Kernel-Based Copula Processes

Sebastian Jaimungal[1] and Eddie K.H. Ng[2]

[1]Department of Statistics, University of Toronto
sebastian.jaimungal@utoronto.ca
[2]The Edward S. Rogers Sr. Department of Electrical and Computer Engineering, University of Toronto
eddie@psi.toronto.edu

Abstract. Kernel-based Copula Processes (KCPs), a new versatile tool for analyzing multiple time-series, are proposed here as a unifying framework to model the interdependency across multiple time-series and the long-range dependency within an individual time-series. KCPs build on the celebrated theory of copula which allows for the modeling of complex interdependence structure, while leveraging the power of kernel methods for efficient learning and parsimonious model specification. Specifically, KCPs can be viewed as a generalization of the Gaussian processes enabling non-Gaussian predictions to be made. Such non-Gaussian features are extremely important in a variety of application areas. As one application, we consider temperature series from weather stations across the US. Not only are KCPs found to have modeled the heteroskedasticity of the individual temperature changes well, the KCPs also successfully discovered the interdependencies among different stations. Such results are beneficial for weather derivatives trading and risk management, for example.

Keywords: Copula, Kernel Methods, Gaussian Processes, Time-Series Analysis, Heteroskedasticity, Maximum Likelihood Estimation, Financial Derivatives, Risk Management.

1 Introduction

Gaussian processes (GPs) [3] have enjoyed wide acceptance in many machine learning applications as a non-linear regression and classification tool. GPs cleverly exploited the closure property of Gaussian random variables, under marginalization and conditioning, for efficient learning and inference. Model selection can be handled naturally under a Bayesian framework, and missing data can be treated effortlessly. The same closure property also limits the GPs to make predictions with Gaussian distributions. However, the world is an inherently non-Gaussian place. The Gaussian prior assumption is simply not valid in many real-world applications such as the modeling of financial data[12], wind-speeds and temperatures, and catastrophic damages (e.g. floods, earth-quakes and forest fires[6] [8]) where it has been shown that risk-management based on Gaussian assumptions severely under-estimates the real-risks and greatly inflates the social and economical costs. The ability to accommodate non-Gaussian features is therefore crucial.

W. Buntine et al. (Eds.): ECML PKDD 2009, Part I, LNAI 5781, pp. 628–643, 2009.

Copula theory [13] on the other hand has been proven to be an effective way to model complex dependencies and distributional behavior by decomposing the joint-distribution into the dependency structure and the individual marginal distributions. However, the applications of copula functions are mostly static while its dynamic extensions are either limited to first-order Markov processes [16] or simply using copula functions as a way to combine multiple time-series [5] [11] [14] [15].

This work proposes a novel time-series analysis model dubbed *Kernel-based Copula Processes* (KCPs), and offers a way to better address the non-Gaussian features of real-world data. KCPs are a powerful union of GP and copula functions. Unlike GPs where a time-series is modeled as a high-dimensional Gaussian distribtuion, KCPs use a copula function to separate the specification of dependency structure and the behavior in the marginal distributions; thus allowing for complex non-Gaussian behavior, while retaining most of the attractiveness of GPs such as efficient learning, inference, and natural missing data handling. For example, in the univariate setting, a high-dimensional elliptical copula function is used to capture any long-range temporal dependencies, while an arbitrary non-Gaussian distribution is used to match the non-Gaussian behavior in the marginal distribution. A kernel function is used along with the elliptical copula to keep the model parsimonious. As will be shown in this work, heteroskedastic processes can also be captured with KCPs through proper design of this kernel function.

Under this proposed framework, we further propose a natural extension for multi-variate time-series which allows multiple time-series with drastically different characteristics to be described under a single model. Further, the modular formulation of multivariate KCPs lends itself naturally to a multi-core/processor environment for parallel computation.

Section 2 presents the details of the KCPs framework in both univariate and multivariate settings. Section 3 presents a synthetic univariate example to illustrate the modeling power enjoyed by the non-Gaussian features of KCPs. Finally, section 4 presents a real-life example of modeling multiple temperature series from weather stations across the United States. A new kernel function is derived especially to accomodate the heteroskedastic nature of the temperature data. The performance of different flavors of KCPs are compared with GPs and GARCH [2].

2 Kernel-Based Copula Processes (KCPs)

This section introduces the basic formulation of copula processes for time-series analysis. Section 2.1 provides the formulation for the univariate time-series case which shows how kernel methods can bring a sense of time or parametric ordering to a high-dimensional copula, thus allowing the modeling of random processes instead of just a collection of random variables. This approach is reminiscent of the formulation of Gaussian processes (GP) as will be discussed in this section. Section 2.2 extends the basic formulation to the multiple time-series case. Section 2.3 presents the estimation methods for KCPs.

2.1 Univariate Time-Series

Consider a one-dimensional function $g(x)$ with a set of noisy observations $\mathbf{y} = \{y_1, ..., y_N\}^T = g(\mathbf{x}) + \varepsilon_n$ at $\mathbf{x} = \{x_1, ..., x_N\}^T$. In performing tasks such as regression or forecasting, the objective is to predict the values of the function, $\mathbf{y}^* = \{y_1^*, ..., y_N^*\}^T$, at $\mathbf{x}^* = \{x_1^*, ..., x_M^*\}^T$ where $\mathbf{x} \bigcap \mathbf{x}^* = \emptyset$. Under the KCP assumption, the joint distribution of the observed and predicted values is given by:

$$\mathbb{P}(\mathbf{Y} \leq \mathbf{y}, \mathbf{Y}^* \leq \mathbf{y}^* | \mathbf{x}, \mathbf{x}^*) = \mathbb{C}(\{F_i(y_i)\}_{i=1}^N, \{F_j(y_j^*)\}_{j=1}^M | \mathbf{x}, \mathbf{x}^*) \tag{1}$$

$$= \mathbb{C}(\{u_i\}_{i=1}^N, \{u_j^*\}_{j=1}^M | \mathbf{x}, \mathbf{x}^*) \tag{2}$$

where $\mathbb{C}(\cdot)$ is the copula function[1] of the process, $F_i(\cdot)$ are the marginal cumulative distribution functions (CDFs), and u_i is the transformed y_i via the CDFs, i.e. $u_i \doteq F_i(y_i)$.

The joint density of $(\mathbf{y}, \mathbf{y}^*)$ can be obtained by differentiating with respect to $\{\mathbf{y}, \mathbf{y}^*\}$:

$$\mathbb{P}(\mathbf{y}, \mathbf{y}^* | \mathbf{x}, \mathbf{x}^*) = \mathfrak{c}(\{F_i(y_i)\}_{i=1}^N, \{F_j(y_j^*)\}_{j=1}^M | \mathbf{x}, \mathbf{x}^*) \prod_{i=1}^N f_i(y_i) \prod_{j=1}^M f_j(y_j^*) \tag{3}$$

where $f_i(\cdot)$ are the marginal PDFs and $\mathfrak{c}(\cdot)$ is the so-called *copula density function* given by the derivative of the copula function with respect to its parameters:

$$\mathfrak{c}(\{u_i\}_{i=1}^N) = \frac{\partial^N \mathbb{C}}{\partial u_1 \partial u_2 \dots \partial u_N} \tag{4}$$

In this formulation, all the marginal distributions are constrained to have the same parametric form. The role of the copula function $\mathbb{C}$ is to capture the entire dependency structure of $\{\mathbf{y}, \mathbf{y}^*\}$. There exists a vast array of parametric or empirical copula functions which allow the specification of complex and heavy-tail dependencies. Nelson [13] provides a wide collection of examples. However, when attention is focused on the elliptical class of copula functions, the dependency structure in the copula function can be specified by a kernel function to achieve parsimonious model specification.

For example, consider a copula process with a *Gaussian copula* function as follows:

$$\mathbb{C}_\Lambda(\{F_i(y_i)\}_{i=1}^N) = \Phi_\Lambda(\{\Phi^{-1}(F_i(y_i))\}_{i=1}^N) \tag{5}$$

[1] The copula function is a joint distribution function of uniform random variables. Sklar's theorem [13] states that given any joint distribution $H(y_1, \cdots, y_N)$ of random variables $\{y_i\}$ and margins $\{F_i(y_i)\}$. There exists a copula function $\mathbb{C}$ such that for all y_i in the respective domain, $H(y_1, \cdots, y_N) = \mathbb{C}(F_1(y_1), \cdots, F_N(y_N))$. If $\{F_i(y_i)\}$ are continuous, then $\mathbb{C}$ is unique. The converse is also true.

where the $\Phi(\cdot)$ and $\Phi_\Lambda(\cdot)$ are the standardized univariate normal CDF and the zero-mean multi-variate CDF (with covariance matrix Λ) respectively. The covariance matrix Λ can be defined with the Gram matrix:

$$\Lambda = \begin{bmatrix} k(x_1, x_1) & \dots & k(x_1, x_N) \\ \vdots & \ddots & \vdots \\ k(x_N, x_1) & \dots & k(x_N, x_N) \end{bmatrix} + \sigma_n^2 \mathbf{I} \tag{6}$$

where σ_n^2 is the noise variance and $k(x_i, x_j)$ is a kernel function describing the covariance[2] between the pairs of random variables y_i and y_j. The radial basis function (RBF),

$$k_{RBF}(x_i, x_j) = \exp\left(-\frac{1}{2h}(x_i - x_j)^2\right) \tag{7}$$

and the Ornstein-Uhlenbeck functions,

$$k_{OU}(x_i, x_j) = \exp\left|-\frac{1}{h}(x_i - x_j)\right| \tag{8}$$

are examples of commonly used and versatile kernel functions. More examples of kernel functions can be found in [3]. In the Appendix, we show how to construct a nonstationary kernel function from a stochastic differential equation (SDE) based on the understanding of the dynamics of the temperature data studied in Section 4.

Notice that if the copula is Gaussian and the marginal distributions are assumed to be normal, then the copula process reduces to the familiar GPs.

2.2 Multivariate Time-Series

In practical problems, random processes rarely exist in logical isolation but rather typically form an intricate web of dependencies. Thus practitioners are often faced with the challenge of analyzing not a univariate time-series, but multiple codependent time-series. To this end, we propose a multivariate KCP framework to capture such complex interdependencies.

In the cases where the daily values from each series are highly related and yet there exists no ordering among them, one natural parametrization of the joint distribution is as follows:

$$\begin{aligned} \mathbb{P}(\mathbf{Y}^{(1)} \leq \mathbf{y}^{(1)}, \dots, \mathbf{Y}^{(M)} \leq \mathbf{y}^{(M)}) = \\ \mathbb{C}_b\left[\mathbb{C}_1\left(\{F_i^{(1)}(y_i^{(1)})\}_{i=1}^{N_1}\right), \dots, \mathbb{C}_M\left(\{F_i^{(M)}(y_i^{(M)})\}_{i=1}^{N_M}\right)\right] \end{aligned} \tag{9}$$

where $\mathbb{C}_b(\cdot)$ is the *binding copula function* connecting M individual time-series together. The individual time-series are modeled as in the single KCP case

[2] For non-Gaussian random variables, the kernel function is not the covariance itself, but rather it dictates the behavior of the covariance.

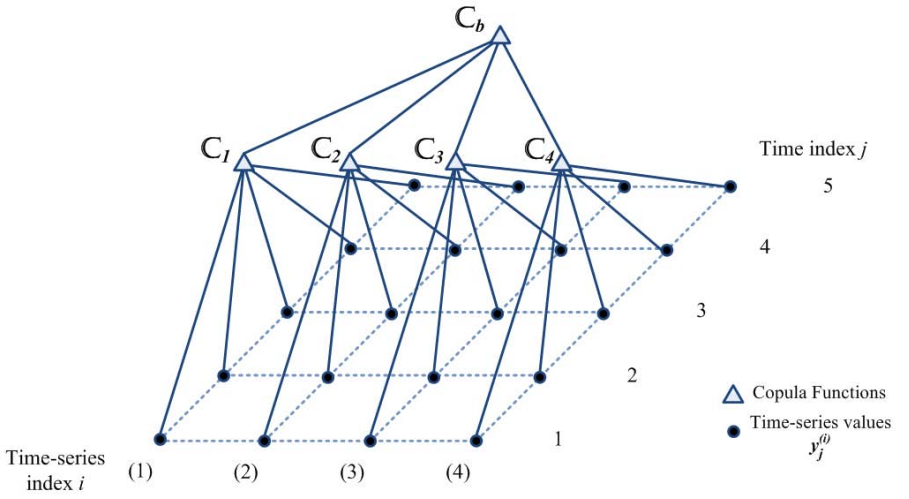

Fig. 1. A schematic illustration of a binding copula linking multiple KCPs. In this case, there are four time-series, each with five time samples. The temporal dependencies of each time-series are first described by the individual KCPs $\{\mathbb{C}_i\}_{i=1}^4$. The inter-time-series dependency is then described by a binding copula $\mathbb{C}_b$.

(Section 2.1), while the dependent structure resides with the binding copula. Figure 2.2 provides a schematic depiction of a binding copula joining multiple copula processes together. The advantage of this configuration is that time-series with greatly different individual dynamics can be captured by the individual copula processes, before they are joined together by the binding copula. For example, one time-series can be the changes in the LIBOR[3] rate while others could be the returns of any stocks that are sensitive to interest-rates, such as utilities companies or companies with high debt load. In such cases, the stock returns would have much higher volatility but shorter term correlation than the changes in the LIBOR rate.

2.3 Learning

The learning of KCPs can be performed by maximizing likelihood. The likelihood function of a generic KCP is given by:

$$\mathcal{L}(\theta) = \mathbb{P}(\mathbf{y}|\mathbf{x},\theta) = \mathfrak{c}(\{F_i(y_i)\}_{i=1}^N) \prod_{i=1}^N f_i(y_i) \tag{10}$$

[3] The London Interbank Offered rate (LIBOR) is a daily reference rate based on the interest rates at which banks offer to lend to other banks [10].

where $\mathbf{y}$ is the set of training data and θ is the set of hyper-parameters of the copula process.

For instance, assuming the choice of Gaussian copula and margins are Student's t-distributed, then the negative log-likelihood function is given as

$$-\log(\mathcal{L}(\theta)) = \frac{1}{2}\log|\Lambda| + \frac{1}{2}\mathbf{z}^T\Lambda\mathbf{z} - \frac{1}{2}\mathbf{z}^T\mathbf{z} - \sum_{i=1}^{N}\log(t(y_i,\upsilon)) \tag{11}$$

where $\mathbf{z}^T = \{z_1 = \Phi^{-1}(T_\upsilon(y_1)), \ldots, z_N = \Phi^{-1}(T_\upsilon(y_N))\}$ is the transformed random vector $z_i \sim \mathcal{N}(0,1)$, and $T_\upsilon(y_i)$ and $t_i(y_i,\upsilon)$ are the univariate Student's t-distribution and density functions (with degree of freedom υ) respectively; and θ are the hyper-parameters of the kernel function. Here, without loss of generality the time-series samples are assumed to be standardized.

The likelihood function for multivariate KCPs can be derived by taking the $N \cdot M - th$ order derivative of Eq. (9) with respect to $y_i^{(j)}$:

$$\begin{aligned}\mathcal{L}(\theta) &= \mathbb{P}(\mathbf{y}^{(1)}, \ldots, \mathbf{y}^{(M)}) \\ &= \frac{\partial^M \mathbb{C}_b}{\partial\mathbb{C}_1 \ldots \partial\mathbb{C}_M} \cdot \left[\frac{\partial^{N_1}\mathbb{C}_1}{\partial u_1^{(1)}\ldots\partial u_{N_1}^{(1)}}\prod_{j=1}^{N_1} f_j^{(1)}(y)\right] \cdot \ldots \\ &\qquad \cdot \left[\frac{\partial^{N_M}\mathbb{C}_M}{\partial u_1^{(M)}\ldots\partial u_{N_M}^{(M)}}\prod_{j=1}^{N_M} f_j^{(M)}(y)\right] \\ &\doteq \mathfrak{c}_b \cdot \mathcal{L}_1(\theta_1) \cdot \ldots \cdot \mathcal{L}_M(\theta_M)\end{aligned} \tag{12}$$

The overall likelihood function factorizes into the product of the binding copula density, $\mathfrak{c}_b$, and the likelihood functions of the copula processes of the individual time-series, $\mathcal{L}_i(\theta_i)$. This modular property allows for two ways to learn the model parameters: (i) all the parameters are learnt at the same time; or (ii) the parameters for each time-series are learnt separately and are fixed while learning the parameters of the binding copula. The latter approach is particular convenient for equity portfolio analysis where individual names are often added and removed as portfolio compositions change.

2.4 Inference

One of the strengths for KCPs (which is also enjoyed by GPs) is that the entire predictive distribution is available during inference, thus predictions are not limited to point prediction, which allows for much more accurate risk management.

The predictive distribution of a univariate KCP is given by:

$$\mathbb{P}(\mathbf{y}^*|\mathbf{y}) = \frac{\mathbb{C}_{\tilde{\Lambda}}(\{F_i(y_i)\}_{i=1}^N, \{F_i(y_j^*)\}_{j=1}^M)}{\mathbb{C}_\Lambda(\{F_i(y_i)\}_{i=1}^N)} \tag{13}$$

where Λ and $\tilde{\Lambda}$ are the Gram matrices of the observations and the combined observations and targets respectively as defined in Eq. 6. The dependency on $\mathbf{x}$ and $\mathbf{x}^*$ is implicit and it is dropped from the notation for clarity of presentation.

The *maximum a-posterior* (MAP) estimator could be used to provide a point estimate of the target values:

$$\bar{\mathbf{y}^*} = \max_{\mathbf{y}^*} \mathbb{P}(\mathbf{y}^*|\mathbf{y}) \tag{14}$$

When the choice of copula functions and marginal distributions result in a symmetric predictive distribution, the MAP estimate will coincide with the mean-squared estimator $\mathbb{E}|_{\mathbb{P}(\mathbf{y}^*|\mathbf{y})}[\mathbf{y}]$.

2.5 Model Selection and Performance Metrics

A model selection method is required to select the best of parametric functions for both the copula and marginal component in KCPs. Given the modularity of multivariate KCPs, model selection can be performed in a modular fashion as well. That is, the best univariate KCP is selected for the individual time series and the associated results are subsequently used to learn the best binding copula. Furthermore, in regression and out-of-sample prediction of stochastic processes, attention must be focused on obtaining an accurate predictive distribution. Point-predictions are meaningless since the realized value from a stochastic process will almost always differ from the prediction. Knowing the entire predictive distribution on the other hand yields a much more complete picture. To this end, conventional performance metrics such as least-square errors between the point prediction and a set of realized values is of little use.

In this study, in addition to the maximized likelihood, the probability-integral-transform (PIT) [7] is employed as a complementary criteria for model selection. The maximized likelihood relates to the predicted density having minimum variance about the observations[4], while PIT measures how well the the predictive distribution corresponds to the true underlying distribution. Diebold *et al.* [7] evaluate PIT graphically, but it can also be evaluated quantitatively with the Anderson-Darling criterion[4] [1]. We will use log-likelihood and likelihood ratios to rank the competing models, while using the PIT with the Anderson-Darling criterion to evaluate the validity of the model on a standalone basis.

In evaluating the goodness-of-fit in the multiple time-series level, a visual comparison of the Gram and pair-wise Pearson's correlation matrices will also be given.

[4] The Anderson-Darling criterion has the added advantage that it produces a score that focuses goodness-of-fit at the tail regions of the distributions.

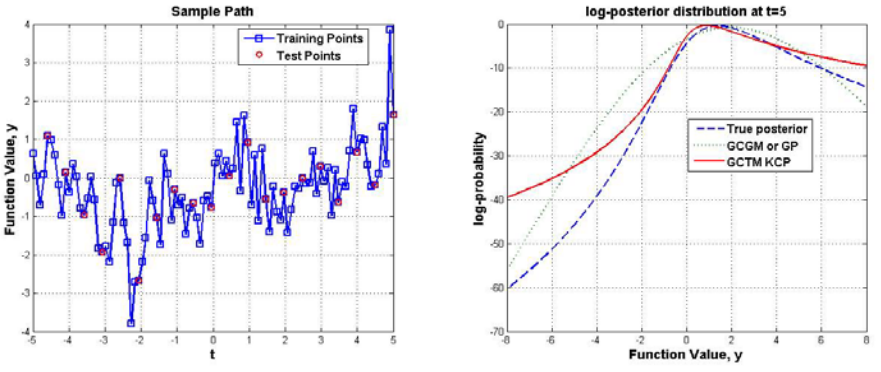

Fig. 2. A synthetic data set generated by a AR(2) process with skew-normal innovations is shown on the left panel. The posterior distribution of the process at $t = 5$ and the corresponding predictive distributions given by GP and GCTM KCP are shown on the right panel.

3 Synthetic Example

In this section a synthetic data set is used to illustrate the KCP in action in its simplest form. Here, a sample path is drawn from an AR(2) process with skew-normal[5] innovations:

$$y_t = 0.1y_{t-1} + 0.05y_{t-1} + 1 + \varepsilon_t \tag{16}$$

where ε_t is a standardized skew-normal with skewness of 0.5.

One hundred samples were taken at regular intervals within the range of $x = [-5, 5]$. Every fifth sample is taken as the test set (i.e. the regression / prediction targets), while the rest of the samples are used as the training set. Then two KCPs: Gaussian copula with Gaussian marginal distribution (GCGM) (or simply a Gaussian process (GP)) and Gaussian copula with t-marginal (GCTM), are learnt using MLE. The Ornstein-Uhlenbeck kernel (8) is used in both cases due to its auto-regressive nature.

[5] The skew-normal distribution can be obtained by taking the limit of the skew-t distribution by Hansen[9] as the degree of freedom η approaches infinity. The full formulation of the standardized skew-t distribution is reproduced here for reference:

$$g(z|\eta,\lambda) = \begin{cases} bc\left(1 + \frac{1}{\eta-2}\left(\frac{bz+a}{1-\lambda}\right)^2\right)^{-(\eta+1)/2} & z < -\frac{a}{b} \\ bc\left(1 + \frac{1}{\eta-2}\left(\frac{bz+a}{1+\lambda}\right)^2\right)^{-(\eta+1)/2} & z \geq -\frac{a}{b} \end{cases} \tag{15}$$

where $2 < \eta < \infty$, and $|\lambda| < 1$. The constants a, b, and c are given by $a = 4\lambda c\left(\frac{\eta-2}{\eta-1}\right)$, $b^2 = 1 + 3\lambda^2 - a^2$, and $c = \frac{\Gamma\left(\frac{\eta+1}{2}\right)}{\sqrt{\pi(\eta-2)}\Gamma\left(\frac{\eta}{2}\right)}$.

To illustrate the superior modeling power of KCPs over GPs, we consider the predictive distribution generated by a GP and a GCTM KCP in Fig. 2. This experiment is a simple example to illustrate the power and flexibility of copula processes simply by changing the marginal distributions from Gaussian to Student's t. Recall a GP is a special case of the GCTM KCP as the degree of freedom parameter $\upsilon \to \infty$. The two particular types of KCPs were chosen to illustrate what a slight departure from GPs framework can achieve.

As depicted, the predictive distribution produced by GCTM KCP is much closer to the true posterior distribution of the AR(2) process than that produced by GP. GP, restricted by its symmetric nature, must over shift its mean to the positive side to compensate for the excess probability mass in that region. This would introduce even greater prediction errors if point predictions are considered. GCTM KCP on the other hand is able to produce an asymmetric predictive distribution based on observed data to much better match the true distribution.

4 Real-Life Application

4.1 Data

To showcase the modeling power of KCPs with multiple time-series, we consider the daily maximum temperatures recorded by the United States Historical Climatology Network (HCN) [**?**]. Data from as far back as the early 1800s from a network of 1219 weather stations are freely downloadable from the Network's website. This study will consider the data from a few arbitrary three-year periods from 156 stations of various states. Missing data is commonplace in such data as equipment is moved and maintained; however, KCPs handle missing data naturally.

The latitude/longtitude and elevation coordinates for each weather station are also known. We will later use this information to help learn the dependency structure.

To focus on the stochastic nature of the data, we first subtract a sinusoidal seasonal trend from that data. A customized sinusoidal function is used for each weather station based on the least-square error criteria. An example of the temperature data with the fitted sinusoidal trend is given in Fig. 3a with the detrended version in Fig. 3b.

4.2 Model

Since there exists no natural ordering among the weather stations, we will take advantage of the modularity of the KCPs and first learn a univariate KCP for each weather station and connect them with a binding copula.

Upon closer examination of the detrended data, some notable features are observed which helped narrow down the modeling choices.

First, the second moment autocorrelation function (ACF), i.e. the ACF of the square of the data, experiences an annual cycle as illustrated in Fig. 3c. This implies the rate of fluctuation, or volatility, of maximum temperature goes through

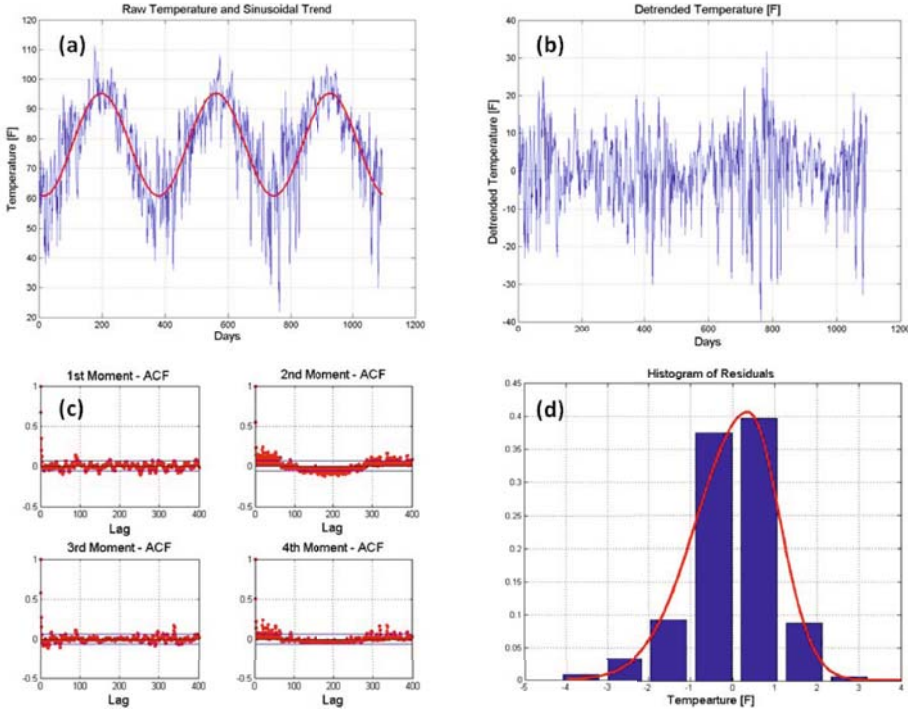

Fig. 3. Weather station TX410120 of Texas. Panel (a) depicts the raw maximum temperature data and the sinusoidal trend; Panel (b) shows the detrended data; Panel (c) shows the auto-correlation function of the first four moments of the detrended data; Panel (d) shows the empirical density and the estimated skew-normal distribution. The second moment ACF in (c) and the detrended temperature in (b) clearly shows the periodic volatility of the residuals.

an annual cycle. On the other hand, the first moment ACF drops off relatively quickly (in a few days time), but this also implies recent temperature trends tend to persist. This provides an important hint in designing or selecting the type of kernel function that would be appropriate for the analysis. In this case, we designed a heteroskedastic kernel function that satisfies the above features:

$$k(t_1, t_2) = \sigma^2 \cdot e^{-\kappa \Delta t} \left[\frac{1}{2\kappa} + \frac{\alpha}{4\kappa^2 + T^{-2}} \left(2\kappa \cdot \sin\left(\frac{\min(t_1, t_2)}{T} + \phi \right) - \frac{1}{T} \cos\left(\frac{\min(t_1, t_2)}{T} + \phi \right) \right) \right] \tag{17}$$

where σ is roughly the average volatility, κ is the inverse of length scale of dependency in days, T is the period ($= 1$ year) and ϕ is the phase of volatility. Please see the Appendix for the derivation. Note the kernel function is non-stationary as the volatility of maximum temperature depends on the day of the year.

Another helpful observed feature in the detrended data, is that a skewed-normal fits the empirical histogram well. Figure 3d shows a sample empirical histogram and the best-fit skew-normal distribution. Although the histograms are not strictly valid[6], it helps us narrow down the choice of marginal distributions.

It is important to note that the above analysis is not necessary for using KCPs. However, it is a simple way to incorporate additional domain knowledge by providing the initial values of the parameters[7] for MLE and as a basis for forming the prior distribution for Bayesian learning.

For this study, five competing models for the univariate time-series were investigated. These are (i) GP, (ii) Gaussian copula with skew-normal marginal (GCSM) KCP, (iii) t-copula with Gaussian marginal (TCGM) KCP, (iv) t-copula with skew-normal marginal (TCSM) KCP, and (v) GARCH(1,1) (with t-innovations) model. The GARCH(1,1) model is included here because it is one of the best-known models for heteroskedastic processes [2][10]. All KCPs above use the heteroskedastic kernel function (17).

The Gaussian- and t-copula are used as a binding copula to describe the interdependencies among the temperature series. In this case the Gaussian RBF kernel (7) is used, with the weighted distance metric:

$$d(i,j)^2 = d_{\texttt{lat}}^2 + d_{\texttt{long}}^2 + w \cdot d_{\texttt{elev}}^2 \tag{18}$$

where $d(i,j)$ is the weighted distance between weather station i and j, d_x are distances along latitude, longitude, and elevation, and w is the weight on the distance along elevation.

4.3 Results

First we examine the results from the univariate KCPs. Figure 4 provides a sample comparison of the detrended data from weather station MD181750 of Maryland. The first column of panels from left to right show the time series, the corresponding ACF of the second moment, the histograms, and the q-q plots. The first row shows the plots of the detrended data while the subsequent rows show the samples generated from the corresponding models after learning on the same set of the detrended data. All KCP models work relatively well in the sense that the annual periodic change in volatilities was recovered as evident in both the time-domain plots and the ACF plots. The GARCH(1,1) model picked up some periodicity in

[6] Histograms are conventionally used to visually evaluate the probability density of a random variables. A random process on the other hand is a collection of random variables. Thus a sample path of a time-series (which is a random process), represents a series of draws from a potentially different distribution. Therefore, collapsing all the samples from a time-series to form a histogram is axiomatically incorrect.

[7] The skewness parameter can be obtained by the maximum likelihood fit of histograms of the residuals as depicted in 3d and the length scale parameter κ in the kernel function (17) can be estimated by $\hat{\kappa} = -\log(\rho|_{\tau=1})$ where $\rho(\tau)$ denotes the ACF for the first moment at lag τ.

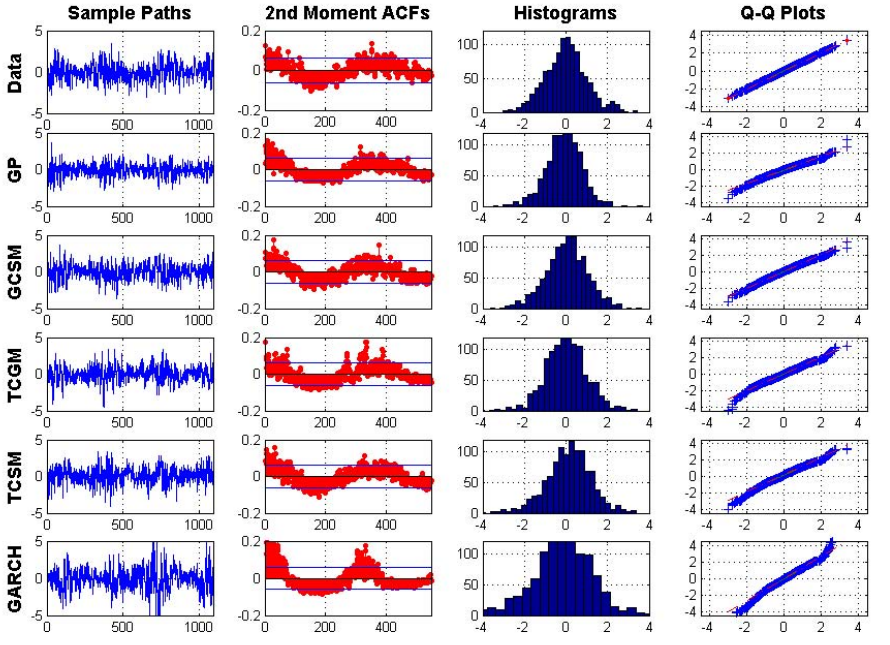

Fig. 4. Comparing stylized facts of the detrended data and sample paths generated by models with MLE parameters based on the weather station MD181750 of Maryland

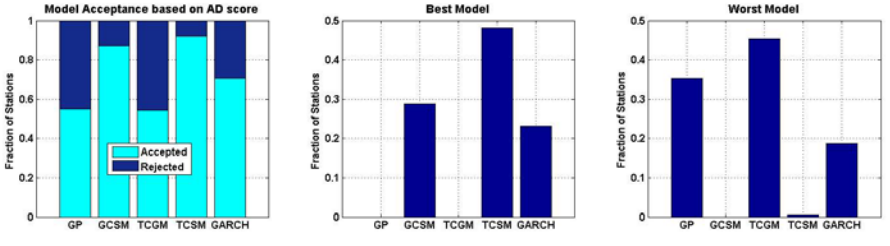

Fig. 5. Univariate model selection results. The left panel shows the fraction of model acceptance/rejection based on the Anderson-Darling criterion; the middle and right panels show the fractions of each models ranked to be the most and the least appropriate model for the univariate temperature series according to the achieved log-likelihood.

volatility as well, but the period was incorrect. Examining the histograms and the q-q plots visually, the TCSM KCP seems the best match with the data.

More generally, using the data from 156 weather stations across the US, the Anderson-Darling test rejects the hypothesis that GP and TCGM KCP are sufficiently good models (with 95% confidence level) in $\sim$43% and $\sim$42% of the cases; whereas, the TCSM KCP was shown to be the best model as it was rejected in only $\sim$7.5% of the cases. The acceptance and rejection frequencies for

Table 1. Medians of pair-wise likelihood ratios among different models and the corresponding critical values for the likelihood ratio tests at 95% confidence level. Note that only the upper triangular part of this table is shown as the table is anti-symmetric.

	p_{GP}	p_{GCSM}	p_{TCGM}	p_{TCSM}	p_{GARCH}
q_{GP}	-	5.41 (3.84) dB	0.0 (3.84) dB	6.05 (5.99) dB	4.6 (5.99) dB
q_{GCSM}	-	-	-5.6 (0.00) dB	0.0 (3.84) dB	-4.6 (3.84) dB
q_{TCGM}	-	-	-	5.8 (3.84) dB	5.1 (3.84) dB
q_{TCSM}	-	-	-	-	-5.0 (0.00) dB
q_{GARCH}	-	-	-	-	-

all models are shown in the left panel of Fig. 5. The middle and the right panels of Fig. 5 show the frequency of each model being the *best* and the *worst* model for each data set as ranked by the maximized likelihoods. The GCSM and TCSM KCPs clearly dominate the GP and TCGM KCP, demonstrating that most of the modeling power (for this data set) comes from the skewness of the marginal distributions while the extra tail-dependency from the t-copula provides an incremental gain. The performance of the GARCH(1,1) model with t-innovations seems to be hit-and-miss with this data set, yet it is quite impressive that it ranks first in about 22% of the time, even though the KCPs use a kernel function that is custom designed for this data set. However, it is important to note that only the maximized likelihoods are used to rank models here. Model complexity (i.e. number of parameters in each model) is not considered. The likelihood ratio test that follows will address this issue by using a significance threshold which depends on the number of model parameters.

Table 1 lists the *median* pair-wise likelihood ratios computed across the data from 156 weather stations for each model-pair, where $\frac{p_x}{q_y}$ denotes the likelihood ratios of model x over model y. The corresponding critical values for the likelihood ratio tests at a 95% confidence level are also listed in parentheses. Note, if a particular pair-wise likelihood ratio $\frac{p_x}{q_y}$ is greater than the corresponding critical value, then model x is preferred over model y with statistical significance. The number of parameters used in each model is also taken into account when computing the critical values, thus balancing the desire for performance and the aversion to model complexity. Here, the likelihood ratios again show the superiority of KCPs over GPs and GARCH(1,1) by allowing different combinations of copulas and marginal distributions for the specific applications. In particular, both GCSM and TCSM are better than GP, TCGM, and GARCH(1,1) with statistical significance. Further, the likelihood ratios for $\frac{p_{TCGM}}{q_{GP}}$ and $\frac{p_{TCSM}}{q_{GCSM}}$ confirm that the t-copula provides statistically insignificant performance gain over the Gaussian copula for this data set. Thus, on average, the GCSM KCP is the most powerful and yet parsimonious model for this application.

Now we shift our attention to the effectiveness of the binding copula in capturing the dependencies of multiple time-series for weather stations located in the same state. Figure 6 shows the Gram matrices of binding copulas and pair-wise correlation matrices of the detrended data after transforming by the distribution induced by the univariate KCPs that was ranked to be the best earlier. Recall

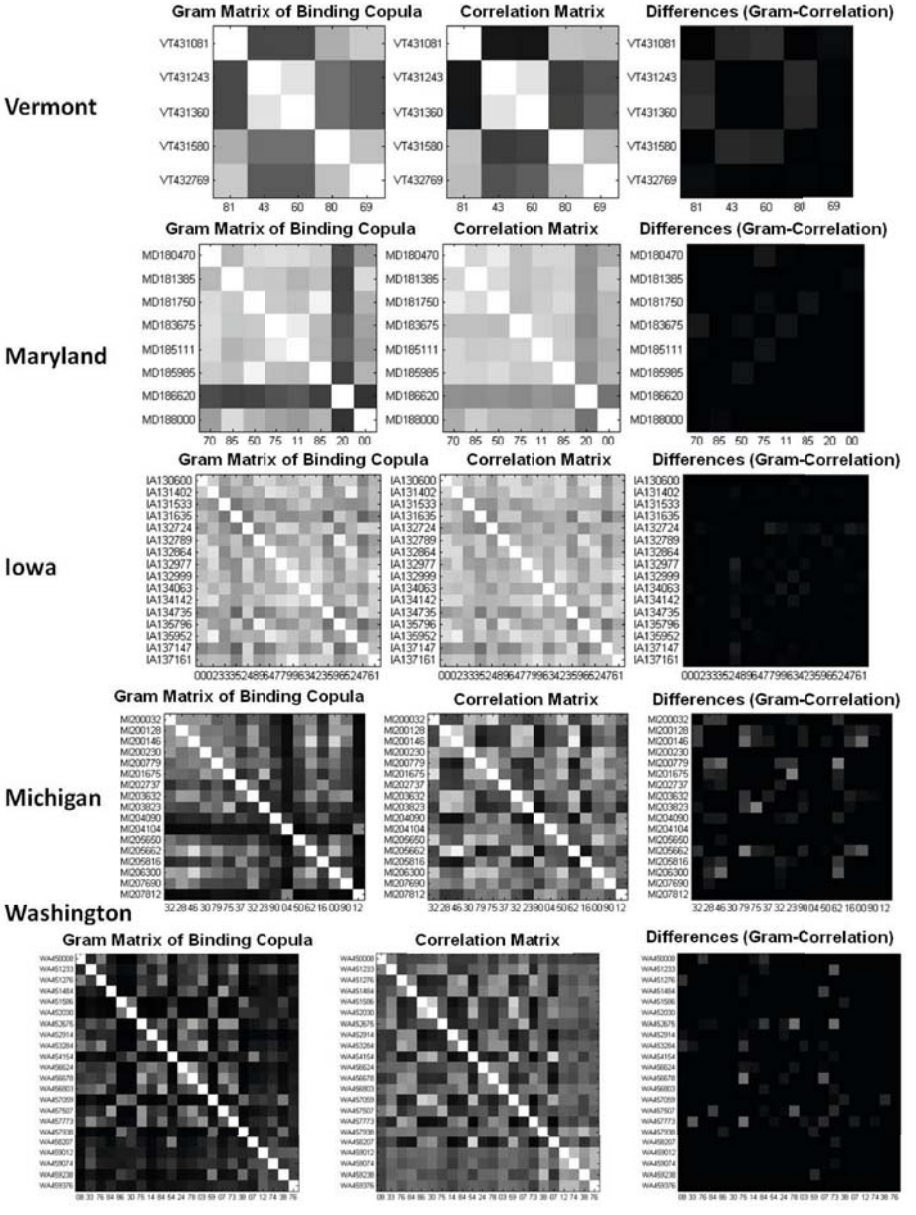

Fig. 6. Sample multivariate KCP results from a few states. The Gram matrix (left) of the binding copula, pairwise correlation (middle), and the difference of the two matrices (right). The IDs of the weather stations are labeled on the vertical axes while only the last two digits of the IDs are shown on the horizontal axes.

that for non-Gaussian random variables, the Gram matrix is not identical to the covariance matrix, but rather it dictates the behavior of the covariance. Nevertheless, the simple t-copula with only geographical distance information was able to recover a substantial amount of interdependency on a visual basis.

5 Conclusion and Future Work

This paper introduced the kernel-based copula processes (KCPs), a novel time-series analysis tool, which inherits features from both the theory of copula and Gaussian processes. The KCPs allows the modeling of non-Gaussian processes in a concise and parsimonious manner by separating the specification of the dependency structure and the margins.

An extension to multivariate time-series was also presented. A binding copula was used to encapsulate the interdependency among time series. This modular approach not only lends itself naturally to implementation on a multi-core / -processor for computational efficiency, it also allows time-series with different lengths and sampling frequencies to be described seamlessly under a single model.

Further research directions include possible sampling or variational methods to accommodate very large data set; the incorporation of a latent process for the kernel hyper-parameters to model processes with stochastic variances; and the application of KCPs in classification problems in similar capacity as GPs.

References

1. Anderson, T.W., Darling, D.A.: A test of goodness of fit. The Journal of American Statistical Association 49(268), 765–769 (1954)
2. Bollerslev, T.: Generalized autoregressive conditional heteroskedasticity. Journal of Econometrics, 31
3. Williams, C.K.I., Rasmussen, C.E.: Gaussian Processes for Machine Learning. The MIT Press, Cambridge (2006)
4. Carney, M., Cunningham, P.: Evaluating density forecast models. tcd-cs-2006- 21. Database, Trinity College Dublin, Department of Computer Science (2006)
5. Chen, X., Fan, Y.: Estimation of copula-based semiparameteric time series models. Journal of Econometrics 130(2), 307–335 (2006)
6. Sandberg, D.V., Alvarado, E., Stewart, G.: Modeling large forest fires as extreme events. Northwest Science, 72
7. Gunther, T.A., Diebold, F.X., Tay, A.S.: Evaluating density forecasts with applications to financial risk management. International Economic Review 39(4), 863–883 (1998)
8. Rosso, R., De Michele, C., Salvadori, G., Kottegoda, N.T.: Extremes in nature: an approach using Copulas. Springer, New York (2007)
9. Hansen, B.E.: Autoregressive conditional density estimation. International Economic Review 35(3), 705–730 (1994)
10. Hull, J.C.: Options, Futures, and Other Derivatives, 5th edn. Prentice-Hall, New Jersey (2003)
11. Lee, T.-H., Long, X.: Copula-based multivariate garch model with uncorrelated dependent errors. Database, UCR Working Paper 2005-16 (2005)

12. Mandelbrot, B.: The variation of certain speculative prices. Journal of Business, 26
13. Nelsen, R.B.: An Introduction to Copulas. Springer Series in Statistics. Springer-Verlag New York, Inc, New York (2006)
14. Fermanian, J.-D., Doukhan, P., Lang, G.: Copulas of a vector-valued stationary weakly dependent process. Database, Working paper CREST (2004)
15. Patton, A.J.: Modelling asymmetric exchange rate dependence. International Economic Review 47(2), 527–556 (2006)
16. Olsen, E.T., Darsow, W.F., Nguyen, B.: Copulas and markov processes. Illinois Journal of Mathematics, pp. 600–642 (1992)
17. Jr. M.J. Menne R.S. Vose Williams, C.N. and D.R. Easterling. United states historical climatology network daily temperature, precipitation, and snow data. Database, The Carbon Dioxide Information Analysis Center, Oak Ridge National Laboratory, U.S. Department of Energy (2006)

Appendix: Heteroskedastic Kernel Function

To derive the heteroskedastic kernel function, consider a random process (or time-series) X_t decomposed into a deterministic trend g_t and a stochastic component with heteroskedastic variance, i.e. $X_t = g_t + y_t$. Assume y_t follows a common Vasicek mean-reverting process:

$$dy_t = -\kappa \cdot (\theta - y_t) \cdot dt + \sigma_t \cdot dW_t \tag{19}$$

where κ is the rate of mean-reversion, θ is the mean-reversion level, σ_u is the variance process of y_t, and W_t is a Wiener process in a usual SDE setup. This generalized Vasicek process has solution:

$$y_t = y_0 \cdot e^{-\kappa t} + \theta \cdot (1 - e^{-\kappa t}) + \int_0^t \sigma_u e^{-\kappa(t-u)} dW_u \tag{20}$$

Thus the auto-covariance of X_t is given by

$$\begin{aligned} Cov[X_{t_1}, X_{t_2}] &= \mathbb{E}\left[\left(\int_0^{t_1} \sigma_u e^{-\kappa(t_1-u)} dW_u\right)\left(\int_0^{t_2} \sigma_u e^{-\kappa(t_2-u)} dW_u\right)\right] \\ &= e^{-\kappa(t_1+t_2)} \int_0^{\min(t_1,t_2)} \sigma_u^2 e^{-2\kappa u} du \end{aligned} \tag{21}$$

Notably, that this kernel function is not a function of the difference between t_1 and t_2 and is therefore non-stationary. Now we must specify the form of the variance process. For the detrended temperature data, the variance process is periodic, consequently, we assume it takes on the following form

$$\sigma_t^2 = \sigma^2 \cdot \left(\alpha \cdot \sin\left(\frac{t}{T} + \phi\right)\right) \tag{22}$$

where σ, α, T, and ϕ are all constants. Substituting (22) into (21), simplifying gives the final form of the kernel function found in (17).