A new nonparametric model of maximum-entropy (MaxEnt) copula density function is proposed, which offers the following advantages: (i) it is valid for mixed random vector. By ‘mixed’, we mean the method works for any combination of discrete or continuous variables in a fully automated manner; (ii) it yields a bonafide density estimate with interpretable parameters. By ‘bonafide’, we mean the estimate guarantees to be a non-negative function, integrates to 1; and (iii) it plays a unifying role in our understanding of a large class of statistical methods for mixed \((X, Y)\). Our approach utilises modern machinery of nonparametric statistics to represent and approximate log-copula density function via LP-Fourier transform. Several real-data examples are also provided to explore the key theoretical and practical implications of the theory.

1. Copula statistical learning

Copulas are the ‘bridge’ between the univariate and the multivariate statistics world, with applications in a wide variety of science and engineering fields – from economics to finance to marketing to healthcare. Because of the ubiquity of copula in empirical research, it is becoming necessary to develop a general theory that can unify and simplify the copula learning process. In this paper, we present a new class specially-designed nonparametric maximum-entropy (MaxEnt) copula model that offers the following advantages: first, it yields a bonafide (smooth, non-negative, and integrates to 1) copula density estimate with interpretable parameters that provide insights into the nature of the dependence between the random variables \((X, Y)\). Second, the method is data-type agnostic – which is to say that it automatically (self) adapts to mixed-data types (any combination of discrete, continuous or even categorical). Third, and most notably, our copula-theoretic framework subsumes and unifies a wide range of statistical learning methods using a common mathematical notation – unlocking deep, surprising connections and insights, which were previously unknown. In the development of our theory and algorithms, the LP-Fourier method of copula modelling (which was initiated by Mukhopadhyay and Parzen 2020) plays an indispensable role.
2. Self-adaptive nonparametric models

We introduce two new classes of maximum-entropy (MaxEnt) copula density models. But before diving into technical details, it will be instructive to review some basic definitions and concepts related to copula.

2.1. Background concepts and notation

2.1.1. Sklar’s Copula representation theory (Sklar 1959)

The joint cumulative distribution function (cdf) of any pair of random variables \((X, Y)\)

\[
F_{X,Y}(x, y) = \Pr(X \leq x, Y \leq y), \quad \text{for } (x, y) \in \mathcal{R}^2
\]
can be decomposed as a function of the marginal cdfs \(F_X\) and \(F_Y\)

\[
F_{X,Y}(x, y) = \text{Cop}_{X,Y} \left( F_X(x), F_Y(y) \right), \quad \text{for } (x, y) \in \mathcal{R}^2
\]

where \(\text{Cop}_{X,Y}\) denotes a copula distribution function with uniform marginals. To set the stage, we start with the continuous marginals case, which will be generalised later to allow mixed- \((X, Y)\). Taking derivative of Equation (1), we get

\[
f_{X,Y}(x, y) = f_X(x)f_Y(y) \text{cop}_{X,Y} \left( F_X(x), F_Y(y) \right), \quad \text{for } (x, y) \in \mathcal{R}^2
\]

which decouples the joint density into the marginals and the copula. One can rewrite Equation (2) to represent copula as a ‘normalised’ joint density function

\[
\text{cop}_{X,Y} \left( F_X(x), F_Y(y) \right) := \text{dep}_{X,Y} (x, y) = \frac{f_{X,Y}(x, y)}{f_X(x)f_Y(y)},
\]

which is also known as the dependence function, pioneered by Hoeffding (1940). To make (3) a proper density function (i.e. one that integrates to one) we perform quantile transformation by substituting \(F_X(x) = u\) and \(F_Y(y) = v\):

\[
\text{cop}_{X,Y}(u, v) = \text{dep}_{X,Y}(F_X^{-1}(u), F_Y^{-1}(v)) = \frac{f_{X,Y}(F_X^{-1}(u), F_Y^{-1}(v))}{f_X(F_X^{-1}(u))f_Y(F_Y^{-1}(v))}, \quad (u, v) \in [0, 1]^2.
\]

We are now ready to extend this copula density concept to the mixed \((X, Y)\) case.

2.1.2. Pre-Copula: conditional comparison density (Parzen and Mukhopadhyay 2013)

Before we introduce the generalised copula density, we need to introduce a new concept – conditional comparison density (CCD). For a continuous \(X\), CCD is defined as

\[
d(u; X | Y = y) = \frac{f_{X|Y}(F_X^{-1}(u) | y)}{f_X(F_X^{-1}(u))}, \quad 0 < u < 1.
\]

For \(Y\) discrete, we represent it using probability mass function (pmf):

\[
d(v; Y | X = x) = \frac{p_{Y|X}(Q_Y(v) | x)}{p_Y(Q_Y(v))} = \frac{\Pr(Y = Q_Y(v) | X = x)}{\Pr(Y = Q_Y(v))}, \quad 0 < v < 1.
\]
where \( Q_Y(v) \) is the quantile function of \( Y \). It is easy to see that the CCDs (4) and (5) are proper densities in the sense that
\[
\int_0^1 d(u; X, X \mid Y = y) \, du = \int_0^1 d(v; Y, Y \mid X = x) \, dv = 1.
\]

### 2.1.3. Generalised Copula representation theory (Mukhopadhyay and Parzen 2020)

For the mixed case, when \( Y \) is discrete and \( X \) is continuous the joint density of (3) is defined by either side of the following identity:
\[
\Pr(Y \mid X = x) f_X(x) = f_{X \mid Y}(x \mid y) \Pr(Y = y).
\]
This can be rewritten as the ratios of conditionals and their respective marginals:
\[
\text{Pre-Bayes' Rule} : \frac{\Pr(Y = y \mid X = x)}{\Pr(Y = y)} = \frac{f_{X \mid Y}(x \mid y)}{f_X(x)}.
\] (6)

This formula (6) can be interpreted as the slices of the mixed copula density, since
\[
\text{cop}_{X,Y} \left( F_X(x), F_Y(y) \right) = \frac{f_{X \mid Y}(x \mid y)}{f_X(x)} = \frac{\Pr(Y = y \mid X = x)}{\Pr(Y = y)}.
\] (7)

Substituting \( F_X(x) = u \) and \( F_Y(y) = v \), we get the following definition of the generalised copula density in terms of conditional comparison density (CCD):
\[
\text{cop}_{X,Y}(u, v) = d(u; X, X \mid Y = Q(v; Y)) = d(v; Y, Y \mid X = Q(u; X)), \quad 0 < u, v < 1.
\] (8)

Bayes’ theorem ensures the equality of two CCDs, with copula being the common value. Equipped with this fundamentals, we now develop the nonparametric theory of MaxEnt copula modelling.

### 2.2. Log-bilinear model

An exponential Fourier series representation of copula density function is given. The reasons for entertaining an exponential model for copula is motivated from two different perspectives.

#### 2.2.1. The problem of unboundedness

One peculiar aspect of copula density function is that it can be unbounded at the corners of the unit square. In fact, many common parametric copula families – Gaussian, Clayton, Gumbel, etc. – tend to infinity at the boundaries. So naturally the question arises: How to develop suitable approximation methods that can accommodate a broader class of copula density shapes, including the unbounded ones? The first key insight: logarithm of the copula density function is far more convenient to approximate (due to its well-behaved nature) than the original density itself. We thus express the logarithm of copula density \( \log \text{cop}_{X,Y} \) in the Fourier series – instead of doing canonical \( L_2 \) approximation, which expands \( \text{cop}_{X,Y} \) directly in an orthogonal series (Mukhopadhyay and Parzen 2020). Accordingly, for densities with rapidly changing tails, 'log-Fourier' method leads to an improved estimate that
is less wiggly and more parsimonious than the $L_2$-orthogonal series model. In addition, the resulting exponential form guarantees the non-negativity of the estimated density function.

2.2.1.1. Choice of Orthonormal Basis. To expand log-copula density function, we choose the LP-family of polynomials (see Appendix A.1), which are especially suited to approximate functions of mixed random variables. In particular, we approximate log cop$_{X,Y}$ by expanding it in the tensor-product of LP-bases $\{S_j \otimes S_k\}$, which are orthonormal with respect to the empirical-product measure $\{\tilde{F}_X \otimes \tilde{F}_Y\}$. LP-bases’ appeal lies in its ability to approximate the quirky shapes of mixed-copula functions in a completely automated way, see Figure 1. Consequently, it provides a unified way to develop nonparametric smoothing algorithms that simultaneously hold for mixed data types.

Definition 2.1: The exponential copula model admits the following LP-expansion:

$$\text{cop}_\theta(u, v; X, Y) = \frac{1}{Z_\theta} \exp \left\{ \sum_{j,k>0} \theta_{jk} S_j(u; X) S_k(v; Y) \right\},$$

where $Z_\theta$ is the normalisation factor that ensures cop$_\theta$ is a proper density

$$Z_\theta = \int_0^1 \int_0^1 \exp \left\{ \sum_{j,k>0} \theta_{jk} S_j(u; X) S_k(v; Y) \right\} \, du \, dv.$$

We refer (9) as the log-bilinear copula model.

2.2.2. The maximum entropy principle

Another justification for choosing the exponential model comes from the principle of maximum entropy (MaxEnt), pioneered by E. T. Jaynes (1957). The maxent principle defines a unique probability distribution by maximising the entropy $H(\text{cop}) = -\int \text{cop}_{X,Y} \log \text{cop}_{X,Y}$ under the normalisation constraint $\int \text{cop}_{X,Y} = 1$ and the following LP-co-moment conditions:

$$\mathbb{E}_{\text{cop}_\theta} [S_j(U; X)S_k(V; Y)] = \text{LP}_{jk}.$$  (10)

LP-co-means are orthogonal ‘moments’ of copula, which can be estimated by

$$\widetilde{\text{LP}}_{jk} = \mathbb{E}_{\text{Cop}} [S_j(U; X)S_k(V; Y)] = \frac{1}{n} \sum_{i=1}^n S_j(\tilde{F}_X(x_i); X) S_k(\tilde{F}_Y(y_i); Y).$$  (11)

Applying calculus of variations, one can show that the maxent constrained optimisation problem leads to the exponential (9) form. The usefulness of Jaynes’ maximum entropy principle lies in providing a constructive mechanism to uniquely identify a probability distribution that is maximally non-committal (flattest possible) with regard to all unspecified information beyond the given constraints.
Figure 1. MaxEnt LP-copula density estimates for (a) kidney fitness (age vs tot: both continuous marginals), (b) PLOS data (title length vs number of authors: both discrete marginals), (c) horseshoe crab data (number of satellites vs width: mixed discrete-continuous marginals) and (d) challenger space shuttle data (temperature vs number of O-ring failures: mixed continuous-discrete marginals).

2.2.3. Estimation

We fit a truncated exponential series estimator of copula density

$$\text{cop}_\theta(u, v; X, Y) = \frac{1}{Z_\theta} \exp \left\{ \sum_{j=1}^{m_1} \sum_{k=1}^{m_2} \theta_{jk} S_j(u; X) S_k(v; Y) \right\}.$$ 

The task of finding the maximum likelihood estimates (MLE) of $\theta$ boils down to solving the following sets of equations for $j = 1, \ldots, m_1$ and $k = 1, \ldots, m_2$:

$$\frac{\partial \log \text{cop}_\theta}{\partial \theta_{jk}} = \frac{\partial \log Z_\theta}{\partial \theta_{jk}} - \frac{1}{n} \sum_{i=1}^{n} S_j(\tilde{F}_X(x_i); X) S_k(\tilde{F}_Y(y_i); Y) = 0. \quad (12)$$
Note that the derivative of the log-partition function is equal to the expectation of the LP-co-mean functions:

$$\frac{\partial \log Z_\theta}{\partial \theta_{jk}} = \mathbb{E}_{\text{cop}_\theta}[S_j(U; X)S_k(V; Y)].$$ (13)

Replacing (13) and (11) into (12) implies that the MLE of MaxEnt model is same as the method of moment satisfying the following moment conditions:

$$\int \int_{[0,1]^2} S_j(u; X)S_k(v; Y) \text{cop}_\theta(u, v; X, Y) \, du \, dv = \hat{\Lambda} P_{jk}.$$ 

At this point, one can apply any convex optimisation routine (e.g. Newton’s method, gradient descent, stochastic gradient descent, etc.) to solve for $\hat{\theta}$.

2.2.4. Asymptotic

Let the sequence of $m_1$ and $m_2$ increase with sample size with an appropriate rate $\frac{(m_1,m_2)^3}{n} \to 0$ as $n \to \infty$. Then, under certain suitable regularity conditions, the exponential cop is a consistent estimate in the sense of Kullback–Leibler distance; see Barron and Sheu (1991) for more details.

2.2.5. Determining informative constraints

Jayne’s maximum entropy principle assumes that a proper set of constraints (i.e. sufficient statistic functions) are given to the modeller, one that captures the phenomena under study. This assumption may be legitimate for studying thermodynamic experiments in statistical mechanics or for specifying prior distribution in Bayesian analysis, but certainly not for building empirical models.

Which comes first: a parametric model or sufficient statistics? After all, the identification of significant components (sufficient statistics) is a prerequisite for constructing a legitimate probability model from the data (Mukhopadhyay, Parzen, and Lahiri 2012). Therefore the question of how to judiciously design and select the constraints from data seems inescapable for nonparametrically learning maxent copula density function from data; also see Appendix A.3, which discusses the ‘two cultures’ of maxent modelling. We address this issue as follows: (i) compute $\hat{\Lambda} P_{jk}$ using the formula equation (11); (ii) sort them in descending order based on their magnitude (absolute value); (iii) compute the penalised ordered sum of squares

$$\text{PenSum}(q) = \text{Sum of squares of top } q \text{ LP comeans} - \frac{\gamma_n}{n} q.$$ 

For AIC penalty choose $\gamma_n = 2$, for BIC choose $\gamma_n = \log n$, etc. Further details can be found in Mukhopadhyay and Parzen (2020, Section 4.3). (iv) Find the $q$ that maximises the PenSum$(q)$. Store the selected indices $(j, k)$ in the set $\mathcal{I}$. (v) Carry out maxent optimisation routine based only on the selected LP-sufficient statistics-based constraints:

$$\left\{ S_j(u; X)S_k(v; Y) \right\}, \quad (j, k) \in \mathcal{I}.$$ 

This pruning strategy guards against overfitting. Finally, return the estimated reduced-order (with effective dimension $|\mathcal{I}|$) maxent copula model.
Remark 2.1 (Nonparametric MaxEnt): The proposed nonparametric maxent mechanism produces a copula density estimate, which is flexible (can adapt to the ‘shape of the data’ without making risky a-priori assumptions) and yet possesses a compact analytical form.

2.3. Log-linear model

We provide a second parameterisation of copula density.

Definition 2.2: The log-linear orthogonal expansion of LP-copula is given by

$$\text{cop}_\mu(u, v; X, Y) = \exp \left\{ \mu_0 + \sum_{k>0} \mu_k \phi_k(u; X) \psi_k(v; Y) \right\}, \quad 0 < u, v < 1,$$

(14)

We call the parameters of this model ‘log-linear LP-correlations’ that satisfy for $k > 0$

$$\mu_k = \int \int_{[0,1]^2} \phi_k(u; X) \psi_k(v; Y) \log \text{cop}_\mu(u, v; X, Y) \, du \, dv,$$

2.3.1. Connection

Two fundamental representations, namely the log-bilinear (9) and loglinear (14) copula models, share some interesting connections. To see that perform singular value decomposition (SVD) of the $\Theta$-matrix whose $(j,k)$th entry is $\theta_{jk}$:

$$\Theta = U\Omega V^T,$$

$u_{ij}$ and $v_{ij}$ are the elements of the singular vectors with singular values $\mu_1 \geq \mu_2 \geq \cdots \geq 0$. Then the spectral bases can be expressed as the linear combinations of the LP-polynomials:

$$\phi_k(u; X) = \sum_j u_{jk} S_j(u; X)$$

(15)

$$\psi_k(u; Y) = \sum_l v_{lk} S_l(v; Y).$$

(16)

Hence, the LP-spectral functions (15)–(16) satisfy the following orthonormality conditions:

$$\int \phi_k(u; X) \, du = \int \psi_k(v; Y) \, dv = 0$$

$$\int \phi_j(u; X) \psi_k(u; Y) \, du = \delta_{jk}, \quad \text{for } j \neq k.$$

2.4. A few examples

We demonstrate the flexibility of the LP-copula models using real data examples.

Example 2.1 (Kidney fitness data, Efron and Hastie 2016, Section 1.1): It contains measurements on $n = 157$ healthy volunteers (potential donors). For each volunteer, we have
their age (in years) and a composite measure ‘tot’ of overall function of kidney function. To understand the relationship between age and tot, we estimate the copula:

$$\hat{\text{cop}}_{X,Y}(u, v) = \exp \left\{ -0.40S_1(u; X)S_1(v; Y) + 0.18S_2(u; X)S_2(v; Y) - 0.12 \right\},$$
displayed in Figure 1(a). At the global scale, the shape of the copula density indicates a prominent negative ($\hat{\theta}_{11} = -0.40$) association between age and tot. Moreover, at the local scale, significant heterogeneity of the strength of dependence is clearly visible, as captured by the nonlinear asymmetric copula: the correlation between age and tot is quite high for older (say, $> 70$) donors, compared to younger ones. This allows us to gain refined insights into how kidney function declines with age.

**Example 2.2 (PLOS data, both discrete marginals):** It contains information on $n = 878$ journal articles published in PLOS Medicine between 2011 and 2015. For each article, two variables were extracted: length of the title and the number of authors. The dataset is available in the R-package **dobsom**. The checkerboard-shaped estimated discrete copula

$$\hat{\text{cop}}_{X,Y}(u, v) = \exp \left\{ 0.42S_1(u; X)S_1(v; Y) + 0.10S_2(u; X)S_2(v; Y) - 0.07S_2(u; X)S_2(v; Y) - 0.12 \right\},$$
is shown in Figure 1(b), which shows a strong positive nonlinear association. In particular, the sharp lower tail, around the $(0, 0)$, indicates that the smaller values of $(X, Y)$ have a greater tendency to occur together than the larger ones.

**Example 2.3 (Horseshoe Crabs data, mixed marginals):** The study consists of $n = 173$ nesting horseshoe crabs (Agresti 2013). For each female crab in the study, we have its carapace width (cm) and number of male crabs residing nearby her nest. The goal of the study is to investigate whether carapace width affects number of male satellites for the female horseshoe crabs. If so, how—what is the shape of the copula dependence function? The estimated copula, shown in Figure 1(c), is given by

$$\hat{\text{cop}}_{X,Y}(u, v) = \exp \left\{ 0.375S_1(u; X)S_1(v; Y) - 0.077 \right\}.$$ 

This indicates a significant positive linear correlation between the width and number of satellites of a female crab.

**Example 2.4 (1986 Challenger Shuttle O-Ring data):** On January 28, 1986, just after 73 seconds into the flight, Challenger space shuttle broke apart, killing all seven crew members on board. The purpose of this study is to investigate whether the ambient temperature during the launch was related to the damage of shuttle’s O-rings. For that we have $n = 23$ previous shuttle missions data, consisting of launch temperatures (degrees F), and number of damaged O-rings (out of 6). The estimated LP-maxent copula density

$$\hat{\text{cop}}_{X,Y}(u, v) = \exp \left\{ -0.37S_1(u; X)S_1(v; Y) - 0.27S_3(u; X)S_1(v; Y) - 0.14 \right\}$$
is displayed in panel (d) and shows a strong negative association between the temperature at the launch and the number of damaged O-rings. Moreover, the sharp peak of the copula density around the edge $(0, 1)$ further implies that cold temperatures can excessively increase the risk of failure of the o-rings.
3. Applications to statistical modelling

The scope of the general theory of the preceding section goes far beyond simply a tool for nonparametric copula approximation. In this section, we show how one can derive a large class of applied statistical methods in a unified manner by suitably reformulating them in terms of the LP-maxent copula model. In doing so, we also provide statistical interpretations of LP-maxent parameters under different data modelling tasks.

3.1. Goodman’s association model for categorical data

Categorical data analysis will be viewed through the lens of LP-copula modelling. Let \( X \) and \( Y \) denote two discrete categorical variables; \( X \) with \( I \) categories and \( Y \) with \( J \) categories. The data are summarised in an \( I \times J \) contingency table \( \mathbf{F} \); \( f_{kl} \) is the observed cell count in row \( k \) and column \( l \) of \( \mathbf{F} \) and \( n = f_{++} \) is the total frequency. The row and column totals are denoted as \( f_{k+} \) and \( f_{+l} \). The observed joint \( \Pr(X = k, Y = l) \) is denoted by \( \tilde{p}_{kl} = f_{kl}/f_{++} \); the respective row and column marginals are given by \( \tilde{p}_{k+} = f_{k+}/f_{++} \) and \( \tilde{p}_{+l} = f_{+l}/f_{++} \).

3.1.1. LP log-linear model

We specialise our general copula model (14) for two-way contingency tables. The discrete LP-copula for the \( I \times J \) table is given by

\[
\text{cop}(F_X(k), F_Y(l)) = \exp \left( \mu_0 + \sum_{j=1}^{m} \mu_j \phi_{jk} \psi_{jl} \right),
\]

where we abbreviate the row and column scores \( \phi_{j}(F_X(k)) = \phi_{jk} \) and \( \psi_{j}(F_Y(l)) = \psi_{jl} \) for \( k = 1, 2, \ldots, I \) and \( l = 1, 2, \ldots, J \). The number of components \( m \leq M = \min(I - 1, J - 1) \); we call the log-linear model (17) ‘saturated’ (or ‘dense’) when we have \( m = M \) components. The non-increasing sequence of model parameters \( \mu_j \)’s is called ‘intrinsic association parameters’ that satisfy

\[
\mu_j = \sum_{k=1}^{I} \sum_{l=1}^{J} \left( \log p_{kl} \right) p_{k+} p_{+l} \phi_{jk} \psi_{jl}, \quad \text{for} \ j = 1, \ldots, m.
\]

Note that the discrete LP-row and column scores, by design, satisfy (for \( j \neq j' \)):

\[
\sum_{k=1}^{I} \phi_{jk} p_{k+} = \sum_{l=1}^{J} \psi_{jl} p_{+l} = 0 \quad (19)
\]

\[
\sum_{k=1}^{I} \phi_{jk}^2 p_{k+} = \sum_{l=1}^{J} \psi_{jl}^2 p_{+l} = 1 \quad (20)
\]

\[
\sum_{k=1}^{I} \phi_{jk} \phi_{j'k} p_{k+} = \sum_{l=1}^{J} \phi_{jl} \psi_{j'l} p_{+l} = 0. \quad (21)
\]
3.1.2. Interpretation

It is clear from (18) that the parameters $\mu_j$’s are fundamentally different from the standard Pearsonian-type correlation $\text{Cor}(\phi_j, \psi_j) = \mathbb{E}[\phi_j \psi_j]$, due to (19)–(21):

$$\rho_j = \sum_{k=1}^{I} \sum_{l=1}^{J} p_{kl} \phi_{jk} \psi_{jl}, \quad \text{for } j = 1, \ldots, m. \quad (22)$$

The coefficients of the LP-MaxEnt-copula expansion for contingency tables carry a special interpretation in terms of log-odds-ratio. To see this we start by examining the $2 \times 2$ case.

3.1.3. The $2 \times 2$ contingency table

Applying (18) for two-by-two tables, we have

$$\mu_1 = \sum_{k=0}^{1} \sum_{l=0}^{1} (\log p_{kl}) p_{k+} p_{+l} \phi_{1k} \psi_{1l}. \quad (23)$$

Note that for dichotomous $X$ the LP-spectral basis $\phi_1(F_X(x))$ is equal to $T_1(x; F_X)$. Consequently, we have the following explicit formula for $\phi_1$ and $\psi_1$:

$$\phi_1(F_X(x)) = \frac{x - p_{2+}}{\sqrt{p_{1+} p_{2+}}} \quad (24)$$

$$\psi_1(F_Y(y)) = \frac{y - p_{+2}}{\sqrt{p_{1+} p_{+2}}} \quad (25)$$

Substituting this into (23) yields the following important result.

**Theorem 3.1:** For 2-by-2 contingency tables, the estimate of the statistical parameter $\mu_1$ of the maxent LP-copula model

$$\text{cop}_\mu(u, v; X, Y) = e^{\mu_0 + \mu_1 \phi_1(u; X) \psi_1(v; Y)}$$

can be expressed as follows:

$$\widehat{\mu}_1 = \log \left[ \frac{\tilde{p}_{11} \tilde{p}_{22}}{\tilde{p}_{12} \tilde{p}_{21}} \right] \left( \frac{\tilde{p}_{1+} \tilde{p}_{+1} \tilde{p}_{2+} \tilde{p}_{+2}}{\tilde{p}_{1+} \tilde{p}_{+1} \tilde{p}_{2+} \tilde{p}_{+2}} \right)^{1/2}, \quad (26)$$

where the part inside the square bracket is the sample log-odds-ratio.

**Remark 3.1 (Significance of Theorem 3.1):** We have derived the log-odds-ratio statistic from first principles using a copula-theoretic framework. To the best of our knowledge, no other study has discovered this connection; see also of Goodman (1991, Equation (16)) and Gilula, Krieger, and Ritov (1988). In fact, one can view Theorem 3.1 as a special case of the much more general result described next.

**Theorem 3.2:** For an $I \times J$ table, consider a two-by-two subtable with rows $k$ and $k'$ and columns $l$ and $l'$. Then the logarithm of odds-ratio $\vartheta_{kl,kl'}$ is connected with the intrinsic association parameters $\mu_j$ in the following way:

$$\log \vartheta_{kl,kl'} = \sum_{j=1}^{M} \mu_j (\phi_{jk} - \phi_{jk'}) (\psi_{jl} - \psi_{jl'}). \quad (27)$$
To deduce (26) from (27), verify the following, utilising the LP-basis formulae (24)–(25)

\[
\phi_{11} - \phi_{10} := \phi_1(F_X(1)) - \phi_1(F_X(0)) = \frac{p_1^+ + p_2^+}{\sqrt{p_1^+ p_2^+}} = (p_1^+ p_2^+)^{-1/2}
\]

\[
\psi_{11} - \psi_{10} := \psi_1(F_Y(1)) - \psi_1(F_Y(0)) = \frac{p_1^+ + p_2^+}{\sqrt{p_1^+ p_2^+}} = (p_1^+ p_2^+)^{-1/2}.
\]

3.1.4. Reproducing Goodman’s association model

Our discrete copula-based categorical data model (17) expresses the logarithm of ‘dependence-ratios’

\[
\text{cop}(F_X(k), F_Y(l)) = \frac{p_{kl}}{p_{k+} p_{+l}}
\]

as a linear combination of LP-orthonormal row and column scores satisfying (19)–(21). The copula-dependence ratio (28) measures the strength of association between the k-th row category and l-th column category. To make the connection even more explicit, rewrite (17) for two-way contingency tables as follows:

\[
\log p_{kl} = \mu_0 + \mu_R^k + \mu_C^l + \sum_{j=1}^{m} \mu_j \phi_j \psi_{jl},
\]

where \(\mu_R^k\) denotes the logarithm of row marginal \(\log p_{k+}\) and \(\mu_C^l\) denotes the logarithm of column marginal \(\log p_{+l}\). Goodman (1991) called this model (29) a ‘weighted association model’ where weights are marginal row and column proportions. He used the term ‘association model’ (to distinguish it from correlation (22) based model) as it studies the relationship between rows and columns using odds ratio.

**Remark 3.2:** Log-linear models are a powerful statistical tool for categorical data analysis (Agresti 2013). Here we have provided a contemporary unified view of loglinear modelling for contingency tables from discrete LP-copula viewpoint. This newfound connection might open up new avenues of research.

3.2. Logratio biplot: graphical exploratory analysis

We describe a graphical exploratory tool – logratio biplot, which allows a quick visual understanding of the relationship between the categorical variables X and Y. In the following, we describe the process of constructing logratio biplot from the LP-copula model (17).

3.2.1. Copula-based algorithm

Construct two scatter plots based on the top two dominant components of the LP-copula model: the first one is associated with the row categories, formed by the points \((\mu_1 \phi_{1k}, \mu_2 \phi_{2k})\) for \(k = 1, \ldots, I\); and the second one is associated with the column categories, formed by the points \((\mu_1 \psi_{1l}, \mu_2 \psi_{2l})\) for \(l = 1, \ldots, J\). Logratio biplot is a two-dimensional display obtained by overlaying these two scatter plots—the prefix ‘bi’ refers to the fact that it shares a common set of axes for both the rows and columns categories.
3.2.2. Interpretation

Here we offer an intuitive explanation of the logratio biplot from the copula perspective. We start by recalling the definition of conditional comparison density (CCD; see Equations (6)–(7)), as the copula-slice. For fixed $X = k$, logratio-embedding coordinates $(\mu_j \phi_{jk})$ can be viewed as the LP-Fourier coefficients of $d(F_Y(y); Y, Y | X = k)$, since

$$
d(F_Y(y); Y, Y | X = k) = \exp \left\{ \mu_0 + \sum_{j=1}^{m} (\mu_j \phi_{jk}) \psi_{jl} \right\},
$$

Similarly, the logratio coordinates $(\mu_j \psi_{jl})$ for fixed $Y = l$ can be interpreted as the LP-expansion coefficients of $d(F_X(x); X, X | Y = l)$. Hence, the logratio biplot can alternatively be viewed as follows: (i) estimate the discrete LP-copula density; (ii) extract the copula slice $\hat{d}(u; Y, Y | X = k)$ along with its LP coefficients $(\mu_1 \phi_{1k}, \mu_2 \phi_{2k})$; (iii) similarly, get the estimated $\hat{d}(v; X, X | Y = l)$ – the copula slice at $Y = F_Y(l)$ along with its LP coefficients $(\mu_1 \psi_{1l}, \mu_2 \psi_{2l})$; (iv) hence, the logratio biplot (see Figure 2b) measures the association between the row and column categories $X = k$ and $Y = l$ by measuring the similarity between the ‘shapes’ of $\hat{d}(u; Y, Y | X = k)$ and $\hat{d}(v; X, X | Y = l)$ through their LP-Fourier coefficients.

Remark 3.3 (Historical Significance): The following remarks are pertinent: (i) log-ratio map traditionally taught and practiced using matrix-algebra (Greenacre 2018). This is in sharp contrast with our approach, which has provided a statistical synthesis of log-ratio biplot from a new copula-theoretic viewpoint. To the best of author’s knowledge, this is the first work that established such a connection. (ii) Logratio biplot has some important differences with the correspondence analysis pioneered by the French statistician Jean-Paul Benzécri; for more details, see Goodman (1991) and Benzécri (1991). However, in practice, these two methods often lead to very similar conclusions (e.g. contrast Figures 2b and 8).

Figure 2. 1970 draft lottery: (a) piecewise-constant nonlinear LP-copula density estimate and (b) two-dimensional logratio biplot that essentially captures all the useful information.
Example 3.1 (1970 Vietnam War-era US Draft Lottery (Fienberg 1971)): All eligible American men aged 19–26 were drafted through a lottery system in 1970 to fill the needs of the country’s armed forces. In 1970, the US conducted a draft lottery to determine the order (risk) of induction. The results of the draft are given to us in the form of a $12 \times 3$ contingency table (see Table 4 in the appendix): rows are months of the year from January to December, and columns denote three categories of risk of being drafted – high, medium, and low. The question is of interest whether the lottery was fairly conducted; in other words, is there any association between the two categories of $12 \times 3$ table? The discrete ’staircase-shaped’ LP-copula estimate is shown in the Figure 2(a), whose explicit form is given below:

$$\tilde{\text{cop}}_{X,Y}(u, v; X, Y) = \exp \left\{ 0.26 \phi_1(u; X) \psi_1(v; Y) + 0.18 \phi_2(u; X) \psi_2(v; Y) - 0.052 \right\}.$$ 

We now overlay the scatter plots $(0.26 \phi_{1k}, 0.18 \phi_{2k})$ for $k = 1, \ldots, 12$ and $(0.26 \psi_{1l}, 0.18 \psi_{2l})$ for $l = 1, 2, 3$ to construct the logratio biplot, as displayed in Figure 2. This easy-to-interpret two-dimensional graph captures the essential dependence pattern between the row (month: in blue dots) and the column (risk category: in red triangles) variables.

3.3. Loglinear modelling of large sparse contingency tables

It has been known for a long time that classical maximum likelihood-based log-linear models break down when applied to large sparse contingency tables with many zero cells, see Fienberg and Rinaldo (2007). Here we discuss a new maxent copula-based smooth method for fitting a parsimonious log-linear model to sparse contingency tables.

Example 3.2 (Zelterman data): The dataset (Zelterman 1987, Table 1) is summarised as a $28 \times 26$ cross-classified table that reports monthly salary and number of years of experience since bachelor’s degree of $n = 129$ women employed as mathematicians or statisticians. The table is extremely sparse – 86% cells are empty! See Figure 9 of Appendix A.5.

3.3.1. A parsimonious model

The estimated smooth log-linear LP-copula model for the Zelterman data is given by

$$\tilde{\text{cop}}_{X,Y}(u, v) = \exp \left\{ 0.52 \phi_1(u; X) \psi_1(v; Y) + 0.37 \phi_2(u; X) \psi_2(v; Y) + 0.18 \phi_3(u; X) \psi_3(v; Y) - 0.27 \right\},$$

displayed in Figure 3. This shows a strong positive correlation between salary and number of years of experience. However, the most notable aspect is the effective model dimension, which can be viewed as the intrinsic degrees of freedom (df). Our LP-maxent approach distills a compressed representation with reduced numbers of parameters that yields a smooth estimates: only requiring $m = 3$ components to capture the pattern in the data – a radical compression with negligible information loss! Contrast this with the dimension of the saturated loglinear model: $(28 - 1) \times (26 - 1) = 675$ – a case of a severely overparameterised non-smooth model with inflated degrees of freedom, which leads to an inaccurate goodness-of-fit test for checking independence between rows and columns. More on this in Section 3.5.2.
3.3.2. Smoothing ordered contingency tables

The nonparametric maximum likelihood-based cell probability estimates \( \tilde{p}_{X,Y}(k,l) = f_{kl}/n \) are very noisy and unreliable for sparse contingency tables. By sparse, we mean tables with a large number of cells relative to the number of observations.

Using Sklar’s representation theorem, one can simply estimate the joint probability \( \hat{p}_{X,Y}(k,l) \) by multiplying the empirical product pmf \( \tilde{p}_X(k)\tilde{p}_Y(l) \) with the smoothed LP-copula. In other words, the copula can be viewed as a data-adaptive bivariate discrete density-sharpening function that corrects the independent product-density to estimate the cell probabilities.

\[
d\text{Kernel}(k,l) = \hat{c}_{P_{X,Y}}(\tilde{F}_X(k),\tilde{F}_Y(l)), \quad \text{for } k = 1, \ldots, I; \ l = 1, \ldots, J. \tag{30}
\]

where the discrete-kernel function satisfies

\[
\frac{1}{n^2} \sum_{k=1}^{I} \sum_{l=1}^{J} d\text{Kernel}(k,l)f_{k+l} = 1.
\]

This approach can be generalised for any bivariate discrete distribution; see next section.

Remark 3.4: For a comprehensive literature on traditional kernel-based nonparametric smoothing methods for sparse contingency tables, readers are encouraged to consult Simonoff (1985, 1995) and references therein.

3.4. Modelling bivariate discrete distributions

The topic of nonparametric smoothing for multivariate discrete distributions has received far less attention than the continuous one. Two significant contributions in this direction
include: Aitchison and Aitken (1976) and Simonoff (1983). In what follows, we discuss a new LP-copula-based procedure for modelling correlated discrete random variables.

**Example 3.3 (Shunter accident data, Arbous and Kerrich 1951):** As a motivating example, consider the following data: we are given the number of accidents incurred by \( n = 122 \) shunters in two consecutive year periods, namely 1937–1942 and 1943–1947. To save space, we display the bivariate discrete data in a contingency table format, see Table 1.

**Algorithm.** The main steps of our analysis are described below.

**Step 1. Modelling marginal distributions.** We start by looking at the marginal distributions of \( X \) and \( Y \). As seen in Figure 4, negative binomial distributions provide excellent fit. To fix the notation, by \( G_{\mu,\phi} = \text{NB}(y; \mu, \phi) \), we mean the following probability distribution:

\[
\text{NB}(y; \mu, \phi) = \left( \frac{y + \phi - 1}{y} \right) \left( \frac{\mu}{\mu + \phi} \right)^y \left( \frac{\phi}{\mu + \phi} \right)^\phi, \quad y \in \mathbb{N},
\]

where \( \mathbb{E}(X) = \mu \) and \( \text{Var}(X) = \mu + \frac{\mu^2}{\phi} \). Using the method of MLE, we get: \( X \sim G_1 = \text{NB}(x; \hat{\mu} = 0.97, \hat{\phi} = 3.60) \), and \( Y \sim G_2 = \text{NB}(y; \hat{\mu} = 4.30, \hat{\phi} = 1.27) \).

**Step 2. Generalised copula density.** The probability of bivariate distribution at \((x, y)\) can be written as follows (generalising Sklar’s Theorem):

\[
\text{Pr}(X = x, Y = y) := p_{X,Y}(x, y) = g_1(x)g_2(y) \cdot \text{cop}_{X,Y}(G_1(x), G_2(y)), \quad (31)
\]

where generalised log-copula density admits the following decomposition:

\[
\log \left\{ \text{cop}_{X,Y}(G_1(x), G_2(y)) \right\} = \sum_{j=1}^{m_1} \sum_{k=1}^{m_2} \theta_{jk} T_j(x; G_1) T_k(y; G_2) - \log Z_\theta. \quad (32)
\]

It is important to note that the set of LP-basis functions \( \{T_j(x; G_1)\} \) and \( \{T_k(y; G_2)\} \) are specially designed for the parametric marginals \( G_1 \) and \( G_2 \), obeying the following weighted orthonormality conditions:

\[
\sum_x g_1(x) T_j(x; G_1) = 0, \quad \text{and} \quad \sum_x g_1(x) T_j(x; G_1) T_k(x; G_1) = \delta_{jk};
\]

\[
\sum_x g_2(x) T_j(x; G_2) = 0, \quad \text{and} \quad \sum_x g_2(x) T_j(x; G_2) T_k(x; G_2) = \delta_{jk}.
\]
Figure 4. Accident data. Top panel: marginal modelling – comparing the observed empirical pmf with the fitted negative binomial (NB) distribution. Bottom panel shows the estimated smooth joint pmf, which is obtained by ‘sharpening’ the product of marginal densities (top row) using the LP-smooth copula (Equation 34).
We call them gLP-basis, to distinguish them from the earlier empirical LP-polynomial systems \( \{ T_j(x;\tilde{F}_X) \} \) and \( \{ T_k(y;\tilde{F}_Y) \} \); see Appendix A.1.

**Remark 3.5 (Generalised copula as density-sharpening function):** The generalised copula
\[
\text{cop}_{X,Y}(G_1(x), G_2(y)) = \frac{p_{X,Y}(x,y)}{g_1(x)g_2(y)},
\]
acts as a bivariate ‘density sharpening function’ in (31) that corrects the possibly misspecified \( g_1(x)g_2(y) \). This is very much in the spirit of Mukhopadhyay (2021, 2022). It is also instructive to contrast our generalised copula (33) with the usual definition of copula (c.f. Section 2.1):
\[
\text{cop}_{X,Y}(F_X(x), F_Y(y)) = \frac{p_{X,Y}(x,y)}{p_X(x)p_Y(y)},
\]
which requires correct specification of the marginals \( p_X(x) \) and \( p_Y(y) \).

**Step 3:** Exploratory goodness-of-fit. The estimated log-bilinear LP-copula is
\[
\hat{\text{cop}}_{X,Y}(u,v) = \exp \left\{ 0.287S_1(u; G_1)S_1(v; G_2) - 0.043 \right\},
\]
There are three important conclusions that can be drawn from this non-uniform copula density estimate: (i) goodness of fit diagnostic: the independence model (product of parametric marginals) \( g_{\perp}(x,y) = g_1(x)g_2(y) \) is not adequate for the data. (ii) Nature of discrepancy: the presence of significant \( \hat{\theta}_{11} = 0.287 \) in the model (34) implies that the tentative independence model should be updated by incorporating the strong (positive) ‘linear’ correlation between \( X \) and \( Y \). (iii) Nonparametric repair: how to update the initial \( g_{\perp}(x,y) \) to construct a ‘better’ model? Equation (31) gives the general updating rule, which simply says: copula provides the necessary bivariate-correction function to reduce the ‘gap’ between the starting misspecified model \( g_{\perp}(x,y) \) and the true unknown distribution \( p_{X,Y}(x,y) \). (iv) In contrast to unsmoothed empirical multilinear copulas (Genest, Nešlehová, and Rémillard 2014), our method produces smoothed and compactly parametrisable \( \hat{\text{cop}}(u,v) \) for discrete data.

**Step 4:** LP-smoothed probability estimation. The bottom panel Figure 4 shows the final smooth probability estimate \( \hat{p}_{X,Y}(x,y) \), computed by substituting (34) into (31). Also compare Tables 5 and 6 of Appendix A.5.

**Remark 3.6:** Our procedure fits a ‘hybrid’ model: a nonparametrically corrected (through copula) multivariate parametric density estimate. One can use any parametric distribution instead of a negative binomial. The algorithm remains fully automatic, irrespective of the choice of parametric marginals \( G_1 \) and \( G_2 \), which makes it a universal procedure.

### 3.5. Mutual information

Mutual information (MI) is a fundamental quantity in Statistics and Machine Learning, with wide-ranging applications from neuroscience to physics to biology. For continuous
random variables \((X, Y)\), mutual information is defined as

\[
\text{MI}(X, Y) = \int \int f_{X,Y}(x,y) \log \frac{f_{X,Y}(x,y)}{f_X(x)f_Y(y)} \, dx \, dy.
\] (35)

Among nonparametric MI estimators, \(k\)-nearest-neighbour and kernel-density-based methods (Moon, Rajagopalan, and Lall 1995; Kraskov, Stögbauer, and Grassberger 2004; Zeng, Xia, and Tong 2018) are undoubtedly the most popular ones. Here we are concerned with a slightly general problem of developing a flexible MI estimation algorithm that is: (D1) applicable for mixed\(^6\) \((X, Y)\); (D2) robust in the presence of noise; and, (D3) invariant under monotone transformations.\(^7\) To achieve this goal, we start by rewriting MI (35) using copula:

\[
\text{MI}(X, Y) = \int_{[0,1]^2} \text{cop}_{X,Y}(u,v) \log \text{cop}_{X,Y}(u,v) \, du \, dv.
\] (36)

The next theorem presents an elegant closed-form expression for MI in terms of LP-copula parameters, which allows a fast and efficient estimation algorithm.

**Theorem 3.3:** Let \((X, Y)\) be a mixed-pair of random variables. Under the LP log-bilinear copula model (9), the mutual information between \(X\) and \(Y\) has the following representation in terms of LP-co-mean parameters \(LP_{jk}\) and maximum entropy coefficients \(\theta_{jk}\)

\[
\text{MI}_\theta(X, Y) = \sum \sum_{j,k>0} \theta_{jk} \text{LP}_{jk} - \log Z_{\theta}.
\] (37)

**Proof:** Express mutual information as

\[
\text{MI}_\theta(X, Y) = \mathbb{E}_{X,Y}[^{\text{log cop}_\theta}] = \sum \sum_{j,k} \theta_{jk} \mathbb{E}_{X,Y}[S_j(U; X)S_k(V; Y)] - \log Z_{\theta}.
\]

The first equality follows from (36) and the second one from (9). Complete the proof by replacing by \(LP_{jk}\) by \(\mathbb{E}[S_j(U; X)S_k(V; Y)]\) by virtue of (10). As a practical consequence, we have the following efficient and direct MI-estimator, satisfying D1–D3:

\[
\hat{\text{MI}}_\theta(X, Y) = \sum \sum_{j,k>0} \hat{\theta}_{jk} \tilde{LP}_{jk} - \log Z_{\theta}.
\] (38)

3.5.1. **Bootstrap inference**

Bootstrap provides a convenient way to estimate the standard error of the estimate (38). Perform bootstrap sampling, i.e. sample \(n\) pairs of \((x_i, y_i)\) with replacement and compute \(\hat{\text{MI}}\). Repeat the process, say, \(B = 500\) times to get the sampling distribution of the statistic. Finally, return the standard error of the bootstrap sampling distribution along with 95% percentile-confidence interval.
Table 2. The data table on malocclusion of the teeth in infants were obtained by M. Hellman and reported in the classic paper by Yates (1934, p. 230).

| Normal teeth | Malocclusion |
|--------------|-------------|
| Breast-fed   | 4           | 16          |
| Bottle-fed   | 1           | 21          |

3.5.1.1. Continuous \((X, Y)\) example. Consider the kidney fitness data, discussed in Example 2.1. The LP-copula-based (using \(m = 4\)) method yields: \(\hat{\text{MI}} = 0.230(\pm 0.021)\). To understand how precise is the estimate, we have reported the bootstrap standard error in parentheses.

Remark 3.7: MI (36) measures the departure of copula density from uniformity. This is because, MI can be viewed as the Kullback–Leibler (KL) divergence between copula and the uniform density: \(\text{MI}(X, Y) = \text{KL}(\text{cop}; U_{[0,1]^2})\). A few immediate consequences: (i) MI is always nonnegative, i.e. \(\text{MI}(X, Y) \geq 0\), and equality holds if and only if variables are independent. Moreover, the stronger the dependence between two variables, the larger the MI. (ii) MI is also invariant under different marginalisations. Two additional applications of MI (for categorical data and feature selection problems) are presented below.

3.5.2. Application 1: \((X, Y)\) discrete: smooth-\(G^2\) statistic

Given \(n\) independent samples from an \(I \times J\) contingency table, the \(G^2\)-test of goodness-of-fit, also known as the log-likelihood ratio test, is defined as

\[
G^2(X, Y) = 2n \sum_{k=1}^{I} \sum_{l=1}^{J} \tilde{p}_{kl} \log \frac{\tilde{p}_{kl}}{\hat{p}_k \hat{p}_l},
\]

which under the null hypothesis of independence has asymptotic \(\chi^2_{(I-1)(J-1)}\) distribution. From (39), one can immediately conclude the following.

Theorem 3.4: The \(G^2\) log-likelihood ratio statistic can be viewed as the raw nonparametric MI-estimate

\[
G^2(X, Y)/2n = \tilde{\text{MI}}(X, Y),
\]

where \(\tilde{\text{MI}}(X, Y)\) is obtained by replacing the unknown distributions in (35) with their empirical estimates.

Example 3.4 (Hellman’s Infant data, Yates 1934): We verify the identity (40) for the following \(2 \times 2\) Table 2, which shows cross-tabulation of \(n = 42\) infants based on whether the infant was breast-fed or bottle-fed.

The estimated LP-copula density (shown in Figure 11 of Appendix A.5) is given by

\[
\text{cop}(u, v; X, Y) = \exp \left\{ 0.234 S_1(u; X)S_1(v; Y) - 0.03 \right\}.
\]

The empirical MI estimate is given by \(2n \times \tilde{\text{MI}}(X, Y) = 2.50\), with the \(p\)-value 0.12 computed using the asymptotic null distribution \(\chi^2_1\). This exactly matches with the \(G^2\)-statistic value; one may use the R-function \(\text{GTest}\).
The problem arises when we try to apply $G^2$-test for large sparse tables, and it is not hard to see why; the adequacy of asymptotic $\chi^2_{(I-1)(J-1)}$ distribution depends both on the sample size $n$ and the number of cells $p = IJ$. Koehler (1986) showed that the approximation completely breaks down when $n/p < 5$, leading to erroneous statistical inference due to significant loss of power; see Appendix A.4. The following example demonstrates this.

**Example 3.5 (Zelterman Data Continued):** Log-likelihood ratio $G^2$-test produces $p$-value 1, firmly concluding the independence between salary and years of experience. This directly contradicts our analysis of Section 3.3, where we found a clear positive dependence between these two variables. Why $G^2$-test was unable to detect that effect? Because it is based on chi-square approximation with degrees of freedom $(28-1) \times (26-1) = 675$. This inflated degrees of freedom completely ruined the power of the test. To address this problem, we recommend the following smoothed version:

\[
\text{Smooth}\,G^2(X, Y)/2n = \hat{M}(X, Y),
\]

where $\hat{M}(X, Y)$ is computed based on the LP-bilinear copula model:

\[
\hat{c}(u, v; X, Y) = \exp \left\{ 0.50 S_1(u; Y) S_1(v; X) + 0.18 S_2(u; Y) S_2(v; X) - 0.167 \right\},
\]

Smooth-$G^2$ analysis (with $df = 2$) generates $p$-value $2.48 \times 10^{-11}$, thereby successfully detecting the association. The crucial aspect of our approach lies in its ability to provide a reduced dimensional parametrisation of copula density. For the Zelterman data, we need just two components (i.e. the effective degrees of freedom is 2) to capture the pattern.

**Remark 3.8 (Discrete variables with many categories):** Discrete distributions over large domains routinely arise in large-scale biomedical data such as diagnosis codes, drug compounds and genotypes (Seok and Kang 2015). The method proposed here can be used to jointly model such random variables.

### 3.5.3. Application 2: $(X, Y)$ mixed: feature importance score

We consider the two-sample feature selection problem where $Y$ is a binary response variable, and $X$ is a predictor variable that can be either discrete or continuous.

**Example 3.6 (Chronic Kidney Disease data):** The goal of this study is to investigate whether kidney function is related to red blood cell count (RBC). We have a sample of $n = 203$ participants, among whom 79 have chronic kidney disease (ckd) and another 124 are non-ckd. $Y$ denotes the kidney disease status and $X$ denotes the measurements on RBC (unit in million cells per cubic millimeter of blood). The estimated LP-copula density

\[
\hat{c}(u, v) = \exp \left\{ -0.76 S_1(u; Y) S_1(v; X) + 0.18 S_1(u; Y) S_2(v; X) - 0.19 S_1(u; Y) S_4(v; X) - 0.33 \right\},
\]

is shown in Figure 5. We propose mutual information-based feature importance measure based on the formula (38); this yields $\hat{M}(Y, X) = 0.36$ with $p$-value almost zero, strongly indicating that RBC is an importance risk-factor related to kidney dysfunction.

What additional insights can we glean from this copula? To answer that question, let’s focus our attention on the copula-slice for $u \in [0.61, 1]$. This segment of the copula is essentially the conditional comparison density $d(v; X, X | Y = 1)$ (see Section 2.1),
Figure 5. CKD data: The top panel shows the estimated copula density. The bottom panel shows the two-sample boxplots and the conditional comparison density (CCD) $d(v; X, X \mid Y = 1)$. The piecewise constant shape of estimated CCD is not an aberration of our nonparametric approximation method; it reflects the inherent discreteness of the feature $X$ (red blood cell count).

which can be easily derived by substituting $S_1(\tilde{P}_Y(1); Y) = \sqrt{\frac{1-\tilde{\mu}}{\tilde{\mu}}} = 1.25$ into (43), where $\tilde{\mu} = 79/203 = 0.389$:

$$\hat{d}(v; X, X \mid Y = 1) = \exp \left\{ -0.95S_1(v; X) + 0.23S_2(v; X) - 0.24S_4(v; X) - 0.33 \right\}.$$

A few remarks on the interpretation of the above formula:
• **Distributional effect-size**: Bearing in mind Equations (4), (7), note that \( d(v; X, X | Y = 1) \) compares two densities: \( f_{X | Y = 1}(x) \) with \( f_X(x) \), thereby capturing the distributional difference. This has advantages over traditional two-sample feature importance statistic (e.g. Student’s \( t \) or Wilcoxon statistic) that can only measure differences in location or mean.

• **Explainability**: The estimated \( \hat{d}(v; X, X | Y = 1) \) involves three significant LP-components of \( X \); the presence of the first-order ‘linear’ \( S_1(v; X) \) indicates location-difference; the second-order ‘quadratic’ \( S_2(v; X) \) indicates scale-difference; and fourth-order ‘quartic’ \( S_4(v; X) \) indicates the presence of tail-difference in the two RBC-distributions. In addition, the negative sign of the linear effect \( \hat{\theta}_{11} = -0.95 \) implies reduced mean level of RBC in the ckd-population. Medically, this makes complete sense, since a dysfunctional kidney cannot produce enough Erythropoietin (EPO) hormone, which causes the RBC to drop.

### 3.6. Nonparametric copula-logistic regression

We describe a new copula-based nonparametric logistic regression model. The key result is given by the following theorem, which provides a first-principle derivation of a robust nonlinear generalisation of the classical linear logistic regression model.

**Theorem 3.5**: Let \( \mu = \Pr(Y = 1) \) and \( \mu(x) = \Pr(Y = 1 | X = x) \). Then we have the following expression for the logit (log-odds) probability model:

\[
\text{logit}\{\mu(x)\} = \log \left( \frac{\mu(x)}{1 - \mu(x)} \right) = \alpha_0 + \sum_j \alpha_j T_j(x; F_X),
\]

where \( \alpha_0 = \text{logit}(\mu) \) and \( \alpha_j = \frac{\theta_{1j}}{\sqrt{\mu(1-\mu)}} \).

**Proof**: The proof consists of four main steps.

**Step 1.** To begin with, notice that for \( Y \) binary and \( X \) continuous, the general log-bilinear LP-copula density function (9) reduces to the following form:

\[
\text{cop}_\theta(u, v; X, Y) = \frac{1}{Z_\theta} \exp \left\{ \sum_j \theta_{1j} S_j(u; X) S_1(v; Y) \right\},
\]

since we can construct at most \( 2 - 1 = 1 \) LP-basis function for binary \( Y \).

**Step 2.** Apply copula-based Bayes Theorem (Equation 8) and express the conditional comparison densities as follows:

\[
d_1(x) \equiv d(F_X(x); X, X | Y = 1) = \frac{\mu(x)}{\mu},
\]

and also,

\[
d_0(x) \equiv d(F_X(x); X, X | Y = 0) = \frac{1 - \mu(x)}{1 - \mu}.
\]
Taking logarithm of the ratio of (46) and (47), we get the following important identity:

\[
\log \left( \frac{\mu(x)}{1 - \mu(x)} \right) = \log \left( \frac{\mu}{1 - \mu} \right) + \log d_1(x) - \log d_0(x). \tag{48}
\]

**Step 3.** From (45), one can deduce the following orthonormal expansion of maxent-conditional copula slices \( \log d_1 \) and \( \log d_0 \):

\[
\log d_1(x) = \sum_j \left( \theta_j T_1(1; F_Y) \right) T_j(x; F_X) - \log Z_\theta \tag{49}
\]

\[
\log d_0(x) = \sum_j \left( \theta_j T_1(0; F_Y) \right) T_j(x; F_X) - \log Z_\theta \tag{50}
\]

**Step 4.** Substituting (49) and (50) into (48) we get

\[
\log \left( \frac{\mu(x)}{1 - \mu(x)} \right) = \log \left( \frac{\mu}{1 - \mu} \right) + \sum_j \left\{ \frac{\theta_j}{\sqrt{\mu(1 - \mu)}} \right\} T_j(x; F_X),
\]

since for binary \( Y \) we have (see Appendix A.1):

\[
T_1(1; F_Y) - T_1(0; F_Y) = \frac{1 - \mu}{\sqrt{\mu(1 - \mu)}} + \frac{\mu}{\sqrt{\mu(1 - \mu)}} = \frac{1}{\sqrt{\mu(1 - \mu)}}.
\]

Substitute \( \alpha_0 = \logit(\mu) \) and \( \alpha_j = \frac{\theta_j}{\sqrt{\mu(1 - \mu)}} \) to complete the proof. \( \blacksquare \)

### 3.6.1. High-dimensional Copula-based additives logistic regression

Generalise the univariate copula-logistic regression model (44) to the high-dimensional case as follows:

\[
\logit \left( \mu(x) \right) = \alpha_0 + \sum_{j=1}^{p} h_j(x_j). \tag{51}
\]

Nonparametrically approximate the unknown smooth \( h_j \)'s by LP-polynomial series of \( X_j \)

\[
h_j(x_j) = \sum_{k=1}^{m} \alpha_{jk} T_k(x_j; F_{X_j}), \quad \text{for } j = 1, \ldots, p. \tag{52}
\]

**Remark 3.9 (Estimation and Computation):** A sparse nonparametric LP-additive model of the form (51)–(52) can be estimated by penalised regression techniques (lasso, elastic net, etc.), whose implementation is remarkably easy using glmnet R-function (Friedman, Hastie, and Tibshirani 2010):

\[
\text{glmnet}(y \sim T_X, \text{family = binomial})
\]

where \( T_X \) simply is the column-wise stacked \([T_{X_1} \mid \cdots \mid T_{X_p}]\) LP-feature matrix. Another advantage of this formulation is that a large body of already existing theoretical work (see the monograph Hastie, Tibshirani, and Wainwright 2015) on \( \ell_1 \)-regularised logistic regression model can be directly used to study the properties of (52).
Figure 6. UCI credit card data: We demonstrate the predictive performance and explainability of the LP-copula-based additive logistic regression model. The boxplots of the accuracy (AUC) are shown in the top left panel. On average, LP-logistic regression provides a 12% boost in the accuracy. The LS-plot is shown in the top left. For easy interpretation, we display the scaled the LS-plot: $\left( \hat{\alpha}_j^{(1)}, \hat{\alpha}_j^{(2)} \right)$. We can see a tight cluster around the origin $(0, 0)$, which indicates that most of the variables are irrelevant for prediction (sparsity-assumption). Also, it is evident that the majority of the features have either location or scale information (differences), the only exception being the variable pay_0 – which denotes the repayment status of the last two months ($-1 = \text{pay duly}, 1 = \text{payment delay for one month}, 2 = \text{payment delay for two months}, \text{and so on}$). This is further illustrated using three variables, as marked in the LS-plot; see also the two-sample density estimates shown at the bottom panel.

Example 3.7 (UCI Credit Card data): The dataset is available in the UCI Machine Learning Repository. It contains records of $n = 30,000$ cardholders from an important Taiwan-based bank. For each customer, we have a response variable $Y$ denoting: default payment status ($\text{Yes} = 1, \text{No} = 0$), along with $p = 23$ predictor variables (e.g. gender, education, age, history of past payment, etc.). We randomly partition the data into training and test sets, with an 80–20 split, repeated 100 times. We measure the prediction accuracy using AUC (the area under the ROC curve). Figure 6 compares two kinds of lasso-logistic regressions: (i) usual version: based on feature matrix $X$; and (ii) LP-copula version: based on feature matrix $T_X$. As we can see, LP-copula based additive logistic regression classifier
significantly outperforms the classical logistic regression model. To gain further insight into the nature of impact of each variable, we plot the lasso-smoothed location and scale coefficients:

\[ \text{LS - Feature plot : } (\hat{\alpha}_{j1}, \hat{\alpha}_{j2}), \text{ for } j = 1, \ldots, p. \]

L-stands for location and S-stands for scale. The purpose of the LS-feature plot is to characterize 'how' each feature impacts the classification task. For example, consider the three variables pay_0, limit_balance, and pay_6, shown in the bottom panel Figure 6. Each one of them contains unique discriminatory information: pay_0 has location as well as scale information, hence it appeared at the top-right of the LS-plot. The variable limit_balance mainly shows location differences, whereas the variable pay_6 shows contrasting scale in the two populations. In short, LS-plot explains 'why and how' each variable is important using a compact diagram, which is easy to interpret by researchers and practitioners.

4. Conclusion: Copula-based statistical learning

This paper makes the following contributions: (i) we introduce modern statistical theory and principles for maximum entropy copula density estimation that is self-adaptive for the mixed \((X,Y)\) – described in Section 2. (ii) Our general copula-based formulation provides a unifying framework of data analysis from which one can systematically distill a number of fundamental statistical methods by revealing some completely unexpected connections between them. The importance of our theory in applied and theoretical statistics is highlighted in Section 3, taking examples from different sub-fields of statistics: Log-linear analysis of categorical data, logratio biplot, smoothing large sparse contingency tables, mutual information, smooth-\(G^2\) statistic, feature selection, and copula-based logistic regression. We hope that this new perspective on copula modelling will offer more effective ways of developing united statistical algorithms for mixed-\((X,Y)\).

Dedication: two legends from two different cultures

This paper is dedicated to the birth centenary of E. T. Jaynes (1922–1998), the originator of the maximum entropy principle.

I also like to dedicate this paper to the memory of Leo Goodman (1928–2020) – a transformative legend of categorical data analysis, who passed away on December 22, 2020, at the age of 92 due to COVID-19.

This paper is inspired in part by the author’s intention to demonstrate how these two modelling philosophies can be connected and united in some ways. This is achieved by employing a new nonparametric representation theory of generalized copula density.

Notes

1. since the second derivative of the log-partition function is a s positive semi-definite covariance matrix
2. See Mukhopadhyay and Parzen (2020) for a parallel result on the LP-orthogonal series copula model.
3. Compare our equation (18) with Equation (34) of Goodman (1996).
4. Goodman (1996, p. 410) calls it 'Pearson ratios'.
5. A similar philosophy was proposed in Mukhopadhyay (2017) for univariate continuous distribution case.
6. Reliably estimating MI for mixed case is notoriously challenging task (Gao, Kannan, Oh, and Viswanath 2017).
7. This is essential to make the analysis less sensitive to various types of data preprocessing, which is done routinely in applications like bioinformatics, astronomy, and neuroscience.
8. In 1935, Samuel Wilks introduced log-likelihood ratio test as an alternative to Pearson’s chi-square test. In our notation, Pearson proposed \( \int \text{cop}^2_{X,Y}(u,v) \) and Wilks proposed \( 2 \times \int \text{cop}_{X,Y}(u,v) \log \text{cop}_{X,Y} \) – both are conceptually equivalent: measuring how much the copula density deviates from the uniformity.

**Disclosure statement**

No potential conflict of interest was reported by the author.

**Supplementary Material**

The supplement contains additional details on computation, methods, and numerical simulations.

**ORCID**

Subhadeep Mukhopadhyay [http://orcid.org/0000-0002-6841-0132](http://orcid.org/0000-0002-6841-0132)

**References**

Agresti, A. (2013), *Categorical Data Analysis* (3rd ed.). Hoboken (NJ): John Wiley & Sons.

Aitchison, J., and Aitken, C.G. (1976), ‘Multivariate Binary Discrimination by the Kernel Method’, *Biometrika*, 63, 413–420.

Arbous, A.G., and Kerrich, J. (1951), ‘Accident Statistics and the Concept of Accident-proneness’, *Biometrics*, 7, 340–432.

Barron, A.R., and Sheu, C. (1991), ‘Approximation of Density Functions by Sequences of Exponential Families’, *Annals of Statistics*, 19, 1347–1369.

Benzecri, J.P. (1991), ‘Comment on Leo Goodman’s “Models, and Graphical Displays in the Analysis of Cross-Classified Data”’, *Journal of the American Statistical Association*, 86, 1112–1115.

Efron, B., and Hastie, T. (2016), *Computer Age Statistical Inference* (Vol. 5). Cambridge: Cambridge University Press.

Fienberg, S.E. (1971), ‘Randomization and Social Affairs: the 1970 Draft Lottery’, *Science*, 171, 255–261.

Fienberg, S.E., and Rinaldo, A. (2007), ‘Three Centuries of Categorical Data Analysis: Log-linear Models and Maximum Likelihood Estimation’, *Journal of Statistical Planning and Inference*, 137, 3430–3445.

Friedman, J., Hastie, T., and Tibshirani, R. (2010), ‘Regularization Paths for Generalized Linear Models via Coordinate Descent’, *Journal of Statistical Software*, 33, 1.

Gao, W., Kannan, S., Oh, S., and Viswanath, P. (2017), ‘Estimating Mutual Information for Discrete-Continuous Mixtures’, Advances in Neural Information Processing Systems, Long Beach, CA, USA, pp. 5986–5997.

Genest, C., Nešlehová, J., and Rémillard, B. (2014), ‘On the Empirical Multilinear Copula Process for Count Data’, *Bernoulli*, 20, 1344–1371.

Gilula, Z., Krieger, A.M., and Ritov, Y. (1988), ‘Ordinal Association in Contingency Tables: Some Interpretive Aspects’, *Journal of the American Statistical Association*, 83, 540–545.

Goodman, L.A. (1991), ‘Measures, Models, and Graphical Displays in the Analysis of Cross-classified Data (with Discussion)’, *Journal of the American Statistical Association*, 86, 1085–1111.
Goodman, L.A. (1996), ‘A Single General Method for the Analysis of Cross-classified Data: Reconciliation and Synthesis of Some Methods of Pearson, Yule, and Fisher, and Also Some Methods of Correspondence Analysis and Association Analysis’, *Journal of the American Statistical Association*, 91, 408–428.

Greenacre, M. (2018), *Compositional Data Analysis in Practice*. Boca Raton (FL): CRC Press.

Hastie, T., Tibshirani, R., and Wainwright, M. (2015), *Statistical Learning with Sparsity*, Monographs on Statistics and Applied Probability, New York, NY, 143.

Hoeffding, W. (1940), ‘Masstabinvariante Korrelationstheorie’, *Schriften des Mathematischen Seminars und des Instituts für Angewandte Mathematik der Universität Berlin*, 5, 179–233.

Jaynes, E.T. (1957), ‘Information Theory and Statistical Mechanics’, *Physical Review*, 106, 620.

Koehler, K.J. (1986), ‘Goodness-of-fit Tests for Log-linear Models in Sparse Contingency Tables’, *Journal of the American Statistical Association*, 81, 483–493.

Kraskov, A., Stögbauer, H., and Grassberger, P. (2004), ‘Estimating Mutual Information’, *Physical Review E*, 69, 066138.

Moon, Y.-I., Rajagopalan, B., and Lall, U. (1995), ‘Estimation of Mutual Information Using Kernel Density Estimators’, *Physical Review E*, 52, 2318.

Mukhopadhyay, S. (2017), ‘Large-Scale Mode Identification and Data-Driven Sciences’, *Electronic Journal of Statistics*, 11, 215–240.

Mukhopadhyay, S. (2021), ‘Density Sharpening: Principles and Applications to Discrete Data Analysis’, Technical Report, pp. 1–51, arXiv:2108.07372.

Mukhopadhyay, S. (2022), ‘Modelplasticity and Abductive Decision Making’, Technical Report, pp. 1–27, arXiv:2203.03040.

Mukhopadhyay, S., and Parzen, E. (2020), ‘Nonparametric Universal Copula Modeling’, *Applied Stochastic Models in Business and Industry, Special Issue on “Data Science”*, 36, 77–94.

Mukhopadhyay, S., Parzen, E., and Lahiri, S. (2012), ‘From Data to Constraints’, *Bayesian Inference And Maximum Entropy Methods In Science And Engineering: 31st International Workshop, Waterloo, Canada*, 1443, 32–39.

Parzen, E., and Mukhopadhyay, S. (2013), ‘United Statistical Algorithms, LP-Comoment, Copula Density, Nonparametric Modeling’, in *59th ISI World Statistics Congress (WSC)*, Hong Kong, pp. 4719–4724.

Seok, J., and Kang, Y.S. (2015), ‘Mutual Information Between Discrete Variables with Many Categories Using Recursive Adaptive Partitioning’, *Scientific Reports*, 5, 1–10.

Simonoff, J.S. (1983), ‘A Penalty Function Approach to Smoothing Large Sparse Contingency Tables’, *The Annals of Statistics*, 11, 208–218.

Simonoff, J.S. (1985), ‘An Improved Goodness-of-fit Statistic for Sparse Multinomials’, *Journal of the American Statistical Association*, 80, 671–677.

Simonoff, J.S. (1995), ‘Smoothing Categorical Data’, *Journal of Statistical Planning and Inference*, 47, 41–69.

Sklar, M. (1959), ‘Fonctions De Répartition à N Dimensions Et Leurs Marges’, *Publ. Inst. Statistique Univ. Paris*, 8, 229–231.

Yates, F. (1934), ‘Contingency Tables Involving Small Numbers and the $\chi^2$ Test’, *Supplement to the Journal of the Royal Statistical Society*, 1, 217–235.

Zelterman, D. (1987), ‘Goodness-of-fit Tests for Large Sparse Multinomial Distributions’, *Journal of the American Statistical Association*, 82, 624–629.

Zeng, X., Xia, Y., and Tong, H. (2018), ‘Jackknife Approach to the Estimation of Mutual Information’, *Proceedings of the National Academy of Sciences*, 115, 9956–9961.