Copula Based Trainable Calibration Error Estimator of Multi-Label Classification with Label Interdependencies

Arkapal Panda[1] **Utpal Garain**[1]
[1]Computer Vision and Pattern Recognition Unit,
Indian Statistical Institute

Abstract

A key challenge in calibrating Multi-Label Classification(MLC) problems is to consider the interdependencies among labels. To address this, in this research we propose an unbiased, differentiable, trainable calibration error estimator for MLC problems by using Copula. Unlike other methods for calibrating MLC tasks that focus on marginal calibration, this novel estimator takes label interdependencies into account and enables us to tackle the strictest notion of calibration that is canonical calibration. To design the estimator, we begin by leveraging the kernel trick to construct a continuous distribution from the discrete label space. Then we take a semiparametric approach to construct the estimator where the marginals are modeled non-parametrically and the Copula is modeled parametrically. Theoretically we show that our estimator is unbiased and converges to true L^p calibration error. We also use our estimator as a regularizer at the time of training and observe that it reduces calibration error on test datasets significantly. Experiments on a well established dataset endorses our claims.

1 INTRODUCTION

In recent years, Deep Neural Networks (DNNs) have demonstrated outstanding performance in both multi-class and multi-label classification tasks. However, studies (Guo et al., 2017) have revealed that despite increase in the performance in terms of accuracy, modern DNNs suffers from predicting overconfident probability estimates, making DNNs less calibrated. A DNN is perfectly calibrated if the predicted probability p, constitutes that the particular class will be correctly predicted $100 * p\%$ of time. For example, in fields like medical imaging, disease detection, autonomous driving etc. the predicted probability can be used to determine whether human intervention is needed or not. Poorly calibrated DNNs in safety critical tasks like this is going to raise false sense of confidence which might lead to catastrophic consequences. Therefore, not only accuracy, calibration too is of utmost importance at the time of applying DNNs in safety critical tasks.

Although, there are various methods (Naeini et al., 2015; Kumar et al., 2018) available in literature which addresses the problem of calibration but they are mostly limited to binary and multi-class classification tasks. Only a few studies (Cheng and Vasconcelos, 2024) have been conducted to address the problem of calibration in multi-label tasks. However, Multi-label Classification (MLC) (Liu et al., 2017) plays a crucial role in diverse applications, ranging from medical imaging to autonomous driving. In a conventional classification task, the presumption that every instance is linked to a single label is incongruent with real-world situations, as instances are frequently linked to many labels. Also, it has been observed that the labels in multi-label datasets are often interdependent. For example, in an image of chest Xray, there might be multiple indications. Complex interdependencies between labels in multi-label problems present extra challenges, especially in the calibration of the model, where predicting accurate probability estimates help in downstream tasks like decision making and uncertainty estimation. A range of studies related to MLC (Liu and Tsang, 2015; Shen et al., 2017) has highlighted that the assumption of independence causes degradation in performance. The notion of calibration for multi-label tasks has mostly been restricted to marginal(label

Proceedings of the 28[th] International Conference on Artificial Intelligence and Statistics (AISTATS) 2025, Mai Khao, Thailand. PMLR: Volume 258.

wise) (Kull et al., 2019) calibration where each label is calibrated separately. This is a very strong definition of calibration and ignores the role of label dependence in multi-label tasks. One of the most strict notion of calibration is canonical calibration (Kull and Flach, 2015; Bröcker, 2009), requires the whole output vector of predicted probabilities to be calibrated, thus takes into account the interdependence among labels. (Cheng and Vasconcelos, 2024) has made an attempt to calibrate the predicted probabilities of MLC tasks but didn't consider interdependence of labels. Existing calibration error estimation methods for binary and multi-class settings, such as BBQ (Naeini et al., 2015), Mix-n-Match (Zhang et al., 2020), and ECE^{KDE} (Popordanoska et al., 2022), are not designed to incorporate label interdependencies and often suffer from statistical and computational limitations. BBQ estimator is asymptotically inconsistent in many situations; Mix-n-Match, although consistent but it is difficult to track it in higher dimensions; ECE^{KDE} is consistent and can be used with any loss function as a regularizer at the time of training but it can only handle binary and multi-class classification tasks. In this work, we address this gap by introducing a novel calibration error estimator specifically designed to handle label interdependencies in MLC tasks, which we incorporate as an auxiliary loss during the optimization process. In particular, to model the joint distribution of the labels we resort to the observation made by Sklar (Sklar, 1996).

Sklar's Observation: Sklar (Sklar, 1996) proposed that univariate marginals and the multivariate dependence structure can be modeled separately, and the dependence information can be captured by using a Copula (Nelsen, 2006). Therefore, to incorporate dependence information and effectively model the joint distribution of labels in MLC tasks, Copulas should be considered (Nelsen, 2006).

The key contribution of this paper are as follows: **1.** We leverage kernel trick to construct continuous distributions of the discrete label space because Sklar's theorem cannot be applied in discrete distributions. **2.** We develop an estimator of canonical calibration error for MLC tasks which takes label dependence into account. To achieve this, we adopted a semi-parametric approach, modeling the copula parametrically while handling the marginal distributions non-parametrically. **3.** Theoretically we show that our estimator is consistent, unbiased and piece-wise differentiable. **4.** We use our estimator as an auxiliary loss along with the traditional loss function during training and observed that it serves as a resularizer and produces better-calibrated probability estimates for MLC tasks. Experimental results on several datasets endorse our claim.

2 RELATED WORK

Copula Modeling: Copulas (Nelsen, 2006) has long been used in variety of applications especially in finance (Cherubini, 2004), survival analysis (E. Shemyakin and Youn, 2006), econometrics (Patton, 2012) and has gained a lot of success. In machine learning literature copula is widely used due to it's tractability and ability to consider dependencies among different labels. Copulas are used in fields like multi-agent learning (Wang et al., 2021), kernel methods (Póczos et al., 2012), graphical methods (Czado and Nagler, 2022), properties of random sequences (Wilson and Ghahramani, 2010), multi-label classification tasks (Liu, 2019). However, it has not yet been used in calibration or uncertainty estimation.

Calibration: The literature in multi-class calibration can be divided into two parts mainly: post-hoc calibration and trainable calibration strategies. Post-hoc calibration techniques does not need any access to the model architecture. It learns a mapping function from a held out dataset and passes the predicted probabilities through that function. Some studies regarding post-hoc calibration include Histogram Binning (Zadrozny and Elkan, 2001), Temperature Scaling (Ding et al., 2021), BBQ (Naeini et al., 2015), Dirichlet Calibration (Kull et al., 2019) etc. Trainable calibration strategies generally include a differentiable calibration error estimator into the training process. MMCE (Kumar et al., 2018) used a kernel based surrogate loss; (Müller et al., 2019) has used weighted average of labels instead of hard targets during training and showed improvement in calibration. But this methods are not designed to handle calibration error in multi-label problems where label interdependency is prevalent. To calibrate MLC problems Cheng(2024) (Cheng and Vasconcelos, 2024) proposed a strictly proper asymmetric loss function to handle label imbalance; Chen(2024) (Chen et al., 2024) proposed DCLR algorithm which learns dynamic instance level and prototype level similarities for each label. But these methods use conventional calibration error metrics to estimate the calibration error.

3 PRELIMINARIES

Let $(\Omega, \mathcal{F}, P)$ be a probability space where Ω represents the whole space, $\mathcal{F}$ is the sigma-algebra and P is the associated probability measure. Let $\mathcal{X} = \mathbb{R}^n$

and $\mathcal{Y} = \{0,1\}^K$(K is assumed as the number of labels throughout this work). Let h be the multi-label classification model such that $h : \mathcal{X} \rightarrow [0,1]^K$. So we can write $h = (h_1, \ldots, h_K)$. Let $\mathbf{x} : \Omega \rightarrow \mathcal{X}$ and $\mathbf{y} : \Omega \rightarrow \mathcal{Y}$ be the random variables. $D(H), R(H)$ be the domain and range of any function H. If $\mathbf{a}, \mathbf{b} \in \mathbb{R}^n$ where $\mathbf{a} = (a_1, \ldots, a_n); \mathbf{b} = (b_1, \ldots, b_n)$ and $a_i \leq b_i; \forall i \in \{1, \ldots, n\}$ then a n-rectangle is defined by $[\mathbf{a}, \mathbf{b}] = [a_1, b_1] \times \cdots \times [a_n, b_n]$. The set of vertices of a n-rectangle is defined by $\mathcal{V} = \{v_1, \ldots, v_n | \forall i, v_i = a_i \, or \, b_i\}$. Below, we follow (Nelsen, 2006) for the definitions that we need to understand copula properly.

Definition 1 (Calibration Error (Naeini et al., 2015))**.** *For a model h, the L^p calibration error is defined as follows:*

$$CE_p(h) = \left(E\left[\left| \left| E[\mathbf{y}|h(\mathbf{x})] - h(\mathbf{x}) \right| \right|_p^p \right] \right)^{\frac{1}{p}} \quad (1)$$

Definition 2 (H-volume)**.** *Let $R_1, \ldots, R_n$ be non-empty subsets of $\bar{\mathbb{R}}$, and let H be a function such that $D(H) = R_1 \times \cdots \times R_n$. Then the H volume of a rectangle $\mathcal{R} = [\mathbf{a}, \mathbf{b}]$, whose vertices are all in $D(H)$, is given by:*

$$V_H(\mathcal{R}) = \sum_{\mathbf{v} \in \mathcal{V}} r(\mathbf{v}) H(\mathbf{v}).$$

The function r is defined as follows:

$$r(\mathbf{v}) = \begin{cases} 1 & when \; \left| \{i : v_i = a_i\} \right| \equiv 0 (mod \; 2) \\ -1 & o.w. \end{cases}$$

Definition 3 (n-increasing)**.** *A real-valued function $H : \mathbb{R}^n \rightarrow \mathbb{R}$ will be called n-increasing if $V_H(\mathcal{R}) \geq 0$ for every n-rectangle $\mathcal{R}$ with vertices in $D(H)$.*

Definition 4. *A real function H with domain $R_1 \times \cdots \times R_n$ with a_i as the least element of each R_i, is said to be grounded if $H(\mathbf{v}) = 0$ for all $\mathbf{v} \in D(H)$ such that $v_i = a_i$ for one i at least.*

Definition 5 (Margin of H)**.** *Assume that each R_i is bounded and non-empty. So, $\forall i, R_i$ has a maximum b_i. Then margin of H is going to be the function $H_i(u) = H(b_1, \ldots, b_{i-1}, u, b_{i+1}, \ldots, b_n)$. For higher dimensional margin, less number of dimensions are fixed in H. For i dimensional margin we call it i-margin of H.*

Definition 6 (n-Copula)**.** *A n-copula(n-dimensional copula) is a function $C : [0,1]^n \rightarrow [0,1]$ with the following properties:*

- *The function C is both grounded and n-increasing.*
- *The function C has marginal distributions C_i for $i \in 1, \ldots, n$, where each $C_i(u)$ satisfies $C_i(u) = u$ for $u \in [0,1]$.*

It can be noted from definition 6 that every i-margin of a Copula is a i-Copula.

Now we shall state Sklar's Theorem.

Theorem 1 (Sklar (Sklar, 1996))**.** *Let $\mathbf{X} = (X_1, X_2, \ldots, X_n)$ be a random vector with joint cumulative distribution function (CDF):*

$$F_{\mathbf{X}}(x_1, \ldots, x_n) = P(X_1 \leq x_1, X_2 \leq x_2, \ldots, X_n \leq x_n)$$

where the marginal CDFs of X_is be:

$$F_i(x_i) = P(X_i \leq x_i), \quad for \; i = 1, 2, \ldots, n.$$

Then there exists a n-dimensional copula function $C : [0,1]^n \rightarrow [0,1]$ such that:

$$F_{\mathbf{X}}(x_1, x_2, \ldots, x_n) = C(F_1(x_1), F_2(x_2), \ldots, F_n(x_n)) \quad (2)$$

where $(x_1, \ldots, x_n) \in [-\infty, \infty]$. If all marginals F_is are continuous then C is unique otherwise if some marginals are discrete then C is uniquely determined on $R(F_1) \times \ldots \times R(F_n)$.
Conversely, if C is an n-copula and $F_1, \ldots, F_n$ are univariate CDFs, then the function F defined in Equation 2 serves as the joint CDF with marginals $F_1, \ldots, F_n$.

Sklar's theorem enables the independent modeling of marginal distributions and the dependency structure between the marginals by using a copula. By selecting an appropriate copula, one can effectively capture and summarize the entire dependence among the marginals.

Additionally, we can easily derive the joint probability density function(PDF) from copula. By assuming that F has partial derivatives the joint PDF is defined as

$$\begin{aligned} f(x_1, \ldots, x_n) &= \frac{\partial C(F_1(x_1), \ldots, F_n(x_n))}{\partial F_1 \cdots \partial F_n} \prod_{i=1}^{n} f_i(x_i) \\ &= c(F_1(x_1), \ldots, F_n(x_n)) \prod_{i=1}^{n} f_i(x_i) \end{aligned} \quad (3)$$

The function $c(F_1(x_1), \ldots, F_n(x_n))$ represents the copula density, while f_i denotes the marginal density function of x_i.

4 METHODOLOGY

Let, $\mathbf{y} = \{y_1, \ldots, y_K\}$. As we are dealing with MLC problems, $h(\mathbf{x})$ and the conditional expectation in equation 1 are vectors of dimension K. In order

to empirically compute $E[\mathbf{y}|h(\mathbf{x})]$, we need to estimate pdf of $(y_i, h_1(\mathbf{x}), \ldots, h_K(\mathbf{x})), \forall i \in \{1, \ldots, K\}$ which we can get by marginalizing the joint cdf of $(y_1, \ldots, y_K, h_1(\mathbf{x}), \ldots, h_K(\mathbf{x}))$ over all $y_j : j \in \{1, \ldots, K\} \backslash i$. Unlike other (Cheng and Vasconcelos, 2024; Chen et al., 2024) works, where calibration error is estimated by treating each label as independent binary outcomes, our method inherits information of label dependence through marginalization. This is because the joint distribution of $(y_1, \ldots, y_K, h_1(\mathbf{x}), \ldots, h_K(\mathbf{x}))$ contains the information of label dependence.

In this study, we employ a semiparametric method to estimate CE_p, where we use a parametric model for the copula while estimating the marginal CDFs nonparametrically. It is essential for the estimator to be both consistent and differentiable. The proposed estimator is defined as:

$$\widehat{CE_p(h)^p} = \frac{1}{n} \sum_{j=1}^{n} \left(\left[\left| \left| \widehat{E[\mathbf{y}|h(\mathbf{x})]} \Big|_{\mathbf{x}_j} - h(\mathbf{x}_j) \right| \right|_p^p \right] \right) \quad (4)$$

where $\widehat{E[\mathbf{y}|h(\mathbf{x})]}\Big|_{\mathbf{x}_j}$ is $\widehat{E[\mathbf{y}|h(\mathbf{x})]}$ evaluated at $h(\mathbf{x}) = h(\mathbf{x}_j)$. Directly getting an empirical form of equation 4 is not possible as the functional form of $E[\mathbf{y}|h(\mathbf{x})]$ is not tractable. In the sections below, we will provide details about how to use copula and Skalr's theorem to construct a empirically tractable form that can be used during the training process as an auxiliary loss.

4.1 Constructing Continuous distribution of Label Space

As we are dealing with MLC problems, the label space is boolean valued vector and Sklar's theorem cannot be applied in discrete variables. So, we need to construct a continuous distribution to replace the discrete one. Let p_j be the probability mass function of y_j for $j \in 1, \ldots, K$. Following Tsukahara (2005) (Tsukahara, 2005), we employ an uniform kernel density function $\frac{p_j(y_j)}{2b}$ for an observation y_j with a small bandwidth b to obtain a continuous distribution from discrete label space . By taking $b = 0.5$, the CDF of the continuous random variable z_j which is transformed from the discrete random variable y_j is given by

$$\mathfrak{F}_j(z_j) = \begin{cases} 0 & \text{if } z_j < -0.5 \\ (p_j(0)(z_j + 0.5)) & \text{if } -0.5 \leq z_j \leq 0.5 \\ (p_j(0) + p_j(1)(z_j - 0.5)) & \text{if } 0.5 \leq z_j \leq 1 \\ 1 & \text{if } z_j > 1.5 \end{cases} \quad (5)$$

4.2 Copula Modeling

We denote the marginal CDF of $h_j(\mathbf{x}), j \in \{1, \ldots, K\}$ by F_j. From equation 1 we have $E[\mathbf{z}|h(\mathbf{x})] = \big[E[z_1|h(\mathbf{x})], \ldots, E[z_K|h(\mathbf{x})]\big]$. Now, by Skalr's theorem the joint CDF of $(z_1, \ldots, z_K, h_1(\mathbf{x}), \ldots, h_K(\mathbf{x}))$ can be expressed by a $2K$ copula $C(\mathfrak{F}_1(z_1), \ldots, \mathfrak{F}_K(z_K), F_1(h_1(\mathbf{x})), \ldots, F_K(h_K(\mathbf{x})))$. To get an analytical expression for $E[z_j|h(\mathbf{x})]$, we examine the $(K+1)$-dimensional marginal distribution of C over the variables $(z_j, h_1(\mathbf{x}), \ldots, h_K(\mathbf{x}))$. This margin will naturally inherit the information of label interdependencies from the $2K$-copula C.

From equation 3 the joint cdf of $(z_j|h(\mathbf{x}))$ can be written as

$$f_{z_j}(z_j|h(\mathbf{x})) = \frac{\lambda(z_j) c(\mathfrak{F}_j(z_j), F_1(h_1(\mathbf{x})), \ldots, F_K(h_K(\mathbf{x}))}{c_h(F_1(h_1(\mathbf{x})), \ldots, F_K(h_K(\mathbf{x}))} \quad (6)$$

$\lambda(z_j)$ is the PDF of z_j and

$$c_h(F_1(h_1(\mathbf{x})), \ldots, F_K(h_K(\mathbf{x})) = \frac{\partial C(1, F_1(h_1(\mathbf{x})), \ldots, F_n(h_K(\mathbf{x})))}{\partial F_1 \cdots \partial F_n}$$

is the copula density function of $h_1(\mathbf{x}), \ldots, h_K(\mathbf{x})$. Now, we derive the conditional expectation of z_j given $h(\mathbf{x})$

$$\begin{aligned} & E[z_j|h(\mathbf{x})] \\ &= \int_{-\infty}^{\infty} z_j f_{z_j}(z_j|h(\mathbf{x})) dz_j \\ &= \int_{-\infty}^{\infty} z_j \frac{\lambda(z_j) c(\mathfrak{F}_j(z_j), F_1(h_1(\mathbf{x})), \ldots, F_K(h_K(\mathbf{x}))}{c_h(F_1(h_1(\mathbf{x})), \ldots, F_K(h_K(\mathbf{x}))} dz_j \quad (7) \\ &= \frac{\alpha_j(F_1(h_1(\mathbf{x})), \ldots, F_K(h_K(\mathbf{x}))}{c_h(F_1(h_1(\mathbf{x})), \ldots, F_K(h_K(\mathbf{x}))} \\ &= E[z_j \phi(\mathfrak{F}_j(z_j), F_1(h_1(\mathbf{x})), \ldots, F_K(h_K(\mathbf{x})))] \end{aligned}$$

where

$$\alpha_j(F_1(h_1(\mathbf{x})), \ldots, F_K(h_K(\mathbf{x})) = E[z_j c(\mathfrak{F}_j(z_j), F_1(h_1(\mathbf{x})), \ldots, F_K(h_K(\mathbf{x}))]$$

and

$$\phi(\mathfrak{F}_j(z_j), F_1(h_1(\mathbf{x})), \ldots, F_K(h_K(\mathbf{x}))) = \frac{c(\mathfrak{F}_j(z_j), F_1(h_1(\mathbf{x})), \ldots, F_K(h_K(\mathbf{x}))}{c_h(F_1(h_1(\mathbf{x})), \ldots, F_K(h_K(\mathbf{x}))}.$$

From the derivation above we can see that conditional expectation can be formulated form copula density function.

Given the estimators $\hat{\phi}, \hat{\mathfrak{F}}_j$ and $\hat{F_1}, \ldots, \hat{F_K}$; $E[z_j|h(\mathbf{x})]$ can be estimated by $\widehat{E[z_j|h(\mathbf{x})]} = E[z_j\hat{\phi}(\hat{\mathfrak{F}}_j(z_j), \hat{F_1}(h_1(\mathbf{x})), \ldots, \hat{F_K}(h_K(\mathbf{x})))]$. So the estimator of $E[\mathbf{z}|h(\mathbf{x})]$ can be defined as $\widehat{E[\mathbf{z}|h(\mathbf{x})]} = \big[\widehat{E[z_1|h(\mathbf{x})]}, \ldots, \widehat{E[z_K|h(\mathbf{x})]}\big]$. To estimate ϕ we need to estimate copula densities c and c_h. We take a semiparametric approach towards this which is described below.

4.3 Estimators of $\mathfrak{F}_j$ and F_j

Let $(z_1^{(i)}, \ldots, z_K^{(i)}, h_1(\mathbf{x}^{(i)}), \ldots, h_K(\mathbf{x}^{(i)}))\Big|_{i=1}^{N}$ be N iid samples. The probability mass function $p_j(0)$ can be estimated by $\hat{p}_j(0) = \frac{\sum_{i=1}^{N} \mathbb{1}(y_j^{(i)}=0)}{N}$ and $p_j(1)$ can be estimated by $\hat{p}_j(1) = 1 - \hat{p}_j(0)$. Then the estimator of $\mathfrak{F}_j(z_j)$ can be written as

$$\hat{\mathfrak{F}}_j(z_j) = \begin{cases} 0 & \text{if } z_j < -0.5 \\ \hat{p}_j(0)(z_j + 0.5) & \text{if } -0.5 \leq z_j \leq 0.5 \\ \hat{p}_j(0) + \hat{p}_j(1)(z_j - 0.5) & \text{if } 0.5 \leq z_j \leq 1.5 \\ 1 & \text{if } z_j > 1.5 \end{cases} \tag{8}$$

The estimator of $F_j, j \in \{1, \ldots, K\}$, $\hat{F}_j$, can be estimated by the empirical CDF of $h_j(\mathbf{x})$. It can be defined as follows:

$$\hat{F}_j(h_j(\mathbf{x})) = \frac{\sum_{i=1}^{N} \mathbb{1}(h_j(\mathbf{x}) \leq h_j(\mathbf{x}^{(i)}))}{N} \tag{9}$$

As $N \to \infty$, $\hat{F}_j(h_j(\mathbf{x}))$ converges to true CDF by Glivenko-Cantelli theorem (Talagrand, 1987).

4.4 Estimation of Copula

To estimate copula we will take a parametric approach. Let us assume that the $2K$-copula density is part of a parametric family characterized by the parameter θ. That means the copula density function can be written as $c(\mathfrak{F}_1(z_1), \ldots, \mathfrak{F}_K(z_K), F_1(h_1(\mathbf{x})), \ldots, F_K(h_K(\mathbf{x})); \theta)$, $\theta \in \mathbb{R}^m$). Let θ^* be the true parameter of copula θ. We could have used Maximum Likelihood Estimator(MLE) but empirically estimating the MLE of θ is not possible directly as $\mathfrak{F}_j, F_j;\ j \in \{1, \ldots K\}$ is not tractable empirically. To tackle this problem we resort to Maximum Pseduo-Likelihood Estimator(MPLE) (Genest et al., 1995; Tsukahara, 2005), $\hat{\theta}$, which is defined as:

$$\hat{\theta} = \arg\max_{\theta} \log\Big[c(\mathfrak{F}_1(z_1), \ldots, \mathfrak{F}_K(z_K), F_1(h_1(\mathbf{x})), \ldots, F_K(h_K(\mathbf{x})); \theta)\Big] \tag{10}$$

Let, $\hat{\theta}_j$ be the MPLE of $(K+1)$ margin on variables $(z_j, h_1(\mathbf{x}), \ldots, h_K(\mathbf{x}))$ whose corresponding true copula parameter is θ_j^* and $\hat{\theta}_j$, $\theta_j^* \in \mathbb{R}^d$

Tsukahara(2005) (Tsukahara, 2005) has showed that

$$\hat{\theta}_j - \theta_j^* = \frac{\sum_{i=1}^{N} \psi_i}{N} + o(N^{-\frac{1}{2}})$$

where $\psi_i = \psi(\mathfrak{F}_j(z_j^{(i)}), F_1(h_1(\mathbf{x}^{(i)})), \ldots, F_K(h_K(\mathbf{x}^{(i)}))$; $\theta_j^*)$, a random vector on θ_j^* such that $E(\psi) = (0, \ldots, 0)$ and ψ is a L^2 function. So, from this we can deduce that $\hat{\theta}_j$ is an unbiased estimator when $n \to \infty$ and consistent too.

Notations: To make our analysis easy to understand we introduce a few notations

$\dot{c} = \left(\frac{\partial c}{\partial \theta_j^{(1)}}, \ldots, \frac{\partial c}{\partial \theta_j^{(d)}}\right)^T$; $c_b = \frac{\partial c}{\partial u_b}, b \in \{1, \ldots, K+1\}$;

$c_{h,b} = \frac{\partial c_h}{\partial u_b}$ and $\alpha_{j,b} = \frac{\partial \alpha_j}{\partial u_b}, b \in \{1, \ldots, K\}$;

$\dot{c_h} = \left(\frac{\partial c_h}{\partial \theta_j^{(1)}}, \ldots, \frac{\partial c_h}{\partial \theta_j^{(d)}}\right)^T$; $\dot{\alpha_j} = \left(\frac{\partial \alpha_j}{\partial \theta_j^{(1)}}, \ldots, \frac{\partial \alpha_j}{\partial \theta_j^{(d)}}\right)^T$;

5 UNBIASEDNESS AND CONSISTENCY OF $\widehat{E[\mathbf{z}|h(\mathbf{x})]}$

To prove the unbiasdness and consistency of $\widehat{E[\mathbf{z}|h(\mathbf{x})]}$ we need to show that same for $\widehat{E[z_j|h(\mathbf{x})]}$ for all $j \in \{1, \ldots, K\}$. So we need to analyze $(K+1)$ margins of c on variables $(z_j, h_1(\mathbf{x}), \ldots, h_K(\mathbf{x}))$. Below we will state some assumptions related to continuity and the existence of moments of the conditional expectation upto some order. In future we might try to extend this work without these assumptions.

Assumption 1. *1. $\dot{c}$ and $\dot{c}_a$ are continuous.*

2. $E|z_j| \leq \infty$

3. The square of the weighted expectation of z_j where weights are c or c_h is finite.

Keeping these assumptions in our hand below we will state theorems below. The proofs are available in the appendix.

Observe that, when the dimension of input is 1, $E[z_j|h(\mathbf{x})] = \alpha_j(F_1(h_1(\mathbf{x})); \theta^*) = E[z_j c(\mathfrak{F}_j(z_j), F_1(h_1(\mathbf{x})); \theta^*)]$. That implies $\widehat{E[z_j|h(\mathbf{x})]} = \frac{\sum_{i=1}^{N} z_j^{(i)} c(\hat{\mathfrak{F}}_j(z_j^{(i)}), \hat{F}_1(h_1(\mathbf{x})); \theta^*)}{N}$.

Lemma 1. *For $j \in \{1, \ldots K\}$, $\hat{\theta}_j = \theta_j^* + o(N^{-\frac{1}{2}})$ and*

when assumption 3 *holds we have that*

$$\widehat{E[z_j|h(\mathbf{x})]} - \frac{\sum_{i=1}^{N} z_j^{(i)} c(\hat{\mathfrak{F}}_j(z_j^{(i)}), \hat{F}_1(h_1(\mathbf{x})); \theta_j^*)}{N}$$
$$= \frac{1}{N}\sum_{i=1}^{N} z_j^{(i)}(\hat{\mathfrak{F}}_j(z_j^{(i)} - \mathfrak{F}_j(z_j^{(i)}) c_1(\mathfrak{F}_j(z_j^{(i)}), F_1(h_1(\mathbf{x})); \theta_j^*)$$
$$+ (\hat{F}_1(h_1(\mathbf{x})) - F_1(h_1(\mathbf{x})))\alpha_{j,1}(F_1(h_1(\mathbf{x}))$$
$$+ (\hat{\theta}_j - \theta_j^*)\dot{\alpha}_j(F_1(h_1(\mathbf{x}))) + o(N^{\frac{1}{2}})$$

Now, we present the theorem below based on lemma 1

Theorem 2. *When the input data has dimension* 1 *and assumptions* 1, 2 *is satisfied along with the conditions of lemma 1* $\widehat{E[z_j|h(\mathbf{x})]}$ *is a consistent and unbiased estimator of* $E[z_j|h(\mathbf{x})]$.

The generalized version of lemma 1 when the input data can have any dimensions will be

Lemma 2. *Under assumptions* 1, 2 *and* 3 *and the conditions of theorem 2 we have that*

$$\widehat{E[z_j|h(\mathbf{x})]} - E[z_j|h(\mathbf{x})] =$$
$$= \frac{1}{N}\sum_{i=1}^{N}\Big((z_j^{(i)}(\hat{\mathfrak{F}}_j(z_j^{(i)} - \mathfrak{F}_j(z_j^{(i)}) c(\mathfrak{F}_j(z_j^{(i)}), F_1(h_1(\mathbf{x})),$$
$$\ldots, F_K(h_K(\mathbf{x})); \theta_j^*) - E[z_j|h(\mathbf{x})]\sum_{a=1}^{K}(\mathbb{1}(h_a(\mathbf{x}^{(i)}) \leq$$
$$h_a(\mathbf{x})) - F_a(h_a(\mathbf{x})))c_{h,a}(F_1(h_1(\mathbf{x})), \ldots, F_K(h_K(\mathbf{x})))$$
$$+ \psi_i' \dot{c}_h(F_1(h_1(\mathbf{x})), \ldots, F_K(h_K(\mathbf{x})); \theta_j^*)\Big) / \Big(c_h(F_1(h_1(\mathbf{x}))$$
$$, \ldots, F_K(h_K(\mathbf{x})); \theta_j^*)\Big)$$

Theorem 3. *For input data of any dimension, given that the conditions of lemma 2 is satisfied,* $\widehat{E[z_j|h(\mathbf{x})]}$ *is an unbiased an consistent estimator of* $E[z_j|h(\mathbf{x})]$. *Thus* $\widehat{E[\mathbf{z}|h(\mathbf{x})]}$ *is an unbiased and consistent estimator of* $E[\mathbf{z}|h(\mathbf{x})]$

In theorem 2 and 3 we have derived that the semi-parametric estimator of $E[\mathbf{z}|h(\mathbf{x})]$ is unbiased and consistent. So, to estimate the L^p calibration error we can now directly use $\widehat{E[\mathbf{z}|h(\mathbf{x})]}$ as the empirical form of $\widehat{E[\mathbf{z}|h(\mathbf{x})]}$ is available to us through theorem 2 and 3.

6 TRAINABLE CALIBRATION ERROR

Most of the loss functions that are used to train MLC tasks are designed to achieve consistency in terms of Bayesian risk optimization. They are not guaranteed to achieve the same for calibration. As we are interested in developing a model h which not only accurate but also well calibrated, during optimization we use $CE(h)$ jointly along with a loss function. In other words, $CE(h)$ is going to act as an auxiliary loss in the joint risk minimization task. We do it by bounding the $CE(h)$ by B and formulate the following optimization problem: $h = \arg\min_{h\in\mathcal{H}} \text{Risk}(h)$ s.t. $CE(h) < B$ with the associated lagrangian:

$$h = \arg\min_{h\in\mathcal{H}}\Big(\text{Risk}(h) + \beta \cdot CE(h)\Big) \tag{11}$$

$\mathcal{H}$ is the class of all possible models. We follow Popordanoska(2022) (Popordanoska et al., 2022) and derive the functional form of the optimization function in case of Mean Squared Error(MSE) loss as a first instantiation of our framework for MLC problems. We have $MSE(h) = E(\mathbf{z} - h(\mathbf{x}))^2$. We jointly optimize it in conjunction with L^2 calibration error:

$$h = \arg\min_{h\in\mathcal{H}}\Big(\text{Risk}(h) + \beta \cdot CE_2(h)^2\Big)$$
$$= \arg\min_{h\in\mathcal{H}}\Big(\text{Risk}(h) + \gamma \cdot E\big[E[\mathbf{z}|h(\mathbf{x})]^2\big]\Big) \tag{12}$$

where $\gamma = \frac{\beta}{\beta+1}$. The estimator of $E\big[E[z_j|h(\mathbf{x})]^2\big]$ can be written as :

$$\widehat{E\big[E[z_j|h(\mathbf{x})]^2\big]}$$
$$= \frac{1}{N}\sum_{l=1}^{N}\frac{\Big(\sum_{i=1}^{N} z_j^{(i)} c(\hat{\mathfrak{F}}_j(z_j^{(i)}), \hat{F}(h(\mathbf{x}^{(l)})))\Big)^2}{\Big(c_h(\hat{F}(h(\mathbf{x}^{(l)}))\Big)^2} \tag{13}$$

we will use the estimator as equation 13 in our experiment.

7 EXPERIMENTS AND RESULTS

We have used Gaussian copula as the desired copula in our experiment. The Gaussian copula is derived from a multivariate normal distribution and is expressed in terms of the CDFs of the standard normal distribution. Functional form of a Gaussian copula for an d-dimensional random vector $(X_1, \ldots, X_d)$ is given by:

$$C\left(v_1, \ldots, v_d; \sum\right) = \Phi_{\sum}\left(\phi^{-1}(v_1), \ldots, \phi^{-1}(v_d)\right).$$

where:

- $\Phi_{\sum}$ is the d-dimensional multivariate normal distribution with a zero mean vector and correlation matrix $\sum \in [-1, 1]$.
- ϕ^{-1} is the inverse CDF of a standard normal distribution.

- $v_i = F_i(X_i)$ are the marginal CDFs.

The density function can be written as:

$$c(v_1, \ldots, v_d; Z) = \det\left(\sum\right)^{-1/2} \exp\left(-\frac{1}{2}\xi'\left(\sum^{-1} - I_d\right)\xi\right)$$

where:

- $\xi = \left(\phi^{-1}(v_1), \ldots, \phi^{-1}(v_d)\right)'$ is the transformed variable in the standard normal domain.
- I_d is the identity matrix of dimension d.

In our experiments, PyTorch, SciKitLearn and SciPy are used to implement the models and the estimators. They are trained using a NVIDIA RTX A5000 GPU. Our implementations can be found in the following link: `https://github.com/highonai/CDEMLC.git`

For our experiments we have used ResNet50 (He et al., 2016) along with the famous PASCAL-VOC (Everingham et al., 2012) dataset. The dataset has 20 labels and each image contains one or more than one label. We have used the standard train, validation and test dataset the is provided with the original data. We have run our model for 100 epochs and chose the best model. We have compared our results with other optimization functions like Cross Entropy, Focal Loss etc. and found that our method, which we name as Copula Density Estimation for MLC(CDEMLC), is able to produce better calibrated results. The results of the experiment can be found below:

Dataset	CE	FL	CDEMLC
PASCAL-VOC	0.03819 $\pm$0.0048	0.2034 $\pm$0.036	0.0012 $\pm$0.0005

8 CONCLUSION

In our work, we have proposed a novel calibration error estimator for MLC tasks which considers label dependencies. We have proved the unbiasedness and consistency of our estimator. We have taken a semiparametric approach where the copula is estimated parametrically and the marginal CDFs are estimated nonparametrically. In future we want to estimate the copula non-parametrically too as non-parametric estimation gives much more flexibility compared to parametric estimation.

References

Bröcker, J. (2009). Reliability, sufficiency, and the decomposition of proper scores. *Quarterly Journal of the Royal Meteorological Society: A journal of the atmospheric sciences, applied meteorology and physical oceanography*, 135(643):1512–1519.

Chen, T., Wang, W., Pu, T., Qin, J., Yang, Z., Liu, J., and Lin, L. (2024). Dynamic correlation learning and regularization for multi-label confidence calibration. *IEEE Transactions on Image Processing*.

Cheng, J. and Vasconcelos, N. (2024). Towards calibrated multi-label deep neural networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 27589–27599.

Cherubini, U. (2004). Copula methods in finance. *John Wiley & Sons google schola*, 2:949–956.

Czado, C. and Nagler, T. (2022). Vine copula based modeling. *Annual Review of Statistics and Its Application*, 9(1):453–477.

Ding, Z., Han, X., Liu, P., and Niethammer, M. (2021). Local temperature scaling for probability calibration. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6889–6899.

E. Shemyakin, A. and Youn, H. (2006). Copula models of joint last survivor analysis. *Applied Stochastic Models in Business and Industry*, 22(2):211–224.

Everingham, M., Van Gool, L., Williams, C., Winn, J., and Zisserman, A. (2012). The pascal visual object classes challenge 2012 (voc2012) results. 2012 http://www. pascal-network. org/challenges. In *VOC/voc2012/workshop/index. html*.

Feller, W. (1991). *An introduction to probability theory and its applications, Volume 2*, volume 81. John Wiley & Sons.

Genest, C., Ghoudi, K., and Rivest, L.-P. (1995). A semiparametric estimation procedure of dependence parameters in multivariate families of distributions. *Biometrika*, 82(3):543–552.

Guo, C., Pleiss, G., Sun, Y., and Weinberger, K. Q. (2017). On calibration of modern neural networks. In *International conference on machine learning*, pages 1321–1330. PMLR.

He, K., Zhang, X., Ren, S., and Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778.

Kull, M. and Flach, P. (2015). Novel decompositions of proper scoring rules for classification: Score adjustment as precursor to calibration. In *Machine Learning and Knowledge Discovery in Databases: Euro-*

pean Conference, ECML PKDD 2015, Porto, Portugal, September 7-11, 2015, Proceedings, Part I 15, pages 68–85. Springer.

Kull, M., Perello Nieto, M., Kängsepp, M., Silva Filho, T., Song, H., and Flach, P. (2019). Beyond temperature scaling: Obtaining well-calibrated multi-class probabilities with dirichlet calibration. *Advances in neural information processing systems*, 32.

Kumar, A., Sarawagi, S., and Jain, U. (2018). Trainable calibration measures for neural networks from kernel mean embeddings. In *International Conference on Machine Learning*, pages 2805–2814. PMLR.

Liu, W. (2019). Copula multi-label learning. *Advances in Neural Information Processing Systems*, 32.

Liu, W. and Tsang, I. (2015). On the optimality of classifier chain for multi-label classification. *Advances in Neural Information Processing Systems*, 28.

Liu, W., Tsang, I. W., Klaus-Robert, M., et al. (2017). An easy-to-hard learning paradigm for multiple classes and multiple labels. *Journal of Machine Learning Research*, 18(94):1–38.

Mann, H. B. and Wald, A. (1943). On stochastic limit and order relationships. *The Annals of Mathematical Statistics*, 14(3):217–226.

Müller, R., Kornblith, S., and Hinton, G. E. (2019). When does label smoothing help? *Advances in neural information processing systems*, 32.

Naeini, M. P., Cooper, G., and Hauskrecht, M. (2015). Obtaining well calibrated probabilities using bayesian binning. In *Proceedings of the AAAI conference on artificial intelligence*, volume 29.

Nelsen, R. B. (2006). An introduction to copulas.

Patton, A. J. (2012). A review of copula models for economic time series. *Journal of Multivariate Analysis*, 110:4–18.

Póczos, B., Ghahramani, Z., and Schneider, J. (2012). Copula-based kernel dependency measures. *arXiv preprint arXiv:1206.4682*.

Popordanoska, T., Sayer, R., and Blaschko, M. (2022). A consistent and differentiable lp canonical calibration error estimator. *Advances in Neural Information Processing Systems*, 35:7933–7946.

Shen, X., Liu, W., Tsang, I. W., Sun, Q.-S., and Ong, Y.-S. (2017). Multilabel prediction via cross-view search. *IEEE transactions on neural networks and learning systems*, 29(9):4324–4338.

Sklar, A. (1996). Random variables, distribution functions, and copulas: a personal look backward and forward. *Lecture notes-monograph series*, pages 1–14.

Talagrand, M. (1987). The glivenko-cantelli problem. *The Annals of Probability*, pages 837–870.

Tsukahara, H. (2005). Semiparametric estimation in copula models. *Canadian Journal of Statistics*, 33(3):357–375.

Wang, H., Yu, L., Cao, Z., and Ermon, S. (2021). Multi-agent imitation learning with copulas. In *Machine Learning and Knowledge Discovery in Databases. Research Track: European Conference, ECML PKDD 2021, Bilbao, Spain, September 13–17, 2021, Proceedings, Part I 21*, pages 139–156. Springer.

Wilson, A. G. and Ghahramani, Z. (2010). Copula processes. *Advances in Neural Information Processing Systems*, 23.

Zadrozny, B. and Elkan, C. (2001). Obtaining calibrated probability estimates from decision trees and naive bayesian classifiers. In *Icml*, volume 1, pages 609–616.

Zhang, J., Kailkhura, B., and Han, T. Y.-J. (2020). Mix-n-match: Ensemble and compositional methods for uncertainty calibration in deep learning. In *International conference on machine learning*, pages 11117–11128. PMLR.

Checklist

1. For all models and algorithms presented, check if you include:
   (a) A clear description of the mathematical setting, assumptions, algorithm, and/or model. Yes
   (b) An analysis of the properties and complexity (time, space, sample size) of any algorithm. Not Applicable
   (c) (Optional) Anonymized source code, with specification of all dependencies, including external libraries.

2. For any theoretical claim, check if you include:
   (a) Statements of the full set of assumptions of all theoretical results. Yes
   (b) Complete proofs of all theoretical results. Yes. **Provided in the appendix section**
   (c) Clear explanations of any assumptions. Yes

3. For all figures and tables that present empirical results, check if you include:
   (a) The code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL). Yes(`https://github.com/highonai/CDEMLC.git`)
   (b) All the training details (e.g., data splits, hyperparameters, how they were chosen). Not Applicable. The splits are already defined in the main dataset.
   (c) A clear definition of the specific measure or statistics and error bars (e.g., with respect to the random seed after running experiments multiple times). Yes
   (d) A description of the computing infrastructure used. (e.g., type of GPUs, internal cluster, or cloud provider). Yes

4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets, check if you include:
   (a) Citations of the creator If your work uses existing assets. Not Applicable
   (b) The license information of the assets, if applicable. Not Applicable
   (c) New assets either in the supplemental material or as a URL, if applicable. Not Applicable
   (d) Information about consent from data providers/curators. Not Applicable
   (e) Discussion of sensible content if applicable, e.g., personally identifiable information or offensive content. Not Applicable

5. If you used crowdsourcing or conducted research with human subjects, check if you include:
   (a) The full text of instructions given to participants and screenshots. Not Applicable
   (b) Descriptions of potential participant risks, with links to Institutional Review Board (IRB) approvals if applicable. Not Applicable
   (c) The estimated hourly wage paid to participants and the total amount spent on participant compensation. Not Applicable

APPENDIX

Assumptions

To make sure our proof are tractable we make some assumptions. These assumptions are very basic and any predictive model follows these assumptions easily.

Assumption 2.

$$\hat{F}_j(h_j(\mathbf{x})) = \frac{\sum_{i=1}^{N} \mathbb{1}(h_j(\mathbf{x}^{(i)}) \leq h_j(\mathbf{x}))}{N} + o(N^{-\frac{1}{2}}) \tag{14}$$

where $j \in \{1, \ldots, K\}$

Assumption 3. *For* $j \in \{1, \ldots, K\}$

$$\hat{\theta}_j - \theta_j^* = \frac{\sum_{i=1}^{N} \psi_i}{N} + o(N^{-\frac{1}{2}}) \tag{15}$$

where $\psi_i = \psi(\mathfrak{F}_j(z_j^{(i)}), F_1(h_1(\mathbf{x}^{(i)})), \ldots, F_K(h_K(\mathbf{x}^{(i)})); \theta_j^*)$ *is a random vector such that* $E(\psi) = (0, \ldots, 0)$ *and* $E||\psi||_2^2 < \infty$

Assumption 1 and 2 are due to Tsukahara(2005) (Tsukahara, 2005).

Assumption 4. *1.* $\dot{c}$ *and* $\dot{c}_b$ *are continuous* $\forall b \in \{1, \ldots, K+1\}$

2. $E|z_j| < \infty$
3. $E[z_j c_b]^2 < \infty$ *and* $E[z_j c]^2 < \infty$ *where* $b \in \{1, \ldots, K+1\}$ *and* $j \in \{1, \ldots, K\}$
4. $E[z_j \frac{\partial c(\mathfrak{F}_j(z_j), F_1(h_1(\mathbf{x})), \ldots, F_K(h_K(\mathbf{x}); \theta_j^*))}{\partial \theta_{j'}}] < \infty$ *for all* j' *and* $j \in \{1, \ldots, K\}$

We follow the style of proofs by Liu (2019) and Nelsen (2006) for the following proofs

Missing Proofs

Proof of Lemma 1

Proof. For $h(\mathbf{x}) = h_1(\mathbf{x})$ we begin with the taylor expansion of the conditional expectation is

$$\widehat{E[z_j|h_1(\mathbf{x})]} = \frac{\sum_{i=1}^{N} z_j^{(i)} c(\hat{\mathfrak{F}}_j(z_j^{(i)}), \hat{F}_1(h_1(\mathbf{x})); \theta_j^*)}{N} + \sigma_1 + \sigma_2 + \sigma_3$$

σ_i's are defined below

$$\sigma_1 = \frac{1}{N} \sum_{i=1}^{N} z_j^{(i)} (\hat{\mathfrak{F}}_j(z_j^{(i)}) - \mathfrak{F}_j(z_j^{(i)}) c_1(\tilde{u}_{i,j}, \tilde{u}_1; \tilde{\theta}_j)$$

$$\sigma_2 = \frac{1}{N} \sum_{i=1}^{N} z_j^{(i)} (\hat{F}_j(h_1(\mathbf{x})) - F_j(h_1(\mathbf{x}))) c_2(\tilde{u}_{i,j}, \tilde{u}_1; \tilde{\theta}_j)$$

$$\sigma_3 = \frac{1}{N} \sum_{i=1}^{N} z_i^{(i)} (\hat{\theta}_j - \theta_j^*)^T \dot{c}(\tilde{u}_{i,j}, \tilde{u}_1; \tilde{\theta}_j)$$

$\tilde{u}_{i,j} = \mathfrak{F}_j(z_j^{(i)}) + t(\hat{\mathfrak{F}}_j(z_j^{(i)}) - \mathfrak{F}_j(z_j^{(i)}))$; $\tilde{u}_1 = F_1(x_1) + t(\hat{F}_1(x_1) - F_1(x_1))$; $\tilde{\theta}_j = \theta_j^* + t(\hat{\theta}_j - \theta_j^*)$ for some small t. σ_1 can be written as

$$\begin{aligned}\sigma_1 = \frac{1}{N} \sum_{i=1}^{N} z_j^{(i)} (\hat{\mathfrak{F}}_j(z_j^{(i)}) - \mathfrak{F}_j(z_j^{(i)}) c_1(F_j(z_j^{(i)}), F_1(x_1); \theta_j^*) \\ + \frac{1}{N} \sum_{i=1}^{N} z_j^{(i)} \left(\hat{F}_j(z_j^{(i)}) - F_j(z_j^{(i)})\right) \left(c_1(\tilde{u}_{i,j}, \tilde{u}_1; \tilde{\theta}_j) - c_1(F_j(z_j^{(i)}), F_1(x_1); \theta_j^*)\right)\end{aligned} \tag{16}$$

the second term in the right hand side of equation 16 can be written as

$$\frac{1}{N}\sum_{i=1}^{N} z_j^{(i)} \left(\hat{F}_j(z_j^{(i)}) - F_j(z_j^{(i)})\right) \left(c_1(\tilde{u}_{i,j}, \tilde{u}_1; \tilde{\theta}_j) - c_1(F_j(z_j^{(i)}), F_1(x_1); \theta_j^*)\right)$$
$$\leq \frac{1}{N}\sum_{i=1}^{N} \left|z_j^{(i)}\right| \sup_{z_j} \left|\hat{F}_j(z_j^{(i)}) - F_j(z_j^{(i)})\right| \sup_{i} \left|c_1(\tilde{u}_{i,j}, \tilde{u}_1; \tilde{\theta}_j) - c_1(F_j(z_j^{(i)}), F_1(x_1); \theta_j^*)\right| \tag{17}$$

Using assumption 4.1 and continuous mapping theorem (Mann and Wald, 1943) we get $\left|c_1(\tilde{u}_{i,j}, \tilde{u}_1; \tilde{\theta}_j) - c_1(F_j(z_j^{(i)}), F_1(x_1); \theta_j^*)\right| = O(1)$. Additionally assumption 4.2 ensures $E|z_j| < \infty$.
Finally by applying D'Moivre's theorem (Feller, 1991) and Weak Law of Large Numbers(WLLN) on $\left|\hat{F}_j(z_j^{(i)}) - F_j(z_j^{(i)})\right|$ we get that

$$\frac{1}{N}\sum_{i=1}^{N} \left|z_j^{(i)}\right| \sup_{z_j} \left|\hat{F}_j(z_j^{(i)}) - F_j(z_j^{(i)})\right| \sup_{i} \left|c_1(\tilde{u}_{i,j}, \tilde{u}_1; \tilde{\theta}_j) - c_1(F_j(z_j^{(i)}), F_1(x_1); \theta_j^*)\right|$$
$$\leq O(1)O(N^{-1/2})O(1) = O(N^{-1/2})$$

Similarly this can be done for σ_2 and σ_3. Then by using WLLN again we get the proof of our theorem.

□

Proof of Theorem 1

Proof. From Lemma 1 we have that :

$$\begin{aligned} E\left[\widehat{E[z_j|h(\mathbf{x})]}\right] &= E\left[\frac{\sum_{i=1}^{N} z_j^{(i)} c(\hat{\mathfrak{F}}_j(z_j^{(i)}), \hat{F}_1(h_1(\mathbf{x})); \theta_j^*)}{N}\right] \\ &+ E\left[\frac{1}{N}\sum_{i=1}^{N} z_j^{(i)} (\hat{\mathfrak{F}}_j(z_j^{(i)}) - \mathfrak{F}_j(z_j^{(i)})) c_1(\mathfrak{F}_j(z_j^{(i)}), F_1(h_1(\mathbf{x})); \theta_j^*)\right] \\ &+ E\left[(\hat{F}_1(h_1(\mathbf{x})) - F_1(h_1(\mathbf{x})))\alpha_{j,1}(F_1(h_1(\mathbf{x})))\right] \\ &+ E\left[(\hat{\theta}_j - \theta_j^*)\dot{\alpha}_j(F_1(h_1(\mathbf{x})))\right] \end{aligned} \tag{18}$$

Now,

$$\begin{aligned} E\left[\frac{\sum_{i=1}^{N} z_j^{(i)} c(\hat{\mathfrak{F}}_j(z_j^{(i)}), \hat{F}_1(h_1(\mathbf{x})); \theta_j^*)}{N}\right] &= E\left[z_j^{(i)} c(\hat{\mathfrak{F}}_j(z_j^{(i)}), \hat{F}_1(h_1(\mathbf{x})); \theta_j^*)\right] \\ &= E[z_j|h(\mathbf{x})]. \end{aligned}$$

For $-0.5 < z_j < 0.5$ we get that

$$\begin{aligned} &E\left[z_j^{(i)} (\hat{\mathfrak{F}}_j(z_j^{(i)}) - \mathfrak{F}_j(z_j^{(i)})) c_1(\mathfrak{F}_j(z_j^{(i)}), F_1(h_1(\mathbf{x})); \theta_j^*)\right] \\ &= E\left[z_j^{(i)} \left(p_j(0)(z_j + 0.5) - \mathfrak{F}_j(z_j^{(i)})\right) c_1(\mathfrak{F}_j(z_j^{(i)}), F_1(h_1(\mathbf{x})); \theta_j^*)\right] = 0. \end{aligned}$$

Similarly, for $0.5 < z_j < 1.5$ we get

$$E\left[z_j^{(i)} (\hat{\mathfrak{F}}_j(z_j^{(i)}) - \mathfrak{F}_j(z_j^{(i)})) c_1(\mathfrak{F}_j(z_j^{(i)}), F_1(h_1(\mathbf{x})); \theta_j^*)\right] = 0$$

By assumption 4.1 and 4.2 the 3rd and 4th term of the right hand side of equation 18 equals to 0. So $\widehat{E[z_j|h(\mathbf{x})]}$ is an unbiased estimator. Also by using assumption 4.3 and WLLN we know that $\widehat{E[z_j|h(\mathbf{x})]}$ is a consistent estimator.

□

Proof of Lemma 2:

Proof. For the general case let $F(h(\mathbf{x})) = (F_1(h_1(\mathbf{x})), \ldots, F_K(h_K(\mathbf{x})))$ and $E[z_j|h(\mathbf{x})] = \frac{\alpha_j(F(h(\mathbf{x}));\theta_j^*)}{c_h(F(h(\mathbf{x}));\theta_j^*)}$.
Similar to lemma 1, we estimate the numerator by $\hat{\alpha}_j(\hat{F}(h(\mathbf{x}));\hat{\theta}_j) = \frac{1}{N}\sum_{i=1}^{N} z_j^{(i)} c(\hat{\mathfrak{F}}_j(z_j^{(i)}), \hat{F}(h(\mathbf{x}));\hat{\theta}_j)$. Now, by using assumptions 2,3 and 4 and conditions of lemma 1 we get

$$\hat{\alpha}_j(\hat{F}(h(\mathbf{x}));\hat{\theta}_j) - \alpha_j(F(h(\mathbf{x}));\theta_j^*) = \frac{1}{N}\sum_{i=1}^{N} \Psi_i(h(\mathbf{x});\theta_j^*) + O(N^{-\frac{1}{2}}) \tag{19}$$

where

$$\begin{aligned}\Psi_i(h(\mathbf{x});\theta_j^*) = {} & z_j^{(i)}(\widehat{\mathfrak{F}}_j(z_j^{(i)}) - \mathfrak{F}_j(z_j^{(i)}))c_1(\widehat{\mathfrak{F}}_j(z_j^{(i)}), F(h(\mathbf{x}));\theta_j^*) \\ & + \sum_{a=1}^{K}\Big(\mathbb{1}(h_a(\mathbf{x}^{(i)}) \leq h_a(\mathbf{x})) - F_a(h_a(\mathbf{x}))\Big)\alpha_{j,a}(F(h(\mathbf{x}));\theta_j^*) + \psi_i'\dot{\alpha_j}(F(h(\mathbf{x}));\theta_j^*).\end{aligned}$$

Similarly the denominator $c_h(F(h(\mathbf{x}));\theta_j^*)$ can be estimate by

$$\hat{c}_h(\hat{F}(h(\mathbf{x}));\hat{\theta}_j) = \frac{1}{N}\sum_{i=1}^{N} c(\hat{\mathfrak{F}}_j(z_j^{(i)}), \hat{F}(h(\mathbf{x}));\hat{\theta}_j)$$

Again, in the similar way we get

$$\hat{c}_h(\hat{F}(h(\mathbf{x}));\hat{\theta}_j) - c_h(F(h(\mathbf{x}));\theta_j^*) = \frac{1}{N}\sum_{i=1}^{N} \Omega_i(h(\mathbf{x});\theta_j^*) + O(N^{-\frac{1}{2}}) \tag{20}$$

where

$$\begin{aligned}\Omega_i(h(\mathbf{x});\theta_j^*) = \sum_{a=1}^{K}\Big(\mathbb{1}(h_a(\mathbf{x}^{(i)}) \leq & h_a(\mathbf{x})) - F_a(h_a(\mathbf{x})))c_{h,a}(F_1(h_1(\mathbf{x})), \ldots, F_K(h_K(\mathbf{x}))) \\ & + \psi_i'\dot{c}_h(F_1(h_1(\mathbf{x})), \ldots, F_K(h_K(\mathbf{x}));\theta_j^*)\Big)\end{aligned}$$

□

From equation 19 and equation 20 and under the conditions of lemma 2, Theorem 3 follows directly.