Dynamic Copula Networks for Modeling Real-valued Time Series

Elad Eban
The Hebrew University
School of Computer Science

Gideon Rothschild
UCSF Center for
Integrative Neuroscience

Adi Mizrahi
The Hebrew University
Institute of Life Sciences

Israel Nelken
The Hebrew University
Institute of Life Sciences

Gal Elidan
The Hebrew University
Department of Statistics

Abstract

Probabilistic modeling of temporal phenomena is of central importance in a variety of fields ranging from neuroscience to economics to speech recognition. While the task has received extensive attention in recent decades, learning temporal models for multivariate real-valued data that is non-Gaussian is still a formidable challenge. Recently, the power of copulas, a framework for representing complex multi-modal and heavy-tailed distributions, was fused with the formalism of Bayesian networks to allow for flexible modeling of high-dimensional distributions. In this work we introduce Dynamic Copula Bayesian Networks (DCBNs), a generalization aimed at capturing the distribution of rich temporal sequences. We apply our model to three markedly different real-life domains and demonstrate substantial quantitative and qualitative advantages.

1 Introduction

Probabilistic modeling of temporal phenomena is of great interest in diverse fields ranging from computational biology to economics to speech recognition. A central challenge in such modeling is the daunting dimension of the data. For example, even a simple EEG signal may include many thousands of time points. The sequential nature of such phenomena, however, often allows us to make realistic simplifying assumptions. First, a periodic behavior is often observed. Second, reasonable Markovian assumptions allow us to construct models over *entire* sequences via local building blocks that are limited to a finite horizon. Using these assumptions, general purpose tools such as Markov chains, Hidden Markov models, and Kalman filters are widely used with great success in a profusion of temporal applications.

Dynamic Bayesian networks (DBNs) [Dean and Kanazawa, 1989] are a temporal extension of the widely used framework of Bayesian networks [Pearl, 1988]. Like all probabilistic graphical models, DBNs rely on local building blocks that capture the close range behavior of random variables (both within each time "slice" and across time). Importantly, the framework generalizes many of the commonly used temporal constructs. Yet, despite wide empirical success, DBNs are susceptible to critical limitations. Most notably, while DBNs are in principle applicable to real-valued domains, practical considerations when tackling such scenarios almost always force us to use simple parametric building blocks. In fact, the overwhelming majority of continuous DBNs rely on a simple Gaussian representation.

Obviously, the Gaussian assumption, while mathematically convenient, is often unrealistic. Neural activity levels, for example, are characterized by noisy and lengthy "rest" periods and rare extreme events (spikes), resulting in a heavy-tailed distribution; Financial daily changes measurements are often characterized by normally distributed stable periods intermingled with economic upheavals whose erratic behavior is far from Gaussian; Sensor measurements of human activity (e.g. walking, running) often has a highly peaked behavior due the nature of the activity performed and/or our physical limitations. Our goal is to model such complex non-Gaussian phenomena.

To cope with continuous non-Gaussian challenges in a non-temporal context, Elidan [2010] recently proposed the Copula BN model (CBN) that fuses the frameworks of the statistical copula and BNs. Briefly, copulas allows us to model complex real-valued distributions by separating the choice of the (possibly nonparametric) univariate marginals and the dependence function that "couples" them into a coherent

Appearing in Proceedings of the 16^{th} International Conference on Artificial Intelligence and Statistics (AISTATS) 2013, Scottsdale, AZ, USA. Volume 31 of JMLR: W&CP 31.

joint distribution. Using the language of probabilistic graphical models, the CBN model extends the idea to high-dimension and in practice leads to substantial quantitative and qualitative gains (see [Elidan, 2010] for more details).

In this work we extend CBNs to work in the temporal domain in the same way that DBNs extend non-temporal BNs, and formulate the Dynamic Copula Bayesian Network (DCBN) model. Although the extension is straightforward theoretically, its practical merits are substantial. In particular, with little computational effort, the model allows us to capture complex multivariate temporal phenomena whose behavior is quite far from that of a multivariate Gaussian distribution.

We apply our model to the three markedly different temporal settings discussed above: neural activity levels, EMG physical action measurements, and financial daily return data. We demonstrate a significant quantitative advantage in all cases in terms of the quality of the multivariate density learned, relative to temporal and non-temporal Gaussian baselines. In the action dataset, where a discriminative task may also be of interest, we also show a marked improvement in terms of predictive ability. Finally, for the neural activity data, we also provide a qualitative evaluation that enhances the ability of our model to faithfully capture real-life complex phenomena.

The rest of the paper is organized as follows: following a description of the necessary background material in Section 2, in Section 3 we describe the DCBN temporal model and briefly explain how the model is automatically learned from data. We demonstrate the practical merit of the model in Section 4, and end with concluding remarks in Section 5.

2 Background

In this section we briefly review the framework of copulas and the recently introduced Copula BN model [Elidan, 2010]. Let $\mathcal{X} = \{X_1, \ldots, X_N\}$ be a finite set of scalar real-valued random variables and let $F_{\mathcal{X}}(\mathbf{x}) \equiv P(X_1 \leq x_1, \ldots, X_n \leq x_N)$ be a (cumulative) distribution over $\mathcal{X}$, with lower case letters denoting assignment to variables. For compactness, we use $F_i(x_i) \equiv F_{X_i}(x_i) = P(X_i \leq x_i, X_{\mathcal{X}/X_i} = \infty)$, and for density functions we similarly use $f_i(x_i) \equiv f_{X_i}(x_i)$. When there is no ambiguity we sometimes abuse notation and use $F(x_i) \equiv F_{X_i}(x_i)$, and similarly for densities and for sets of variables.

2.1 Copulas

A copula function [Sklar, 1959] links marginal distributions to form a multivariate one. Formally,

Definition 2.1 *Let $U_1, \ldots, U_N$ be real random variables marginally uniformly distributed on $[0, 1]$. A copula function $C : [0, 1]^N \to [0, 1]$ is a joint distribution*

$$C_\theta(u_1, \ldots, u_N) = P(U_1 \leq u_1, \ldots, U_N \leq u_N),$$

where θ are the parameters of the copula function.

Sklar's seminal theorem states that *any* joint distribution $F_{\mathcal{X}}(\mathbf{x})$ can be represented as a copula function C of its univariate marginals

$$F_{\mathcal{X}}(\mathbf{x}) = C(F_1(x_1), \ldots, F_N(x_N)).$$

When the univariate marginals are continuous, C is uniquely defined. The constructive converse, which is of central interest from a modeling perspective, is also true: *any* copula function taking *any* marginal distributions $\{F_i(x_i)\}$ as its arguments, defines a valid joint distribution with marginals $\{F_i(x_i)\}$. Thus, copulas are "distribution generating" functions that allow us to separate the choice of the univariate marginals and that of the dependence structure, encoded in the copula function C. Importantly, this flexibility often results in a construction that is beneficial in practice.

Assuming C has Nth order partial derivatives (true almost everywhere when continuous), the joint density can be derived from the copula function using the derivative chain rule

$$\begin{aligned} f(\mathbf{x}) &= \frac{\partial^N C(F_1(x_1), \ldots, F_N(x_N))}{\partial F_1(x_1) \ldots \partial F_N(x_N)} \prod_i f_i(x_i) \\ &\equiv c_\theta(F_1(x_1), \ldots, F_N(x_N)) \prod_i f_i(x_i), \quad (1) \end{aligned}$$

where $c(\cdot)$ is called the *copula density*.

Example 2.1: The Gaussian copula is undoubtedly the most commonly used copula family with applications ranging from mainstream financial risk assessment to climatology applications. Its distribution is defined as

$$C_\Sigma(\{U_i\}) = \Phi_\Sigma\left(\Phi^{-1}(U_1), \ldots, \Phi^{-1}(U_N)\right), \quad (2)$$

where Σ is a correlation matrix, Φ is the standard normal distribution, and Φ_Σ is a zero mean normal distribution with correlation matrix Σ.

Figure 1 exemplifies the flexibility that comes with this seemingly limited elliptical copula family. Shown are samples from this copula using two different marginals. As can be seen, a variety of markedly different and multi-modal distributions can be constructed. Generally, and without any added computational difficulty, we can mix and match *any* marginals with *any* copula function to form a valid joint distribution. ∎

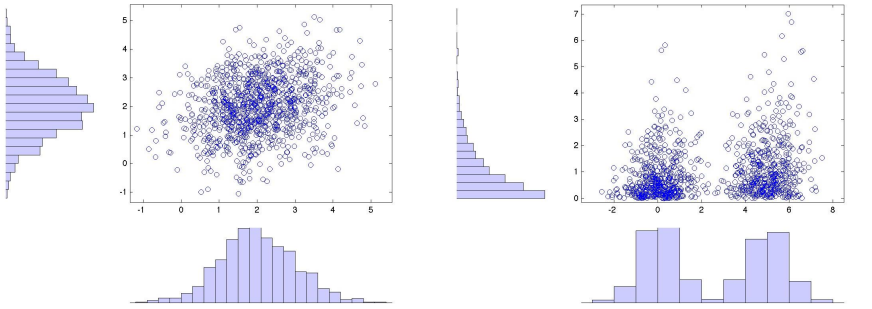


Figure 1: Samples from the bivariate Gaussian copula with correlation $\theta = 0.25$. (left) with unit variance Gaussian marginals; (right) with a mixture of Gaussian and Gamma marginals.

Normal marginals

Gaussian Mix and Gamma marginals

2.2 Bayesian networks and Copula Bayesian Networks

Let $\mathcal{G}$ be a directed acyclic graph (DAG) whose nodes correspond to the set of random variables $\mathcal{X} = \{X_1, \ldots, X_N\}$, and let $\mathbf{Pa}_i = \{\mathbf{Pa}_{i1}, \ldots, \mathbf{Pa}_{ik_i}\}$ be the parents of X_i in $\mathcal{G}$. A Bayesian network (BN) [Pearl, 1988] is used to represent a joint density over $\mathcal{X}$ using a qualitative graph structure and quantitative parameters that define local conditional densities. The graph $\mathcal{G}$ encodes the independence statements $I(\mathcal{G}) = \{(X_i \perp \mathit{NonDesc}_i \mid \mathbf{Pa}_i)\}$, where $\perp$ denotes the independence relationship, and $\mathit{NonDesc}_i$ are nodes that are not descendants of X_i in $\mathcal{G}$. It is easy to show that if $I(\mathcal{G})$ hold, then the joint density decomposes into a product of local conditional densities $f_{\mathcal{X}}(\mathbf{x}) = \prod_i f_i(X_i \mid \mathbf{Pa}_i)$. Conversely, any such product of local conditional densities defines a valid joint density where $I(\mathcal{G})$ hold.

We now describe the recently introduced CBN model that fuses the BN and copula formalisms:

Definition 2.2: A Copula Bayesian Network (CBN) is a triplet $\mathcal{C} = (\mathcal{G}, \Theta_C, \Theta_f)$ that defines $f_{\mathcal{X}}(\mathbf{x})$. $\mathcal{G}$ encodes the independencies $(X_i \perp \mathit{NonDesc}_i \mid \mathbf{Pa}_i)$, assumed to hold in $f_{\mathcal{X}}(\mathbf{x})$. Θ_C is a set of local copula functions $C_i(F(x_i), F(\mathbf{pa}_{i1}), \ldots, F(\mathbf{pa}_{ik_i}))$ that are associated with the nodes of $\mathcal{G}$ that have at least one parent. In addition, Θ_f is the set of parameters representing the marginal densities $f_i(x_i)$ (and distributions $F_i(x_i)$). The joint density $f_{\mathcal{X}}(\mathbf{x})$ is defined as

$$f_{\mathcal{X}}(\mathbf{x}) = \prod_{i=1}^{N} R_{c_i}\left(F(x_i), F(\mathbf{pa}_{i1}), \ldots, F(\mathbf{pa}_{ik_i})\right) f_i(x_i),$$

where, if X_i has at least one parent in the graph $\mathcal{G}$, the term $R_{c_i}\left(F(x_i), F(\mathbf{pa}_{i1}), \ldots, F(\mathbf{pa}_{ik_i})\right)$ is defined as

$$R_{c_i}(\cdot) \equiv \frac{c_i(F(x_i), F(\mathbf{pa}_{i1}), \ldots, F(\mathbf{pa}_{ik_i}))}{\frac{\partial^K C_i(1, F(\mathbf{pa}_{i1}), \ldots, F(\mathbf{pa}_{ik_i}))}{\partial F(\mathbf{pa}_{i1}) \ldots \partial F(\mathbf{pa}_{ik_i})}}.$$

When X_i has no parents in $\mathcal{G}$, $R_{c_i}(\cdot) \equiv 1$. ∎

The term $R_{c_i}\left(F(x_i), F(\mathbf{pa}_{i1}), \ldots, F(\mathbf{pa}_{ik_i})\right) f_i(x_i)$ is always a valid conditional density, namely $f(x_i \mid \mathbf{pa}_i)$, and can be easily computed. In particular, when the copula density $c(\cdot)$ has an explicit form, so does $R_{c_i}(\cdot)$ since it involves derivatives of a lesser order.

Elidan [2010] showed that a CBN defines a valid joint density so that, like other graphical models, a CBN takes advantage of the independence assumptions to represent $f_{\mathcal{X}}(\mathbf{x})$ compactly. Differently, as in copulas, when the graph is tree structured, the univariate marginals of a CBN are exactly $f_i(x_i)$. For more general structures, the marginals can be skewed, though only slightly so in practice. Empirically, CBNs offers significant advantages in terms of generalization ability (see Elidan [2010] for more details).

3 Dynamic Copula Bayesian Network

A dynamic Bayesian network (DBN) [Dean and Kanazawa, 1989] is a powerful probabilistic model that generalizes BNs, allowing us to represent and reason about the dynamics of complex structured distributions. Briefly, using a standard temporal Markovian assumption, a joint distribution is defined via a *template* Bayesian network that encodes the structured probability of variables $\mathcal{X}^t$ at time t given other variables at time t, as well as those of the preceding time point $t-1$. Like standard BNs, the representation is quite general in that it relies on a black-box representation of each variable given its parents in the model. See Figure 2 for an illustration.

However, when the domain is real-valued, computational considerations make the use of complex local representation infeasible. As noted, the overwhelming majority of continuous dynamic Bayesian networks rely on the linear Gaussian parameterization. Our goal is to overcome this limitation by building on the power of copulas. To do so, we generalize the recently introduced CBN model that fuses the copula and BN formalisms, in the same way that DBNs generalize BNs to capture temporal dynamics.

3.1 The DCBN Model

Let $\mathcal{X}$ be a set of *template* random variables and let $\mathcal{X}^{(0:T)} = [\mathcal{X}^{(0)}, X^{(1)} \ldots, \mathcal{X}^{(T)}]$ be a replication of this set for the different times $0, \ldots, T$. Like a DBN, a DCBN is used to represent a joint distribution over $\mathcal{X}^{(0:T)}$. Differently than a DBN, the representation relies on a copula-based building block. Formally,

Definition 3.1 *A Dynamic Copula Bayesian Network (DCBN) is a 5-tuple* $\mathcal{DC} = (\mathcal{G}_0, \Theta_{C_0}, \mathcal{G}_{\rightarrow}, \Theta_{C_{\rightarrow}}, \Theta_f)$*:*

- $\mathcal{G}_0$ *is a DAG that encodes independencies over* $\mathcal{X}^{(0)}$ *(as described in Section 2.2)*
- Θ_{C_0} *are local copula densities* $c_i(F(x_i), \{F(\mathbf{pa}_{ik}\})$ *defined over the random variables* $X_i \in \mathcal{X}^{(0)}$ *and their parents in* $\mathcal{G}_0$*.*
- $\mathcal{G}_{\rightarrow}$ *is a DAG defined over a two-replication set of random variables* $\mathcal{X}'$ *and* $\mathcal{X}''$*.* $\mathcal{G}_{\rightarrow}$ *is constrained so that no (backward) edges from variables in* $\mathcal{X}''$ *to variables in* $\mathcal{X}'$ *are allowed.*
- $\Theta_{C_{\rightarrow}}$ *are a set of local copula densities defined over variables* $X_i \in \mathcal{X}''$ *only and their parents in* $\mathcal{G}_{\rightarrow}$*, which may be in* $\mathcal{X}''$ *as well as in* $\mathcal{X}'$*.*
- Θ_f *are the parameters describing the univariate marginal distributions (and densities) of all the template random variables* $\mathcal{X}$*.*

Given a temporal sequence $\mathbf{x}^{(0:T)}$*, the joint density of the entire temporal sequence is then defined as*

$$f_{\mathcal{X}^{(0:T)}}(\mathbf{x}^{(0:T)}) = f_{\mathcal{X}^{(0)}}(\mathbf{x}^{(0)}) \prod_{t=0}^{T-1} f_{\mathcal{X}^{(t:t+1)}}(\mathbf{x}^{(t+1)}|\mathbf{x}^{(t)}),$$

where the initial density *is defined as:*

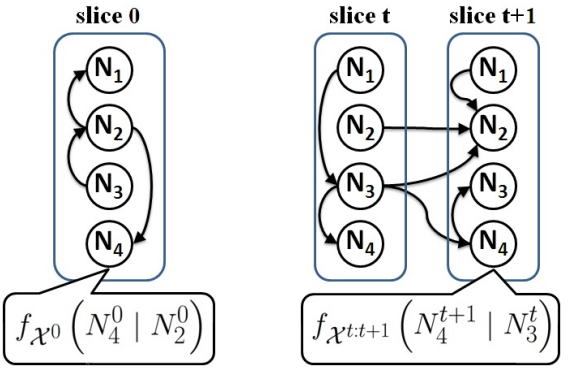
$$f_{\mathcal{X}^{(0)}}(\mathbf{x}^{(0)}) \equiv \prod_{i \in \mathcal{X}} R^{(0)}_{c_i}(F(x_i), \{F(\mathbf{Pa}_i{}^{\mathcal{G}_0})\}) f_i(x_i),$$

and the copula ratio $R^{(0)}_{c_i}$ *is as defined in Section 2.2 with the parameters* θ_{C_0} *and the assignment* $\mathbf{x}^{(0)}$*. The* transition density *is defined as*

$$f_{\mathcal{X}^{(t:t+1)}}(\mathbf{x}^{(t+1)}|\mathbf{x}^{(t)}) \equiv \prod_{i \in \mathcal{X}} R_{c_i}(F(x_i^{(t)}), \{F(\mathbf{Pa}_i{}^{\mathcal{G}_{\rightarrow}})\}) f_i(x_i^{(t)}),$$

where R_{c_i} *is defined as above but with the parameterization defined by* $\theta_{C_{\rightarrow}}$ *that is* not *dependent on* t*.*

Thus, the DCBN model generalizes the CBN model to the temporal domain in that it allows us to separate the choice of the univariate marginals and that of the dependency structure:



(a) Representation of $\mathcal{G}_0$ (b) Representation of $\mathcal{G}_{\rightarrow}$

Figure 2: An illustration of a dynamic structured probabilistic graphical model (DBN or DCBN)

Lemma 3.1 : *Let* $\mathcal{X}$ *be a set of real-valued random variables. Given any positive univariate marginal densities* θ_f*, and any set of local copula functions* θ_{C_0} *and* $\theta_{C_{\rightarrow}}$*, the DCBN model defined above represents a valid joint density over* $\mathcal{X}^{(0:T)}$*.*

The proof is similar to the proof of Elidan [2010] for the CBN model. The modeling flexibility that the construction offers allows us to learn powerful structured temporal models. In particular, as we demonstrate in Section 4, the explicit control over the univariate marginals that the framework offers leads to quantitative and qualitative advantages.

3.2 Learning the Copula Parameters

As is standard for probabilistic graphical models, the decomposable form of the joint density defined by the DCBN model facilitates relatively efficient structure learning and estimation using standard machinery.

Given a complete training data set $\mathcal{D}$ of M sequences of length T_m each, the log-likelihood of the DCBN model $\mathcal{DC}$ can be written as

$$\ell(\mathcal{D} : \mathcal{DC}) \quad = \quad \sum_{m=1}^{M} \log \left(f_{\mathcal{X}_m^{(0:T_m)}}(\mathbf{x}^{(0:T_m)}[m]) \right) \quad (3)$$

where $\mathbf{x}^{(0:T_m)}[m]$ is used to denote the assignments to the variables in the mth sequence. As in the case of DBNs, Eq. (3) can (after some rearrangement of terms), be written as a decomposable sum of the families (variables and their parents) in $\mathcal{G}_0$ and $\mathcal{G}_{\rightarrow}$.

The only complication, from an estimation perspective, is that univariate marginals $\{F_i(x_i)\}$ are shared across all terms, hindering our ability to perform decomposable estimation. To overcome this difficulty, we adopt the common solution in the copula community

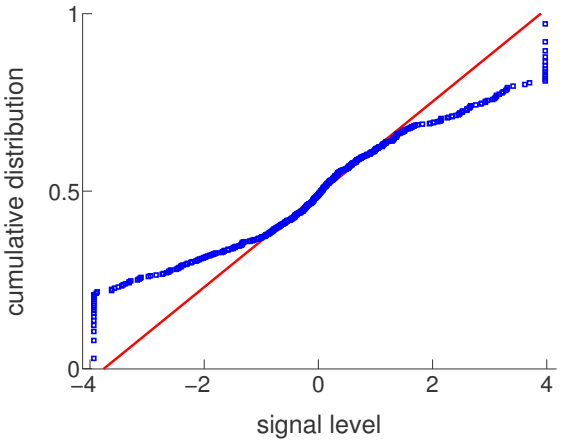


Figure 3: The (typical) cumulative distribution of one of the **EMG** sensor measurements. The solid red lines shows the best Gaussian fit.

(with appealing theoretical and practical properties) of estimating the marginals first [Joe and Xu, 1996]. Given the marginals, the parameters of each local copula can be estimated independently of the others, using a closed form where possible (e.g., for the Gaussian copula) or, for example, a conjugate gradient procedure. In both cases, estimation is carried out by concatenating statistics across all time slices. See Koller and Friedman [2009] for details on how this is carried out in the context of standard DBNs.

3.3 Structure Learning

Very briefly, to learn the structure of both $\mathcal{G}_0$ and $\mathcal{G}_{\rightarrow}$, we use a standard greedy search that applies local structure modifications (e.g., add/delete/reverse edge) based on a model selection score. In this work we use the Bayesian Information Criterion (BIC) of Schwarz [1978] that, like other scores, balances the likelihood and the complexity of the model:

$$\text{score}(\mathcal{G} \ : \ \mathcal{D}) = \ell(\mathcal{D} : \hat{\theta}, \mathcal{G}) - 0.5 \log(K) |\Theta_{\mathcal{G}}|, \quad (4)$$

where $\hat{\theta}$ are the maximum-likelihood parameters, K is the number of samples and $|\Theta_{\mathcal{G}}|$ is the number of free parameters associated with the graph structure $\mathcal{G}$.

When learning a DCBN (and similarly for a standard DBNs), the decomposition of the likelihood allows us to use the BIC score to learn $\mathcal{G}_0$ and $\mathcal{G}_{\rightarrow}$ separately. When learning $\mathcal{G}_0$, K is the number of sequences , and when learning $\mathcal{G}_{\rightarrow}$, K is the overall number of transitions in all sequences. To cope with local maxima during the search we also use a TABU list and random restarts [Glover and Laguna, 1993]. See Koller and Friedman [2009] for details on this standard learning procedure for temporal models.

3.4 Univariate Marginal Estimation

The central strength of the copula-based representation is that it allows us to separate the choice of the univariate marginal distribution and that of the dependency copula functions. Thus, we can use nonparametric marginal estimation which, in the univariate case, is typically extremely accurate and robust. Perhaps the most common choice is to use kernel density estimation with a Gaussian kernel (see, for example, [Bowman and Azzalini, 1997]).

Unfortunately, in some cases the Gaussian kernel may not be sufficiently powerful. For example, the distribution of the EMG activity measurements in **EMG** dataset that we use in our experimental evaluation is quite heavy-tailed as can be seen in Figure 3. A standard Gaussian (solid red line) is clearly a bad fit for such a distribution. Obviously, nonparametric kernel estimation with a Gaussian kernel is significantly more accurate than a naive Gaussian fit. However, it is still exponentially concentrated around the training data and, as confirmed in preliminary experiments, has poor generalization performance.

Instead, since in choosing the marginal distribution we are not constrained in any way, we use a simpler histogram representation that is robust to outliers. Concretely, let $\{x_i\}_1^n$ be a given set of training samples of the random variable X_i and use $L = \min\{x_i\}$ and $R = \max\{x_i\}$ to denote the minimum and maximum values, respectively. We partition the interval $[L, R]$ into K equal bins of width $w = \frac{R-L}{K}$. To account for outliers in the test data, we add a bin at $[L-w, L]$ and one at $[R, R+w]$, and assign a single pseudo sample to each. Using N_j to denote the number of samples that falls in the interval that corresponds to the jth bin, the density in that interval is then set to $(N_j)/(n+2)$, and zero elsewhere.

4 Experimental Evaluation

In this section we demonstrate the merit of the **DCBN** model for capturing real-valued temporal phenomena in three markedly different real-life scenarios.

In principle, we can use any family of local copulas in the model and even mix different copula families without significant computational difficulty. However, for simplicity, and to emphasize the generic power of the **DCBN** model, in all the experiments below we instantiate the model with the simple and commonly used Gaussian copula defined in Eq. (2). To model the univariate marginal distribution of each variable, we use the nonparametric histogram described in Section 3.4, using 50 bins. We compare our **DCBN** model to two baselines: a temporal model where a full multivariate

Gaussian is used to model the conditional probability of *all* neurons at time t, given *all* neurons at time $t - \Delta$ (**tMVG**); A standard dynamic Bayesian network (**DBN**) with a linear Gaussian representation, so that each variable is modeled as a normal distribution centered around a linear combination of its parents with a variable dependent variance $X_i|X_{Par_i} \sim N(\beta_0 + \beta X_{Par_i}, \sigma_i)$. The structure of both the **DBN** model and our **DCBN** model was learned using the same greedy structure search procedure guided by the Bayesian Information Criterion [Schwarz, 1978] score, as described in Section 3.3.

We consider the following datasets:

- **Neural**. Calcium concentration (a proxy for neural activity) of neurons of a mice auditory cortex loaded with calcium fluoresces indicator. The dataset includes measurements of 5-17 neurons in 13 recording sessions each with 24,00-57,000 time points. In each experiment, we corrected a DC drift in the signal by subtracting from each sequence a linear term. Specifically, for a sequence $s(t) : t \in \{1 \dots T\}$, denote by $\hat{s}$ the least square linear approximation of s. The output of our preprocessing is simply $s_{out}(t) = s(t) - \hat{s}(t)$. Finally, due to the density of the signal, modeling neural activity at time t given time $t - 1$ is a trivial problem. Thus, for all models during *both* train and test time, we use models where $\mathbf{X}(t)$ is conditioned on $\mathbf{X}(t - \Delta)$. We use $\Delta = \frac{1}{2}$ seconds which is sensible biologically (results were similar for other values, see supplementary results).
- **EMG.** The EMG physical action dataset from the UCI repository. This data consists of EMG recordings of four subjects while performing ten normal and ten aggressive activities. An EMG apparatus with skin-surface electrodes placed on the subjects' limbs was used to record $\sim 10,000$ measurements in eight locations. Since different recordings can be of completely different scales (e.g., of different subjects), the measurements of each variable in each recording was centered and its variance was normalized to 1.
- **DOW.** Daily changes of the stocks comprising the Dow Jones 30 index. The data includes 1508 daily adjusted changes aver a period of five years (2001-2005). To avoid arbitrary imputation, two stocks not included as part of the index in all of these days were excluded (KFT,TRV).

For all datasets, when splitting the data into train and test instances, we maintain temporal coherence: For the **Neural** and **DOW** datasets, the entire sequence of measurements was split into five *consecutive* segments of equal length. In each fold, one segment is held out for testing while the others are used for training. In the **EMG** domain, the natural division is by the subject carrying out the activity so that in each experiment we train on the activity of three subjects and test on that of the fourth held-out subject.

4.1 Quantitative Likelihood Evaluation

We start with a quantitative evaluation of the different models on all three datasets in terms of log-probability (log-loss) per instance performance on held out test data. Figure 4 (left) show the performance of the models on the **DOW** dataset as a function of the maximal number of parents allowed for each variable during structure search. As a further point of reference, we also shows the performance of a non-temporal multivariate Gaussian model (**MVG**). The benefit from temporal modeling, and the consistent advantage of our copula-based **DCBN** model over the **DBN** model across the entire range is evident. Note that an advantage of 3 bits, for example, translate into each test instance being $2^3 = 8$ times more likely, so that the advantages shown are substantial.

The superiority our **DCBN** over the **tMVG** model is particularly striking when we consider the complexities of the two models. For N stocks, the **tMVG** model requires $\frac{2N(2N-1)}{2} + 2N$ parameters. In contrast, the **DCBN** model requires at most $2 \times N \times (K + 1)$ parameters, where K is the maximal number of parents allowed. Interestingly, our **DCBN** model is superior to **tMVG** even when $K = 0$, and the advantage grows substantially as more parents are allowed. While this may sound surprising at first, the explanation is quite simple. Intuitively, the nonparametric marginals provide an extremely accurate estimate of the univariate distribution thus boosting the ability of the model to capture the overall joint distribution. At the same time, the parameters of the univariate marginals do not take place in the joint learning process. This is particularly appealing since learning can be carried out both efficiently and robustly.

Figure 4 (left) is typical and is qualitatively similar for *all* 13 experiments of the **Neural** dataset and 18/20 of the **EMG** dataset experiments (see supplementary material). To get a broad view of the performance for these datasets across experiments, Figure 4 (center and right) compare the performance of our **DCBN** model to that of a standard **DBN** across all experiments and all model complexities (number of maximal parents allows). To put all experiments on approximately the same scale, the units of both axes are in bits/instance improvement over the independence (no parent) Gaussian model. Impressively, in the over-

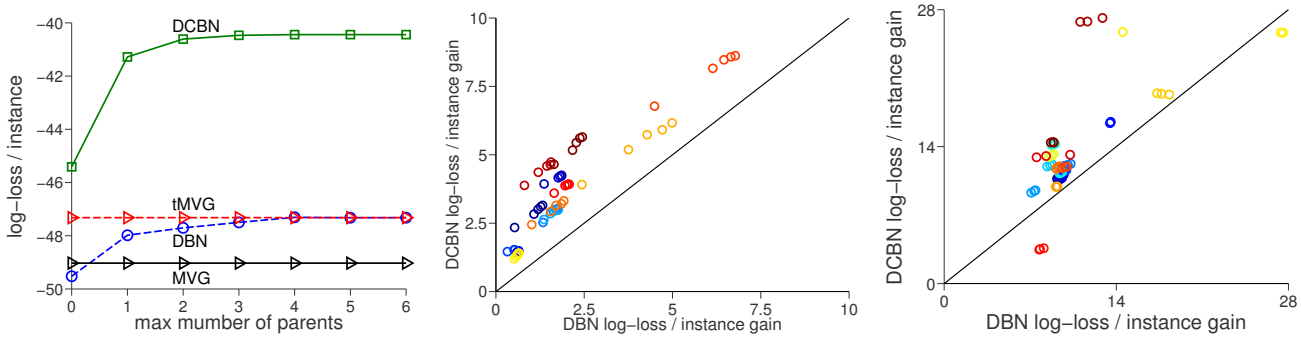


Figure 4: (left) **DOW** dataset average test log-loss per instance of the different models over 5 folds (y-axis) as a function of the maximal number of parents allowed in the model. Compared are our copula-based **DCBN** model, a standard linear Gaussian **DBN** model and a full multivariate Gaussian model (**tMVG**). Also shown is a non-temporal **MVG** model. The structure of both the **DBN** and **DCBN** models was learned using the same greedy procedure and the BIC score. (center) **Neural** and (right) **EMG** average log-loss per instance improvement relative to the full independent Gaussian model. Each point corresponds to a specific experiment and a specific bound on the number of parents in the network. The colors correspond to different experiments.

whelming majority of cases our model leads to substantial gains in test set performance. Specifically, our model is superior in all of the 65 **Neural** experiments and 55/60 of **EMG** experiments.

4.2 Qualitative Neuronal Assessment

As exemplified in Figure 1, the strength of the copula representation is that it can faithfully capture a complex distribution. Thus, aside from the quantitative advantage reported above, we would expect a copula-based model to better capture the underlying qualitative phenomenon. We now demonstrate that this is the case for the **Neural** data. Concretely, in contrast to a standard **DBN**, we show that our model allows us to "identify" the biologically meaningful spiking events.

We start by evaluating the likelihood of the models in the physiologically relevant periods around spiking events. To do so, we use the procedure described in Rothschild et al. [2010] to identify spike-evoked transients. Since our model is defined over *all* neurons at each time point, we consider firing of any of the neurons to be an *event* of interest. We can then measure performance (relative to the independent Gaussian model), as a function of the temporal distance from the event. A typical result is shown Figure 6, where the advantage of our **DCBN** model over the **DBN** model is evident (similar results for the other 12 neuronal experiments are provided in the supplementary material).

Importantly, our **DCBN** model "identifies" the salient regions, although this information was not available at training time, and offers the greatest advantage in the vicinity of the event. The explanation is simple: near an event, the activity of neurons is more correlated and the advantage over an independent model manifests more clearly. In contrast, the **DBN** model is not able to capture the more complex distribution and, in many cases, degrades in performance in that region. This should not come as a surprise given that fact that the **DBN** model has Gaussian marginals that cannot capture heavy-tailed behavior.

Our results also shed light on an ongoing biological debate [Schneidman et al., 2006, Ganmor et al., 2011]. It was suggested [Schneidman et al., 2006] that pairwise interactions dominate neuronal network dynamics and higher-order interactions account for only 10% of the overall information. We can use our DCBN model to evaluate the contribution of higher degree interactions in terms of test likelihood performance. The result is summarized in Figure 5. As expected, the contribution of the first parent in the network (corresponding to pairwise interactions) is the largest and accounts for approximately 65% of the information gain. At the same time, the third and forth order interactions alone contributing close to 30% of the gain. This suggests that higher-order interactions play an important role in neuronal activity. We leave further biological investigation of these findings to future work.

4.3 EMG - Discriminative Assessment

The EMG dataset involves a variety of different activities, and naturally lends itself to the discriminative task of activity recognition. Given a test sequence, we simply predict the activity to be the one that corresponds to the model whose posterior probability is

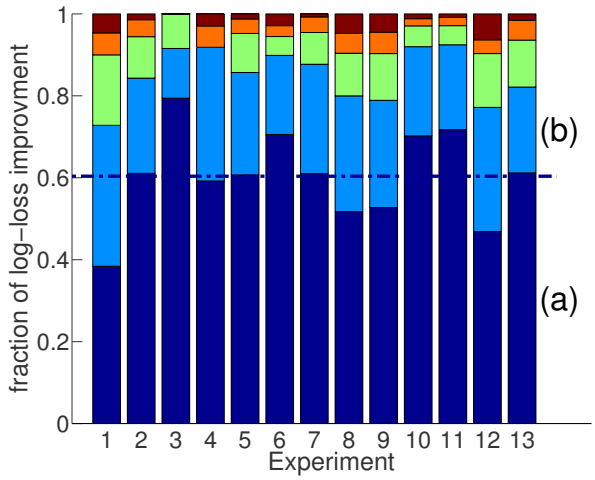


Figure 5: Log-loss/instance improvement of our **DCBN** model as a fraction of the information gain of a 5-parent model relative to the independence baseline. The lower bars (a) show the improvement when allowing one parent for each variable, corresponding to pairwise dependencies. The second layer (b) correspond to the *additional* improvement when allowing up to two parents, and so on. The dotted line shows the first level average across experiments.

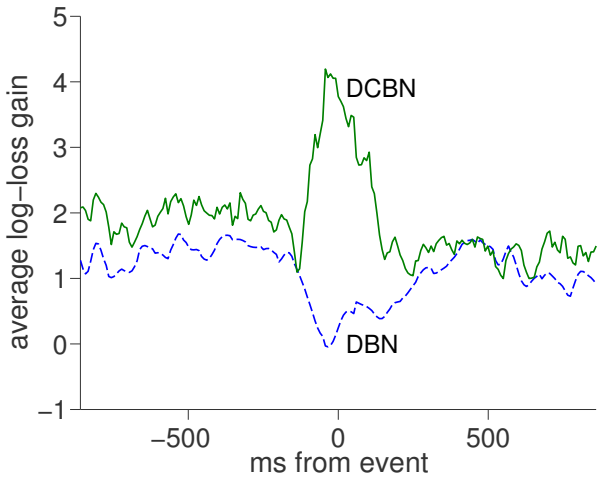


Figure 6: Average test set improvement in log-loss/instance relative to the independent Gaussian model (y-axis) as a function of the distance from a spike event (x-axis) for the **Neural** dataset. Shown is the improvement of the **DCBN** and **DBN** models with up to 4 parents per variable

maximized:

$$\hat{\alpha} = arg\max_{\alpha} \Pr(M_\alpha|seq) \propto \Pr(seq|M_\alpha),$$

where M_α is the model learned for the activity α. We similarly use the temporal linear Gaussian **DBN** as a generative baseline. We also compare to a discriminative baseline. Since there are only 4 recordings for every activity, we can only consider a simple kNN classifier (K=3), applied to first and second moments of the variables (we also tried kNN with a Fourier basis representation of the sequences with worse results).

When attempting to identify one of twenty activities, the performance of all methods is comparable and unimpressive and averages around 30% (with a small advantage to our **DCBN** model). In this case, a model with no parents achieves similar performance. Intuitively, the similarity between some activities is simply too high for greater differences between the models to manifest. Indeed, the experimental setting in which the data was collected was described as including 10 aggressive and 10 normal activities. When we consider this coarser-grained binary classification problem, the differences between the models are more evident: while the nearest neighbor and the **DBN** baseline are able to achieve a solid 85% classification accuracy, our **DCBN** model offers an impressive 5% improvement in predictive accuracy.

5 Conclusions and Future Work

In this work we introduced Dynamic Copula Bayesian Networks (DCBN), a model aimed at coping with the challenge of modeling the joint temporal behavior of multiple real-valued variables. Using the statistical framework of copulas, our model allows us to separately choose the univariate marginal representation and the joint dependence function, thus generalizing the recently introduced copula BN model. Importantly, our model offers great flexibility in accurately capturing real-valued non-Gaussian temporal phenomena.

We applied our generic construction to three markedly different real-life domains. In all cases, our generic model offers consistent and significant quantitative advantages, even when using only straightforward univariate marginal histograms and the simple Gaussian copula. We also demonstrated the merit of the model in terms of the ability to qualitatively capture physically meaningful aspects of the domain.

More generally, the DCBN model allows for the mix-and-match of any univariate marginal representation and an expressive range of copula families without any computational difficulty. Obviously, and in contrast to temporal modeling using standard DBNs, this amounts to substantial modeling flexibility. Thus, our model opens the door for numerous novel applications in a variety of complex domains where, for computational reasons, only relatively simple (e.g., Gaussian) representations were explored.

Acknowledgements

The research was supported by the Gatsby Charitable Foundation, Intel Cooperation, and the Israel Science Foundation. The authors thank Yair Weiss, Uri Heinemann, Daniel Zoran, and Ofer Meshi for useful comments in various stages of the preparation of this paper.

References

A. Bowman and A. Azzalini. *Applied Smoothing Techniques for Data Analysis*. Oxford University Press, 1997.

T. Dean and K. Kanazawa. A model for reasoning about persistence and causation. *Computational Intelligence*, 5:142–150, 1989.

G. Elidan. Copula bayesian networks. *Advances in Neural Information Processing Systems*, 24, 2010.

E. Ganmor, R. Segev, and E. Schneidman. The architecture of functional interaction networks in the retina. *The Journal of Neuroscience*, 31(8): 3044–3054, 2011.

F. Glover and M. Laguna. Tabu search. In C. Reeves, editor, *Modern Heuristic Techniques for Combinatorial Problems*, Oxford, England, 1993. Blackwell Scientific Publishing.

H. Joe and J. Xu. The estimation method of inference functions for margins for multivariate models. Technical Report 166, Department of Statistics, University of British Columbia, 1996.

D. Koller and N. Friedman. *Probabilistic Graphical Models: Principles and Techniques*. The MIT Press, 2009.

J. Pearl. *Probabilistic Reasoning in Intelligent Systems*. Morgan Kauffman, 1988.

G. Rothschild, I. Nelken, and A. Mizrahi. Functional organization and population dynamics in the mouse primary auditory cortex. *Nature neuroscience*, 13 (3):353–360, 2010.

E. Schneidman, M. Berry, R. Segev, and W. Bialek. Weak pairwise correlations imply strongly correlated network states in a neural population. *Nature*, 440(7087):1007–1012, 2006.

G. Schwarz. Estimating the dimension of a model. *Annals of Statistics*, 6:461–464, 1978.

A. Sklar. Fonctions de repartition a n dimensions et leurs marges. *Publications de l'Institut de Statistique de L'Universite de Paris*, 8:229–231, 1959.