OPTIMAL KERNELS FOR NONSTATIONARY SPECTRAL ESTIMATION

Akbar M. Sayeed, *Student Member, IEEE,* and Douglas L. Jones, *Member, IEEE*

Coordinated Science Laboratory
University of Illinois at Urbana–Champaign
1308 West Main Street
Urbana, IL 61801*

E-mail: akbar@csl.uiuc.edu
d-jones@csl.uiuc.edu

Tel: (217) 244-6384
Fax: (217) 244-1642

IEEE Transactions on Signal Processing, vol. 43, no. 2, pp. 478–491, February 1995.

Abstract

Current theories of a time-varying spectrum of a nonstationary process all involve, either by definition or by difficulties in estimation, an assumption that the signal statistics vary slowly over time. This restrictive quasi-stationarity assumption limits the use of existing estimation techniques to a small class of nonstationary processes. We overcome this limitation by deriving a statistically optimal kernel, within Cohen's class of time-frequency representations (TFRs), for estimating the Wigner-Ville Spectrum of a nonstationary process. We also solve the related problem of minimum mean-squared error estimation of an arbitrary bilinear TFR of a realization of a process from a correlated observation. Both optimal time-frequency invariant and time-frequency varying kernels are derived. It is proven that, in the presence of any additive noise, optimal performance requires a nontrivial kernel, and that optimal estimation may require smoothing filters very different from those based on a quasi-stationarity assumption. Examples confirm that the optimal estimators often yield tremendous improvements in performance over existing methods. In particular, the ability of the optimal kernel to suppress interference is quite remarkable, thus making the proposed framework potentially useful for interference suppression via time-frequency filtering.

*This work was supported by the National Science Foundation under Grant No. MIP 90-12747, the Joint Services Electronics Program under Grant No. N00014-90-J-1270, and the Schlumberger Foundation.

1 Introduction

Spectral analysis is of fundamental importance in the analysis and processing of wide-sense stationary random processes; the power spectral density (PSD) has an immediate physical interpretation as a spectral distribution of power and plays a central role in linear filtering, prediction and estimation. However, highly nonstationary signals arise in many applications, such as acoustic, speech, and biological signals. Thus, there is a need for extending the techniques of classical spectral analysis to nonstationary processes.

Nonstationary Spectrum Estimation. A number of extensions for the definition of a nonstationary spectrum have been proposed, none of which is completely satisfactory. (Loynes has in fact argued that a satisfactory extension may not exist at all [1].) More notable ones are the evolutionary spectrum (ES) proposed by Priestley [2] for the class of oscillatory processes, and the Wigner-Ville spectrum (WVS) proposed by Martin [3] for the class of harmonizable processes. Both definitions are based on second-order statistics and reduce to the PSD in the stationary case. One attractive feature of the ES is that it is always nonnegative, which is consistent with the usual interpretation of energy spectrum, whereas the WVS is not necessarily so. However, there are quite a few important advantages of the WVS over the ES:

- The ES is not unique, whereas the WVS is.
- There is no assumption of "slowly time-varying characteristics" in the theory of the WVS, whereas this assumption, except for the definition, underlies almost the entire development of the theory of the ES.
- There is no simple way, in general, of computing the ES from the correlation function, whereas the WVS is explicitly defined in terms of the correlation function.

The WVS also has some other desirable properties, as discussed in [3, 4], which the ES does not necessarily have. For these reasons, we adopt the WVS as our definition for the nonstationary spectrum and address the problem of estimating it from a realization of the process. But before we present our approach, we briefly discuss the limitations of the existing estimation techniques, which require a new approach to overcome.

The problem of estimating the PSD from a single realization of a stationary process involves the well known bias-variance trade-off; smoothing reduces the variance of the estimate but introduces bias. This trade-off should be optimized in some sense. Ergodicity plays a central role for stationary processes in that ensemble-averages can be replaced by time-averages. In the case of nonstationary processes, the situation is complicated by the fact that the concept of ergodicity, in its true form, does not exist anymore, since time-averaging smooths out the nonstationary structure of the process. To overcome this problem, the quasi-stationarity assumption is usually invoked; that is, it is assumed that the characteristics of the process are changing slowly with time so that locally, at any particular time t_o, the process can be approximated by a stationary process over some finite interval $T_s(t_o)$ around t_o. The problem then reduces to PSD estimation over $T_s(t_o)$. Implicit is the crucial assumption that the process also decorrelates fast enough so that a reasonable estimate of the spectrum over $T_s(t_o)$ can be obtained.

As we mentioned before, this quasi-stationarity assumption is effectively embedded in the definition of the ES because Priestley, with the problem of estimation in mind, develops his theory for the class of semi-stationary processes whose characteristics are changing slowly with time [2]. Although the *definition* of WVS does not involve any such assumptions, all current techniques for *estimating* the WVS invoke the quasi-stationarity assumption [4, 5, 6]. Clearly, the quasi-stationarity assumption is not valid in general. For example, there can exist spectrally stationary processes whose characteristics do not change with frequency, or other classes of processes whose characteristics remain constant along certain directions or curves in the time-frequency (TF) plane. This is simply due to the fact that the class of nonstationary processes is richer than the class of stationary processes. Clearly, TF analysis, being primarily concerned with nonstationary processes, needs to incorporate all these various kinds of processes; assumptions like quasi-stationarity are far too restrictive to provide a satisfactory theory.

So, the question we address is: given the statistics of the process, and without making the quasi-stationarity assumption, what is the "best" estimator of the nonstationary spectrum which optimizes the bias-variance trade-off in some sense? For reasons discussed earlier, we begin with the WVS as our definition of the nonstationary spectrum. As the class of estimators we choose Cohen's class of time-frequency representations (TFRs), which is described in the next section.[1] Since Cohen's class is completely characterized by a kernel, the question is equivalent to the "best" choice of kernel. In this context, (7) yields a useful interpretation of the estimator; the WD of the realization x, the "empirical" WD, is smoothed by the kernel Φ, which may vary with time and frequency, to produce an estimate of the WVS.

As we have already noted, the concept of ergodicity does not hold for nonstationary processes. We are confronted with the bias-variance trade-off, but unlike the case of PSD estimation where independent smoothing in the time and frequency directions suffices, the direction of smoothing in the TF plane becomes important in the nonstationary case. Intuitively, at each point (t, f) in the TF plane, the kernel Φ should average over some region, $G_{(t,f)}$, over which the characteristics of the process are essentially constant. The shape, size and orientation of $G_{(t,f)}$ are the crucial parameters and depend on the structure of the WVS. If we allow Φ to vary with time and frequency, then at each (t, f), its support should correspond to $G_{(t,f)}$. On the other hand, if we aim to design Φ so that it does not vary with time and frequency, then its support should correspond to some sort of an average of the $\{G_{(t,f)}\}$ over (t, f). We will refer to the former as the "local" and the latter as the "global" kernel. Clearly, if the shape and area of the $\{G_{(t,f)}\}$ vary substantially over the TF support of the WVS, a local kernel would be more appropriate.

To capture these intuitive notions about the form of the kernel, we consider minimum mean-squared error (mmse) estimation of the WVS. Recall that our primary objective is to optimize the bias-variance trade-off, and since *mean-squared error* $(mse) = variance + bias^2$, mmse estimation is clearly a reasonable way of doing so. The global kernel is obtained by minimizing the integrated mse, whereas the local kernel is obtained by minimizing the mse at each value of (t, f). The support of the optimal kernel obtained in this manner is then an estimate of the regions $\{G_{(t,f)}\}$; we use this information to illustrate the importance of the *direction* of smoothing in the TF plane, and that smoothing in the time and/or frequency direction(s) is not always appropriate.

Optimal TFR Estimation. So far we have discussed the problem of estimating the WVS from a realization, but our proposed framework naturally leads to another estimation problem as well. Instead of estimating the spectrum, what if we are interested in estimating a particular TFR of a realization from a noise-corrupted version of that realization? This is a plausible scenario, for example, for extracting the TFR of a random signal, characterized by a finite set of random parameters, from a noisy observation of a realization. More generally, we consider mmse estimation of an arbitrary TFR, characterized by a kernel ϕ_r, of a realization of a process from the corresponding realization of a correlated process. The class of estimators is again Cohen's class characterized by the kernel ϕ, which may (local) or may not (global) vary with time and frequency. We assume that the kernel ϕ_r, which we shall refer to as the reference kernel and

[1] We note that Martin and Flandrin [4], and Amin [5, 6] also consider the same class of estimators. Other choices of the class of estimators are also possible; refer to the remarks in the concluding section.

which may also vary with time and frequency, yields useful TFRs for all realizations of the process. We shall refer to this problem as the "TFR estimation" problem.

As we have already mentioned, all existing techniques for nonstationary spectral estimation assume quasi-stationarity [4, 7, 5, 6, 8]. In particular, Kayhan, El-Jaroudi and Chapparo [7], and Riedel [8] have proposed techniques for estimating the ES from a realization. On the other hand, Martin and Flandrin [4], and Amin [5, 6] have addressed the problem of estimating the WVS using TFRs from Cohen's class; the focus has primarily been on smoothed-pseudo-Wigner distributions (SPWDs) because they allow independent time and frequency smoothing owing to separable kernels. In [4], Martin and Flandrin propose SPWDs as a class of estimators, and in [5] Amin has proposed approximating arbitrary time-frequency kernels by SPWD kernels for nonstationary spectral estimation. Thus, we primarily restrict the comparison of our optimal kernels to SPWD kernels. We note in passing that in [6], the Born-Jordan kernel [9] was shown to be optimal in the sense of minimizing average variance for white noise processes. However, the effect of averaging on *bias* was not taken into account in [6], and it seems unlikely that a kernel optimal for white noise will perform satisfactorily for nonstationary processes.

We now present an outline of the paper. In the next section we define the WVS and describe our class of estimators. In section 3 we discuss the global WVS and TFR estimation problems, and in section 4 we solve the corresponding local problems. Some particular cases of global WVS estimation are presented in section 5 to show that the optimal kernel solutions are intuitively satisfying. Section 6 illustrates the superiority of the proposed scheme to existing methods through examples. Section 7 highlights the significance of the results and the limitations.

2 The Class of Estimators

The WVS of a random process, X, is defined as

$$WV_X(t,f) = E\{\int X(t+\tau/2)X^*(t-\tau/2)e^{-i2\pi f\tau}d\tau\} = \int R_X(t+\tau/2, t-\tau/2)e^{-i2\pi f\tau}d\tau \ , \tag{1}$$

where E denotes the expectation operator and $R_X(u,v) = E\{X(u)X^*(v)\}$ is the correlation function of X. All unlabeled integrals go from $-\infty$ to ∞. The integral inside the expectation operator is a stochastic integral, formally the Wigner distribution (WD) of X, and will be interpreted as a mean-square (m.s.) integral. The interchange of expectation and integration in the second equation is justified if the above-mentioned stochastic integral exists in the m.s. sense; a necessary and sufficient condition for its existence being

$$\int\int E\{q_X(t,\tau_1)q_X^*(t,\tau_2)\}e^{-i2\pi f(\tau_1-\tau_2)}d\tau_1 d\tau_2 \ < \infty \text{ for all } (t,f) \ , \tag{2}$$

where $q_X(t,\tau) = X(t+\tau/2)X^*(t-\tau/2)$. In addition, we restrict ourselves to harmonizable processes [3], the processes for which the two-dimensional (2D) Fourier transform of $R_X(t,s)$ exists; that is,

$$R_X(t,s) = \int\int e^{i2\pi t\mu}e^{-i2\pi s\nu}F_X(\mu,\nu)d\mu d\nu \ . \tag{3}$$

The WVS can then be equivalently expressed in terms of $F_X(\mu,\nu)$, the spectral correlation function [3], as

$$WV_X(t,f) = \int F_X(f+\theta/2, f-\theta/2)e^{i2\pi\theta t}d\theta \ . \tag{4}$$

Note that the WVS contains all the information in the correlation functions because both can be recovered from WV_X via a Fourier transform. Another useful, equivalent, representation of the correlation functions is in terms of the expected ambiguity function of the process (see (6)).

We are concerned with the estimation of the WVS from a realization of the process. Our class of estimators is Cohen's class [9] of bilinear time-frequency representations (TFRs). Although Cohen's class has been defined for deterministic signals, it will become clear from the following discussion that it can also be used for estimation of the WVS.

For a given deterministic signal, x, a particular TFR from Cohen's class can be written as [9]

$$P_x(t,f;\phi) = \int\int A_x(\theta,\tau)\phi(\theta,\tau)e^{-i2\pi f\tau}e^{i2\pi\theta t}d\theta d\tau \ , \tag{5}$$

where $A_x(\theta,\tau)$ is the ambiguity function (AF) of the signal x, defined as

$$A_x(\theta,\tau) = \int x(u+\tau/2)x^*(u-\tau/2)e^{-i2\pi\theta u}du \ , \tag{6}$$

and $\phi(\theta,\tau)$ is the 2D kernel which completely characterizes the particular TFR P_x; the kernel may explicitly depend on time and frequency. P_x can equivalently be expressed as

$$P_x(t,f;\Phi) = \int\int W_x(t',f')\Phi(t-t',f-f')dt'df' = (W_x ** \Phi)(t,f) \ , \tag{7}$$

where '$**$' denotes 2D convolution, W_x is the WD of the signal x, and Φ is the 2D Fourier transform of ϕ; $\Phi(u,\nu) = \mathcal{F}_{\theta\to -u}\mathcal{F}_{\tau\to\nu}\phi(\theta,\tau)$.

The interpretation of Cohen's class as a class of estimators for the WVS becomes clear from the following equivalent expression for P_x:

$$P_x(t,f;\Pi)=\int\left\{\int x(u+\tau/2)x^*(u-\tau/2)\Pi(t-u,\tau)du\right\}e^{-i2\pi f\tau}d\tau=\int \hat{R}_x(t+\tau/2,t-\tau/2)e^{-i2\pi f\tau}d\tau \ , \tag{8}$$

where $\Pi(u,\tau) = \mathcal{F}_{\theta\to -u}\phi(\theta,\tau)$ is the representation of the kernel in the (t,τ) domain. If x denotes a realization of a random process, then the inner integral in (8) represents an estimate, formed by the kernel Π, of the correlation function, $\hat{R}_X$. This estimate is then used to form an estimate of the WVS.

Finally, we note that if x denotes a realization of a random process, then all the integrals defined in (5)-(8) become stochastic integrals and will be interpreted as m.s. integrals. We assume that the kernel ϕ is chosen such that the existence of the WD as a m.s. integral implies the existence of P_X as a m.s. integral.

3 The Global Problem

In this section we formulate and solve the global WVS and TFR estimation problems, in which the kernel is not allowed to vary with time and frequency, and discuss the implications of the solutions. From now on, we will use uppercase letters to denote random processes and variables, and lowercase letters for realizations, deterministic signals, and constants.

3.1 Global WVS estimation

Recall from the introduction that the objective here is to optimally estimate the WVS of a process from an observed realization of a correlated process; the observed realization, for example, may be a noise-corrupted realization of the process whose WVS is desired. We now formulate the problem.

Let $X(t)$ and $Y(t)$, $t \in T \subset \mathbb{R}$, be two random processes defined on the same probability space. Y denotes the process whose WVS, WV_Y, is to be estimated from a realization x of X; X can either be Y or a process correlated in some way to Y, such as a noise-corrupted version of it. We assume that both X and

Y possess finite fourth-order moments and that $T \subset \mathbb{R}$ is a finite interval. This, in particular, implies that both processes have finite energy; that is,

$$\int_T E\{|X(t)|^2\}dt < \infty \ , \tag{9}$$

and similarly for Y. This is not a restrictive assumption because in practice we will always be dealing with finite observation intervals, and all realizations will be of finite energy almost surely. Let ϕ denote the kernel which characterizes our estimate $P_x(\phi)$ of WV_Y. As motivated in the introduction, we are interested in mmse estimation, and since our objective is to design a global kernel, the problem is formulated as

$$\phi_{opt} = \arg\inf_{\phi} E\left\{\iint_T |P_X(t,f;\phi) - WV_Y(t,f)|^2 dtdf\right\} \ . \tag{10}$$

That is, ϕ_{opt} minimizes the integrated mean-squared error between P_X and WV_Y. Using Parseval's theorem and assuming that the integral in (10) exists in the m.s. sense, we arrive at the following equivalent formulation:

$$\phi_{opt} = \arg\inf_{\phi} \iint E\{|A_X(\theta,\tau)\phi(\theta,\tau) - EA_Y(\theta,\tau)|^2\}d\theta d\tau \ . \tag{11}$$

The integrand in the above equation is a nonnegative quantity for all (θ,τ), so the infimum of the integral is equivalent to obtaining the infimum of the integrand for each value of (θ,τ). Thus we have

$$\phi_{opt}(\theta,\tau) = \arg\inf_{\phi(\theta,\tau)} E\{|A_X(\theta,\tau)\phi(\theta,\tau) - EA_Y(\theta,\tau)|^2\}, \quad (\theta,\tau) \in \mathbb{R}^2 \ . \tag{12}$$

The solution is quite clear now. For each value of (θ,τ), $A_X(\theta,\tau)$ is a second-order random variable, $EA_Y(\theta,\tau)$ is a constant, and (12) says that $\phi_{opt}(\theta,\tau)$ should be chosen so that $A_X(\theta,\tau)\phi_{opt}(\theta,\tau)$ is the linear mmse estimate of $EA_Y(\theta,\tau)$ given $A_X(\theta,\tau)$. By the orthogonality principle [10], ϕ_{opt} must satisfy

$$B_X(\theta,\tau)\phi_{opt}(\theta,\tau) = B_{YX}(\theta,\tau), \ (\theta,\tau) \in \mathbb{R}^2 \ , \tag{13}$$

where

$$B_X(\theta,\tau) = E\{|A_X(\theta,\tau)|^2\} \geq 0 \ , \tag{14}$$

and

$$B_{YX}(\theta,\tau) = EA_Y(\theta,\tau)EA_X^*(\theta,\tau) \ . \tag{15}$$

Also define B_Y as

$$B_Y(\theta,\tau) = |EA_Y(\theta,\tau)|^2 \geq 0 \ , \tag{16}$$

and the support of B_{YX} as

$$S_{YX} = \overline{\{(\theta,\tau) \in \mathbb{R}^2 : |B_{YX}(\theta,\tau)| > 0\}} \ , \tag{17}$$

where overbar denotes closure with respect to the Euclidean norm on $\mathbb{R}^2$. Similarly define S_X, S_Y as the supports of B_X, B_Y, respectively. Note that $S_{YX} \subseteq S_X \cap S_Y$, which follows from the Cauchy-Schwarz (CS) inequality.

From (13) we note that ϕ_{opt} can be explicitly obtained by inverting (13) for each (θ,τ). If B_X is bounded away from zero over S_{YX}, that is $B_X(\theta,\tau) \geq \alpha > 0$ for (θ,τ) in S_{YX}, then ϕ_{opt} is bounded. But ϕ_{opt} may be unbounded if B_X is not bounded away from zero over S_{YX}, and in that case a bounded approximation to ϕ_{opt} can be obtained as discussed in Appendix A. However, it can be shown that if the observed process X includes some independent, additive white Gaussian noise, then B_X is bounded away from zero, and hence ϕ_{opt} is bounded [11]. Thus, in most cases of interest, the following proposition characterizes the globally optimal kernel.

Proposition 1. Let B_X, B_Y, B_{YX} and S_{YX} be as defined in (14), (16), (15) and (17), respectively. Then, the globally optimal kernel, ϕ_{opt}, solving (10), is given by

$$\phi_{opt}(\theta,\tau) = \begin{cases} k_{YX}(\theta,\tau) & \text{if } (\theta,\tau) \in S_{YX} \\ 0 & \text{otherwise} \end{cases} , \tag{18}$$

where

$$k_{YX}(\theta,\tau) = \frac{B_{YX}(\theta,\tau)}{B_X(\theta,\tau)} , \tag{19}$$

and the corresponding minimum mean-squared error is given by

$$mmse = \int_{S_Y \backslash S_{YX}} B_Y(\theta,\tau) d\theta d\tau + \int_{S_{YX}} \left[B_Y(\theta,\tau) - \frac{|B_{YX}(\theta,\tau)|^2}{B_X(\theta,\tau)} \right] d\theta d\tau \ . \tag{20}$$

Proof: (18) and (19) follow immediately from (13) by noting that that since $S_{YX} \subseteq S_X$, $(\theta,\tau) \in \mathbb{R}^2 \backslash S_{YX} \Rightarrow \phi_{opt}(\theta,\tau) = 0$. The expression for minimum mse is obtained by simply substituting ϕ_{opt} for ϕ in

$$mse(\phi) = \int E\{|A_X(\theta,\tau)\phi(\theta,\tau) - EA_Y(\theta,\tau)|^2\} d\theta d\tau \tag{21}$$

and taking into account the supports of the various terms.

Corollary. If $X = Y$, that is, the observed process is noise free, then

$$k_{YX}(\theta,\tau) = \frac{|EA_X(\theta,\tau)|^2}{E\{|A_X(\theta,\tau)|^2\}} \ . \tag{22}$$

The expression for ϕ_{opt} in (18) is quite informative about its support. Let $\breve{S}_X$, $\breve{S}_Y$ denote the supports of EA_X and EA_Y, respectively, and S_ϕ the support of ϕ_{opt}. From (17) and (15), it is clear that $S_{YX} = \breve{S}_X \cap \breve{S}_Y$. Thus, from (18) and (19), it follows that $S_\phi = \breve{S}_X \cap \breve{S}_Y$ for the case when ϕ_{opt} is bounded (when ϕ_{opt} is not bounded, $S_{\hat{\phi}} \approx \breve{S}_X \cap \breve{S}_Y$). For most cases of observation noise in X (independent, additive or multiplicative noise, for example), $\breve{S}_Y \subseteq \breve{S}_X$. Thus, essentially, $S_\phi = \breve{S}_Y$. This says that the support of the optimal kernel, $\phi_{opt}(\theta,\tau)$, is matched to that of the expected AF of the process Y whose WVS is to be estimated. This is the first indication that the optimal kernel possesses the desired characteristics as discussed in the introduction. We will return to this discussion in section 5 when we consider some specific cases.

Another interpretation of the global solution can be obtained as follows. Let

$$\phi_1(\theta,\tau) = \frac{EA_X^*(\theta,\tau)}{E\{|A_X(\theta,\tau)|^2\}}. \tag{23}$$

Then, $\phi_{opt} = \phi_1 EA_Y$. This implies that

$$P_x(t,f;\phi_{opt}) = WV_Y(t,f) ** P_x(t,f;\phi_1) \ . \tag{24}$$

That is, the globally optimal estimate of WV_Y, based on the observed realization x, is WV_Y itself, convolved with a TFR of x generated by ϕ_1.

Arbitrary Nonstationary Spectral Estimation. So far we have discussed optimal estimation of the WVS, which is formally the expected value of the WD of the process (see (1)). Now suppose that instead of estimating the WVS, we are interested in estimating an arbitrary nonstationary spectrum, defined as the expected value of an arbitrary TFR characterized by a kernel $\phi_o(\theta,\tau)$; that is, we want to estimate $E\{P_Y(t,f;\phi_o)\}$. In this case, the optimization problem becomes

$$\phi_{opt} = \arg\inf_{\phi} E\left\{\iint_T |P_X(t,f;\phi) - EP_Y(t,f;\phi_o)|^2 dtdf\right\} \ , \tag{25}$$

and ϕ_{opt} is characterized by

$$B_X(\theta,\tau)\phi_{opt}(\theta,\tau) = B_{YX}(\theta,\tau)\phi_o(\theta,\tau), \quad (\theta,\tau) \in \mathbb{R}^2 \ , \tag{26}$$

that is, $\phi_{opt} = \phi_{WVS}\phi_o = k_{YX}\phi_o$, where $\phi_{WVS} = k_{YX}$ is the optimal kernel for WVS estimation given by (18). For example, if we choose the Rihaczek distribution ($\phi_o(\theta,\tau) = e^{i\pi\theta\tau}$) [9] for defining the nonstationary spectrum, the optimal kernel is $\phi_{opt}(\theta,\tau) = e^{i\pi\theta\tau}B_{YX}(\theta,\tau)/B_X(\theta,\tau)$.

From the above discussion we see that given the kernel, ϕ_o, defining the nonstationary spectrum, the optimal kernel is completely characterized by $k_{YX} = B_{YX}/B_X$. Clearly, based on the observed realization, x, and without any other information, the simplest and most intuitive estimate of the spectrum, $E\{P_Y(t,f;\phi_o)\}$, is the "empirical" TFR, $P_x(t,f;\phi_o)$; for example, the simplest estimate of the WVS is the WD of the realization. On the other hand, the optimal estimate is $P_x(t,f;k_{YX}\phi_o)$. Thus, k_{YX} characterizes the averaging (filtering) done on the empirical TFR by the optimal kernel to yield the optimal estimate. A natural question is whether the empirical TFR itself is an adequate or optimal estimate in certain situations ("no-averaging" scenarios)? In Appendix B (Proposition B1) we show that in almost all nontrivial cases the empirical TFR is not an optimal estimate; that is, $k_{YX}(\theta,\tau)$ is not identically a constant, and thus effects some averaging for optimal estimation.

An important scenario of nonstationary spectrum estimation is when X is a noise-corrupted version of Y. In this case, Proposition B1 makes a strong statement: it says that if the noise is independent of the process Y, then the empirical TFR is *never* an optimal estimate unless there is no noise; that is, $X = Y$. In addition to that, the process itself must be more or less degenerate: a deterministic signal scaled by a random variable. A special case of this scenario is that of estimating a particular TFR, say the WD, of a deterministic signal from an observation which has additive Gaussian noise in it. In this case, an argument similar to the proof of Proposition B1 shows that smoothing of the empirical WD is needed unless there is no noise.

Also, it is shown in Proposition B2 that perfect estimation (mmse=0, an example of no averaging) is possible if and only if Y is essentially a deterministic signal scaled by a random variable, and X is the same deterministic signal scaled by a possibly different random variable.

The broad implication of the discussion in Appendix B is that, in almost all cases of interest, some averaging is necessary in order to optimally estimate the nonstationary spectrum of a process from a realization; the empirical TFR would be too noisy an estimate.

We note in passing that any side constraints on the kernel that constrain $\phi(\theta,\tau)$ to have certain fixed values for some regions in the (θ,τ) domain, like the time, frequency, or energy marginal constraints, can easily be incorporated because of the simple "least-squares" nature of the optimal kernel solution; simply set $\phi_{opt}(\theta,\tau)$ to the constrained values over the constrained regions and solve for the optimal values, as in Proposition 1, over the remaining regions.

We finish our discussion of global WVS estimation by generalizing the results to the case of multiple independent realizations. Suppose that we have M independent realizations, $x_1, x_2, \cdots x_M$, of X available to us. The most intuitive estimate of the nonstationary spectrum, $EP_Y(\phi_o)$, is the empirical average of the empirical TFRs, $P_{x_i}(\phi_o)$;

$$\frac{1}{M}\sum_{k=1}^{M} P_{x_i}(\phi_o) \ , \tag{27}$$

whereas we are interested in finding the kernel ϕ_{opt}^M which, when applied to the empirical average (27), results in a mmse estimate. Thus, equivalently, the kernel acts on the empirical average of the observed ambiguity functions, $\tilde{A}_x = \frac{1}{M}\sum_{k=1}^{M} A_{x_i}$, and it can be easily verified that the optimal kernel is given by

$$\phi_{opt}^M(\theta,\tau) = \frac{EA_Y(\theta,\tau)E\tilde{A}_X^*(\theta,\tau)}{E\{|\tilde{A}_X^*(\theta,\tau)|^2\}}\phi_o(\theta,\tau) = \frac{MEA_Y(\theta,\tau)EA_X^*(\theta,\tau)}{E\{|A_X(\theta,\tau)|^2\}+(M-1)|EA_X(\theta,\tau)|^2}\phi_o(\theta,\tau) \ . \tag{28}$$

3.2 Global TFR estimation

As mentioned in the introduction, in the TFR estimation problem we are interested in estimating a particular TFR of a realization of a process from the corresponding realization of a correlated process. For example, we may wish to estimate the TFR of a signal in the presence of additive noise and nonstationary interference. More specifically, let $X(t)$ and $Y(t)$, $t \in T \subset \mathbb{R}$, be two random processes defined on the same probability space. We again assume that both X and Y possess finite fourth-order moments and that T is a finite interval. Let ϕ_r denote the reference kernel, which we assume to have been chosen to produce useful TFRs for all realizations of Y. X is the observed process, which is correlated with Y. Recall that a typical scenario for this problem is where Y represents a random signal characterized by a finite set of random parameters, X represents a noisy version of Y, and the objective is to undo the effects of noise to yield a good estimate of $P_y(\phi_r)$ from the noisy observation x.

We are again interested in mmse estimation of $P_y(\phi_r)$ by $P_x(\phi)$, and since we want to design a global kernel, the problem is formulated as

$$\phi_{opt} = \arg\inf_{\phi} E\left\{\iint_T |P_X(t,f;\phi)-P_Y(t,f;\phi_r)|^2 dtdf\right\} \ . \tag{29}$$

The above problem is very similar to the WVS estimation problem in (10), the only difference being that in (29) what we are trying to estimate, $P_y(\phi_r)$, is a random function as opposed to the deterministic function WV_Y in (10). Proceeding similarly as in the previous section, we arrive at the following characterization of ϕ_{opt}:

$$B_X(\theta,\tau)\phi_{opt}(\theta,\tau) = \hat{B}_{YX}(\theta,\tau)\phi_r(\theta,\tau), \quad (\theta,\tau) \in \mathbb{R}^2 \ , \tag{30}$$

where B_X is defined as before in (14) and

$$\hat{B}_{YX}(\theta,\tau) = E\{A_Y(\theta,\tau)A_X^*(\theta,\tau)\} \ . \tag{31}$$

Also define $\hat{B}_Y$ as

$$\hat{B}_Y(\theta,\tau) = E\{|A_Y(\theta,\tau)|^2\} \ . \tag{32}$$

We thus have the following proposition, which characterizes the optimal kernel.

Proposition 2. Let B_X, $\hat{B}_{YX}$ and $\hat{B}_Y$ be defined as in (14), (31) and (32), respectively. Then, the globally optimal kernel, ϕ_{opt}, solving (29), is given by

$$\phi_{opt}(\theta,\tau) = \begin{cases} k_{YX}(\theta,\tau)\phi_r(\theta,\tau) & \text{if } (\theta,\tau) \in \hat{S}_{YX} \\ 0 & \text{otherwise} \end{cases} \ , \tag{33}$$

where $\hat{S}_{YX}$ is the support of $\hat{B}_{YX}$ and

$$k_{YX}(\theta,\tau) = \frac{\hat{B}_{YX}(\theta,\tau)}{B_X(\theta,\tau)} \ . \tag{34}$$

The mmse is given by

$$mmse = \int_{\hat{S}_Y \backslash \hat{S}_{YX}} |\phi_r(\theta,\tau)|^2 \hat{B}_Y(\theta,\tau) d\theta d\tau + \int_{\hat{S}_{YX}} |\phi_r(\theta,\tau)|^2 \left[\hat{B}_Y(\theta,\tau) - \frac{|\hat{B}_{YX}(\theta,\tau)|^2}{B_X(\theta,\tau)} \right] d\theta d\tau \ , \tag{35}$$

where $\hat{S}_Y$ is the support of $\hat{B}_Y$.

Proof: Similar to the proof of Proposition 1.

We note that ϕ_{opt} in (33) may not be bounded, in which case a bounded approximation can be obtained in a similar way as discussed in Appendix A.

It is worth noting the similarity of ϕ_{opt} to the optimal Wiener filter for linear mmse estimation of a reference stationary process from an observed process: $H(f) = S_{YX}(f)/S_X(f)$, where $H(f)$ is the transfer function of the optimal Wiener filter, $S_{YX}(f)$ is the cross-power spectral density of the processes, and $S_X(f)$ is the (auto-)power spectral density of the observed process. To see the similarity, first note that ϕ_{opt} is simply a weighted version of the reference kernel ϕ_r, and thus the weighting k_{YX} completely characterizes it, given ϕ_r. And as evident from (31) and (14), $\hat{B}_{YX}$ is simply the zero-lag cross-correlation between the ambiguity functions A_Y and A_X, and B_X is the zero-lag auto-correlation of A_X.

The expression for ϕ_{opt} in (33) also shows that the support of the optimal kernel is more or less concentrated in that region of the (θ,τ) plane where the cross-correlation $\hat{B}_{YX}$ is significant. This also makes intuitive sense because we are interested in those components of X which are strongly correlated with Y and wish to attenuate those which are not. Actually the support of ϕ_{opt}, S_ϕ, satisfies, $S_\phi \subset S_X \cap \hat{S}_Y$. These observations again suggest that the optimal direction of smoothing in the TF plane depends on the structure of the processes and need not be along the time and/or frequency directions. This property of the optimal kernel will become more apparent when we discuss some specific cases in the section 5 and examples in section 6.

In the TFR estimation problem, the simplest and most intuitive estimate of the reference TFR, $P_y(t,f;\phi_r)$, based on the observed realization, x, is the empirical TFR, $P_x(t,f;\phi_r)$. The optimal estimate is $P_x(t,f;k_{YX}\phi_r)$ and thus, just as we argued in the previous section, k_{YX} defined in (34) completely characterizes the optimal averaging done on the empirical TFR to yield the optimal estimate. Again, parallel to the Proposition B1 we show in Appendix B (Proposition B3) that the empirical TFR is almost never optimal; some averaging is always needed in general.

We mention that Altes [12] has discussed a similar problem of estimating the signal from a spectrogram of its noisy observation. Since a signal can be recovered (within a complex constant factor) from its AF, he essentially addresses the problem of mmse estimation of the AF of the signal from a spectrogram of a noisy observation of the signal.

4 The Local Problem

In this section we solve the local estimation problems, in which the kernel ϕ is allowed to vary with time and frequency in order to better track the nonstationary structure of the processes.

4.1 Local WVS estimation

Let X and Y be two random processes defined the same way as in section 3.1. The local WVS estimation problem is formulated as

$$\phi_{opt}^{(t,f)} = \arg\inf_{\phi} E\{|P_X(t,f;\phi) - WV_Y(t,f)|^2\} \ , \tag{36}$$

where $t \in T$ and the superscript (t,f) denotes the possible dependence of the kernel on time and frequency. This is again a problem of linear mmse estimation. First note that for each value of (t,f), $P_X(t,f;\phi \in L_2(\mathbb{R}^2))$ belongs to a Hilbert space $\mathcal{H}$ of second-order random variables, defined on the underlying probability space. The random ambiguity function, $A_X(\theta,\tau)$, generates a subspace M_X of $\mathcal{H}$, as defined by $P_X(t,f;\phi)$ (see (5)):

$$M_X = \{P_X(t,f;\phi) : \phi \in L_2(\mathbb{R}^2)\} \ . \tag{37}$$

In the local problem (36), the orthogonal projection of $WV_Y(t,f)$ onto $\overline{M}_X$ is desired.[2] Thus, by the orthogonality principle and using the expression for $P_X(t,f;\phi)$ in (5), $\phi_{opt}^{(t,f)}$ is characterized by

$$E\{[P_X(t,f;\phi_{opt}^{(t,f)}) - WV_Y(t,f)]A_X^*(\theta',\tau')\} = 0, \text{ for all } (\theta',\tau') \ . \tag{38}$$

Expressing P_X as in (5) and WV_Y in terms of EA_Y, we can write (38) as

$$\int_{\mathbb{R}^2} E\{A_X(\theta,\tau)A_X^*(\theta',\tau')\}e^{i2\pi(\theta t-\tau f)}\phi_{opt}^{(t,f)}(\theta,\tau)d\theta d\tau = WV_Y(t,f)E\{A_X^*(\theta',\tau')\}, \text{ for all } (\theta',\tau') \ . \tag{39}$$

The above linear equation characterizes the locally optimal kernel, $\phi_{opt}^{(t,f)}$, from which we can gain some insight about the optimal solution. The linear equation is of the form $\mathbf{A}_{(t,f)}\phi = WV_Y(t,f)b$ where $\mathbf{A} : L_2(\mathbb{R}^2) \to L_2(\mathbb{R}^2)$ is a linear operator and $b = EA_X^* \in L_2(\mathbb{R}^2)$. Thus, the solution can be formally written as

$$\phi_{opt}^{(t,f)} = WV_Y(t,f)\mathbf{A}_{(t,f)}^{\dagger}b \ , \tag{40}$$

where the superscript '†' denotes the pseudo-inverse. This implies that

$$P_x(t,f;\phi_{opt}^{(t,f)}) = WV_Y(t,f)P_x(t,f;\mathbf{A}_{(t,f)}^{\dagger}b) \ . \tag{41}$$

That is, for each value of (t,f), the optimal estimate of $WV_Y(t,f)$ based on the observed realization x is $WV_Y(t,f)$ itself, scaled by a constant which is the value of the TFR of x generated by $\phi = \mathbf{A}_{(t,f)}^{\dagger}b$. Thus, the TF support of the optimal estimate $P_x(t,f;\phi_{opt}^{(t,f)})$ is contained in the support of $WV_Y(t,f)$. Note that if we constrain ϕ to be independent of (t,f), then (39) yields the global solution (13) by multiplying both sides with $e^{-i2\pi(\theta' t-\tau' f)}$ and integrating over (t,f).

4.2 Local TFR estimation

Let the random processes X and Y be defined in the same way as in section 3.2. Similarly to local WVS estimation, we formulate the related problem of local TFR estimation as

$$\phi_{opt}^{(t,f)} = \arg\inf_{\phi} E\{|P_X(t,f;\phi) - P_Y(t,f;\phi_r^{(t,f)})|^2\} \ , \quad t \in T \ , \tag{42}$$

Note that the reference kernel, $\phi_r \in L_2(\mathbb{R}^2)$, may also vary with time and frequency in this case.[3] Apart from that, (42) is very similar to (36), except that in this case the quantity to be estimated, $P_y(t,f;\phi_r)$,

[2] Since we are projecting onto $\overline{M}_X$, ϕ_{opt} may not be in $L_2(\mathbb{R}^2)$, in general.

[3] Although $\phi_r \in L_2(\mathbb{R}^2)$ strictly precludes the consideration of the WD as the reference TFR, in almost all cases the essential support of the AF realizations will be finite, in which case the WD can effectively be characterized by a unity kernel on that support (and zero everywhere else), making it an admissible reference TFR.

is a random variable as opposed to the constant $WV_Y(t,f)$ in (36). Proceeding similarly to the local WVS estimation solution, we arrive at the following characterization of the optimal kernel

$$\int_{\mathbb{R}^2} E\{A_X(\theta,\tau)A_X^*(\theta',\tau')\}e^{i2\pi(\theta t-\tau f)}\phi_{opt}^{(t,f)}(\theta,\tau)d\theta d\tau = \int_{\mathbb{R}^2} E\{A_Y(\theta,\tau)A_X^*(\theta',\tau')\}e^{i2\pi(\theta t-\tau f)}\phi_r^{(t,f)}(\theta,\tau)d\theta d\tau \quad \text{for all } (\theta',\tau') \ . \tag{43}$$

The above equation is again a linear equation of the form $\mathbf{A}_{(t,f)}\phi = \mathbf{B}_{(t,f)}\phi_r$, where $\mathbf{A}_{(t,f)}, \mathbf{B}_{(t,f)} : L_2(\mathbb{R}^2) \to L_2(\mathbb{R}^2)$ are linear operators. In this case also, if we constrain ϕ and ϕ_r to be independent of (t,f), then (43) yields the global solution (30).

We note a few things about the two local solutions. Comparing (39) with (13), and (43) with (30), we note that the local solutions require much more computation and much more statistical information about the processes X and Y. Even if we have the required statistics, (39) and (43) involve tensors, which makes the local solutions computationally intensive. However, since the cost functionals in (42) and (36) are quadratic in the kernel ϕ, if we know the required statistics we can use any standard quadratic minimization algorithm to find $\phi_{opt}^{(t,f)}$.

5 Some Special Cases

In this section, we consider globally optimal WVS estimation for three special classes of processes, the motivation being to check whether or not the optimal kernel solutions are intuitively satisfying. The three classes of processes that we consider are: temporally stationary processes, spectrally stationary processes, and processes whose Karhunen-Loève (KL) eigenfunctions are linear chirps. In all cases, we assume zero-mean, complex Gaussian processes, which results in the following moment decomposition [13]

$$\begin{aligned} E\{X(t_1)X^*(t_2)X^*(t_3)X(t_4)\} &= R_X(t_1,t_2)R_X(t_4,t_3) + R_X(t_1,t_3)R_X(t_4,t_2) \\ &\quad + E\{X(t_1)X(t_4)\}E\{X^*(t_3)X^*(t_4)\} \ . \end{aligned} \tag{44}$$

To simplify computation, we also assume that $\{X(t),\ t \in T\}$ and $\{e^{i\theta}X(t),\ t \in T\}$ are identically distributed for all $\theta \in \mathbb{R}$ (circular Gaussian), which results in the third term in (44) being identically zero (Grettenberg's Theorem) [13].

5.1 Temporally stationary processes

Let X be a temporally stationary process. Then, the correlation function becomes $R_X(t,s) = R_X(t-s)$, and the WVS reduces to the PSD, S_X:

$$WV_X(t,f) = \int R_X(\tau)e^{-i2\pi f\tau}d\tau = S_X(f) \ . \tag{45}$$

In the case of a finite observation interval $[0,T]$, for T sufficiently large and the decorrelation time of R_X much smaller than T, EA_X can be approximated as

$$EA_X(\theta,\tau) \approx \begin{cases} TR_X(\tau) & \text{if } \theta = 0 \\ 0 & \text{otherwise} \end{cases} \ . \tag{46}$$

Recalling that the support of ϕ_{opt} is the same as that of EA_X, we note from (46) that the optimal kernel is effectively one-dimensional. Its variation along τ is characterized by $w_{opt}(\tau) = \phi_{opt}(0,\tau)$, which is given by

$$w_{opt}(\tau) = \frac{|EA_X(0,\tau)|^2}{|EA_X(0,\tau)|^2 + \operatorname{var}\{A_X(0,\tau)\}} \approx \frac{|R_X(\tau)|^2}{|R_X(\tau)|^2 + \frac{1}{T}\int S_X^2(\mu)d\mu} \ , \tag{47}$$

where $\tau \in [-2T, 2T]$. The expression for w_{opt} in (47) is exactly the expression for the optimal window for mmse estimation of the PSD of X [14]. Note that, in this case, the optimal kernel does uniform time-averaging over the entire interval $[0, T]$; that is, $\Pi(t, \tau) \approx \frac{1}{T} w_{opt}(\tau), t \in [0, T], \tau \in [-2T, 2T]$ in (8), neglecting end effects. This makes intuitive sense because the process is temporally stationary. Smoothing in the frequency direction is governed by the Fourier transform of w_{opt} via $\Phi(t, f) = \mathcal{F}_{\tau \to f} \Pi(t, \tau)$ (see (7)).

5.2 Spectrally stationary processes

The TF dual of the class of temporally stationary processes is the class of spectrally stationary processes whose 2D spectral correlation function, F_X, is only a function of the difference of its arguments; that is, $F_X(\mu, \nu) = F_X(\mu - \nu)$. In this case, the WVS reduces to the "power temporal density (PTD)", Q_X:

$$WV_X(t, f) = \int F_X(\theta) e^{i2\pi\theta t} d\theta = Q_X(t) \ . \tag{48}$$

In the case of finite observation bandwidth, $f \in [0, B]$, if B is sufficiently large and the decorrelation bandwidth of F_X is much smaller than B, again we find that the optimal kernel is essentially one-dimensional and is characterized by the function $H_{opt}(\theta) = \phi_{opt}(\theta, 0)$ given by

$$H_{opt}(\theta) \approx \frac{|F_X(\theta)|^2}{|F_X(\theta)|^2 + \frac{1}{B} \int Q_X^2(u) du} \ , \tag{49}$$

where $\theta \in [-2B, 2B]$. In this case we note that the optimal kernel does frequency-averaging over the entire bandwidth $[0, B]$, which is consistent with the fact that the process is spectrally stationary.

5.3 Chirp processes

We now consider a class of processes for which the optimal solution clearly demonstrates the inadequacy of existing methods for nonstationary spectral estimation. We consider processes whose KL expansion [15] is characterized by eigenfunctions which are linear chirps, all having the same chirp rate. Such a process would arise by modulating a stationary process by $e^{i\alpha t^2}$. Let $X(t), \ t \in [0, T]$, be a process whose correlation function R_X admits the eigenexpansion

$$R_X(t, s) = \sum_{k=1}^{K} \lambda_k \varphi_k(t) \varphi_k^*(s) \ , \quad (t, s) \in [0, T]^2 \ , \tag{50}$$

where

$$\varphi_k(t) = \frac{1}{\sqrt{T}} e^{i(2\pi f_k t + \alpha t^2)}, \ k = 1 \cdots K, \ \alpha \in \mathbb{R}, \ t \in [0, T] \ , \tag{51}$$

$f_k = \frac{m}{T}$ for some integer m and the f_k's are distinct. Using (50) and (1), the WVS is given by

$$WV_X(t, f) = \sum_{k=1}^{K} \lambda_k W_{\varphi_k}(t, f) \ , \tag{52}$$

where, for T large and t not close to 0 or T, $W_{\varphi_k}(t, f)$ is concentrated along the ridges $\delta(f - f_k - \frac{\alpha t}{\pi})$ in the TF plane. Similarly, the expected AF is given by

$$EA_X(\theta, \tau) = \sum_{k=1}^{K} \lambda_k A_{\varphi_k}(\theta, \tau) \ , \tag{53}$$

where

$$A_{\varphi_k}(\theta,\tau) = e^{i2\pi f_k\tau}e^{-i\pi\theta T}\text{sinc}(\pi T(\theta - \frac{\alpha\tau}{\pi}))\ , \tag{54}$$

which, for T large enough, is highly concentrated along $\delta(\theta - \frac{\alpha\tau}{\pi})$. Again, since the support of ϕ_{opt} is the same as that of EA_X, we note from (53) and (54) that, in this case, for T long enough, the support of ϕ_{opt} is essentially concentrated along the line characterized by $\delta(\theta - \frac{\alpha\tau}{\pi})$. Thus, the optimal kernel does TF smoothing along the chirp direction, which makes intuitive sense because the characteristics of the process remain more or less constant along that direction.

This case clearly demonstrates the need for smoothing in *arbitrary* directions in the TF plane depending on the TF structure of the process. Smoothing kernels proposed in the past [4, 5, 6] do not possess this property.

6 Examples

In this section we present some examples to illustrate the superiority of the proposed technique to existing methods. For WVS estimation, the performance of the globally optimal estimator is compared with that of a smoothed-pseudo-Wigner (SPW) estimator proposed by Martin and Flandrin [4], which uses a length $2M-1$ rectangular window for time-smoothing and the Fourier transform of a length $2N-1$ rectangular window for frequency-smoothing. Normalization is chosen such that $\phi_{SPW}(0,0) = 1$. For TFR estimation, in addition to a SPW estimator, the optimal estimator is compared to a matched-filter spectrogram in which the kernel is matched to a characteristic component of the desired signal. In each case, 128 time and frequency samples are taken for discretization.

WVS estimation of a chirp process. Let X be a complex Gaussian process which has a KL-like expansion in terms of Gaussian chirps,

$$X(t) = \sum_{k=1}^{K} Z_k\varphi_k(t) \ \text{ in m.s.,} \quad t \in [0,T]\ , \tag{55}$$

where the Z_k's are uncorrelated, zero-mean, complex Gaussian random variables with $E\{|Z_k|^2\} = \lambda_k$, and $\varphi_k(t) = e^{-(\alpha - i\beta)(t-t_k)^2 + i2\pi f_k t}$. $K=3$ in this example. The WVS, given by (52), is estimated from a noisy realization, the noise being additive, zero-mean, white, complex Gaussian such that $SNR_{max} = 10\log\{\max\{E|X(t)|^2\}/\sigma^2\} = 3dB$, where σ^2 is the variance of noise. We chose the parameters $M=6$, $N=11$ for the SPW estimator, which were approximately optimized by trial and error. The results are shown in figure 1. Clearly, the SPW estimate, whose kernel smooths along the time and frequency directions, is quite different from the true spectrum, whereas the optimal kernel, which is matched to the characteristics of the process, yields a much more accurate estimate. Also, the mean-squared error of the SPW estimate is about 4 times larger than that of the optimal estimate.

Interference suppression: deterministic signal with narrowband noise. This example demonstrates the ability of the optimal kernel to suppress interference. The desired signal Y is a deterministic Gaussian chirp $s(t) = e^{-(\alpha - i\beta)t^2}$, $t \in [0,T]$, and the observation is corrupted by narrowband noise, $N(t) = Ae^{iB}e^{i\Omega t}$, where A, B and Ω are uniformly distributed over $[a1, a2]$, $[-\pi,\pi]$, and $[\omega_1, \omega_2]$, respectively. The parameters a_1, a_2, ω_1 and ω_2 are chosen to make $SNR_{max} = 0dB$, and the normalized (after discretization) bandwidth of noise $BW = \pi/20$ radians. The objective is to extract the WD of the chirp from a noise-corrupted observation. In this case, the optimal TFR is compared to a SPW estimator ($M=2$, $N=16$) and a matched spectrogram in which the window is matched to the Gaussian chirp signal; that is, $\phi_{spect} = |A_s|/A_s(0,0)$. Figure 2 shows the results. The SPW kernel, smoothing along time and frequency directions, is more matched to the narrowband interference than to the chirp signal, and thus yields a TFR which is dominated by the interfering noise. The spectrogram, despite being matched to the desired chirp signal, does not do a good job in suppressing the interference. The optimal kernel, on the other hand, is not only matched to the desired signal, but also has a "notch" in the region of the (θ, τ) plane where A_s and A_N intersect, and thus yields a very accurate estimate in which the interference has been almost completely suppressed.

TFR estimation of BPSK signal with white noise and narrowband co-channel interference. In this example $Y(t) = \sum_k Z_k e^{i\omega_o t} s_o(t - kT_o)$, $t \in [0,T]$, is a BPSK signal where the Z_k's are independent Bernoulli random variables taking on the values $\{-1, 1\}$ with equal probability, and s_o is a rectangular pulse of width $T' \leq T_o$. The observed signal is $X = Y + N + N_1$, where N_1 is complex, white Gaussian noise with variance $\sigma^2_{N_1}$, and N is narrowband co-channel noise, as in the previous example, with center frequency ω_o, the carrier frequency of the BPSK signal. Since Y is a multi-component signal, choosing the WD as the reference TFR is not appropriate because of the significant cross-terms inherent in the WD. Since all of the components in Y are TF translates of the baseband pulse s_o, we choose ϕ_r corresponding to the spectrogram matched to s_o; that is, $\phi_r = |A_{s_o}|/A_{s_o}(0,0)$. The parameters of noise are chosen so that the *SNR* between Y and the narrowband noise, N, is $0dB$ and that between Y and N_1 is $8dB$, making the overall *SINR* a little below $0dB$. Again, the performance of the optimal estimator is compared with that of a SPW estimator ($M = 6$, $N = 8$) and the matched-filter spectrogram ($\phi_{spect} = \phi_r$). Note that in this case the characteristics of all processes are aligned along time and/or frequency directions, and thus the SPW estimator has the potential of performing well. However, there is one caveat; since the characteristics of both signal and noise are somewhat similar, matching the kernel to the signal also matches it to noise. The optimal kernel, however, uses the information about noise to optimize the matching. The results are shown in figure 3. The optimal estimate is almost perfect, whereas both the spectrogram and the SPW estimates are severely affected by the presence of noise.

7 Conclusions

We have addressed two important TF estimation problems: the problem of estimating the WVS of a random process from a corrupted realization, and the related problem of estimating an arbitrary bilinear TFR of a realization from a correlated observation. For the former, all existing techniques are based on the assumption that the process statistics change slowly with time, which limits their use to a small class of nonstationary processes. We overcome this limitation by deriving a kernel within Cohen's class of TFRs which is optimal in a mean-square sense. For the latter problem, which has never been addressed before, we obtain a similar optimal kernel. Both optimal time-frequency invariant and time-frequency varying kernels are derived. Using the nature of the optimal kernel, it is proven that, in the presence of any additive independent noise, optimal performance requires a nontrivial kernel, and that optimal estimation may require smoothing filters very different from those based on a quasi-stationarity assumption. Examples confirm that for a large class of processes the optimal estimators often yield great improvements in performance over existing *ad hoc* methods.

The main limitation of the proposed estimation techniques is that certain second and fourth-order statistics are needed to compute the kernels. Those statistics can be computed if adequate models are available for the processes, or can be estimated if multiple realizations are available. In certain applications involving rotating machinery, for example, in which failures with nonstationary signatures need to be detected, the periodic nature of the signal statistics lends itself to collecting multiple realizations. However, in the case of a single realization, the next major research issue is to form an estimate of the statistics in order to design the optimal kernel to process the realization.

Another question that needs to be answered is the choice of the reference kernel, ϕ_r, for TFR estimation when the signal realizations are multi-component. As done in example 3, if the various components have a similar TF structure, then some sort of a matched kernel may be used.

In the TFR estimation framework, the ability of the optimal kernel to suppress interference is particularly remarkable. Thus, it can also serve as a framework for suppressing interfering signals via TF filtering, and could potentially be very useful for detection, estimation and classification of signals corrupted by nonstationary and co-channel interference.

Finally, we make a few comments about possible extensions of the work. First, we note that although we restricted ourselves to Cohen's class for estimating the WVS, it is by no means the only class of estimators for which the mmse estimation problem can be posed. In fact, any class of bilinear signal representations which is characterized (linearly) by a kernel and includes the WD as a member can be used as a class of estimators. An example is the class of time-scale representations proposed by Rioul and Flandrin [16], which may be useful in the case of processes exhibiting a $1/f$-type spectral structure. The corresponding globally optimal kernel will always be characterized by a linear equation; however, the characterization may not always be as simple and explicit as in the case of Cohen's class, and the solution may not take such a simple form. A second possible extension is to apply mmse estimation techniques to estimate particular physical quantities, like the random instantaneous frequency, which are derivable from the TFR. In such problems, the optimal kernels would, in general, be different from the optimal kernels derived in this paper, and may involve more complicated characterizations.

Appendix A

In considering the case when B_X is not bounded away from zero over S_{YX}, we define

$$S_\epsilon = \overline{\{(\theta,\tau) \in S_{YX} : B_X(\theta,\tau) > \epsilon |B_{YX}(\theta,\tau)|\}} \ , \ \epsilon > 0 \ . \tag{56}$$

Note that since $S_{YX} \subseteq S_X$, $\cup_{n=1}^{\infty} S_{(\epsilon=\frac{1}{n})} = S_{YX}$. Also define

$$c = \inf_{S_{YX}} \frac{B_X(\theta,\tau)}{|B_{YX}(\theta,\tau)|} \ \geq 0 \ . \tag{57}$$

If $c > 0$, then we note from (57) and (13) that ϕ_{opt} is bounded and is characterized as in Proposition 1 in section 3.1. On the other hand, if $c = 0$ then k_{YX} defined in (19) is not bounded. In this case we can obtain a bounded approximation, $\hat{\phi}_{opt}$, to ϕ_{opt} by inverting (13) over $S_\epsilon \subset S_{YX}$, $\epsilon > 0$. Again, $\hat{\phi}_{opt}$ is characterized just as ϕ_{opt} is in Proposition 1 by replacing S_{YX} by S_ϵ. Moreover, since $\cup_{n=1}^{\infty} S_{(\epsilon=\frac{1}{n})} = S_{YX}$, we note from (20) that by choosing ϵ small enough, $\hat{\phi}_{opt}$ can be made arbitrarily close to ϕ_{opt} in the sense of achieving the minimum mse.

Appendix B

In this appendix, we show that in the proposed framework, in almost all nontrivial cases, averaging of the empirical TFR is necessary for optimal estimation of the nonstationary spectrum or the reference TFR. We first discuss the spectrum estimation scenario.

Suppose we start with an arbitrary definition of the nonstationary spectrum corresponding to the kernel ϕ_o. Then, from (26) we have $\phi_{opt} = k_{YX}\phi_o$, which implies that

$$P_x(t,f;\phi_{opt}) = K_{YX}(t,f) * *P_x(t,f;\phi_o) \ , \tag{58}$$

where K_{YX} is the 2d Fourier transform of k_{YX}. Thus, the averaging done by ϕ_{opt} is completely characterized by k_{YX} because *relative* to the *definition* of the spectrum, the empirical TFR, $P_x(t,f;\phi_o)$, corresponds to the "non-averaged" estimate. Now, from (58) it is clear that $k_{YX}(\theta,\tau) = 1 \Leftrightarrow K_{YX}(t,f) = \delta(t)\delta(f)$ corresponds to no averaging; that is, using the empirical TFR, $P_x(t,f;\phi_o)$, as an estimate of the spectrum, $E\{P_Y(t,f;\phi_o)\}$. In the other extreme, $k_{YX}(\theta,\tau) = \delta(\theta)\delta(\tau) \Leftrightarrow K_{YX}(t,f) = 1$ corresponds to maximum averaging. Guided by intuition, we argue that in all nontrivial cases, the optimal kernel effects some averaging. Thus, we want to characterize the case when $k_{YX}(\theta,\tau) = d$, for some constant d, which corresponds to no averaging. We recall from our discussion on the support of the optimal kernel that if either EA_X or EA_Y has essentially finite support, then so does k_{YX}; the finite support of k_{YX} itself introduces some averaging in the optimal estimate. However, we want to characterize the cases when $k_{YX} = d$ over its support, S_{YX} (see (17)).

We first give a general characterization of the "no-averaging" scenario. Recall from our derivation of the optimal kernel in section 3.1 that, at each (θ,τ) in S_{YX}, $\phi_{opt}(\theta,\tau)A_X(\theta,\tau)$ is the linear mmse estimate of $EA_Y(\theta,\tau)$ given $A_X(\theta,\tau)$. Then, it follows from the projection theorem [10] that EA_Y can be uniquely decomposed as

$$EA_Y(\theta,\tau) = k_{YX}(\theta,\tau)A_X(\theta,\tau) + N_Y(\theta,\tau) \ , \tag{59}$$

where k_{YX} is defined in (18), and the component $N_Y = EA_Y - k_{YX}A_X$ is orthogonal to A_X; that is, $E\{N_Y(\theta,\tau)A_X^*(\theta,\tau)\} = 0$ for all (θ,τ). Thus, we see from (59) that the optimal kernel effects no averaging; that is, $k_{YX} = d$ over S_{YX}, if and only if the component of EA_Y corresponding to A_X has the same scale factor ($k_{YX} = d$) for all (θ,τ). This should be compared to the decomposition (59) in the general case in which k_{YX} is not constant and thus A_X is scaled differently for different values of (θ,τ). We now characterize the "no-averaging" cases in an important scenario.

Proposition B1. If $X(t) = Y(t) + Z(t)$, $t \in T$, where both Y and Z are zero-mean and independent of each other, then no averaging is needed ($k_{YX} = 1$) if and only if
i) $Z = 0$ almost surely (a.s.); that is, $X = Y$ a.s., and
ii) $Y(t) = Y_o u(t)$ in the mean-square sense, where Y_o is an arbitrary random variable satisfying $|Y_o| =$ *constant* a.s., and $u(t)$ is some unit-energy deterministic function.
Proof: First note from (18) that $k_{YX} = 1$ if and only if $EA_Y EA_X^* = E\{|A_X|^2\}$. After expanding A_X by using $X = Y + Z$, it can be easily verified that $EA_Y EA_X^* = E\{|A_X|^2\}$ if and only if

$$E\{|A_Z|^2\} + [E\{|A_Y|^2\} - |EA_Y|^2] + 4E\{\mathrm{Re}^2(A_{YZ})\} + EA_Y^* EA_Z = 0 \ , \tag{60}$$

where A_{YZ} is the cross-ambiguity function between Y and Z [9]. Since $A_s(0,0) = \|s\|_2^2$, we note that at the origin, $(\theta,\tau) = (0,0)$, all the terms (treating the terms inside the brackets as one) in (60) are nonnegative, and hence must individually be zero (note that $E\{|A_Y|^2\} - |EA_Y|^2 \geq 0$ by the Cauchy Schwarz (CS) inequality). In particular, $E\{|A_Z(0,0)|^2\} = E\{\|Z\|_2^4\} = 0$, which implies that $Z = 0$ a.s. and thus $X = Y$ a.s. The only nonzero term remaining in (60) is $[E\{|A_Y|^2\} - |EA_Y|^2]$ which, by the CS inequality, is zero if and only if $A_Y(\theta,\tau)$ is a deterministic function a.s., that is $A_Y = \lambda A_u$ for some $\lambda > 0$ and some unit-energy deterministic signal u. Now, $A_Y = \lambda A_u$ implies that $R_Y(t_1,t_2) = \lambda u(t_1)u^*(t_2)$, and thus Y admits the KL expansion $Y(t) = Y_o u(t)$ in mean-square, with $E|Y_o|^2 = \lambda$. On the other hand, if Y admits such a KL expansion, then $A_Y = |Y_o|^2 A_u$, which is equal to λA_u a.s. only if $|Y_o|^2 = \lambda$ a.s. This completes the proof.

A special case which implies no averaging is that of perfect estimation; that is, the minimum mse is zero. The following proposition characterizes such a situation.

Proposition B2. In the global nonstationary spectral estimation problem, the minimum mse (see (20)) is zero, if and only if $EA_Y(\theta,\tau) = \alpha A_X(\theta,\tau)$ a.s., for all (θ,τ), for some $\alpha > 0$. Moreover, the above condition is equivalent to X and Y having the mean-square (KL) representation $X(t) = X_o u(t)$ and $Y(t) = Y_o u(t)$, where $u(t)$ is some unit-energy deterministic signal, X_o and Y_o are random variables satisfying $|X_o|$=constant a.s. and $E|Y_o|^2 = \alpha E|X_o|^2$ for some $\alpha > 0$.
Proof: From (20) we note that minimum mse is zero if and only if $S_Y = S_{YX}$ and $|E\{EA_Y A_X^*\}|^2 = E\{|A_X|^2\}|EA_Y|^2$, which, by the CS inequality, is true if and only $EA_Y(\theta,\tau) = \alpha A_X(\theta,\tau)$ a.s., where $\alpha > 0$ because the AF is positive at the origin. Now, suppose that $EA_Y(\theta,\tau) = \alpha A_X(\theta,\tau)$ a.s. Then, since EA_Y is a deterministic function, we see that A_X is a deterministic function a.s., and, from the proof of Proposition B1, we conclude that $X(t) = X_o u(t)$ where u is some unit-energy deterministic signal and X_o is a random variable satisfying $|X_o| = \lambda =$ constant a.s. Then, using the eigenexpansion for R_Y and the expression for EA_Y in terms of it (see (6)), we conclude that $Y(t) = Y_o u(t)$ in m.s. with $E|Y_o|^2 = \alpha E|X_o|^2$. It is easy to see that if X and Y admit such representation then $EA_Y = \alpha A_X$ a.s. This completes the proof.

TFR Estimation. Next, we state the corresponding results in the TFR estimation problem, to characterize the cases in which no averaging is needed to optimally estimate the reference TFR of a realization of a random process from a noisy version of that realization. Note, arguing similarly as in the discussion preceding Proposition B1, that relative to the reference TFR (corresponding to ϕ_r) to be estimated, the averaging done by the estimator is completely characterized by k_{YX}. The case when k_{YX} is identically a constant corresponds to the "no-averaging" case.

First, we just note that in the TFR estimation problem an exactly similar decomposition as in (59) holds for A_Y as well, with a corresponding "no-averaging" interpretation. Moreover, it can be easily verified that perfect estimation (a case of "no-averaging") is possible if and only if $A_X = \alpha A_Y$ a.s., which in particular implies that $R_X(t,s) = \alpha R_Y(t,s)$. Moreover, the following proposition characterizes the "no-averaging" cases in an important scenario which is typical of many situations of interest.

Proposition B3. If $X(t) = Y(t) + Z(t)$, $t \in T$, where both Y and Z are zero-mean and independent of each other, then no averaging is needed ($k_{YX} = 1$) if and only if $Z = 0$ a.s.; that is $X = Y$ a.s.
Proof: Similar to the proof of Proposition B1.

Acknowledgment

The authors would like to thank the reviewers for their useful suggestions and comments which resulted in an improved presentation.

References

[1] R. M. Loynes, "On the concept of spectrum for nonstationary processes", *J. Roy. Statist. Soc., Ser. B*, vol. 30, pp. 1–30, 1968.

[2] M. B. Priestley, "Evolutionary spectra and non-stationary processes", *J. Roy. Statist. Soc., Ser. B*, vol. 27, pp. 204–229, 1965.

[3] W. Martin, "Time-frequency analysis of random signals", in *Proceedings of the IEEE Int. Conf. on Acoust., Speech and Signal Proc. — ICASSP '82*, 1982, pp. 1325–1328.

[4] W. Martin and P. Flandrin, "Wigner-Ville spectral analysis of non-stationary processes", *IEEE Trans. Acoust., Speech and Signal Processing*, vol. 33, no. 6, pp. 1461–1470, December 1985.

[5] M. Amin, "Time-frequency spectrum analysis and estimation for non-stationary processes", in *Time-Frequency Signal Analysis*, Halsted Press-Wiley, 1992, (B. Boashash, Ed.), pp. 208–232.

[6] S. Hearon and M. Amin, "Statistical trade-offs in modern time-frequency kernel design", in *Proceedings of the 25th Asilomar Conference on Signals, Systems, and Computers*, November 1991.

[7] A. S. Kayhan, A. El-Jaroudi, and L. F. Chaparro, "Minimum-variance evolutionary spectral estimation of nonstationary signals", in *Proceedings of the IEEE Int. Conf. on Acoust., Speech and Signal Proc. — ICASSP '91*, 1991, pp. 3165–3168.

[8] K. S. Riedel, "Optimal data-based kernel estimation of evolutionary spectra", *IEEE Tran. Signal Processing*, vol. 41, pp. 2439–2447, July 1993.

[9] L. Cohen, "Time-frequency distributions — a review", *Proceedings of the IEEE*, vol. 77, no. 7, pp. 941–981, July 1989.

[10] D. G. Luenberger, *Optimization by Vector Space Methods*, John Wiley and Sons, Inc., New York, 1969.

[11] A. M. Sayeed, "Optimal kernels for nonstationary spectral estimation", M.S. Thesis, Univeristy of Illinois at Urbana-Champaign, August 1993.

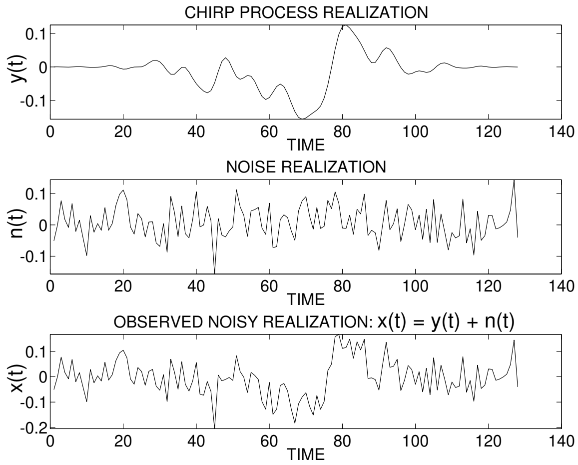
[12] R. A. Altes, "Detection, estimation and classification with spectrograms", *J. Acoust. Soc. Am.*, vol. 67, no. 4, pp. 1232–1246, April 1980.

[13] K. S. Miller, *Complex Stochastic Processes*, Addison-Wesley, 1974.

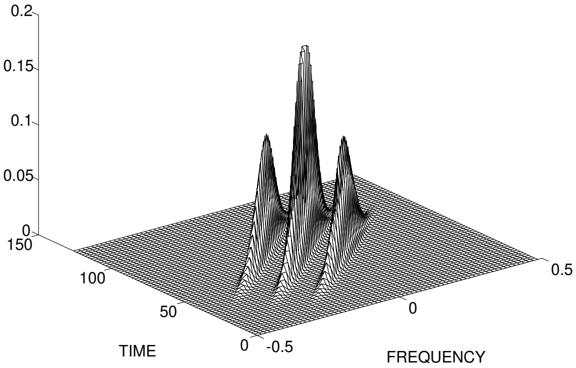
[14] G. M. Jenkins and D. G. Watts, *Spectral Analysis and Its Applications*, Holden-Day, San Francisco, CA, 1968, p. 275.

[15] H. Stark and J. W. Woods, *Probability, Random Processes, and Estimation Theory for Engineers*, Prentice Hall, New Jersey, 1986.

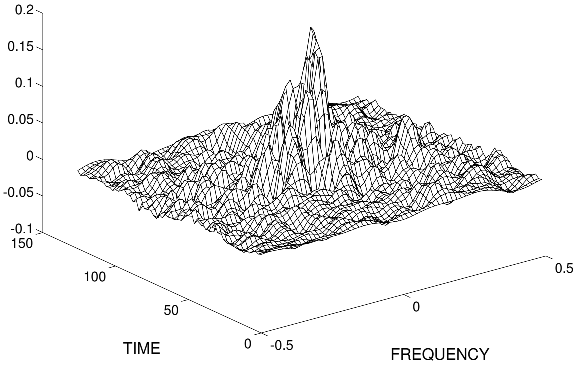
[16] O. Rioul and P. Flandrin, "Time-scale distributions: A general class extending the wavelet transform", *IEEE Trans. Signal Processing*, vol. 46, pp. 1746–1757, May 1992.



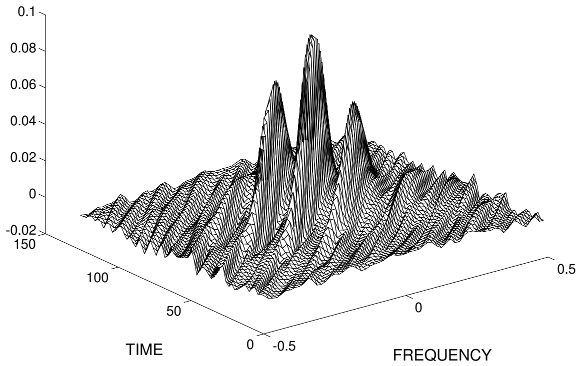
(a)



(b)



(c)



(d)

Figure 1: Estimation of the WVS of a chirp process from a realization corrupted by white noise ($SNR_{max} = 3dB$). (a) Real part of the signal and noise realizations. (b) True WVS of the chirp process. (c) SPW ($M = 6$, $N = 11$) estimate of the WVS. (d) Optimal estimate of the WVS.

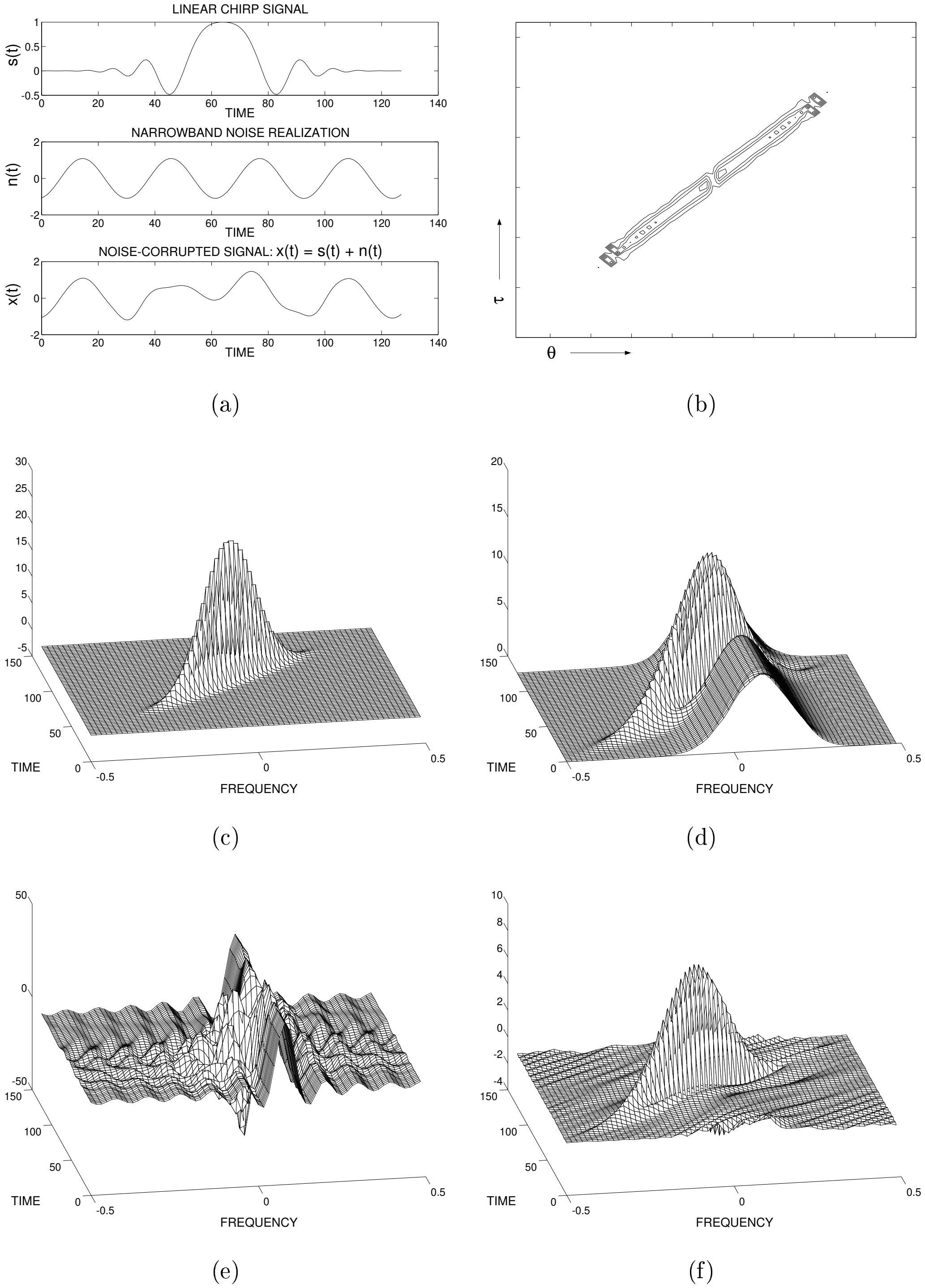
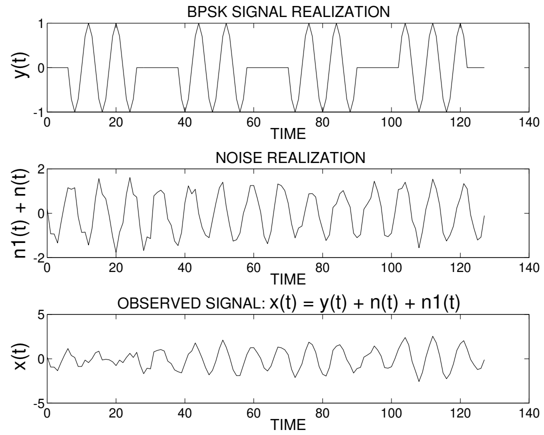
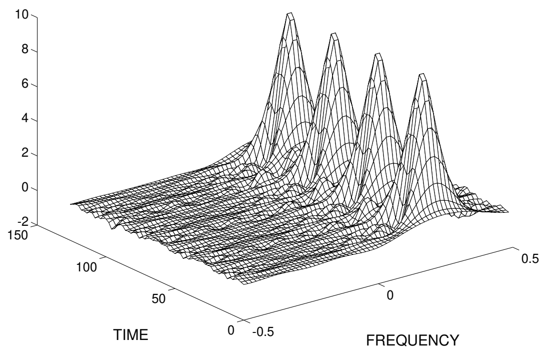


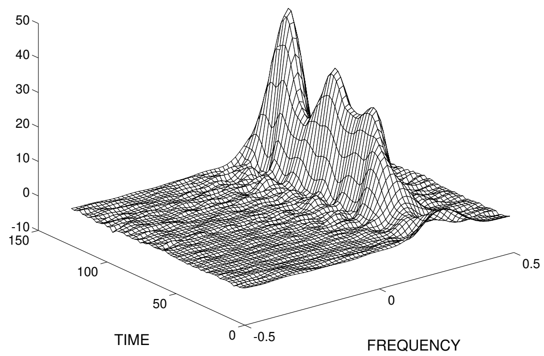
Figure 2: Estimation of the WD of a chirp signal corrupted by narrowband interference ($SNR_{max} = 0dB$). (a) Real part of the signal and noise realizations. (b) Contour plot of the optimal kernel. (c) The WD of the chirp signal. (d) Matched spectrogram of the corrupted signal. (e) SPW ($M = 2$, $N = 16$) estimate of the WD. (f) Optimal estimate of the WD.



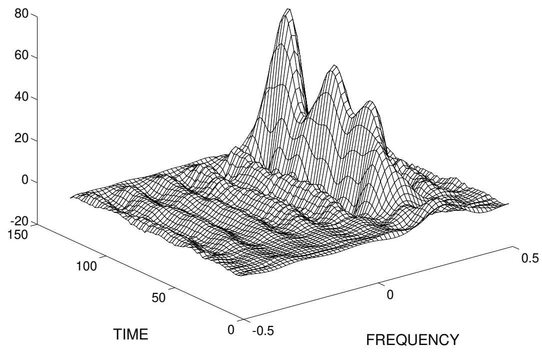
(a)



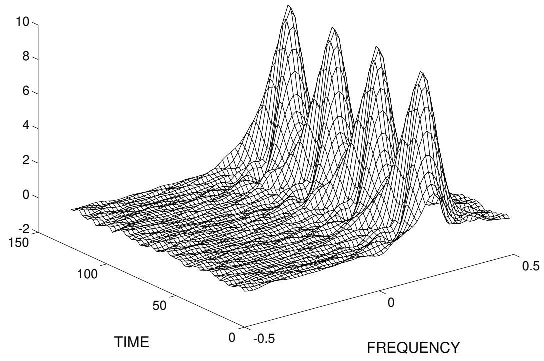
(b)



(c)



(d)



(e)

Figure 3: Estimation of the matched spectrogram of a BPSK signal corrupted by white noise and narrowband co-channel interference ($SINR_{max} = 0dB$). (a) Real part of the signal and noise realizations. (b) Desired TFR: the matched spectrogram (matched to BPSK pulse). (c) Matched spectrogram of the corrupted signal. (d) SPW ($M=6$, $N=8$) estimate of the desired TFR. (e) Optimal estimate of the desired TFR.