

# **Robust Divergence Measures for Time Series Discrimination\***

TA-HSIN LI<sup>†</sup>

Department of Statistics  
Texas A&M University  
College Station, TX 77843-3143

Version 1: September 1995

Version 2: September 1996

Technical Report No. 237

\*This work was supported in part by a grant from Texas Instruments Inc.

<sup>†</sup>Current Address: Department of Mathematical Sciences, IBM T. J. Watson Research Center, Yorktown Height, NY 10598-0218 (e-mail: thl@watson.ibm.com).

## **Abstract**

New divergence measures are introduced for change detection and discrimination of stochastic signals (time series) on the basis of parametric filtering — a technique that combines parametric linear filtering with correlation characterization. The sensitivity of these divergence measures is investigated using local curvatures under additive and multiplicative spectral departure models. It is found that when the time series contain dominant spectral components (such as sharp peaks and notches) of similar characteristics, the new divergence measures are more effective in detecting the dominated spectral deviations than the conventional spectral divergence measures such as the Kolmogorov-Smirnov spectral distance and the Kullback-Leibler spectral divergence. Simulation results are given that illustrate and confirm the theoretical findings.

*Key words and phrases:* Characterization; Discriminant Analysis; Distortion Measure; Filter Bank; Pattern Recognition; Signal Processing; Spectral Analysis

# 1. INTRODUCTION

In discrimination, change detection, and change analysis of stochastic signals (or time series), divergence measures are widely-used statistics that quantify the dissimilarity between the time series in terms of their statistical characteristics [Basseville (1989), Gray *et al.* (1980), Parzen (1992)]. Divergence measures can be derived from different characterizations of time series. Typical characterization methods include parametric probability models [Basseville and Nikiforov (1993)] and nonparametric representations such as Fourier spectral densities [Coates and Diggle (1986), Deshayes and Picard (1986), Parzen (1992)]. In this paper, we introduce some divergence measures that are derived from a recently proposed approach of time series characterization called parametric filtering [Li (1996a)].

To represent the statistical characteristics of time series, the *parametric filtering* (PF) approach combines the interference rejection capability of linear filtering with the correlation characterization property of some parametric filter banks in terms of the output lag-one autocorrelations. The PF approach provides a suitable source for obtaining divergence measures that are (i) universal for time series discrimination and change analysis, (ii) easy to calculate, and (iii) flexible to adjust for effective detection of spectral deviations in the presence of dominant spectral components. The PF divergence measures have been successfully used in speech processing applications to detect acoustic events in continuous-speech waveforms [Li and Gibson (1994), (1996)].

Two filter banks are considered in this paper, the repeated summing and differencing (RSD) and the complex exponential filtering (CEF). The RSD is motivated by the work of Kedem and Slud (1982) in which the higher order zero-crossings obtained from repeated differencing are proposed as features for time series discrimination [see also Kedem (1994)]. The RSD can be regarded as a filter bank of low-pass and high-pass filters in which the bandwidth is determined by the number of times the summing or differencing is repeated. The CEF, on the other hand, is equipped not only with a more flexible bandwidth parameter, but also with a parameter that easily controls the center frequency of the filter and thus allows the analysis to concentrate on the frequency bands of interest that may not be in the high or low frequency regions. For both filters, different time series correlation structures can be discriminated via the *characterization functions* derived from the lag-one autocorrelation of the filtered time series as functions of the filter parameter. New divergence measures result from quantifying the difference between two such characterization functions.

The *dominating effect* of spectral peaks and notches on spectral density based divergence measures, such as the Kolmogorov-Smirnov (KS) spectral distance and the Kullback-Leibler

(KL) spectral divergence, has been noticed in the literature [e.g., Basseville and Nikiforov (1993), Li (1996a), Li and Gibson (1996), Zhang and Taniguchi (1995)]. These dominant narrow band components tend to cause enormous sample variations in the spectral divergence measures, even though they have similar spectral characteristics. As a result, the actual spectral deviations of other (narrow and/or broad band) components may be overshadowed by the sample variations, so that large spectral divergence values may not necessarily correspond to significant spectral changes. This, in change detection applications, leads to high false alarm probabilities, as shown, for example, in Li and Gibson (1996).

To investigate the effectiveness of the new divergence measures in the presence of dominant spectral components, we carry out a sensitivity analysis using the *local curvature approach*. The local curvatures are derived under two general spectral departure models — the additive model and the multiplicative model. Using the local curvatures, we discuss the sensitivity of the CEF divergence measures when a spectrum is contaminated by narrow band peaks and notches. The analysis shows that the new divergence measures can be easily adjusted to balance the efficiency of change detection and robustness against the dominant spectral components. As a result, the new divergence measures are potentially more powerful than the spectral density based divergence measures in detecting spectral deviations in the presence of dominant spectral components. In addition to this, an investigation is carried out on the role of the frequency and bandwidth parameters of the complex exponential filter in determining the sensitivity of the CEF divergence measures. The results may be used as a guide to the selection of these parameters in practice. For a general time series model with a mixed spectrum, strong consistency is established for estimating the divergence measures using their natural sample estimates. Simulation results are given to confirm and demonstrate the theoretical findings.

## 2. PARAMETRIC FILTERING METHODS

Suppose that  $\{X_t\}$  is a real-valued stationary time series with zero mean, unit variance, and autocorrelation function  $r_\tau := E\{X_{t+\tau}X_t\}/E\{X_t^2\}$ . In the following, we summarize two parametric filtering methods that characterize the correlation structure of  $\{X_t\}$ .

### 2.1 Repeated Summing and Differencing

Let  $s_k := \text{sgn}(k)$  and  $X_t(0) := X_t$ . Then, the repeated sums and differences of  $\{X_t\}$  can be obtained recursively from

$$X_t(k) := X_t(k - s_k) + s_k X_{t-1}(k - s_k) \quad (k = \pm 1, \pm 2, \dots), \quad (1)$$

where for a positive  $k$ ,  $\{X_t(k)\}$  represents  $k$  times summing of  $\{X_t\}$ , and for a negative  $k$ , it represents  $|k|$  times differencing of  $\{X_t\}$ . In operator notation, we can write

$$X_t(k) = \mathcal{H}_k(z) X_t \quad (k = 0, \pm 1, \dots),$$

with  $\mathcal{H}_k(z) := (1 + s_k z^{-1})^{|k|}$ , where  $z^{-1}$  is the delay operator such that  $z^{-1} X_t = X_{t-1}$ .

Traditionally, the repeated summing is regarded as a smoothing operation that suppresses high frequency noise, whereas the repeated differencing is used to eliminate low frequency components such as polynomial trends [e.g., Brockwell and Davis (1992)]. The parametric filtering explores an additional use of the repeated sums and differences defined in (1), namely, to characterize the correlation structure of  $\{X_t\}$ . To that end, we consider the lag-one (first-order) autocorrelations

$$\rho(k) := \frac{E\{X_{t+1}(k) X_t(k)\}}{E\{X_t^2(k)\}} \quad (k = 0, \pm 1, \dots) \quad (2)$$

as a *characterization function* of  $\{X_t\}$ .

There is a relationship between  $\rho(k)$  and the *expected zero-crossing rate* of  $\{X_t(k)\}$ , denoted by  $\zeta(k)$ . In fact, under suitable conditions, it may be shown that [Kedem (1994)]

$$\rho(k) = \cos(\pi \zeta(k)).$$

Because the  $\zeta(k)$ 's, known as the higher order crossings (HOC) [Kedem and Slud (1982)], are proportional to the numbers of zero-crossings ( $k = 0$ ), peaks and troughs ( $k = -1$ ), inflection points ( $k = -2$ ), etc., of the original (unfiltered) time series  $\{X_t\}$ , the function  $\rho(k)$  thus indirectly measures these features of  $\{X_t\}$  using the lag-one autocorrelations. The  $\rho(k)$ 's are also referred to by Kedem (1994) as HOC, standing for the higher order *correlations*.

Besides the zero-crossing interpretation, one can also interpret  $\rho(k)$  upon noting the spectral representation

$$\rho(k) = \frac{\int_{-\pi}^{\pi} \cos(\omega) |H_k(\omega)|^2 dF(\omega)}{\int_{-\pi}^{\pi} |H_k(\omega)|^2 dF(\omega)},$$

where  $H_k(\omega) := \mathcal{H}_k(\exp(i\omega))$  is the transfer function of the RSD filter and  $F(\omega)$  is the cumulative spectral distribution function of  $\{X_t\}$ . According to the spectral representation, the inverse cosine transform

$$\psi(k) := \arccos(\rho(k)) \in [0, \pi]$$

defines a measure for the *center of spectral mass* of the filtered time series  $\{X_t(k)\}$  whose power spectrum equals  $|H_k(\omega)|^2 dF(\omega)$ . Due to the filtering, the center of spectral mass is

changed from  $\psi(0)$  to  $\psi(\pm 1), \psi(\pm 2), \dots$ . It is the trajectory of these changes that is explored by the PF approach to characterize the spectrum of  $\{X_t\}$ .

In (1), the repeated differencing corresponds to  $k < 0$ . It is easy to see that as  $k$  decreases, the low frequency contents of  $\{X_t\}$  are increasingly suppressed as a result of the shrinking bandwidth of the high-pass filter; this gives rise to a more oscillatory time series  $\{X_t(k)\}$ , and hence a decreased lag-one autocorrelation. On the other hand, as  $k$  grows, the repeated summing ( $k > 0$ ) increasingly attenuates the high frequency components of  $\{X_t\}$ , leading to a smoother (less oscillatory) time series  $\{X_t(k)\}$ , and thus an increased lag-one autocorrelation. In this way, the  $\rho(k)$ 's constitute a monotone increasing function of  $k$ . More importantly, as  $k$  varies, the *signature* of these lag-one autocorrelations completely determines the correlation structure of  $\{X_t\}$ . This forms the foundation for  $\rho(k)$  to be used as a characterization function of  $\{X_t\}$ . These results are summarized in the following, for proofs see Kedem and Li (1991), and Kedem (1994).

*Proposition 1.* Assume that  $\{X_t\}$  is a zero-mean stationary time series and  $\rho(k)$  is the lag-one autocorrelation of  $\{X_t(k)\}$ . Then, as  $k$  varies, the  $\rho(k)$ 's form a monotone increasing function of  $k$ . The monotonicity is strict unless  $\{X_t\}$  is a pure sinusoid of the form  $X_t = A \cos(\omega_0 t + \phi)$  — in this case,  $\rho(k) = \cos \omega_0 k$  for all  $k$ . Furthermore, the function  $\rho(k)$  is mathematically equivalent to the autocorrelation function  $r_\tau$  of  $\{X_t\}$ .

According to these results, the mapping  $r_\tau \mapsto \rho(k)$  is one-to-one, and it transforms the autocorrelation function of  $\{X_t\}$  into an increasing function of  $k$ . The monotonicity is a desirable graphical feature when  $\rho(k)$  is employed to compare different correlation structures. Finally, we point out that  $\rho(k)$  is sufficient to determine the correlation structure of  $\{X_t\}$  even if it is restricted to  $k \geq 0$  (repeated summing only) or  $k \leq 0$  (repeated differencing only). Including both summing and differencing makes the filter bank practically more versatile in terms of the frequency response.

## 2.2 Complex Exponential Filtering

Though useful in many ways and easy to interpret, the RSD may still lack the flexibility of frequency selection in some applications because it cannot simultaneously suppress the low and high frequency components. This problem may be resolved with the help of the complex exponential filter defined as

$$\mathcal{H}_\alpha(z) := \sum_{j=0}^{\infty} \alpha^{*j} z^{-j} = \frac{1}{1 - \alpha^* z^{-1}},$$

where  $\alpha := \eta \exp(-i\theta)$  is a complex parameter, with  $\eta \in [0, 1)$  and  $\theta \in (-\pi, \pi]$ , and  $\alpha^*$  is the complex conjugate of  $\alpha$ . Applying this filter to  $\{X_t\}$  gives rise to

$$X_t(\alpha) := \mathcal{H}_\alpha(z) X_t = \sum_{j=0}^{\infty} \alpha^{*j} X_{t-j}. \quad (3)$$

The CEF essentially behaves like *band-pass* filtering in which the effective pass band is centered at frequency  $\theta$  with the (half-power) bandwidth  $(1 - \eta)/\sqrt{\eta}$  [Priestley (1981), p. 517]. Based on  $\{X_t(\alpha)\}$ , a characterization function is obtained by considering

$$\gamma_\theta(\eta) := \Re\{e^{-i\theta} \rho(\alpha)\} \quad (0 \leq \eta < 1), \quad (4)$$

where  $\rho(\alpha) := E\{X_{t+1}(\alpha) X_t^*(\alpha)\}/E\{|X_t(\alpha)|^2\}$  is the lag-one autocorrelation of  $\{X_t(\alpha)\}$ . Since it is a demodulated version of  $\rho(\alpha)$ , we refer to  $\gamma_\theta(\eta)$  as the demodulated lag-one autocorrelation of  $\{X_t(\alpha)\}$  [Li (1996a)].

As with the RSD, two essential properties of  $\rho(k)$  are preserved by the CEF characterization function  $\gamma_\theta(\eta)$ , i.e., the equivalence to the autocorrelation function and the monotonicity. The following results, cited from Li (1996a), justify these and some other properties of  $\gamma_\theta(\eta)$  necessary for the development of divergence measures.

*Proposition 2.* With  $\theta \neq (2m + 1)\pi/(2n)$  for any integers  $m$  and  $n$ , the function  $\gamma_\theta(\eta)$  of  $\eta \in [0, 1)$  is mathematically equivalent to the autocorrelation function  $r_\tau$  and hence completely characterizes the correlation structure of  $\{X_t\}$ .

*Proposition 3.* For any fixed  $\theta$ , the function  $\gamma_\theta(\eta)$  is infinitely differentiable with respect to  $\eta \in [0, 1)$ , even if  $\{X_t\}$  has a mixed spectrum.

*Proposition 4.* For any fixed  $\theta$ , the function  $\gamma_\theta(\eta)$  is a non-decreasing function of  $\eta \in [0, 1)$ . The monotonicity is strict unless  $\{X_t\}$  coincides almost surely with a pure sinusoidal sequence of the form  $S_t := A \exp\{i(\omega_0 + \theta)t\} + B \exp\{-i(\omega_0 - \theta)t\}$  for some  $\omega_0 \in [0, \pi]$  — in this case,  $\gamma_\theta(\eta) = \cos \omega_0$  for all  $\eta \in [0, 1)$ .

These results can be summarized as saying that  $\gamma_\theta(\eta)$  transforms the autocorrelation function of  $\{X_t\}$  into a smooth monotone function of  $\eta$  which is unique for almost every  $\theta$  (the set of “bad” values of  $\theta$  has Lebesgue measure zero). In the following, we always assume  $\{X_t\} \neq \{S_t\}$ , so that both  $\gamma_\theta(\eta)$  and  $\rho(k)$  are *strictly* increasing.

To end this section, we note that the domain of  $\gamma_\theta(\eta)$  can be extended to  $\eta \in (-1, 1)$  without loss of the monotonicity. Since  $\gamma_\theta(\eta)$  is symmetric in the sense that  $\gamma_\theta(-\eta) = -\gamma_{\pi-\theta}(\eta)$  [Li (1996a)], it suffices to confine  $\gamma_\theta(\eta)$  in the interval  $\eta \in [0, 1)$ . Further, the monotonicity and boundedness of  $\gamma_\theta(\eta)$  ensure the existence of a finite limiting value,  $\gamma_\theta(1^-)$ , as  $\eta \rightarrow 1^-$ . Therefore, the domain of  $\gamma_\theta(\eta)$  can be extended to  $[0, 1]$  upon regarding  $\gamma_\theta(1^-)$  as the value of

$\gamma_\theta(\eta)$  at  $\eta = 1$ . In the sequel, this *extended domain* of  $\gamma_\theta(\eta)$  is always assumed. Finally we note that  $\gamma_\theta(\eta) = \eta$  for any  $\theta$  if  $\{X_t\}$  is white noise.

### 3. DIVERGENCE MEASURES

Given two time series  $\{X_t^j\}$ , ( $j = 0, 1$ ), let  $\rho^j(k)$  and  $\gamma_\theta^j(\eta)$  denote the corresponding characterization functions from RSD and CEF, respectively. A simple way of quantifying the spectral deviation between  $\{X_t^0\}$  and  $\{X_t^1\}$  is to use divergence measures motivated by the  $L_p$  distance. In fact, for any  $p > 0$ , one may define

$$L_p(\rho^1, \rho^0) := \sum_{k=-\infty}^{\infty} w_k |\rho^1(k) - \rho^0(k)|^p, \quad (5)$$

where  $w_k \geq 0$  is a suitable weight function so that  $L_p(\rho^1, \rho^0) = 0$  implies  $\rho^1(k) = \rho^0(k)$  at least for  $k \geq 0$  or  $k \leq 0$ . Moreover, let  $\Theta \subseteq [0, \pi]$  be a set of frequencies, and let  $\eta_a := \eta_a(\theta)$  and  $\eta_b := \eta_b(\theta)$  be functions of  $\theta$  such that  $0 \leq \eta_a < \eta_b \leq 1$ ; then, with  $\Omega := \{(\theta, \eta) : \theta \in \Theta, \eta \in [\eta_a, \eta_b]\}$ , one may define

$$L_p(\gamma^1, \gamma^0) := \int_{\Omega} |\gamma_\theta^1(\eta) - \gamma_\theta^0(\eta)|^p d\nu, \quad (6)$$

where  $\nu = \nu(\theta, \eta)$  is a suitable measure on  $\Omega$  such that  $L_p(\gamma^1, \gamma^0) = 0$  implies  $\gamma_\theta^1(\eta) = \gamma_\theta^0(\eta)$  at least in an open subinterval of  $[\eta_a, \eta_b]$  for some  $\theta$  satisfying the condition in Proposition 2. With these requirements satisfied, a necessary and sufficient condition for  $\{X_t^0\}$  and  $\{X_t^1\}$  to have the same correlation structure is that the divergence measures in (5) and (6) be zero. Note that these divergence measures are analogous to the Cramér-von Mises distance defined in terms of the cumulative spectral distribution functions [Deshayes and Picard (1986)].

An alternative design of divergence measures can be based on the *increments* of the characterization functions. For any  $0 \leq \eta_a < \eta_b < 1$ , define

$$d_\theta(\eta) := \begin{cases} \frac{1}{2} \dot{\gamma}_\theta(\eta) & \text{if } \eta \in (\eta_a, \eta_b), \\ \frac{1}{2}(\gamma_\theta(\eta_a) + 1) & \text{if } \eta = \eta_a, \\ \frac{1}{2}(1 - \gamma_\theta(\eta_b)) & \text{if } \eta = \eta_b, \end{cases} \quad (7)$$

where  $\dot{\gamma}_\theta(\eta)$  is the derivative of  $\gamma_\theta(\eta)$  with respect to  $\eta$ . In addition, let  $\mu(\eta)$  be a mixed-type Lebesgue measure, such that  $\mu(d\eta) = d\eta$  for  $\eta \in (\eta_a, \eta_b)$  and  $\mu(\{\eta_a\}) = \mu(\{\eta_b\}) = 1$ . Then, we have the following results.

*Theorem 1.* For any real-valued zero-mean stationary time series  $\{X_t\} \neq \{S_t\}$ , the function  $d_\theta(\eta)$  defines on the interval  $[\eta_a, \eta_b]$  a normalized density (with respect to  $\mu$ ), which is strictly positive within  $(\eta_a, \eta_b)$  and uniquely determines the autocorrelation function  $r_\tau$  of  $\{X_t\}$  for almost every  $\theta \in (-\pi, \pi]$ .

*Proof.* By Propositions 3 and 4, the function  $\gamma_\theta(\eta)$  is differentiable and strictly increasing. This, combined with the fact that  $-1 \leq \gamma_\theta(\eta) \leq 1$ , implies that the function  $d_\theta(\eta)$  is finite, non-negative on  $[\eta_a, \eta_b]$ , and strictly positive in  $(\eta_a, \eta_b)$ . Further, it is easy to show that

$$\int_{[\eta_a, \eta_b]} d_\theta(\eta) d\mu = 1.$$

Therefore  $d_\theta(\eta)$  is a normalized density function on  $[\eta_a, \eta_b]$  with respect to  $\mu$ . It may also be shown that

$$\gamma_\theta(\eta) = 2 \int_{[\eta_a, \eta]} d_\theta(\xi) d\mu - 1 \quad \forall \eta \in (\eta_a, \eta_b),$$

so that  $d_\theta(\eta)$  is equivalent to  $\gamma_\theta(\eta)$ .  $\square$

To further appreciate the role of  $d_\theta(\eta)$  in characterizing correlation structures, assume that  $\{X_t\}$  has a continuous spectrum with *normalized* spectral density function

$$f(\omega) := \sum_{\tau=-\infty}^{\infty} r_\tau e^{-i\tau\omega}.$$

If  $f(\omega)$  is sufficiently smooth with  $f(\theta) > 0$ , it may be shown [Li (1995), (1996a)] that  $\gamma_\theta(\eta) \rightarrow 1$  and  $\dot{\gamma}_\theta(\eta) \rightarrow 1/f(\theta)$  as  $\eta \rightarrow 1^-$ . Therefore, with  $\eta_b \rightarrow 1^-$ , one obtains  $d_\theta(\eta_b) \rightarrow 0$  and

$$\lim_{\eta \rightarrow 1^-} d_\theta(\eta) = \frac{1}{2f(\theta)}. \quad (8)$$

In other words, the limiting behavior of  $d_\theta(\eta)$  as  $\eta \rightarrow 1^-$  determines the spectral density function at  $\theta$ : a higher CEF density  $d_\theta(\eta)$  in the neighborhood of  $\eta = 1$  implies a lower spectral density  $f(\theta)$ . This relationship with the spectral density makes the use of  $d_\theta(\eta)$  more interpretable in the design of divergence measures.

Let  $d_\theta^j(\eta)$ , ( $j = 0, 1$ ), be obtained from  $\{X_t^j\}$  according to (7). Then, one may define the KL-type information divergence [Kullback (1959), Basseville (1989), Parzen (1992)]

$$I(d^1, d^0) := \int_{\Omega} \lambda \left( \frac{d_\theta^1(\eta)}{d_\theta^0(\eta)} \right) d_\theta^0(\eta) d\nu, \quad (9)$$

where  $\lambda(x)$  is a non-negative, smooth, and convex function of  $x > 0$  satisfying  $\lambda(1) = \dot{\lambda}(1) = 0$ , and  $\nu$  is a mixed measure on  $\Omega$  such that  $I(d^1, d^0) = 0$  implies  $\lambda(d_\theta^1(\eta)/d_\theta^0(\eta)) = 0$  (and thus  $d_\theta^1(\eta) = d_\theta^0(\eta)$ ) in the interval  $[\eta_a, \eta_c]$  for some  $\eta_c \in (\eta_a, \eta_b)$  and some  $\theta$  that satisfies the condition in Proposition 2. From (9), one may obtain the  $\chi^2$ -type divergence of index  $\beta$  [Parzen (1992)] by taking  $\lambda(x) = \lambda_\beta(x)$ , where

$$\lambda_\beta(x) := \begin{cases} (1 + \beta)^{-1}(\beta^{-1}x(x^\beta - 1) - x - 1) & \text{if } \beta \neq 0, -1, \\ x \log x - x + 1 & \text{if } \beta = 0, \\ x - \log x - 1 & \text{if } \beta = -1. \end{cases}$$

The Kullback-Leibler information corresponds to  $\beta = -1$ .

Since  $\lambda(x) \geq 0$  for any  $x > 0$  and  $\lambda(x) = 0$  if and only if  $x = 1$ , the information divergence (or  $I$ -divergence) in (9) can be regarded as measuring the departure of the ratio  $d_\theta^1(\eta)/d_\theta^0(\eta)$  from unity by a weighted average of  $\lambda(d_\theta^1(\eta)/d_\theta^0(\eta))$ . This interpretation obviously carries over to *non-normalized* densities [Parzen (1992)] and therefore leads to the following divergence measure:

$$D(d^1, d^0) := I(d^1/d^0, 1) = \int_{\Omega} \lambda\left(\frac{d_\theta^1(\eta)}{d_\theta^0(\eta)}\right) d\nu. \quad (10)$$

We may call  $D(d^1, d^0)$  the  $D$ -divergence.

For any time series  $\{X_t^0\}$  and  $\{X_t^1\}$ , possibly with mixed spectra,  $I(d^1, d^0) = D(d^1, d^0) = 0$  if and only if  $\{X_t^0\}$  and  $\{X_t^1\}$  have an identical correlation structure.

Similar divergence measures can be defined using the increments of  $\rho(k)$ . In fact, for any fixed  $k_0 \geq 0$ , let  $\Delta(k_0) := \frac{1}{2}(1 - \rho(k_0))$  and  $\Delta(k) := \frac{1}{2}(\rho(k+1) - \rho(k))$  for  $k < k_0$ ; then, by Proposition 1,  $\{\Delta(k)\}$  is equivalent to  $\{\rho(k)\}$  with  $\Delta(k) > 0$ . If the spectral density  $f(\theta)$  is smooth with  $f(\pi) > 0$ , then, it can be shown [Kedem and Slud (1982)] that  $\rho(k) \rightarrow -1$  as  $k \rightarrow -\infty$  and hence  $\sum_{k=-\infty}^{k_0} \Delta(k) = 1$ . The spectral deviation between  $\{X_t^0\}$  and  $\{X_t^1\}$  may be measured by

$$I(\Delta^1, \Delta^0) := \sum_{k=-\infty}^{k_0} \lambda\left(\frac{\Delta^1(k)}{\Delta^0(k)}\right) \Delta^0(k), \quad (11)$$

$$D(\Delta^1, \Delta^0) := \sum_{k=-\infty}^{k_0} \lambda\left(\frac{\Delta^1(k)}{\Delta^0(k)}\right). \quad (12)$$

According to Proposition 1, these divergence measures vanish if and only if  $\{X_t^0\}$  and  $\{X_t^1\}$  have the same autocorrelation function.

#### 4. SENSITIVITY ANALYSIS

To investigate the sensitivity of the new divergence measures to different types of spectral deviations, and especially to narrow band contaminations, we carry out an analysis of *local curvatures* [e.g., Dey and Birmiwal (1994)]. A local curvature can be interpreted as the instantaneous acceleration of the divergence measure when it takes off from zero as one time series starts to depart from the other. Although the local curvature is only a partial measurement of sensitivity, its mathematical tractability makes it a standard tool for sensitivity analysis. In the following, we only consider the CEF divergence measures  $I(d^1, d^0)$  and  $D(d^1, d^0)$ . The other measures can be investigated in a similar way, and the results will be reported elsewhere.

## 4.1 General Results

Before discussing the narrow band spectral components which tend to dominate spectral density based divergence measures, let us first consider the following spectral departure models of more generality. The first model, called the *additive departure*, takes the form

$$f_1(\omega) = (1 - \epsilon) f_0(\omega) + \epsilon g(\omega), \quad (13)$$

and the second model, called the *multiplicative departure*, can be written as

$$f_1(\omega) = c(\epsilon) \{f_0(\omega)\}^{1-\epsilon} \{g(\omega)\}^\epsilon. \quad (14)$$

In both models,  $f_j(\omega)$  represents the (normalized) spectral density of  $\{X_t^j\}$ . The function  $g(\omega)$ , also a normalized spectral density, characterizes the “direction” of departure from  $f_0(\omega)$ . For convenience, we denote by  $\{X_t^g\}$  the time series whose spectral density equals  $g(\omega)$  and by  $\{X_t^g(\alpha)\}$  the filtered version of  $\{X_t^g\}$ . The constant  $c(\epsilon)$  in (14) normalizes  $f_1(\omega)$  so that  $\int_{-\pi}^{\pi} f_1(\omega) d\omega = 1$ . The constant  $\epsilon$  in both models represents the severeness of spectral deviation as it varies in  $[0, 1]$ .

With the spectral departure models given by (13) and (14), the local curvature of a divergence measure is defined as the second derivative of the divergence measure with respect to  $\epsilon$  at  $\epsilon = 0$  [e.g., Dey and Birmiwal (1994)]. Because of their dependence on  $f_0$  and  $g$ , we denote the local curvatures of  $I(d^1, d^0)$  and  $D(d^1, d^0)$  by  $C_I(f_0, g)$  and  $C_D(f_0, g)$ , respectively. Note that the first derivatives of these divergence measures at  $\epsilon = 0$  are equal to zero because  $\dot{\lambda}(1) = 0$ . Also note that  $C_I(f_0, g)$  and  $C_D(f_0, g)$  do not depend on  $\epsilon$ .

The local curvatures under the additive model are given in the following theorem, and the derivation can be found in Appendix A.

*Theorem 2.* Under the additive spectral departure model (13), the local curvatures of  $I(d^1, d^0)$  and  $D(d^1, d^0)$  can be expressed respectively as

$$C_I(f_0, g) = \ddot{\lambda}(1) \int_{\Omega} c_{\theta}^g(\eta) d_{\theta}^0(\eta) d\nu, \quad (15)$$

$$C_D(f_0, g) = \ddot{\lambda}(1) \int_{\Omega} c_{\theta}^g(\eta) d\nu, \quad (16)$$

where  $c_{\theta}^g(\eta)$  is defined by

$$c_{\theta}^g(\eta) := \left\{ \frac{\varphi_{\theta}^g(\eta) (d_{\theta}^g(\eta) - d_{\theta}^0(\eta)) + \frac{1}{2} \dot{\varphi}_{\theta}^g(\eta) (\gamma_{\theta}^g(\eta) - \gamma_{\theta}^0(\eta))}{d_{\theta}^0(\eta)} \right\}^2 \quad (17)$$

with  $\varphi_{\theta}^g(\eta) := E\{|X_t^g(\alpha)|^2\}/E\{|X_t^0(\alpha)|^2\}$  being the “noise-to-signal ratio” after the filtering and  $\dot{\varphi}_{\theta}^g(\eta)$  being the derivative of  $\varphi_{\theta}^g(\eta)$  with respect to  $\eta$ .

In the multiplicative model (14), we assume that  $\log(g(\omega)/f_0(\omega))$  is bounded so that

$$h(\omega) := f_0(\omega) \left\{ c - \log \frac{g(\omega)}{f_0(\omega)} \right\} \quad (18)$$

is non-negative and integrable for some constant  $c \geq \sup\{\log(g(\omega)/f_0(\omega))\}$ . Under this assumption, one can regard  $h(\omega)$  as a (possibly non-normalized) spectral density which corresponds to a zero-mean stationary time series  $\{X_t^h\}$ . Similarly, one can define  $X_t^h(\alpha)$ ,  $\gamma_\theta^h(\eta)$ , and  $d_\theta^h(\eta)$ . Note that the last two functions of  $\eta$  do not depend on the variance of  $\{X_t^h\}$ . With this notation, the next theorem, a proof of which can be found in Appendix B, provides the expressions for  $C_I(f_0, g)$  and  $C_D(f_0, g)$ .

*Theorem 3.* Under the multiplicative spectral departure model (14), the local curvatures of  $I(d^1, d^0)$  and  $D(d^1, d^0)$  can be expressed respectively as

$$C_I(f_0, g) = \ddot{\lambda}(1) \int_{\Omega} c_\theta^h(\eta) d_\theta^0(\eta) d\nu, \quad (19)$$

$$C_D(f_0, g) = \ddot{\lambda}(1) \int_{\Omega} c_\theta^h(\eta) d\nu, \quad (20)$$

where

$$c_\theta^h(\eta) := \left\{ \frac{\varphi_\theta^h(\eta) (d_\theta^h(\eta) - d_\theta^0(\eta)) + \frac{1}{2} \dot{\varphi}_\theta^h(\eta) (\gamma_\theta^h(\eta) - \gamma_\theta^0(\eta))}{d_\theta^0(\eta)} \right\}^2 \quad (21)$$

with  $\varphi_\theta^h(\eta) := E\{|X_t^h(\alpha)|^2\}/E\{|X_t^0(\alpha)|^2\}$  representing the relative power of contamination after the filtering and  $\dot{\varphi}_\theta^h(\eta)$  being the derivative of  $\varphi_\theta^h(\eta)$  with respect to  $\eta$ .

In Theorems 2 and 3, we note that the sensitivity of  $I(d^1, d^0)$  and  $D(d^1, d^0)$  is directly related to the dissimilarity between the spectrum  $f_0$  and the direction  $g$  in which the deviation occurs. This dissimilarity, as shown in (17) and (21), is measured explicitly in terms of the CEF characterization functions  $\gamma_\theta(\eta)$  and  $d_\theta(\eta)$ . Via  $\varphi_\theta^g(\eta)$  and  $\varphi_\theta^h(\eta)$ , the sensitivity is also directly related to how effective the filter suppresses the contaminations. It is this dependence that gives the PF-based divergence measures an added feature, as compared to some spectral divergence measures, that can be utilized to balance the robustness and sensitivity to narrow band spectral components.

## 4.2 Sensitivity to Narrow Band Components

Using the general results, we now consider some special cases that are directed toward understanding the sensitivity of the CEF divergence measures  $I(d^1, d^0)$  and  $D(d^1, d^0)$  to narrow band spectral components. For simplicity, we take  $\theta = \{\theta\}$  and let  $\nu(\theta, \cdot) = \mu(\cdot)$ . Under this assumption, the integrals over  $\Omega$  become single integrals with respect to  $\mu$  for  $\eta \in [\eta_a, \eta_b]$ .

### A. Narrow Band Spectral Peaks

To investigate the sensitivity to sharp spectral peaks, consider the additive departure model (13) with

$$g(\omega) = \begin{cases} \frac{1}{2}\Delta^{-1} & \text{if } |\omega| \in (\omega_0 - \frac{1}{2}\Delta, \omega_0 + \frac{1}{2}\Delta), \\ 0 & \text{otherwise.} \end{cases} \quad (22)$$

Assuming  $0 < \Delta \ll 1$ , this choice of  $g(\omega)$  represents the situation where  $\{X_t^0\}$  is “contaminated” by an additive narrow band noise of bandwidth  $\Delta$  centered at frequency  $\omega_0$ . Because  $g(\omega)$  is normalized, the following sensitivity analysis should be considered as one with respect to the *bandwidth*, rather than the power, of the narrow band component.

As a corollary to Theorem 2, we have the following result, a proof of which is given in Appendix C.

*Corollary 1.* Let  $g(\omega)$  in the additive spectral departure model (13) be given by (22). If  $|\theta - \omega_0| = \mathcal{O}(1 - \eta_b)$  and  $1 - \eta_b = \mathcal{O}(\Delta^{(1-\varepsilon)/4})$  for some  $\varepsilon \in (0, 1)$ , then the local curvatures of  $I(d^1, d^0)$  and  $D(d^1, d^0)$  can be expressed as  $C_I(f_0, g) = \mathcal{O}(\Delta^{-3(1-\varepsilon)/4})$  and  $C_D(f_0, g) = \mathcal{O}(\Delta^{-3(1-\varepsilon)/4})$ , where  $\Delta > 0$  is small.

According to this result, the sensitivity of  $I(d^1, d^0)$  and  $D(d^1, d^0)$  to the spectral peak at  $\omega_0$  is governed by  $\theta$  and  $\eta_b$  — the closer is  $\eta_b$  to unity and  $\theta$  to  $\omega_0$ , the more sensitive are the divergence measures. Therefore, if the narrow band component is to be rejected, one may choose  $1 - \eta_b \approx \mathcal{O}(1)$  (or  $\varepsilon \approx 1$ ), so that  $C_I(f_0, g)$  and  $C_D(f_0, g)$  are *bounded* regardless of how small  $\Delta$  may be; if the narrow band component is to be detected, one may choose  $1 - \eta_b \approx \mathcal{O}(\Delta^{1/4})$  (or  $\varepsilon \approx 0$ ), so that the divergence measures are more responsive to the presence of  $g(\omega)$ .

This is in contrast with the KL spectral divergence whose local curvature is always *unbounded* as  $\Delta$  approaches zero. In other words, the KL spectral divergence lacks robustness to the narrow band spectral peaks of contamination. To see this, we note that the KL spectral divergence takes the form [e.g., Parzen (1992)]

$$D_{KL}(f_1, f_0) := \int_{-\pi}^{\pi} \lambda_{-1} \left( \frac{f_1(\omega)}{f_0(\omega)} \right) d\omega. \quad (23)$$

Since  $\ddot{\lambda}_{-1}(1) = 1$ , it is easy to show that, under the additive model (13), the local curvature of  $D_{KL}(f_1, f_0)$  can be written as

$$C_{KL}(f_0, g) = \int_{-\pi}^{\pi} \left( \frac{g(\omega)}{f_0(\omega)} - 1 \right)^2 d\omega.$$

For the particular choice of  $g(\omega)$  in (22), the local curvature becomes

$$C_{KL}(f_0, g) = \mathcal{O}(\Delta^{-1}),$$

which grows without bound as  $\Delta \rightarrow 0$ .

According to Corollary 1, the sensitivity ratio of the CEF divergence measures versus the KL spectral divergence is  $\mathcal{O}(\Delta^{(3\epsilon+1)/4})$ . This can be used as an indicator of efficiency when  $I(d^1, d^0)$  and  $D(d^1, d^0)$  are compared with  $D_{KL}(f_1, f_0)$  in detecting narrow band components. Unlike the KL spectral divergence, however, the efficiency of  $I(d^1, d^0)$  and  $D(d^1, d^0)$  can be balanced with the robustness by properly selecting  $\theta$  and  $\eta_b$ .

### B. Impact of $\theta$ and $\eta_b$ on Sensitivity

The previous discussion reveals that the sensitivity of  $I(d^1, d^0)$  and  $D(d^1, d^0)$  to a spectral peak at  $\omega_0$  is directly related to how close  $\eta_b$  is to unity and  $\theta$  is to  $\omega_0$ . In the following, this relationship is further elaborated under the assumption that  $\{X_t^0\}$  is white noise with  $f_0(\omega) = 1/(2\pi)$  and  $g(\omega)$  takes its limiting form  $g(\omega) = \frac{1}{2}\{\delta(\omega + \omega_0) + \delta(\omega - \omega_0)\}$  as  $\Delta \rightarrow 0$  (i.e., the case of white noise plus a sinusoid), where  $\delta(\cdot)$  is the Dirac delta. In this case,  $C_I(f_0, g)$  and  $C_D(f_0, g)$  coincide with  $\tilde{C}_I(f_0, g)$  and  $\tilde{C}_D(f_0, g)$  in (C.1) and (C.2).

Under the white noise assumption, it is easy to show that  $\gamma_\theta^0(\eta) = \eta$ ,  $d_\theta^0(\eta) = \frac{1}{2}$  for  $\eta \in (\eta_a, \eta_b)$ ,  $d_\theta^0(\eta_a) = \frac{1}{2}(\eta_a + 1)$ , and  $d_\theta^0(\eta_b) = \frac{1}{2}(1 - \eta_b)$ . It may also be shown that  $E\{|X_t^0(\alpha)|^2\} = 1/(1 - \eta^2)$ , and hence  $\varphi_\theta^g(\eta) = \frac{1}{2}(1 - \eta^2)\{G_\eta(\omega_0 + \theta) + G_\eta(\omega_0 - \theta)\}$  (see Appendix C). Combining these leads to

$$C_I(f_0, g) = \frac{1}{2}\ddot{\lambda}(1) \left\{ \int_{\eta_a}^{\eta_b} c_\theta^g(\eta) d\eta + c_\theta^g(\eta_a)(\eta_a + 1) + c_\theta^g(\eta_b)(1 - \eta_b) \right\}, \quad (24)$$

$$C_D(f_0, g) = \ddot{\lambda}(1) \left\{ \int_{\eta_a}^{\eta_b} c_\theta^g(\eta) d\eta + c_\theta^g(\eta_a) + c_\theta^g(\eta_b) \right\}, \quad (25)$$

where

$$c_\theta^g(\eta) = \{\varphi_\theta^g(\eta) (2\kappa_\theta(\eta) \sin \omega_0 \sin \theta - 1) + \dot{\varphi}_\theta^g(\eta) (\gamma_\theta^g(\eta) - \eta)\}^2$$

for  $\eta \in (\eta_a, \eta_b)$  and

$$c_\theta^g(\eta_a) = \left\{ \frac{\varphi_\theta^g(\eta_a) (\gamma_\theta^g(\eta_a) - \eta_a)}{\eta_a + 1} \right\}^2, \quad c_\theta^g(\eta_b) = \left\{ \frac{\varphi_\theta^g(\eta_b) (\eta_b - \gamma_\theta^g(\eta_b))}{1 - \eta_b} \right\}^2.$$

Note that the difference between  $\gamma_\theta^g(\eta)$  and  $\gamma_\theta^0(\eta) = \eta$  at the end-points ( $\eta = \eta_a, \eta_b$ ) has a greater impact on  $C_D(f_0, g)$  than on  $C_I(f_0, g)$  as a result of the different weighting in the first two terms of (24) and (25). The larger weights in  $C_D(f_0, g)$  lead to a greater local curvature and hence a higher degree of sensitivity to the spectral peak.

Assuming  $\ddot{\lambda}(1) = 1$ , which holds when  $\lambda(x) = \lambda_{-1}(x)$ , the (standardized) local curvatures in (24) and (25) are calculated by numerical integration for various values of  $\omega_0$ ,  $\theta$ , and  $\eta_b$ . The results are plotted in Figure 1 as functions of  $\theta$ . As we can see from Figure 1, when  $\eta_b$  is fixed,

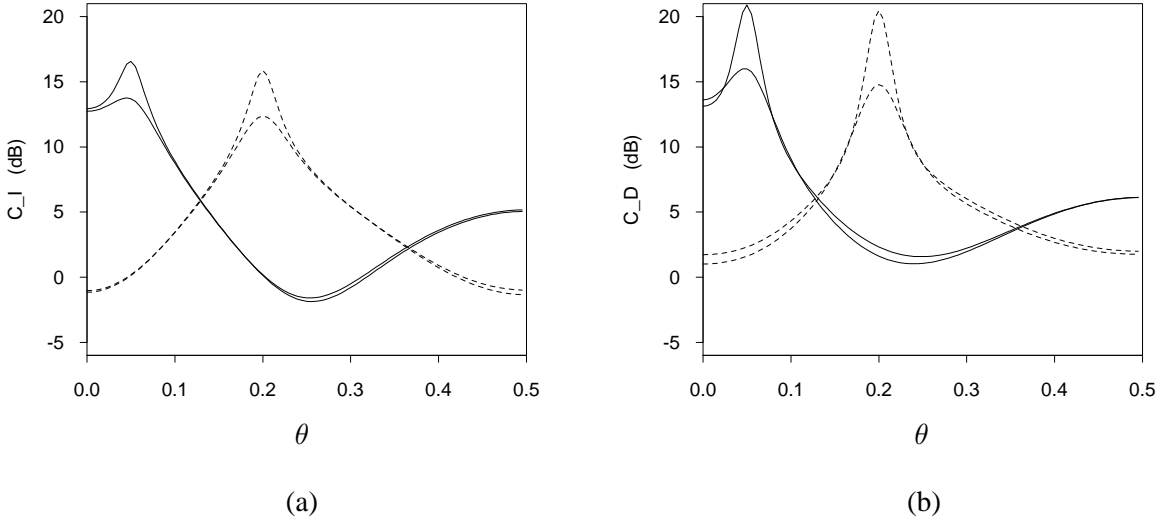


Figure 1: The local curvatures (in decibels) as functions of  $\theta$  ( $\times 2\pi$ ) in the case of white noise plus a sinusoid. (a) Plot of  $C_I(f_0, g)$ . (b) Plot of  $C_D(f_0, g)$ . Solid curves are for  $\omega_0 = 0.05 \times 2\pi$  and dashed curves for  $\omega_0 = 0.20 \times 2\pi$ . For each value of  $\omega_0$ , the curve with a sharper peak corresponds to  $\eta_b = 0.9$  and the other to  $\eta_b = 0.8$ . In all cases,  $\eta_a = 0$ .

the sensitivity of both  $C_I(f_0, g)$  and  $C_D(f_0, g)$  increases as  $\theta$  approaches  $\omega_0$  and decreases as  $\theta$  moves away from  $\omega_0$ . For a fixed  $\theta$  in the vicinity of  $\omega_0$ , the closer is  $\eta$  to unity the more sensitive are the divergence measures. These observations clearly suggest that one take a large value of  $\eta_b$  and place  $\theta$  as close as possible to where the narrow band component may occur if that component is to be detected. On the other hand, if a robust divergence measure is desired to resist the contamination of a narrow band component, one should either move  $\theta$  away from the contaminated frequency band or use a smaller value of  $\eta_b$ . With the flexible choice of  $\theta$  and  $\eta$ , both objectives can be achieved with the CEF divergence measures.

Comparing Figure 1(a) with Figure 1(b) confirms that  $D(d^1, d^0)$  is more sensitive than  $I(d^1, d^0)$  to the narrow band component, as indicated by the fact that  $C_D(f_0, g)$  is greater than  $C_I(f_0, g)$  for the same parameters. The sharper peaks in the graph of  $C_D(f_0, g)$  near  $\theta = \omega_0$  implies that  $D(d^1, d^0)$  has a better *frequency localization property*. Whereas  $I(d^1, d^0)$  is more robust against the narrow band component,  $D(d^1, d^0)$  seems preferable for detecting the narrow band component.

### C. Narrow Band Spectral Notches

A similar investigation can be carried out for the sensitivity to spectral notches. To this end, consider the multiplicative mixture (14). For simplicity, let  $\{X_t^0\}$  be white noise so that

$f_0(\omega) = 1/(2\pi)$ . Assume further that

$$g(\omega) = \begin{cases} g_\Delta & \text{if } |\omega| \in (\omega_0 + \frac{1}{2}\Delta, \omega_0 - \frac{1}{2}\Delta), \\ \frac{1 - 2\Delta g_\Delta}{2\pi - 2\Delta} & \text{otherwise,} \end{cases} \quad (26)$$

where

$$g_\Delta := \frac{1}{2\Delta + 2(\pi - \Delta) \exp(\pi\Delta^{-1})}.$$

Since  $g_\Delta \rightarrow 0$  as  $\Delta \rightarrow 0$ , this model gives rise to a narrow band spectral notch of bandwidth  $\Delta$  in the vicinity of  $\pm\omega_0$ .

To obtain the spectral density  $h(\omega)$  in (18), let  $c := \log(2\pi g_\Delta) + \pi\Delta^{-1}$ . With this choice, it is easy to verify that  $h(\omega) = \frac{1}{2}\Delta^{-1}$  if  $|\omega| \in (\omega_0 + \frac{1}{2}\Delta, \omega_0 - \frac{1}{2}\Delta)$  and  $h(\omega) = 0$  otherwise. In other words,  $h(\omega)$  coincides with the contamination spectrum given by (22). Upon noting the similarities between the local curvatures under the multiplicative model and those under the additive model, it is clear that if we replace  $g$  with  $h$ , then the sensitivity analysis in Sections 4.2A and 4.2B remains valid for the narrow band spectral notch considered here.

With regard to the KL spectral divergence, it is easy to show that, under the multiplicative model (14), we have

$$C_{KL}(f_0, g) = \int_{-\pi}^{\pi} \left\{ I_{KL}(g, f_0) + \log \frac{g(\omega)}{f_0(\omega)} \right\}^2 d\omega,$$

where  $I_{KL}(g, f_0) := \int_{-\pi}^{\pi} \lambda_{-1}(g(\omega)/f_0(\omega)) f_0(\omega) d\omega$  is the KL information. For the choice of  $g(\omega)$  in (26), it may be shown that  $\log g_\Delta = \mathcal{O}(\Delta^{-1})$ . Therefore,  $\log(g(\omega)/f_0(\omega)) = \mathcal{O}(\Delta^{-1})$  if  $|\omega| \in (\omega_0 + \frac{1}{2}\Delta, \omega_0 - \frac{1}{2}\Delta)$  and  $\log(g(\omega)/f_0(\omega)) = \mathcal{O}(1)$  otherwise. As a result, we obtain  $I_{KL}(g, f_0) = \mathcal{O}(1)$  and  $C_{KL}(f_0, g) = \mathcal{O}(\Delta^{-1})$ . This proves that the KL spectral divergence also lacks robustness against the narrow band spectral notch, as it does against contaminating spectral peaks.

## 5. CONSISTENT ESTIMATION

Given a finite-length time series  $\{X_1, \dots, X_n\}$ , the estimation of  $\rho(k)$  is quite straightforward. A simple estimator  $\hat{\rho}(k)$  may be obtained by taking the ordinary lag-one sample autocorrelation of the  $|k|$ th summation or difference  $\{X_1(k), \dots, X_{n-k}(k)\}$ . To estimate  $\gamma_\theta(\eta)$ , the filtered time series  $\{X_t(\alpha)\}$  can only be approximated by  $\hat{X}_t(\alpha) := \sum_{j=0}^{t-1} \alpha^{*j} X_{t-j}$ , ( $t = 1, \dots, n$ ). In practice,  $\{\hat{X}_t(\alpha)\}$  may be calculated recursively by  $\hat{X}_t(\alpha) = \alpha^* \hat{X}_{t-1}(\alpha) + X_t$ , with  $\hat{X}_0(\alpha) = 0$ . A natural estimator of  $\gamma_\theta(\eta)$  is

$$\hat{\gamma}_\theta(\eta) = \Re\{e^{-i\theta} \hat{\rho}(\alpha)\}, \quad (27)$$

where  $\hat{\rho}(\alpha)$  can be the ordinary lag-one sample autocorrelation of  $\{\hat{X}_1(\alpha), \dots, \hat{X}_n(\alpha)\}$  or the least-squares estimator

$$\hat{\rho}(\alpha) = \frac{\sum_{t=3}^n \{ \hat{X}_t(\alpha) \hat{X}_{t-1}^*(\alpha) + \hat{X}_{t-1}(\alpha) \hat{X}_{t-2}^*(\alpha) \}}{2 \sum_{t=3}^n |\hat{X}_{t-1}(\alpha)|^2}.$$

The latter minimizes the sum of forward and backward prediction errors

$$\sum_{t=3}^n \{ |\hat{X}_t(\alpha) - \rho \hat{X}_{t-1}(\alpha)|^2 + |\hat{X}_{t-2}(\alpha) - \rho^* \hat{X}_{t-1}(\alpha)|^2 \}$$

with respect to  $\rho$ . The least-squares estimator is employed in our simulations.

As the sample size  $n$  grows, it may be shown that both  $\hat{\rho}(k)$  and  $\hat{\gamma}_\theta(\eta)$  are *uniformly* consistent under some mild conditions, and the consistency even holds for mixed-spectrum time series. To be more specific, assume that  $\{X_t\}$  takes the form

$$X_t = \sum_{j=1}^m a_j \cos(\omega_j t + \phi_j) + Z_t, \quad (28)$$

where  $a_j > 0$  and  $\omega_j \in (0, \pi)$  are constants and the  $\phi_j$  are i.i.d. random variables with uniform distribution in  $(-\pi, \pi]$  and independent of  $\{Z_t\}$ . Let  $Z_t := \sum_{j=-\infty}^{\infty} b_j \xi_{t-j}$  be a general linear process, where  $\{b_j\}$  is an absolutely summable sequence of constants and  $\{\xi_t\}$  an i.i.d. sequence of random variables with zero mean and finite variance.

In the literature, results concerning the consistency of sample autocovariances of mixed spectrum processes such as  $\{X_t\}$  in (28) have not been entirely adequate. This issue was clarified recently in Li and Kedem (1993) and in Houdré and Kedem (1995), where it was shown that the sample autocovariances of  $\{X_t\}$  given by (28) are strongly consistent. In Li and Kedem (1993), the consistency is also extended to the sample autocovariances of time series obtained by passing  $\{X_t\}$  in (28) through a linear parametric filter. In this case, the sample autocovariances become functions of the filter parameter and hence *uniform* consistency is often desirable. This has been proven in Li and Kedem (1993). To achieve the uniform consistency, the parametric filter is required to be *uniformly strictly stable*, namely, the parametric impulse response of the filter is bounded in absolute value by a constant sequence of finite first moment.

With regard to the estimation of  $\rho(k)$  and  $\gamma_\theta(\eta)$ , it is fortunate that the RSD filter  $\mathcal{H}_k(z)$  and the CEF filter  $\mathcal{H}_\alpha(z)$  are indeed uniformly strictly stable for any  $|k| \leq K$ ,  $\eta \in [\eta_a, \eta_b] \subset [0, 1)$ , and  $\theta \in (-\pi, \pi]$ . Therefore, it can be shown, as Theorem 1 and Corollary 1 of Li and Kedem (1993), that the sample autocorrelations  $\hat{\rho}(k)$  and  $\hat{\rho}(\alpha)$  converge to  $\rho(k)$  and  $\rho(\alpha)$  not only almost surely but also uniformly in  $k$  and  $\alpha$  with  $|k| \leq K$  and  $|\alpha| \in [\eta_a, \eta_b]$ . Furthermore, since

the derivatives of  $\mathcal{H}_\alpha(z)$  with respect to  $\eta$  result in uniformly strictly stable filters, it follows that any derivative of  $\hat{\rho}(\alpha)$  with respect to  $\eta$  also converges almost surely and uniformly in  $\alpha$  to the corresponding derivative of  $\rho(\alpha)$  [Theorem 2 of Li and Kedem (1993)]. This high-order uniform consistency extends straightforwardly to  $\hat{\gamma}_\theta(\eta)$  in (27). Due to the uniform consistency, the sample estimates  $\hat{\rho}(k)$  and  $\hat{\gamma}_\theta(\eta)$  preserve sufficient information about  $\rho(k)$  and  $\gamma_\theta(\eta)$  as *functions* of  $k$  and  $\alpha$  respectively. The following theorem summarizes these results.

*Theorem 4.* Let  $\{X_t\}$  be the mixed-spectrum time series in (28) and  $\{Z_t\}$  a general linear process. Let  $\hat{\gamma}_\theta^{(j)}(\eta)$  and  $\gamma_\theta^{(j)}(\eta)$  represent the  $j$ th derivatives of  $\hat{\gamma}_\theta(\eta)$  and  $\gamma_\theta(\eta)$  with respect to  $\eta$ . Then, for any interval  $[\eta_a, \eta_b] \subset [0, 1)$ , it follows that

$$\lim_{n \rightarrow \infty} \sup_{\alpha \in \mathcal{A}} |\hat{\gamma}_\theta^{(j)}(\eta) - \gamma_\theta^{(j)}(\eta)| = 0$$

almost surely, where  $\mathcal{A} := \{\alpha : \eta \in [\eta_a, \eta_b], \theta \in (-\pi, \pi]\}$ . Under the same conditions,

$$\lim_{n \rightarrow \infty} \max_{|k| \leq K} |\hat{\rho}(k) - \rho(k)| = 0$$

almost surely for any  $K \geq 0$ .

The estimation of the CEF divergence measures is straightforward. For example, with obvious notation, one may use the natural estimates  $I(\hat{d}^1, \hat{d}^0)$  and  $D(\hat{d}^1, \hat{d}^0)$ . For the RSD divergence measures, one should replace  $\rho(k)$  with  $\hat{\rho}(k)$  and truncate the infinite sums in (11) and (12), for example, with  $k_0 = K$  and  $k \geq -K$ , to obtain the sample estimates  $I_K(\hat{\Delta}^1, \hat{\Delta}^0)$  and  $D_K(\hat{\Delta}^1, \hat{\Delta}^0)$ . Similar estimates can be obtained for the  $L_p$  distances (5) and (6). The following theorem is a direct consequence of Theorem 4.

*Theorem 5.* Let  $\{X_t^0\} \neq \{S_t\}$  and  $\{X_t^1\} \neq \{S_t\}$  be mixed-spectrum time series of the form (28). Then, for any closed subset  $\Omega \subset (-\pi, \pi] \times [0, 1)$ ,

$$\lim_{n \rightarrow \infty} I(\hat{d}^1, \hat{d}^0) = I(d^1, d^0) \quad \text{and} \quad \lim_{n \rightarrow \infty} D(\hat{d}^1, \hat{d}^0) = D(d^1, d^0)$$

almost surely. For any  $K \geq 0$ ,  $I_K(\hat{\Delta}^1, \hat{\Delta}^0)$  and  $D_K(\hat{\Delta}^1, \hat{\Delta}^0)$  are also strongly consistent in estimating the truncated  $I(\Delta^1, \Delta^0)$  and  $D(\Delta^1, \Delta^0)$  respectively. Similar results can be obtained for estimating the  $L_p$  distances.

Asymptotic distributions of  $\hat{\rho}(k)$ ,  $\hat{\gamma}_\theta(\eta)$ , and the related (discretized) divergence measures can be derived using the central limit theorems for the sample autocovariances of mixed-spectrum time series [Li, Kedem, and Yakowitz (1994), Li (1996b)]. In particular, for the mixed-spectrum time series  $\{X_t\}$  in (28), it may be shown that both  $\hat{\rho}(k)$  and  $\hat{\gamma}_\theta(\eta)$  converge in distribution to Gaussian processes whose covariance functions can be expressed explicitly using the spectrum of  $\{X_t\}$ . The derivation, however, is quite involved since in the case of CEF

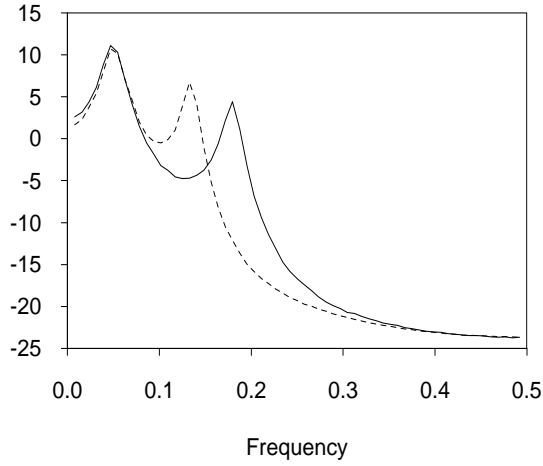


Figure 2: Spectral densities (in decibels) of the time series in Example 1: solid curve for Spectrum I and dashed curve for Spectrum II.

it involves complex-valued multivariate time series whose spectra contain discrete components. We therefore prefer to discuss this problem elsewhere. In the special case of  $\theta = 0$ , the asymptotic normality of  $\hat{\gamma}_0(\eta)$  for fixed  $\eta$  can be found in Li (1996a), where  $\{X_t\}$  is assumed to take the form (28) with  $\{\xi_t\}$  having finite fourth moments.

## 6. NUMERICAL EXAMPLES

In this section, we provide some simulation results to confirm the theoretical findings concerning the sensitivity of the new divergence measures. For simplicity, we concentrate on  $I(d^1, d^0)$ ,  $D(d^1, d^0)$ , and  $L_2(\rho^1, \rho^0)$ . Examples of application to speech processing can be found in Li and Gibson (1996).

*Example 1.* Consider the two spectral densities shown in Figure 2. Note that the first spectral peaks are nearly identical in these spectra. Also note that the spectra take small values and have similar characteristics in the high frequency region. Given two time series with these spectral densities, the objective in this example is to detect the spectral deviations using the divergence measures. This kind of problems can be encountered in on-line change detection of nonstationary time series, including the segmentation of continuous-speech waveforms and the failure monitoring of manufacturing processes [Basseville and Nikiforov (1993), Li and Gibson (1996)].

To evaluate the detection performance, we consider two cases: In Case (i),  $\{X_t^1\}$  and  $\{X_t^0\}$  both have Spectrum I; and in Case (ii),  $\{X_t^1\}$  has Spectrum I and  $\{X_t^0\}$  has Spectrum II. Under the hypothesis testing framework, these cases correspond respectively to the specifications

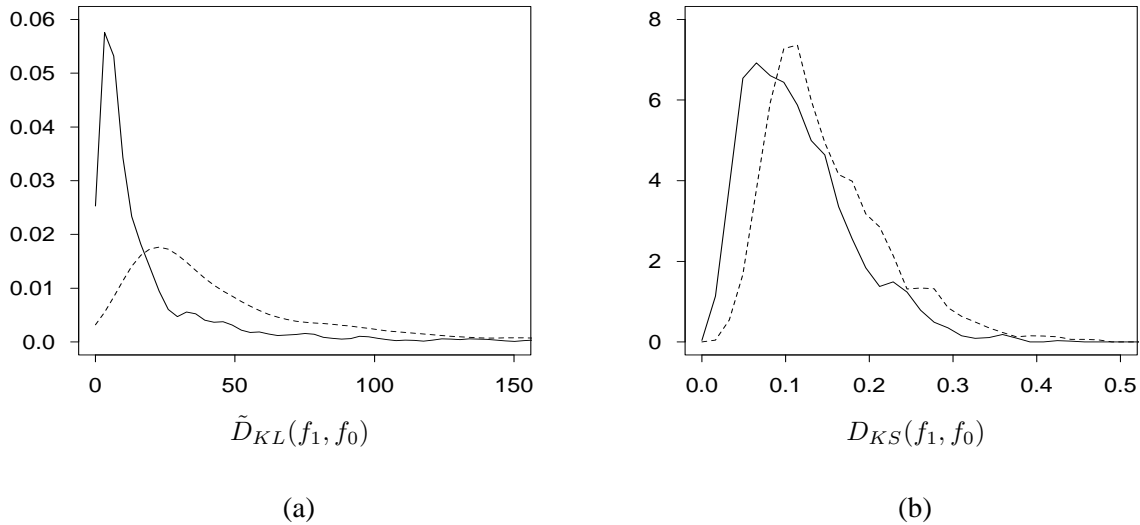


Figure 3: Distribution of the spectral divergence measures for discriminating the time series in Example 1: (a) the symmetrized Kullback-Leibler spectral divergence, (b) the Kolmogorov-Smirnov spectral distance. Solid curves for Case (i) and dashed curves for Case (ii).

of the null (composite) hypothesis that the autocorrelation functions of  $\{X_t^0\}$  and  $\{X_t^1\}$  are identical and of the alternative (composite) hypothesis that the autocorrelation functions are different; divergence measures play the role of test statistics. A good divergence measure should respond to Case (i) with small values and to Case (ii) with large ones. In general, for a divergence measure to be considered reasonable, the distribution of its values in Case (i) should be well separated from that in Case (ii); and the degree of separation determines the detection power of the divergence.

Figure 3(a) presents the distributions of the *symmetrized* KL spectral divergence defined by  $\tilde{D}_{KL}(f_1, f_0) := D_{KL}(f_1, f_0) + D_{KL}(f_0, f_1)$ . The symmetrization is made in order to remove the directionality of the original KL spectral divergence. The distributions in Figure 3(a) are obtained from  $N = 1000$  independent pairs  $\{X_t^0\}$  and  $\{X_t^1\}$  of length  $n = 128$ . For each pair of the time series, their periodograms are smoothed with a rectangular moving window of size  $2L + 1$ , and the smoothed periodograms are used in the calculation of (23), with the integral replaced by a sum over the Fourier frequencies. The window size for periodogram smoothing is experimentally chosen among some candidate values, ranging from  $L = 0$  (no smoothing) up to  $L = 11$  (35% smoothing), so as to obtain a maximal distribution separation as measured by the sum of detection probabilities at significance levels (false alarm probabilities) .01, .05, and .10. In this example, the optimal window size turns out to be  $L = 2$ . The distributions shown in Figure 3(a) are slightly smoothed histograms of the spectral divergence.

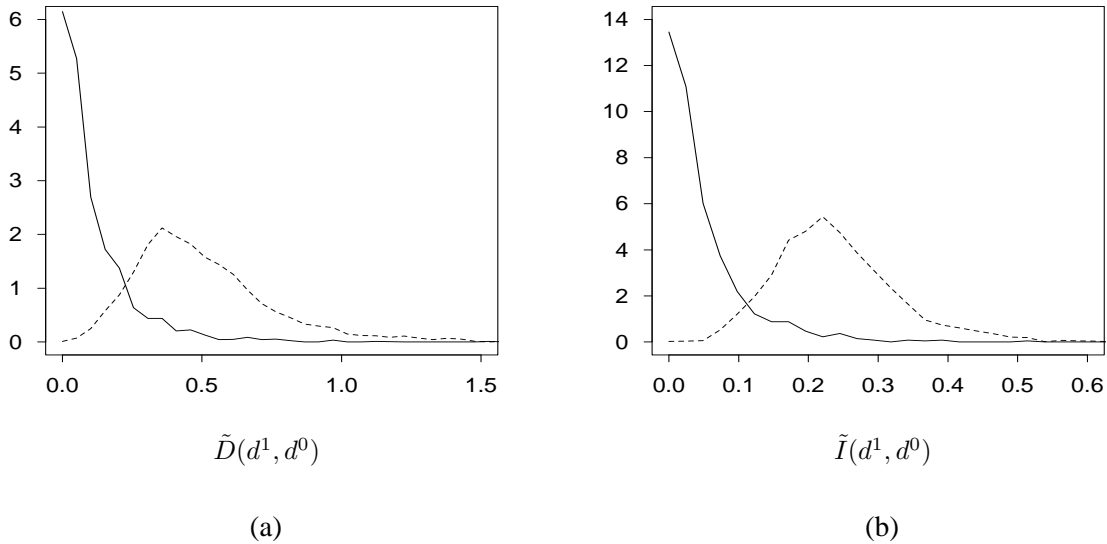


Figure 4: Distribution of the CEF divergence measures for the same data as in Figure 3: (a) the symmetrized  $D$ -divergence, (b) the symmetrized  $I$ -divergence.

A similar plot is shown in Figure 3(b) for the KS spectral distance defined by

$$D_{KS}(f_1, f_0) = \max_{\omega \in (-\pi, \pi]} \{|F_1(\omega) - F_0(\omega)|\},$$

where the  $F_j(\omega)$ , ( $j = 0, 1$ ), are the cumulative spectral distribution functions. In this case, the optimal window size for periodogram smoothing is  $L = 4$ .

As we can see from Figure 3, neither  $D_{KL}(f_1, f_0)$  nor  $D_{KS}(f_1, f_0)$  is very successful in discriminating the times series as a result of the smeared distributions. The dominating effect of spectral peaks and notches in these spectral density based divergence measures is largely responsible for their poor performance [see also Li (1996a), Basseville and Nikiforov (1993)].

For the same data, a much improved performance is achieved by the CEF divergence measures  $D(d^1, d^0)$  and  $I(d^1, d^0)$ . Shown in Figures 4(a) and 4(b) are the distributions of these divergence measures symmetrized in the same way as  $\tilde{D}_{KL}(f_1, f_0)$ . Clearly, the distributions are separated to a much higher degree (if not completely) than those in Figure 3, giving rise to a better detection power. Also, it is not surprising to observe that  $I(d^1, d^0)$  has better separated distributions than  $D(d^1, d^0)$  — according to the discussions in Section 4, the former is more robust to narrow band contaminations.

To obtain the distributions, the CEF divergence measures in (9) and (10) are calculated by replacing the integrals with the sums over a single value of  $\theta$  and  $M$  values of  $\eta$  taking the form  $\eta_i := \cos(\frac{1}{2}\pi(M - i + 1)/M)$ , ( $i = 1, \dots, M$ ). Note that the sampling interval in  $\eta$  decreases as  $\eta$  becomes close to unity. This choice is motivated partly by the relationship in (8) and partly by the necessity of frequency localization. As suggested by the sensitivity analysis in Section 4, the frequency parameter  $\theta$  should be close to where the spectra are suspected to be different

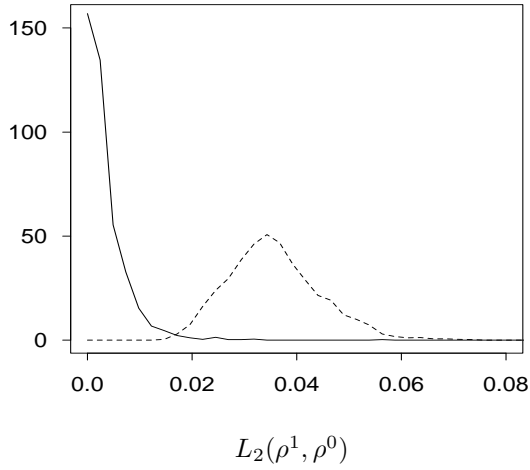


Figure 5: Distribution of the RSD-based  $L_2$  distance for the same data as in Figures 3 and 4.

and away from where the spectral contaminations are expected to occur. Among the candidate values of  $\theta$ , ranging from  $2\pi \times .10$  to  $2\pi \times .30$ , the choice of  $\theta = 2\pi \times .20$  maximizes the sum of experimentally determined detection probabilities at levels .01, .05, and .10. This comes out as no surprise because of the spectral characteristics shown in Figure 2. Since the spectral resolution of the data is  $2\pi/n$  while the bandwidth of the CEF is approximately  $1 - \eta$ , there is no need to take  $M$  greater than 4; otherwise, the smallest bandwidth (determined by  $\eta_M$ ) would fall below the spectral resolution (which equals .049). In Figure 4 are the distributions obtained with  $\theta = 2\pi \times .20$  and  $M = 4$ .

In this example, the RSD divergence  $L_2(\rho^1, \rho^0)$  achieves the best distribution separation and hence the greatest detection power. Shown in Figure 5 are the distributions of  $L_2(\rho^1, \rho^0)$  obtained according to (5) with the restriction that  $w_k = 1$  for  $-K \leq k \leq 0$  and  $w_k = 0$  for  $k > 0$  (i.e., differencing only). The truncation is made at an experimentally selected value of  $K = -3$  that maximizes the sum of detection probabilities at levels .01, .05, and .10. As can be seen, the distributions are almost completely separated. This superior performance can be explained to a large extent by the particular effectiveness of the RSD in suppressing the sample variations of the first spectral peaks.

At three significance levels (false alarm probabilities), Table 1 presents the simulated detection power of the divergence measures discussed above. All the detection thresholds are experimentally determined in Case (i), the null hypothesis, so as to yield the desired levels. Clearly, the CEF and RSD divergence measures are more powerful than the spectral density based divergence measures at all these levels. The performance of  $L_2(\rho^1, \rho^0)$  is especially remarkable, whereas the KS spectral distance gives the poorest performance.

Level	Divergence Measure				
	(a)	(b)	(c)	(d)	(e)
0.10	0.332	0.192	0.871	0.941	1.000
0.05	0.181	0.109	0.621	0.805	1.000
0.01	0.044	0.041	0.150	0.292	0.991

Table 1: Simulated detection power of different divergence measures for time series in Example 1: (a)  $\tilde{D}_{KL}(f_1, f_0)$ ; (b)  $\tilde{D}_{KS}(f_1, f_0)$ ; (c)  $\tilde{D}(d^1, d^0)$ ; (d)  $\tilde{I}(d^1, d^0)$ ; and (e)  $L_2(\rho^1, \rho^0)$ .

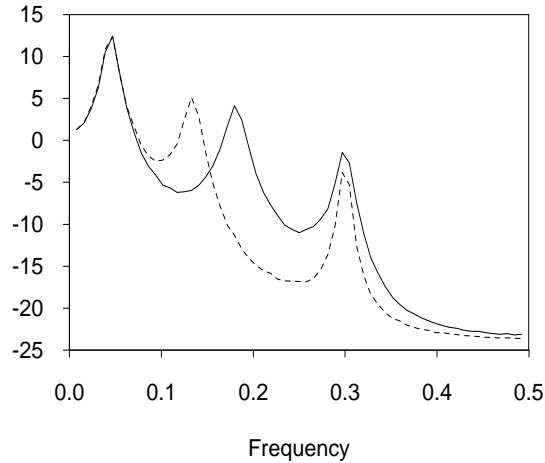


Figure 6: Spectral densities (in decibels) of the time series in Example 2: solid curve for Spectrum I and dashed curve for Spectrum II.

*Example 2.* Consider the spectra shown in Figure 6. These spectra are slightly more complex than those in Figure 2, with similar spectral peaks also in the high frequency region. In this case, the CEF becomes more powerful than the RSD because the former has greater flexibility of frequency selection. As pointed out earlier, the CEF is endowed with a frequency parameter that enables the filter to emphasize a certain frequency band and de-emphasize (but not eliminate) other frequency components. The RSD, on the other hand, cannot “zoom in” to a frequency band in the middle of the frequency domain.

With experimentally determined parameters  $L = 2$ ,  $K = 1$ , and  $\theta = 2\pi \times 0.21$ , as done in Example 1, the simulated detection powers are presented in Table 2. These results are obtained from 1000 independent pairs of time series of length  $n = 128$ . As we can see from Table 2, the CEF divergence measures outperform the RSD divergence  $L_2(\rho^1, \rho^0)$  as well as the spectral density based divergence measures. Again, the KS spectral distance gives the poorest performance. Note that the performance of  $I(d^1, d^0)$  is slightly superior to that of  $D(d^1, d^0)$  due to the same reasons as discussed in Example 1.

Level	Divergence Measure				
	(a)	(b)	(c)	(d)	(e)
0.10	0.592	0.164	0.817	0.885	0.538
0.05	0.329	0.078	0.608	0.714	0.392
0.01	0.090	0.009	0.269	0.288	0.178

Table 2: Simulated detection power of different divergence measures for the time series in Example 2: (a)  $\tilde{D}_{KL}(f_1, f_0)$ ; (b)  $D_{KS}(f_1, f_0)$ ; (c)  $\tilde{D}(d^1, d^0)$ ; (d)  $\tilde{I}(d^1, d^0)$ ; and (e)  $L_2(\rho^1, \rho^0)$ .

## 7. CONCLUDING REMARKS

In this paper, we have introduced some new divergence measures for detecting correlation deviations of time series. Derived from parametric filtering, these divergence measures are shown, both theoretically and experimentally, to be more effective than some traditional spectral density based divergence measures in detecting spectral deviations in the presence of spectral peaks and notches that have similar characteristics. The theoretical analysis is based on the local curvatures of the divergence measures derived under both additive and multiplicative spectral departure models.

As compared with some commonly used filter bank methods [e.g., Lavielle (1993)] in which spectral/correlation analysis is performed independently over certain fixed and non-overlapping frequency bands, the parametric filtering approach combines filtering, which suppresses spectral contaminations, with the characterization property, which ensures the preservation of *complete* spectral/correlation information as a result of the variable bandwidth. If used properly, effective change detection and discrimination of time series can be achieved with minimal loss of robustness.

Future research should be devoted to the further exploration of the statistical properties of the proposed divergence measures such as the asymptotic distribution. Impact of different combinations of the frequency and bandwidth parameters for the CEF should also be investigated.

## Appendix A: Proof of Theorem 2.

Let  $\nabla := d/d\epsilon$  be the differential operator with respect to  $\epsilon$ ; then, upon noting that  $\lambda'(1) = 0$ , it is easy to show that, for  $I(d^1, d^0)$  in (9),

$$\begin{aligned} C_I(f_0, g) &= \nabla^2 I(d^1, d^0) |_{\epsilon=0} \\ &= \ddot{\lambda}(1) \int_{\Omega} \left\{ \frac{\nabla d_{\theta}^1(\eta)}{d_{\theta}^0(\eta)} \right\}^2 \Big|_{\epsilon=0} d_{\theta}^0(\eta) d\nu. \end{aligned} \quad (\text{A.1})$$

Therefore, it suffices to evaluate  $\nabla d_{\theta}^1(\eta)$  at  $\epsilon = 0$ . To that end, let us drop, for convenience, the superscript or subscript ‘‘1’’ whenever it occurs in the following. With this in mind, it may be shown from (7) that

$$\nabla d_{\theta}(\eta) = \begin{cases} \frac{1}{2} \nabla \dot{\gamma}_{\theta}(\eta) & \text{if } \eta \in (\eta_a, \eta_b), \\ \frac{1}{2} \nabla \gamma_{\theta}(\eta_a) & \text{if } \eta = \eta_a, \\ -\frac{1}{2} \nabla \gamma_{\theta}(\eta_b) & \text{if } \eta = \eta_b. \end{cases} \quad (\text{A.2})$$

To evaluate the derivatives of  $\gamma_{\theta}(\eta)$ , consider the spectral representation [Li (1996a)]

$$\gamma_{\theta}(\eta) = \frac{\int_{-\pi}^{\pi} \cos(\omega - \theta) |H_{\alpha}(\omega)|^2 f(\omega) d\omega}{\int_{-\pi}^{\pi} |H_{\alpha}(\omega)|^2 f(\omega) d\omega},$$

where  $H_{\alpha}(\omega) := \mathcal{H}_{\alpha}(\exp(i\omega)) = 1/(1 - \alpha^* \exp(-i\omega))$  is the transfer function of the CEF filter. With  $f(\omega)$  taking the form (13), it follows immediately that

$$\begin{aligned} \nabla \gamma_{\theta}(\eta) &= \frac{\int_{-\pi}^{\pi} \cos(\omega - \theta) |H_{\alpha}(\omega)|^2 (g(\omega) - f_0(\omega)) d\omega}{\int_{-\pi}^{\pi} |H_{\alpha}(\omega)|^2 f(\omega) d\omega} \\ &\quad - \frac{\int_{-\pi}^{\pi} |H_{\alpha}(\omega)|^2 (g(\omega) - f_0(\omega)) d\omega \int_{-\pi}^{\pi} \cos(\omega - \theta) |H_{\alpha}(\omega)|^2 f(\omega) d\omega}{\left\{ \int_{-\pi}^{\pi} |H_{\alpha}(\omega)|^2 f(\omega) d\omega \right\}^2} \\ &= \frac{\int_{-\pi}^{\pi} |H_{\alpha}(\omega)|^2 g(\omega) d\omega}{\int_{-\pi}^{\pi} |H_{\alpha}(\omega)|^2 f(\omega) d\omega} \{ \gamma_{\theta}^g(\eta) - \gamma_{\theta}(\eta) \}. \end{aligned}$$

This, combined with  $f(\omega)|_{\epsilon=0} = f_0(\omega)$  and  $\gamma_{\theta}(\eta)|_{\epsilon=0} = \gamma_{\theta}^0(\eta)$ , leads to

$$\nabla \gamma_{\theta}(\eta) |_{\epsilon=0} = \varphi_{\theta}^g(\eta) \{ \gamma_{\theta}^g(\eta) - \gamma_{\theta}^0(\eta) \}, \quad (\text{A.3})$$

where  $\varphi_{\theta}^g(\eta)$  takes the form

$$\varphi_{\theta}^g(\eta) = \frac{\int_{-\pi}^{\pi} |H_{\alpha}(\omega)|^2 g(\omega) d\omega}{\int_{-\pi}^{\pi} |H_{\alpha}(\omega)|^2 f_0(\omega) d\omega}.$$

Further, it follows from (A.3) that

$$\begin{aligned}\nabla \dot{\gamma}_\theta(\eta) |_{\epsilon=0} &= \frac{d}{d\eta} \nabla \gamma_\theta(\eta) |_{\epsilon=0} \\ &= \varphi_\theta^g(\eta) \{ \dot{\gamma}_\theta^g(\eta) - \dot{\gamma}_\theta^0(\eta) \} + \dot{\varphi}_\theta^g(\eta) \{ \gamma_\theta^g(\eta) - \gamma_\theta^0(\eta) \}.\end{aligned}\quad (\text{A.4})$$

Substituting (A.3) and (A.4) in (A.2) yields

$$\nabla d_\theta(\eta) |_{\epsilon=0} = \varphi_\theta^g(\eta) \{ d_\theta^g(\eta) - d_\theta^0(\eta) \} + \frac{1}{2} \dot{\varphi}_\theta^g(\eta) \{ \gamma_\theta^g(\eta) - \gamma_\theta^0(\eta) \}.$$

This, combined with (A.1) proves (15). A similar argument leads to (16).

### Appendix B: Proof of Theorem 3.

As in the proof of Theorem 2, it is sufficient to evaluate  $\nabla d_\theta^1(\eta)$  at  $\epsilon = 0$  under the multiplicative mixture and substitute the result in (A.1). To this end, let  $h_\epsilon(\omega) := f_1(\omega) \{c - \log(g(\omega)/f_0(\omega))\}$ , where  $f_1(\omega)$  is given by (14); then, with superscript and/or subscript “1” omitted, it may be shown that

$$\begin{aligned}\nabla \gamma_\theta(\eta) &= \frac{\int_{-\pi}^{\pi} \cos(\omega - \theta) |H_\alpha(\omega)|^2 f(\omega) \log(g(\omega)/f_0(\omega)) d\omega}{\int_{-\pi}^{\pi} |H_\alpha(\omega)|^2 f(\omega) d\omega} \\ &\quad - \frac{\int_{-\pi}^{\pi} |H_\alpha(\omega)|^2 f(\omega) \log(g(\omega)/f_0(\omega)) d\omega}{\int_{-\pi}^{\pi} |H_\alpha(\omega)|^2 f(\omega) d\omega} \gamma_\theta(\eta) \\ &= - \frac{\int_{-\pi}^{\pi} |H_\alpha(\omega)|^2 h_\epsilon(\omega) d\omega}{\int_{-\pi}^{\pi} |H_\alpha(\omega)|^2 f(\omega) d\omega} \left\{ \frac{\int_{-\pi}^{\pi} \cos(\omega - \theta) |H_\alpha(\omega)|^2 h_\epsilon(\omega) d\omega}{\int_{-\pi}^{\pi} |H_\alpha(\omega)|^2 h_\epsilon(\omega) d\omega} - \gamma_\theta(\eta) \right\}.\end{aligned}$$

Upon noting that  $h_\epsilon(\omega)|_{\epsilon=0} = h(\omega)$ , one obtains

$$\nabla \gamma_\theta(\eta) |_{\epsilon=0} = - \varphi_\theta^h(\eta) \{ \gamma_\theta^h(\eta) - \gamma_\theta^0(\eta) \},$$

where

$$\varphi_\theta^h(\eta) = \frac{\int_{-\pi}^{\pi} |H_\alpha(\omega)|^2 h(\omega) d\omega}{\int_{-\pi}^{\pi} |H_\alpha(\omega)|^2 f_0(\omega) d\omega}.$$

This expression also leads to

$$\begin{aligned}-\nabla \dot{\gamma}_\theta(\eta) |_{\epsilon=0} &= - \frac{d}{d\eta} \nabla \gamma_\theta(\eta) |_{\epsilon=0} \\ &= \varphi_\theta^h(\eta) \{ \dot{\gamma}_\theta^h(\eta) - \dot{\gamma}_\theta^0(\eta) \} + \dot{\varphi}_\theta^h(\eta) \{ \gamma_\theta^h(\eta) - \gamma_\theta^0(\eta) \}.\end{aligned}$$

The rest of the proof is similar to that of Theorem 2. Note that  $\nabla\gamma_\theta(\eta)$  does not depend on the choice of  $c$  in  $h(\omega)$ , and neither do the local curvatures, because the dependence of  $h(\omega)$  on  $c$  is canceled out in the expression of  $c_\theta^h(\eta)$  in (21).

### Appendix C: Proof of Corollary 1.

Let  $G_\eta(\omega) := (1 - 2\eta \cos \omega + \eta^2)^{-1}$  so that  $|H_\alpha(\omega)|^2 = G_\eta(\omega - \theta)$ . Then, with  $|\theta - \omega_0| = \mathcal{O}(1 - \eta)$ , it may be shown that  $(d^j/d\omega^j) G_\eta(\omega_0 - \theta) = \mathcal{O}((1 - \eta)^{-2(j+1)})$  for  $j \geq 0$ . Further, assuming  $1 - \eta = \mathcal{O}(\Delta^{(1-\varepsilon)/4})$  for some  $\varepsilon \in (0, 1)$ , it follows that

$$\gamma_\theta^g(\eta) = \tilde{\gamma}_\theta^g(\eta) + \mathcal{O}(\Delta(1 - \eta)^{-2}),$$

where  $\tilde{\gamma}_\theta^g(\eta) := (1 - \kappa_\theta(\eta)) \cos(\omega_0 + \theta) + \kappa_\theta(\eta) \cos(\omega_0 - \theta)$  and

$$\kappa_\theta(\eta) := \frac{G_\eta(\omega_0 - \theta)}{G_\eta(\omega_0 + \theta) + G_\eta(\omega_0 - \theta)}.$$

Similarly, it may be shown that  $\dot{\gamma}_\theta^g(\eta) = \dot{\tilde{\gamma}}_\theta^g(\eta) + \mathcal{O}(\Delta(1 - \eta)^{-3})$  and hence

$$d_\theta^g(\eta) = \tilde{d}_\theta^g(\eta) + \mathcal{O}(\Delta(1 - \eta)^{-3}),$$

where  $\tilde{d}_\theta^g(\eta) := \dot{\kappa}_\theta(\eta) \sin \omega_0 \sin \theta$  for  $\eta \in (\eta_a, \eta_b)$ ,  $\tilde{d}_\theta^g(\eta_a) := \frac{1}{2}(\tilde{\gamma}_\theta^g(\eta_a) + 1)$ , and  $\tilde{d}_\theta^g(\eta_b) := \frac{1}{2}(1 - \tilde{\gamma}_\theta^g(\eta_b))$ . Assuming  $f(\omega)$  is sufficiently smooth with  $f(\theta) > 0$ , it follows that  $E\{|X_t^0(\alpha)|^2\} = \mathcal{O}((1 - \eta)^{-1})$ . This implies that

$$\varphi_\theta^g(\eta) = \tilde{\varphi}_\theta^g(\eta) + \mathcal{O}(\Delta(1 - \eta)^{-3}),$$

$$\dot{\varphi}_\theta^g(\eta) = \dot{\tilde{\varphi}}_\theta^g(\eta) + \mathcal{O}(\Delta(1 - \eta)^{-4}),$$

where  $\tilde{\varphi}_\theta^g(\eta) := \frac{1}{2}\{G_\eta(\omega_0 + \theta) + G_\eta(\omega_0 - \theta)\}/E\{|X_t^0(\alpha)|^2\}$ . Further, it may be shown that  $\tilde{d}_\theta^g(\eta) = \mathcal{O}((1 - \eta)^{-1})$ ,  $d_\theta^0(\eta) = \mathcal{O}(1)$ ,  $\tilde{\varphi}_\theta^g(\eta) = \mathcal{O}((1 - \eta)^{-1})$ , and  $\dot{\tilde{\varphi}}_\theta^g(\eta) = \mathcal{O}((1 - \eta)^{-2})$ .

Substituting these results in (15) and (16) leads to

$$C_I(f_0, g) = \tilde{C}_I(f_0, g) + \mathcal{O}(\Delta(1 - \eta_b)^{-3}), \quad (\text{C.1})$$

$$C_D(f_0, g) = \tilde{C}_D(f_0, g) + \mathcal{O}(\Delta(1 - \eta_b)^{-3}), \quad (\text{C.2})$$

where  $\tilde{C}_I(f_0, g)$  and  $\tilde{C}_D(f_0, g)$  are defined by (15) and (16) in terms of  $\tilde{\gamma}_\theta^g(\eta)$ ,  $\tilde{d}_\theta^g(\eta)$ ,  $\tilde{\varphi}_\theta^g(\eta)$ , and  $\dot{\tilde{\varphi}}_\theta^g(\eta)$ . The proof is complete upon noting that  $\tilde{C}_I(f_0, g) = \mathcal{O}((1 - \eta_b)^{-3})$  and  $\tilde{C}_D(f_0, g) = \mathcal{O}((1 - \eta_b)^{-3})$ .

### References

1. Basseville, M. (1989). Distance measures for signal processing and pattern recognition. *Signal Processing*, vol. 18, pp. 349–369.

2. Basseville, M. and Nikiforov I. V. (1993). *Detection of Abrupt Changes: Theory and Applications*, Englewood Cliffs, NJ: Prentice Hall.
3. Brockwell, P. J. and Davis, R. A. (1991). *Time Series: Theory and Methods*, 2nd Ed., New York: Springer.
4. Coates, D. S. and Diggle, P. J. (1986). Tests for comparing two estimated spectral densities. *J. Time Series Anal.*, vol. 7, no. 1, pp. 7–20.
5. Deshayes, J. and Picard, D. (1986). Off-line statistical analysis of change-point models using nonparametric and likelihood methods. *Detection of Abrupt Changes in Signals and Dynamical Systems*, M. Basseville and A. Benveniste, Eds., pp. 103–168, New York: Springer.
6. Dey, D. K. and Birmiwal, L. R. (1994). Robust Bayesian analysis using divergence measures. *Statist. Probab. Letters*, vol. 20, pp. 287–294.
7. Gray, R. M., Buzo, A., Gray, H., and Matsuyama, Y. (1980). Distortion measures for speech processing. *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 28, no. 4, pp. 367–376.
8. Houdré, C. and Kedem, B. (1995). A note on autocovariance estimation in the presence of discrete spectra. *Statist. Probab. Letters*, vol. 24, pp. 1–8.
9. Kedem, B. (1994). *Time Series Analysis by Higher Order Crossings*, Piscataway, NJ: IEEE Press.
10. Kedem, B. and Li, T. H. (1991). Monotone gain, first-order autocorrelation, and zero-crossing rate. *Ann. Statist.*, vol. 91, pp. 1672–1676.
11. Kedem, B. and Slud, E. (1982). Time series discrimination by higher order crossings. *Ann. Statist.*, vol. 10, pp. 786–794.
12. Kullback, S. (1959). *Information Theory and Statistics*. New York: Wiley.
13. Lavielle, M. (1993). Detection of changes in the spectrum of a multidimensional process. *IEEE Trans. Signal Processing*, vol. 41, no. 2, pp. 742–749.
14. Li, T. H. (1995). Time-correlation analysis of nonstationary time series. Tech. Report No. 236, Department of Statistics, Texas A&M University, College Station.

15. Li, T. H. (1996a). Discrimination of time series by parametric filtering. *J. American Statistical Association*, vol. 91, pp. 284–293.
16. Li, T. H. (1996b). Bartlett-type formulas for complex multivariate time series of mixed spectra. *Statist. Probab. Letters*, vol. 28, pp. 259–268.
17. Li, T. H. and J. D. Gibson (1994). Discriminant analysis of speech by parametric filtering. *Proc. 28th Annual Conf. Information Sciences and Systems*, Princeton, NJ, pp. 575–580.
18. Li, T. H. and J. D. Gibson (1996). Speech analysis and segmentation by parametric filtering. *IEEE Trans. Audio, Speech Processing*, vol. 4, no. 3, pp. 203–213.
19. Li, T. H., and Kedem, B. (1993). Strong consistency of the contraction mapping method for frequency estimation. *IEEE Trans. Inform. Theory*, vol. 39, no. 3, pp. 989–998.
20. Li, T. H., Kedem, B., and Yakowitz, S. (1994). Asymptotic normality of sample autocovariances with an application to frequency estimation. *Stochastic Processes and Their Applications*, vol. 52, pp. 329–349.
21. Parzen, E. (1992). Time series, statistics, and information. *New Directions in Time Series Analysis – Part I*. D. Brillinger *et al.* Eds., New York: Springer; pp. 265–286.
22. Priestley, M. B. (1981). *Spectral Analysis and Time Series*, San Diego, CA: Academic Press.
23. Zhang, G. and Taniguchi, M. (1995). Nonparametric approach for discriminant analysis in time series. *J. Nonparametric Statist.*, vol. 5, pp. 91–101.