2012

# Fusing Heterogeneous Data for Detection Under Non-stationary Dependence

Hao He
*Syracuse University*, hhe02@syr.edu

Arun Subramanian
*Syracuse University*, arsubram@syr.edu

Pramod K. Varshney
*Syracuse University*, varshney@syr.edu

Thyagaraju Damarla
*US Army Research Lab*

# Fusing Heterogeneous Data for Detection Under Non-stationary Dependence

Hao He, Arun Subramanian, Pramod K. Varshney
Dept. of Electrical Engineering & Computer Science
Syracuse University, Syracuse, NY 13244 U.S.A.
Email: {hhe02, arsubram, varshney}@syr.edu

Thyagaraju Damarla
US Army Research Lab
2800 Powder Mill Rd., Adelphi, MD 20783 U.S.A.
Email: thyagaraju.damarla.civ@mail.mil

*Abstract*—**In this paper, we consider the problem of detection for dependent, non-stationary signals where the non-stationarity is encoded in the dependence structure. We employ copula theory, which allows for a general parametric characterization of the joint distribution of sensor observations and, hence, allows for a more general description of inter-sensor dependence. We design a copula-based detector using the Neyman-Pearson framework. Our approach involves a sample-wise copula selection scheme, which for a simple hypothesis test, is proved to perform better than previously used single copula selection schemes. We demonstrate the utility of our copula-based approach on simulated data, and also for outdoor sensor data collected by the Army Research Laboratory at the US southwest border.**
**Keywords: Detection, dependence modeling, heterogeneous sensing, model selection, sensor fusion, information fusion.**

## I. INTRODUCTION

Fusion of data from heterogeneous sources of information, observing a certain phenomenon, has been shown to improve the performance of several inference tasks. Two sensors are said to be heterogeneous if their respective observation models cannot be described by the same probability density function (pdf) [1]. Naturally, an information fusion system comprising multi-modal sensors satisfies this definition. However, sensors of the same modality too can be heterogeneous, in the sense defined here, as they may span varied deployment and manufacturing conditions.

In this paper, we consider the design of false-alarm constrained detectors that operate in non-stationary environments. The non-stationarity is assumed to manifest itself as time-varying spatial dependence across the sensors. This is a plausible situation, especially in multi-modal deployments: based on the physics governing the individual modalities, transient phenomena may affect one modality more drastically than the other. This would, therefore, cause the intermodal dependence to fluctuate, but leave the marginal models relatively invariant within the same observation window. In other words, for reasonably short observation windows, the signal from a single modality can be modeled as a quasi-stationary process, an approach that has been used extensively in spectral analysis

and statistical signal processing [2], [3]; modeling cross-sensor dependence, on the other hand, would require a more considered approach.

We use a copula-based approach to model dependence. Copulas are parametric functions that couple univariate marginal distribution functions to the corresponding multivariate distribution function. A copula-based formulation is attractive because observations may exhibit significant nonlinear dependence across sensors, which cannot be adequately characterized by a linear covariance matrix. Many families of copula functions have been defined in the literature to address this issue. Further, while kernel/learning based methods can model nonlinear dependence and are known to converge to the true joint distribution asymptotically, they also suffer from scalability issues stemming from the curse of dimensionality. Copulas are widely used to model stochastic dependence in the fields of econometrics and finance [4] and have been shown to be useful in various signal processing contexts [5]–[8].

In the sections that follow, we develop the above ideas in more detail. Section II discusses the related literature and provides a brief discussion on copula-based inference. The detection problem is formulated in Section III. Copula selection is an important component of any copula based inference task and our approach, described in Section IV, specifically addresses the issue of non-stationarity dependence. We discuss our results in Section V. We compare the performance of our detector to previously used approaches on simulated data and also evaluate its performance on seismic and acoustic data collected by the U.S. Army Research Laboratory at the US southwest border. Concluding remarks and a brief discussion on the directions for future research are provided in Section VI.

## II. BACKGROUND

### A. Previous work

Copula-based approaches for both centralized and distributed signal processing have been studied recently. Iyengar et al. [1] have investigated the general framework of copula-based detection of a phenomenon being observed jointly by heterogeneous sensors. They quantify the performance loss due to copula misspecification and demonstrate that a detector using a copula selection scheme based on area under the receiver operating characteristic (ROC) can provide significant improvement over models assuming independence. Their results on a NIST multibiometric dataset show that the copula

based approach is versatile and can fuse not only heterogeneous sensor measurements, but can also be applied to fuse different algorithms. Sundaresan et al. [9] consider the case of distributed detection and derive the optimum fusion rules for a Neyman-Pearson detector. Sundaresan and Varshney [10] also design and analyze the performance of a copula-based estimation scheme for the localization of a radiation source.

As mentioned above, we also consider seismic-acoustic fusion as an application of the proposed detector. The data collected mimic typical scenarios for surveillance of human activity at border crossings. Signals due to footsteps can be used as surrogates for human presence when monitoring a scene of interest using seismic or acoustic sensors. A copula based detector fusing seismic and acoustic footstep data for personnel detection in indoor environments has been discussed by Iyengar et al. [11]. Their approach combines canonical correlation analysis (CCA) and copula modeling to characterize cross-modal dependence in the frequency domain. Time-frequency analysis using the spectrogram has been shown to effectively characterize footstep information [12]. Copula based approaches fusing indoor seismic data have also been examined for footstep detection [13].

### B. Copula theory

As stated in Section I, copulas are parametric functions that couple univariate marginal distributions to a valid multivariate distribution. They explicitly model the dependence among random variables, which may have arbitrary marginal distributions. Copula theory is an outcome of the work on probabilistic metric spaces [14] and a copula was initially defined, on the unit hypercube, as a joint probability distribution for uniform marginals. Their application to statistical inference is possible largely due to Sklar's theorem, which is stated below without proof [15].

*Theorem 1 (Sklar's Thoerem):* Consider an $m$-dimensional distribution function $F$ with marginal distribution functions $F_1, \ldots, F_m$. Then there exists a copula $C$, such that for all $x_1, \ldots, x_m$ in $[-\infty, \infty]$

$$F(x_1, x_2, \ldots, x_m) = C(F_1(x_1), F_2(x_2), \ldots, F_m(x_m)) \quad (1)$$

If $F_k$ is continuous for $1 \leq k \leq m$, then $C$ is unique, otherwise it is determined uniquely on $RanF_1 \times \ldots \times RanF_m$ where $RanF_k$ is the range of cumulative distribution function (CDF) $F_k$. Conversely, given a copula $C$ and univariate CDFs $F_1, \ldots, F_m$, $F$ as defined in (1) is a valid multivariate CDF with marginals $F_1, \ldots, F_m$.

Note that the arguments of $C$ in (1) are uniformly distributed random variables. As a direct consequence of Sklar's Theorem, for continuous distributions, the joint probability density function (pdf) is obtained by differentiating both sides of (1),

$$f(x_1, \ldots, x_m) = \left( \prod_{i=1}^{m} f_i(x_i) \right) c(F_1(x_1), \ldots, F_m(x_m)) \quad (2)$$

where, $c$ is termed as the copula density and is given by,

$$c(\mathbf{u}) = \frac{\partial^m (C(u_1, \ldots, u_m))}{\partial u_1, \ldots, \partial u_m} \quad (3)$$

where, $u_i = F_i(x_i)$. Several copula functions are defined in the literature, and are constructed to characterize different types of dependence [15], of which the elliptical and Archimedean copulas are widely used. Some of these are listed in Table I. While not explicitly specified in (1) and (2), copula functions contain a *dependence parameter* that quantifies the amount of dependence between the $m$ random variables. We denote the dependence parameter as $\phi$, which, in general, may be a scalar, a vector or a matrix.

An attractive feature of copulas is that nonparametric rank-based measures of dependence, such as Kendall's $\tau$, can be expressed as expectations over the copula distribution. For independent pairs of random variables $(X_1, Y_1)$ and $(X_2, Y_2)$ having the same distribution as $(X, Y)$, concordance is defined as the condition that $(X_1 - X_2)(Y_1 - Y_2) \geq 0$ and discordance is defined as the condition that $(X_1 - X_2)(Y_1 - Y_2) < 0$. Kendall's $\tau$ is defined to be the difference between the probabilities of concordance and discordance:

$$\tau \triangleq \mathbb{P}[(X_1 - X_2)(Y_1 - Y_2) \geq 0] - \mathbb{P}[(X_1 - X_2)(Y_1 - Y_2) < 0].$$

Nelsen has proved the relationship in (4) for a copula, $C$, and random variables $X \sim f_X(x), Y \sim f_Y(y)$ [15, p. 161], i.e.,

$$\tau(\boldsymbol{\phi}) = 4 \, \mathbb{E}[C_\phi(F_X(x), F_Y(y))] - 1. \quad (4)$$

This relationship allows $\tau$ to be expressed in terms of the dependence parameter of the copula, $C$ ($\Sigma$ for the elliptical copulas and $\phi$ for the Archimedean copulas in Table I). For the case of elliptical copulas, parametrized by the matrix $\Sigma = [\rho_{ij}]$,

$$\rho_{ij} = \sin \left( \frac{\pi \tau_{ij}}{2} \right), \quad (5)$$

where $\tau_{i,j}$ is the Kendall's $\tau$ evaluated for the pair $(U_i, U_j) = (F_{X_i}(\cdot), F_{X_j}(\cdot))$. The sample estimate of Kendall's $\tau$, for $N$ observations, can be calculated as the ratio of the difference in the number of concordant pairs, $c$, and discordant pairs, $d$, to the total number of pairs of observations, i.e.,

$$\hat{\tau} = \frac{c - d}{c + d} = \frac{c - d}{\binom{N}{2}} \quad (6)$$

Typically, the value of the dependence parameter is not known *a priori*, and $\phi$ needs to be estimated, e.g., using maximum likelihood estimation (MLE). On the other hand, (6) and (4) imply that Kendall's $\tau$ can be used for calculating computationally efficient estimates of $\phi$.

### III. PROBLEM FORMULATION

Consider a scene or phenomenon being monitored by a sensor suite, consisting of $L$ sensors. The $i$-th sensor, $i = 1, 2, \ldots, L$, makes a set of $N$ measurements, $x_{ij}, j = 1 \ldots, N$. These measurements may represent a time series (with $j$ being the time index), spectral coefficients (with $j$ being the frequency index), or some other feature vector. The vector $\mathbf{x}_j$ denotes the $j$-th measurements at all sensors, i.e., $\mathbf{x}_j = [x_{1j}, x_{2j}, \ldots, x_{Lj}]^\mathsf{T}$. The collective measurements, $\mathbf{x} = [\mathbf{x}_1, \mathbf{x}_2 \ldots, \mathbf{x}_j, \ldots, \mathbf{x}_N]$, are received at a processing unit or fusion center (FC). Based on the *joint* characteristics of $\mathbf{x}$, the FC decides whether a phenomenon is present or absent in

| Copulas | | Parametric Form | | Parameter Range |
|---|---|---|---|---|
| Elliptical copulas | Gaussian | $\Phi_\Sigma(\Phi^{-1}(u_1),\ldots,\Phi^{-1}(u_m)),$ | $\Phi_\Sigma(\mathbf{x}) = \int_0^{\mathbf{x}} \mathcal{N}(\mathbf{x};\mathbf{0},\Sigma)d\mathbf{x},\ \mathbf{x} \in \mathbb{R}^m$ $\Phi^{-1}(u) = \inf_{x\in\mathbb{R}}\{u \le \int_0^x \mathcal{N}(x;0,1)dx\}$ | $\Sigma = [\rho_{ij}], i,j = 1,\ldots,m$ $\rho_{ij} \in [-1,1]$ |
| | Student-$t$ | $t_{\nu,\Sigma}(t_\nu^{-1}(u_1),\ldots,t_\nu^{-1}(u_m)),$ | $t_{\nu,\Sigma}$ : multivariate Student-$t$ CDF $t_\nu^{-1}$ : inverse CDF of univariate Student-$t$ | $\nu$ : degrees of freedom, $\nu \ge 3$ |
| Archimedean copulas | Clayton | $\left(\sum_{i=1}^m u_i^{-\phi} - 1\right)^{-\frac{1}{\phi}}$ | | $\phi \in [-1,\infty)\backslash\{0\}$ |
| | Frank | $-\frac{1}{\phi}\log\left(1 + \frac{\prod_{i=1}^m [\exp\{-\phi u_i\}-1]}{\exp\{-\phi\}-1}\right)$ | | $\phi \in \mathbb{R}\backslash\{0\}$ |
| | Gumbel | $\exp\left\{-\left(\sum_{i=1}^m(-\ln u_i)^\phi\right)^{\frac{1}{\phi}}\right\}$ | | $\phi \in [1,\infty)$ |
| | Independent | $\prod_{i=1}^m u_i$ | | – |

the region of interest and, thus, solves the following hypothesis testing problem:

$$H_0 : f_0(\mathbf{x}) = \prod_{j=1}^N f_0(\mathbf{x}_j)$$
$$H_1 : f_1(\mathbf{x}) = \prod_{j=1}^N f_1(\mathbf{x}_j), \quad (7)$$

where $H_0$ is the null hypothesis that the background process is observed, and $H_1$ is the alternative, i.e., the phenomenon of interest is observed. The pdfs under the null and alternative hypotheses are, respectively, denoted as $f_0$ and $f_1$. In taking the product over all $j$ in (7), we assume that for a given sensor, signals are independent over the index $j$, e.g., over time. However, in general,

$$f_k(\mathbf{x}_j) \ne \prod_{i=1}^L f_k(x_{ij}), \quad k = 0,1 \quad (8)$$

This formulation, therefore, asserts that since the sensors are observing the same phenomenon, at any given instant, sensor measurements need not be independent spatially (across sensors).

Using Sklar's theorem (Section II-B, Theorem 1), the joint densities in (7) can be expressed in terms of the copula densities, $c_0$ and $c_1$, respectively under $H_0$ and $H_1$, as,

$$H_0 : \quad f_0(\mathbf{x}) = \prod_{j=1}^N \left[ \left(\prod_{i=1}^L f_0(x_{ij}|\boldsymbol{\theta}_{0i})\right) \times c_0(u_{1j}^0(\boldsymbol{\theta}_{01}),\ldots,u_{Lj}^0(\boldsymbol{\theta}_{0L})|\boldsymbol{\phi}_0) \right]$$
$$H_1 : \quad f_1(\mathbf{x}) = \prod_{j=1}^N \left[ \left(\prod_{i=1}^L f_1(x_{ij}|\boldsymbol{\theta}_{1i})\right) \times c_1(u_{1j}^1(\boldsymbol{\theta}_{11}),\ldots,u_{Lj}^1(\boldsymbol{\theta}_{1L})|\boldsymbol{\phi}_1) \right]. \quad (8)$$

The copula arguments are the probability integral transforms (PIT) of $x_{ij}$ under hypothesis $H_k$, i.e., for sensor $i$ and measurement $j$,

$$u_{ij}^k(\boldsymbol{\theta}_{ki}) = F_k(x_{ij}|\boldsymbol{\theta}_{ki}) \quad k = 0,1. \quad (9)$$

The quantities $\{\boldsymbol{\theta}_0, \boldsymbol{\theta}_1\}$ and $\{\boldsymbol{\phi}_0, \boldsymbol{\phi}_1\}$ in (8) are, respectively, the marginal density parameters and copula parameters under $\{H_0, H_1\}$. When these parameters are known, the likelihood ratio test (LRT) is the optimal test. Equivalently, we can compare the log-likelihood ratio (LLR) to a threshold $\eta$,

$$T_{\mathsf{LR}}(\mathbf{x}) \underset{H_0}{\overset{H_1}{\gtrless}} \eta, \quad (10)$$

where,

$$T_{\mathsf{LR}}(\mathbf{x}) = \log \frac{f_1(\mathbf{x})}{f_0(\mathbf{x})}$$
$$= \sum_{j=1}^N \sum_{i=1}^L \log \frac{f_1(x_{ij}|\boldsymbol{\theta}_{1i})}{f_0(x_{ij}|\boldsymbol{\theta}_{0i})}$$
$$+ \sum_{j=1}^N \log \frac{c_1(u_{1j}^1(\boldsymbol{\theta}_{11}),\ldots,u_{Lj}^1(\boldsymbol{\theta}_{1L})|\boldsymbol{\phi}_1)}{c_0(u_{1j}^0(\boldsymbol{\theta}_{01}),\ldots,u_{Lj}^0(\boldsymbol{\theta}_{0L})|\boldsymbol{\phi}_0)} \quad (11)$$

These parameters are typically unknown and have to be estimated. Using maximum likelihood (ML) estimates in place of the true parameter values, the test becomes a generalized likelihood ratio test (GLRT) in the Neyman-Pearson framework. From (8) and (9), it is seen that the copula density is also a function of the marginal parameter, $\boldsymbol{\theta}_{ki}$, through the PIT. Thus, ideally, ML estimation of the parameters would require simultaneous maximization of the joint likelihood function over both, the marginal and copula parameters. This is, however, difficult and a consistent two-step estimation procedure is commonly used in copula literature [16]. The two-step maximum likelihood (TSML) procedure first maximizes the individual marginal likelihoods over each $\boldsymbol{\theta}_{ki}$:

$$\widehat{\boldsymbol{\theta}}_{ki} = \arg\max_{\boldsymbol{\theta}_{ki}} \sum_{j=1}^N f_k(x_{ij}|\boldsymbol{\theta}_{ki}) \quad (12)$$

The second step of TSML substitutes $\theta_{ki} = \widehat{\theta}_{ki}$ in (9); the copula likelihood, thus obtained, is then maximized over $\phi_k$, i.e.,

$$\hat{u}_{ij}^k = u_{ij}^k(\widehat{\theta}_{ki}) \tag{13}$$

$$\widehat{\phi}_k = \arg\max_{\phi_k} \sum_{j=1}^{N} c_k(\hat{u}_{1j}^k, \ldots, \hat{u}_{Lj}^k | \phi_k). \tag{14}$$

The GLRT then can be expressed as,

$$T_{\mathsf{GLR}}(\mathbf{x}) \underset{H_0}{\overset{H_1}{\gtrless}} \eta, \tag{15}$$

where,

$$T_{\mathsf{GLR}}(\mathbf{x}) = \sum_{j=1}^{N}\sum_{i=1}^{L} \log \frac{f_1(x_{ij}|\widehat{\theta}_{1i})}{f_0(x_{ij}|\widehat{\theta}_{0i})} \\ + \sum_{j=1}^{N} \log \frac{c_1(u_{1j}^1(\widehat{\theta}_{11}), \ldots, u_{Lj}^1(\widehat{\theta}_{1L})|\widehat{\phi}_1)}{c_0(u_{1j}^0(\widehat{\theta}_{01}), \ldots, u_{Lj}^0(\widehat{\theta}_{0L})|\widehat{\phi}_0)} \tag{16}$$

Alternatively, for the bivariate case ($L = 2$), we can also use the sample estimate of Kendall's $\tau$, defined in (6), to estimate $\phi_k$. Noting that the relation in (6) is invertible, we rewrite the function relationship between $\tau$ and $\phi_k$ in (4), in terms of a function $g_k$ so that,

$$\tau = g_k(\phi_k) \tag{17}$$

$$\Rightarrow \phi_k = g_k^{-1}(\tau). \tag{18}$$

The resultant estimate of $\phi_k$ is given by

$$\widetilde{\phi}_k = g_k^{-1}(\hat{\tau}). \tag{19}$$

Further, since $\hat{\tau}$ is a consistent estimator of $\tau$ [17], $\widetilde{\phi}_k \to \phi_k$ as $N \to \infty$. For finite $N$, using $\widetilde{\phi}_k$ instead of $\widehat{\phi}_k$, in (16), results in a sub-optimal test, but a simpler estimation procedure.

## IV. DETECTION UNDER NON-STATIONARY DEPENDENCE

In Section I, we motivated the need to consider non-stationary dependence. While the preceding section assumes that the family of copulas, $c_0$ and $c_1$, are known, a formulation with non-stationary dependence has to necessarily drop that assumption. In the following discussion, we assume that the background model can be predetermined to some degree: the family of the marginals is known and $c_0$ is known. The more general case of unknown $c_0$ is considered by Iyengar et al. [1], but signal detection for such a scheme would need to be implemented under a training-testing paradigm. However, non-stationarity notwithstanding, the true underlying copula under $H_1$, $\mathfrak{c}$, is typically not known; this "true copula" is usually abstracted as a single copula, but it may, in fact, be a composite of several copulas interacting in an indeterminate fashion, accounting for the non-stationary nature of observations. Due to these complexities, copula selection is an important part of copula based inference and several copula selection methods have been proposed [1], [7], [13]. Our assumptions are stated more precisely as follows:

1) We assume that $f_0$, the marginal density families under $H_0$, are known for each $i = 1, \ldots, L$. The corresponding marginal parameters, $\theta_{0i}$, may be unknown.
2) The $H_0$ copula family, $c_0$, is known but $\phi_0$ may be estimated, if needed. This section, however, assumes, without loss of generality, that $c_0 = 1$, i.e., measurements under $H_0$ are independent across sensors. However, the discussion is valid for any known $c_0$. The independence under the null hypothesis also allows us to simplify our notation; we do not explicitly notate for $H_1$ with respect to the copula function. Therefore, we set

$$c_1(\cdot) \equiv c(\cdot)$$
$$u_{ij}^1(\theta_{ki}) \equiv u_{ij}$$
$$\phi_1 \equiv \phi$$

3) The copula under the alternative, $c_1$, is not known *a priori*. The "best" copula, in the sense of maximum likelihood, is selected from a predefined library of copulas, $\mathcal{C} = \{c_m : m = 1, \ldots, M\}$.

Based on these assumptions, we discuss three detection scenarios: detection with known parameters, detection with unknown parameters, and detection with unknown marginals under $H_1$ and unknown copula parameters.

### A. Detection with known parameters

For some applications, it may be feasible to determine, *a priori*, the value of the copula parameter $\phi_m$ for each $c_m \in \mathcal{C}$. The actual selection of the copula may be done online. For this case the test-statistic is formulated as a modification of (11),

$$T_{\mathsf{LR}}(\mathbf{x}) = \log \frac{f_1(\mathbf{x})}{f_0(\mathbf{x})} \\ = \sum_{j=1}^{N}\sum_{i=1}^{L} \log \frac{f_1(x_{ij}|\theta_{1i})}{f_0(x_{ij}|\theta_{0i})} \\ + \sum_{j=1}^{N} \log c_j^*(u_{1j}, \ldots, u_{Lj}|\phi_j^*), \tag{20}$$

where for each $j$ the maximum copula likelihood is selected from the library, i.e.,

$$c_j^* = \max_{c_m \in \mathcal{C}} \mathcal{C} \tag{21}$$

$$\phi_j^* = \arg c_j^* \tag{22}$$

The key difference here is that previous papers have proposed selecting a single copula for the entire observation window [1], [13], i.e., choose

$$c_{\mathbf{N}}^* = \max_{\substack{c_m \in \mathcal{C} \\ \forall j}} \mathcal{C} \tag{23}$$

$$\phi_{\mathbf{N}}^* = \arg c_{\mathbf{N}}^*. \tag{24}$$

On the other hand, we select the best copula for *each* $j$ adapting to potentially changing dependence structure. Denote the Kullback-Leibler (KL) divergence between the pdfs $g$ and

$h$ as $D(g||h)$. We now prove that selecting the best copula for each $j$, as opposed to a single best copula for all $N$, leads to a smaller KL divergence from the single true copula, $\mathfrak{c}$.

*Proposition 1:* Let $\mathbf{X} \sim f_\mathbf{X}(\mathbf{x}), \mathbf{X} \in \mathbb{R}^{N \times L}$, where,

$$f_\mathbf{X}(\mathbf{x}) = \prod_{j=1}^{N} \left[ \left( \prod_{i=1}^{L} f_{X_i}(x_{ij}|\boldsymbol{\theta}_{1i}) \right) \mathfrak{c}(u_{1j}, \ldots, u_{Lj}|\boldsymbol{\phi}) \right] \quad (25)$$

where $\mathfrak{c}$ is the true copula. For the copula library,

$$\mathcal{C} = \{c_m : m = 1, \ldots, M\}, \quad (26)$$

and selection schemes (21) and (23),

$$D(f_\mathbf{X}||f_{c_j^*}) \leq D(f_\mathbf{X}||f_{c_\mathbf{N}^*}), \quad (27)$$

where $f_{c^*}$ and $f_{c_\mathbf{N}^*}$ are the joint densities for $\mathbf{X}$ under $H_1$ using (21) and (23) respectively.

*Proof:* Consider the case $M = 2$. Choosing $c_1$ over $c_2$ when $c_1 \geq c_2$ is equivalent to the decision rule when copula selection is posed as a decision problem with equally likely copulas. Let $\Omega$ represent the sample space. Let $\Omega_m \subset \Omega$ represent the decision region for $\mathbf{x}$ for which $c_m$ ($m = 1, 2$) is chosen, so that $\Omega_1 \cup \Omega_2 = \Omega$ and $\Omega_1 \cap \Omega_2 = \emptyset$. Denote the product of marginals as $f_p(\mathbf{x}_j)$, i.e.,

$$f_p(\mathbf{x}_j) = \prod_{i=1}^{L} f_{X_i}(x_{ij}|\boldsymbol{\theta}_{1i})$$

Also, define the following sets:

$$\mathcal{J}_1 = \{j : \mathbf{x}_j \in \Omega_1\} \text{ and } \mathcal{J}_2 = \{j : \mathbf{x}_j \in \Omega_2\}.$$

Then,

$$\begin{aligned}
D(f_\mathbf{X}||f_{c_j^*}) &= \int \log \frac{f_\mathbf{X}(\mathbf{x})}{f_{c_j^*}(\mathbf{x})} dF_\mathbf{X} \\
&= \int \log \frac{\prod_{j=1}^{N} f_p(\mathbf{x}_j)\mathfrak{c}(u_{1j}, \ldots, u_{Lj}|\boldsymbol{\phi})}{\prod_{j=1}^{N} f_p(\mathbf{x}_j) \prod_{\mathcal{J}_1} c_1(\cdot|\boldsymbol{\phi}_1) \prod_{\mathcal{J}_2} c_2(\cdot|\boldsymbol{\phi}_2)} dF_\mathbf{X} \\
&= \int \sum_{j=1}^{N} \log \mathfrak{c}(\cdot) dF_\mathbf{X} \\
&\quad - \int \left[ \sum_{\mathcal{J}_1} \log c_1(\cdot) + \sum_{\mathcal{J}_2} \log c_2(\cdot) \right] dF_\mathbf{X}
\end{aligned} \quad (28)$$

The selection criterion in (21), implies that, for the set $\mathcal{J}_2$, $c_2 \geq c_1$. Therefore,

$$\begin{aligned}
\sum_{\mathcal{J}_1} \log c_1(\cdot) + \sum_{\mathcal{J}_2} \log c_2(\cdot) &\geq \sum_{\mathcal{J}_1} \log c_1(\cdot) + \sum_{\mathcal{J}_2} \log c_1(\cdot) \\
&= \sum_{j=1}^{N} \log c_1(\cdot),
\end{aligned} \quad (29)$$

and in a similar manner,

$$\sum_{\mathcal{J}_1} \log c_1(\cdot) + \sum_{\mathcal{J}_2} \log c_2(\cdot) \geq \sum_{j=1}^{N} \log c_2(\cdot) \quad (30)$$

Therefore, depending on whether $c_1$ or $c_2$ was chosen using (23), we can substitute either of the inequalities in (29) or (30) into (28) to get,

$$D(f_\mathbf{X}||f_{c_j^*}) \leq D(f_\mathbf{X}||f_{c_\mathbf{N}^*})$$

This proves (27) for $M = 2$. For $M > 2$ we can successively partition $\Omega_2$ and arrive at a similar result by repeating the above steps. ∎

Proposition 1 implies that a detector using the proposed selection scheme in (21) will suffer a lower loss in detection performance due to copula misspecification [1].

### B. Detection with unknown parameters

With unknown parameters, the statistic in (16) for the composite hypothesis testing problem can be rewritten as,

$$\begin{aligned}
T_{\mathsf{GLR}}(\mathbf{x}) = &\sum_{j=1}^{N} \sum_{i=1}^{L} \log \frac{f_1(x_{ij}|\widehat{\boldsymbol{\theta}}_{1i})}{f_0(x_{ij}|\widehat{\boldsymbol{\theta}}_{0i})} \\
&+ \sum_{j=1}^{N} \log c_j^*(u_{1j}(\widehat{\boldsymbol{\theta}}_{11}), \ldots, u_{Lj}(\widehat{\boldsymbol{\theta}}_{1L})|\widehat{\boldsymbol{\phi}}_j^*),
\end{aligned} \quad (31)$$

where the TSML procedure has been used to obtain estimates of marginal and copula parameters. The copula parameters $\widehat{\boldsymbol{\phi}}_m$ are estimated, over all $N$, for each $c_m \in \mathcal{C}$, so that

$$\mathcal{C} = \{c_m(\widehat{\boldsymbol{\phi}}_m(N)) : m = 1, \ldots, M\} \quad (32)$$
$$c_j^* = \max_{c_m \in \mathcal{C}} \mathcal{C} \quad (33)$$
$$\widehat{\boldsymbol{\phi}}_j^* = \arg c_j^* \quad (34)$$

While this selection method is motivated by the implications of Proposition 1 for the simple hypothesis case, a similar result may not be stated for the composite test. This is because ML estimation requires that all $N$ samples be drawn from the same population; this need not hold true for copula selection from $\mathcal{C}$ with unknown parameters.

The copula parameters can also be estimated using $\hat{\tau}$. The test-statistic is then,

$$\begin{aligned}
T_{\hat{\tau}}(\mathbf{x}) = &\sum_{j=1}^{N} \sum_{i=1}^{L} \log \frac{f_1(x_{ij}|\widehat{\boldsymbol{\theta}}_{1i})}{f_0(x_{ij}|\widehat{\boldsymbol{\theta}}_{0i})} \\
&+ \sum_{j=1}^{N} \log c_j^*(u_{1j}(\widehat{\boldsymbol{\theta}}_{11}), \ldots, u_{Lj}(\widehat{\boldsymbol{\theta}}_{1L})|\widetilde{\boldsymbol{\phi}}_j^*),
\end{aligned} \quad (35)$$

where $\widetilde{\boldsymbol{\phi}}_j^*$ is the estimate of $\boldsymbol{\phi}_j^*$ based on $\hat{\tau}$. Correspondingly,

$$\mathcal{C} = \{c_m(\widetilde{\boldsymbol{\phi}}_m) : m = 1, \ldots, M\} \quad (36)$$
$$c_j^* = \max_{c_m \in \mathcal{C}} \mathcal{C} \quad (37)$$
$$\widetilde{\boldsymbol{\phi}}_j^* = \arg c_j^* \quad (38)$$

### C. Detection with unknown marginals and unknown copula parameters

In many applications, establishing a model under $H_1$ is not feasible. In that case, $f_1$ is determined non-parametrically

and $u_{ij}$ is obtained using the empirical probability integral transform (EPIT). The test statistic is, therefore, expressed as,

$$T_{\text{NPM}}(\mathbf{x}) = \sum_{j=1}^{N} \sum_{i=1}^{L} \log \frac{\hat{f}_1(x_{ij})}{f_0(x_{ij}|\hat{\boldsymbol{\theta}}_{0i})} \tag{39}$$
$$+ \sum_{j=1}^{N} \log c_j^*(\hat{u}_{1j}, \ldots, \hat{u}_{Lj}|\widehat{\boldsymbol{\phi}}_j^*),$$

where $c_j^*$ and associated parameters are selected as indicated in (36), (37) and (38). The uniform random variables in the copula density are evaluated using EPIT,

$$\hat{F}_i(\cdot) = \frac{1}{N} \sum_{j=1}^{N} \mathbb{I}_{x_{ij} < \{\cdot\}} \tag{40}$$

$$\hat{u}_{ij} = \hat{F}_i(x_{ij}) \tag{41}$$

where $\mathbb{I}$ is the indicator function.

The marginal model under $H_1$ is determined through a kernel density estimation procedure. Kernel density estimators [18] provide a smoothed estimate, $\hat{f}_1(x_{ij})$, of the true density. Choosing the correct bandwidth for kernel density estimation is important for an accurate estimate. The kernel bandwidth is chosen using leave-one-out cross-validation. The selected bandwidth, $h^*$, is the minimizer of the cross-validation estimator of risk, $\hat{J}$, for a kernel, $K$. The risk estimator may be easily computed using the approximation [18, p. 136],

$$\hat{J}(h) = \frac{1}{hN^2} \sum_p \sum_q K^* \left( \frac{X_p - X_q}{h} \right)$$
$$+ \frac{2}{Nh} K(0) + \mathcal{O}\left( \frac{1}{N^2} \right), \tag{42}$$

where,

$$K^*(x) = K^{(2)}(x) - 2K(x)$$
$$K^{(2)}(z) = \int K(z-y)K(y)dy.$$

The Gaussian kernel was selected, so that $K(x) = \mathcal{N}(x; 0, 1)$ and $K^{(2)}(z) = \mathcal{N}(z; 0, 2)$. Therefore,

$$h^* = \arg\min_h \hat{J}(h)$$

## V. RESULTS AND DISCUSSION

In this section, we present results when the copula-based tests, discussed in Section IV, are applied to simulated and real data. Our results are presented for a two-sensor case, i.e., $L = 2$. We note, however, that the methods described above apply generally to $L > 2$, as one can construct a multivariate copula using bivariate components [13].

### A. Simulated data

We simulated normal and beta distributed marginals and considered various cases of copula dependence. The marginals and the respective parameters used are tabulated in Table II. For all copula cases we used Kendall's $\tau = 0.2$ to specify dependence. The copula library contains the Gaussian and

TABLE II
DISTRIBUTION OF MARGINALS FOR SIMULATION EXPERIMENTS

| $i$ | $H_0$ | $H_1$ |
|---|---|---|
| 1 | $\mathcal{N}(0,1)$ | $\mathcal{N}(0.1, 1.1)$ |
| 2 | Beta(2.0,2.0) | Beta(2.2,2.2) |

Frank copulas. For all the cases, we compare performances obtained when testing with $T_{\text{GLR}}$ in (31), $T_{\hat{\tau}}$ in (35), GLR using single copula selection and the product rule, i.e., independence assumption. The results presented are averaged over $10^4$ Monte-Carlo trials with 1000 samples per trial.

In Figs. 1,2 and 3 receiver operating characteristics (ROC) for different generating copulas are shown. In Fig. 1, we consider the case in which a $t$ copula is used to generate the data. Note that this represents the case where the true copula is not known and is not included in the copula library. The label non-stationary copula refers to the sample-wise selection scheme proposed in this paper. Fig. 2 represents the case where half of all $x_{ij}$ were simulated with a Gaussian copula and the remaining half consisted of samples generated from the Frank copula. This is, therefore, the case where the copula library also accommodates the generating model. The case of a single generating copula that is also a member of the library is also considered; Fig. 3 shows results when all the data are generated using the Frank copula.

For all simulation scenarios we observe that the GLRT and the test based on $\hat{\tau}$, using our selection scheme perform comparably. Both outperform the single copula selection and product rules. We note that these results represent the unknown parameter case (Section IV-B), for which we were not able to prove that our method would outperform the single copula selection method. An intuition for why we observe this result is that since $\hat{\tau}$ is consistent, for large $N$, $\hat{\tau} \to \tau$. Also, one value of $\tau$ corresponds to different values of $\phi_m = \arg c_m(\cdot), c_m \in \mathcal{C}$. We conjecture that, asymptotically, this is as if the parameter values are known, allowing Proposition 1 to be applicable. This implies that while $\tau$ controls the *amount* of dependence, which remains unchanged for all $N$, different copulas represent the *shape* of the dependence between the data from the two sensors. Verifying this conjecture mathematically is difficult, and will be addressed in our future work.

### B. ARL footstep data

We used the footstep data, made available by the US Army Research Laboratory (ARL), collected at the US southwest border. The dataset consists of raw observations from several sensors of different modalities that were deployed in an outdoor space to record human and animal activity that is typical in perimeter and border surveillance scenarios. The participants in the data collection exercise walked/ran along a predetermined path with sensors laid out along either side of the path. In this paper, we consider copula-based seismic-acoustic fusion.

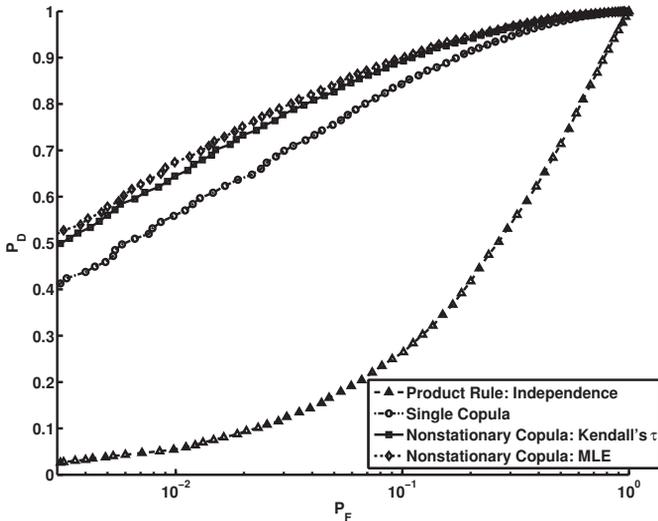Seismic and acoustic time series for activities representing

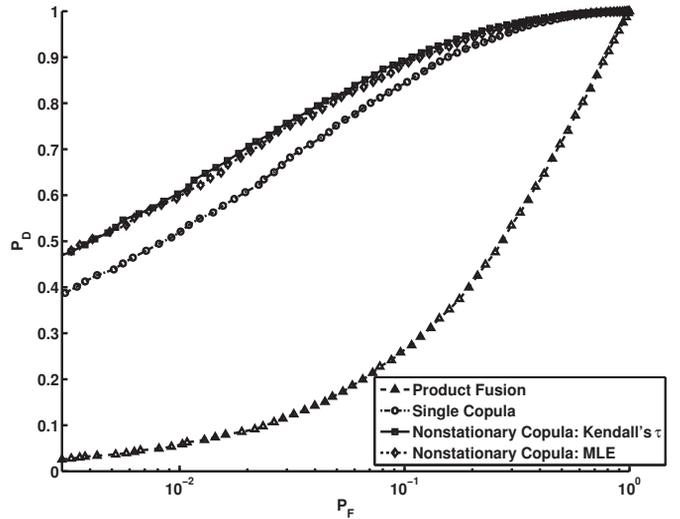Fig. 1. ROCs for $H_1$ data generated using a $t$ copula



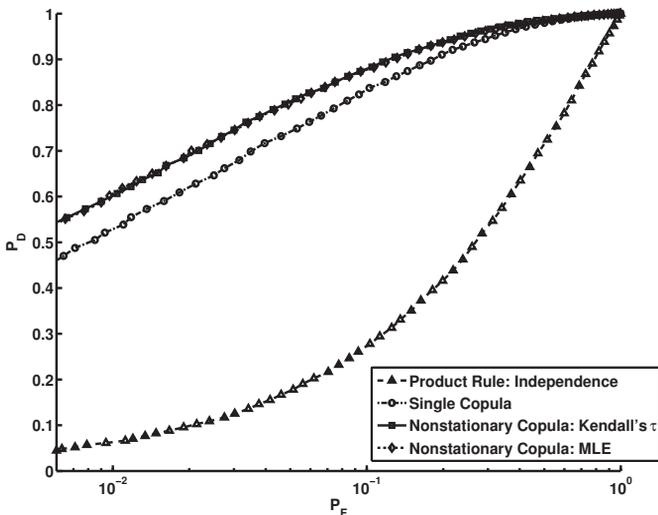Fig. 3. ROCs for $H_1$ data generated using a Frank copula



Fig. 2. ROCs for $H_1$ data generated using Frank and Gaussian copulas

a single person walking, two persons walking and human leading an animal (among other examples) are available in the ARL dataset. Each seismic/acoustic time series contains a leading 60s of background data. We use this as our $H_0$ data. The data are sampled at 10kHz, and are mean centered and oscillatory in nature.

Before applying the copula-based detector, we first pre-process the data. The time series is split into non-overlapping frames of length $T = 512$. This raw time series data is called $x_{Ti}(t)$ where $i = 1, 2$ is the sensor index for the acoustic and seismic modalities respectively, and $t$ is the time index. In keeping with Houston's analysis that Fourier spectra for seismic and acoustic footstep data are more informative than time-domain measurements [12], we set

$$x_{ij} = \sqrt{\mathcal{F}\{x_{Ti}(t)\}^2},$$

where $\mathcal{F}$ is the DFT and $j = 1, \ldots, N = 256$ is the frequency index. Our sensor measurements are, therefore, now

transformed to the frequency domain and the statistics of $\mathbf{x} = [x_{ij}]$ are used as the input to the detector. The copula library consists of Gaussian, Gumbel and Frank copulas. We have observed that due to the interstitial nature of footstep data, including the independence copula (Table I) in the library improves the overall detection performance.

For the ARL dataset, we use the statistic $T_{\text{NPM}}(\mathbf{x})$ in (39). To generate ROCs, we compare the test-statistic to a vector of thresholds. The curve thus generated, for the case when $H_1$ corresponds to one person walking, is shown in Fig. 4. This curve is compared to the ROCs for single copula selection scheme as well as the product rule, i.e., independence assumption for $H_1$. Similar ROCs are obtained for the cases of two persons walking, and man leading an animal and are shown in Fig. 5 and Fig. 6.

For all the three cases, we observe that our proposed method, using the sample-wise copula selection for non-stationary data, outperforms the ROCs corresponding to single copula selection and independence. We further observe that, the two-persons and man-leading-animal cases have a higher probability of detection ($P_D$) for a given probability of false alarm ($P_F$), when compared to the one person case. This is intuitive, since for the two-persons case and man-leading-animal case, we have a higher signal to noise ratio.

## VI. CONCLUSION

In this paper, we have considered a detection problem, with dependent heterogeneous sensor data. We used a copula-based approach to model the inter-sensor dependency and applied our scheme to detect non-stationary phenomena. We considered a specific type of non-stationarity that affects the inter-modal dependence more severely than the individual sensor model. A copula-based approach was used to design detectors for dependent, non-stationary data. We have shown that, for a simple hypothesis testing problem, a sample-wise copula selection scheme performs better than selecting a single copula for the entire observation window. However, a similar
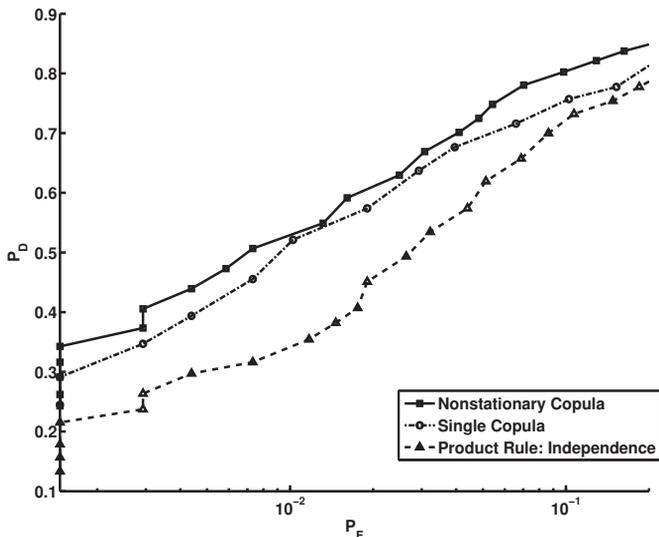
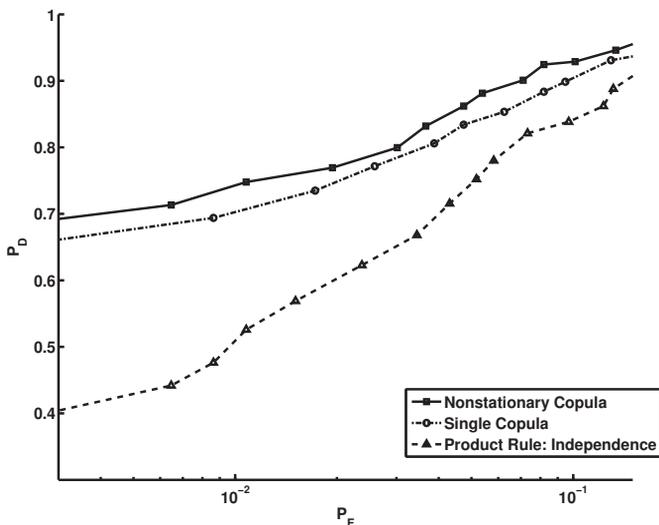Fig. 4. ROCs for the ARL dataset for 1 person vs. background detection.



Fig. 6. ROCs for the ARL dataset for 1 person leading an animal vs. background detection.
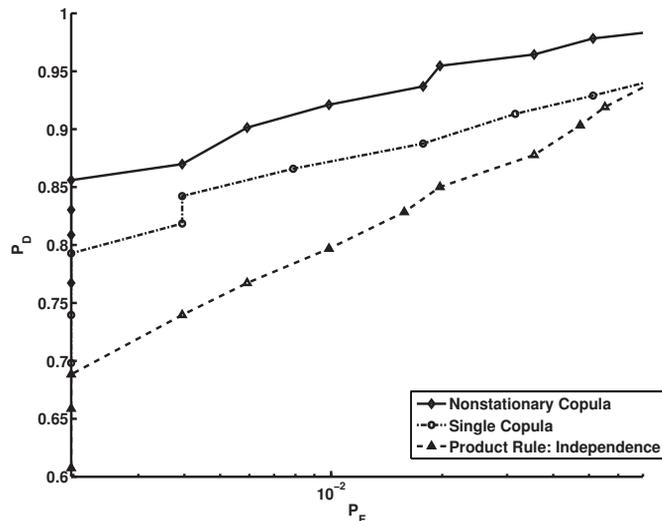


Fig. 5. ROCs for the ARL dataset for 2 persons vs. background detection.

conclusion cannot be formally stated for the GLRT. We note that the copula parameters can be estimated using distribution-free, rank based methods such as Kendall's $\tau$ and observed that using the sample estimate, $\hat{\tau}$, gave comparable performance to MLE-based detection. This motivates future investigation into whether Proposition 1 can be generalized to the case of unknown parameters. Empirical results are encouraging, and support this idea. We note that our method on simulated data performs favorably even when the true copula is not a part of the library. Similarly, results on acoustic and seismic datasets also show that our method yields superior performance.

REFERENCES

[1] S. G. Iyengar, P. K. Varshney, and T. Damarla, "A Parametric Copula Based Framework for Hypotheses Testing using Heterogeneous Data," *IEEE Trans. Signal Process.*, vol. 59, no. 5, pp. 2308 – 2319, May 2011.

[2] W. Martin and P. Flandrin, "Wigner-Ville spectral analysis of nonstationary processes," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 33, no. 6, pp. 1461–1470, 1985.

[3] R. Dwyer and L. Kurz, "Characterizing partition detectors with stationary and quasi-stationary Markov dependent data," *IEEE Trans. Inf. Theory*, vol. 32, no. 4, pp. 471–482, 1986.

[4] U. Cherubini, E. Luciano, and W. Vecchiato, *Copula Methods in Finance*. Wiley, 2004.

[5] M. Davy and A. Doucet, "Copulas: a new insight into positive time-frequency distributions," *IEEE Signal Processing Letters*, vol. 10, no. 7, pp. 215–218, Jul. 2003.

[6] G. Mercier, G. Moser, and S. Serpico, "Conditional copula for change detection on heterogeneous SAR data," in *Proc. IEEE International Geoscience and Remote Sensing Symposium IGARSS 2007*, Jul. 23–28, 2007, pp. 2394–2397.

[7] A. Sundaresan, "Detection and Location Estimation of a Random Signal Source Using Sensor Networks," Ph.D. dissertation, Syracuse University, Syracuse, NY, December 2010.

[8] S. G. Iyengar, "Decision Making with Heterogeneous Sensors - A Copula Based Approach," Ph.D. dissertation, Syracuse University, Syracuse, NY, August 2011.

[9] A. Sundaresan, P. K. Varshney, and N. S. V. Rao, "Copula-Based Fusion of Correlated Decisions," *IEEE Transactions on Aerospace and Electronic Systems*, vol. 47, no. 1, pp. 454–471, 2011.

[10] A. Sundaresan and P. K. Varshney, "Location Estimation of a Random Signal Source Based on Correlated Sensor Observations," *IEEE Transactions on Signal Processing*, vol. 59, no. 2, pp. 787–799, 2011.

[11] S. G. Iyengar, P. K. Varshney, and T. Damarla, "On the Detection of Footsteps Based on Acoustic and Seismic Sensing," in *Conference Record of the Forty-First Asilomar Conference on Signals, Systems and Computers ACSSC 2007*, 4–7 Nov. 2007, pp. 2248–2252.

[12] K. M. Houston and D. P. McGaffigan, "Spectrum analysis techniques for personnel detection using seismic sensors," in *Unattended Ground Sensor Technologies and Applications V*, E. M. Carapezza, Ed., vol. 5090, no. 1. SPIE, 2003, pp. 162–173.

[13] A. Subramanian, A. Sundaresan, and P. K. Varshney, "Fusion for the detection of dependent signals using multivariate copulas," in *Proc. 14th Int Information Fusion (FUSION) Conf*, 2011, pp. 1–8.

[14] B. Schweizer and A. Sklar, *Probabilistic Metric Spaces*. New York: North Holland, 1983.

[15] R. Nelsen, *An Introduction to Copulas*, 2nd ed. New York: Springer, 2006.

[16] P. K. Trivedi and D. M. Zimmer, "Copula Modeling: An Introduction for Practitioners," *Foundations and Trends in Econometrics*, vol. 1, no. 1, pp. 1–111, 2005.

[17] J. D. Gibbons and S. Chakraborti, *Nonparametric Statistical Inference*, 4th ed. New York, NY: Marcel Dekker, 2003.

[18] L. Wasserman, *All of Nonparametric Statistics (Springer Texts in Statistics)*. Secaucus, NJ, USA: Springer-Verlag New York, Inc., 2006.