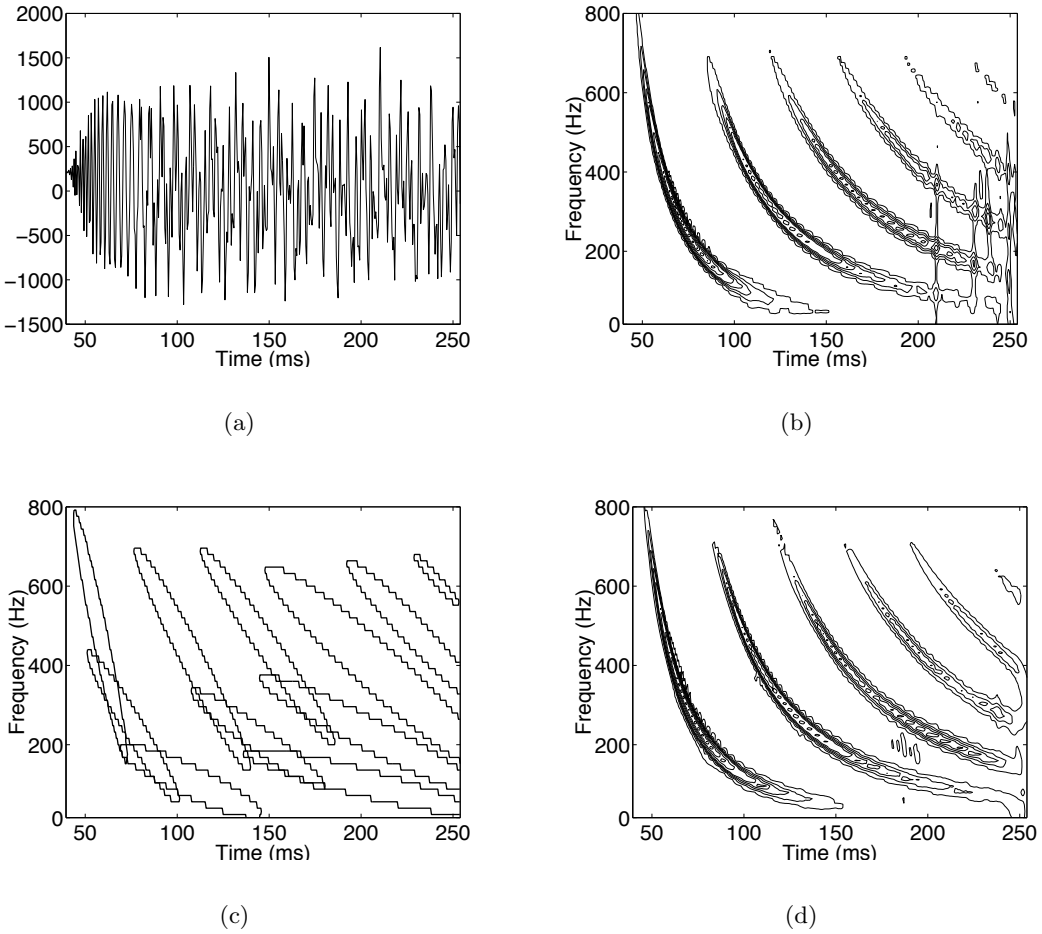


Figure 5: Vibration signal analysis. (a) Vibration signal. (b) The initial distribution. (c) Region allocation. (d) The regionally optimised distribution.