Supplementary materials for “CeCNN: Copula-enhanced convolutional neural networks in joint prediction of refraction error and axial length based on ultra-widefield fundus images”

Contents

A Technical proofs

A.1 Proof of Theorem 2.1

Proof. It is trivial to show that marginally,

$$Pr\{y_2 = 1|\mathcal{X}\} = Pr\{z_2 > 0|\mathcal{X}\} = \mu_2.$$

Then the remaining is to show that

$$Pr\{y_2 = 1|z_1 = z, \mathcal{X}\} = Pr\{z_2 > 0|z_1 = z, \mathcal{X}\}. \tag{1}$$

Recall the conditional probability under the Gaussian copula, the left-hand side of (1) is

$$Pr\{y_2 = 1|z_1 = z, \mathcal{X}\} = \Phi\left(\frac{\Phi^{-1}(\mu_2) + \rho z}{\sqrt{1-\rho^2}}\right).$$

Based on the conditional distribution of multivariate normal, we obtain

$$z_2 | z_1, \mathcal{X} \sim N(\Phi^{-1}(\mu_2) + \rho z_1, 1 - \rho^2).$$

Therefore,

$$Pr\{z_2 > 0 | z_1 = z, \mathcal{X}\} = 1 - \Phi\left(\frac{-\Phi^{-1}(\mu_2) - \rho z}{\sqrt{1-\rho^2}}\right) = \Phi\left(\frac{\Phi^{-1}(\mu_2) + \rho z}{\sqrt{1-\rho^2}}\right).$$

□

A.2 Propositions of the baseline

We provide the following two propositions of the baseline in both the R-C and the R-R tasks. Their proofs are trivial and omitted.

Proposition A.1 (R-R task)**.** *Suppose the fitted CNN $\hat{\mathcal{G}}$ is trained under the empirical loss such that*

$$\hat{\mathcal{G}} = (\hat{g}_1, \hat{g}_2) = \arg \min_{g_1, g_2 \in \mathcal{F}_{cnn}} \mathcal{L}(g_1, g_2) = \sum_{j=1}^{2} \sum_{i=1}^{n} (y_{ij} - g_j(\mathcal{X}_i))^2.$$

Then $\hat{\mathcal{G}}$ is the nonparametric maximum likelihood estimator (MLE) under the distribution

$$(y_1, y_2)^T | \mathcal{X} \sim N_2\{(g_1(\mathcal{X}), g_2(\mathcal{X}))^T, \sigma^2 I_2\},$$

where σ^2 is an arbitrary positive constant.

Proposition A.2 (R-C task)**.** *Suppose the fitted CNN $\hat{\mathcal{G}}$ is trained under the empirical loss such that*

$$\begin{aligned}\hat{\mathcal{G}} = (\hat{g}_1, \hat{g}_2) &= \arg \min_{g_1, g_2 \in \mathcal{F}_{cnn}} \mathcal{L}(g_1, g_2) \\ &= \sum_{i=1}^{n} \{y_{i1} - g_1(\mathcal{X}_i)\}^2 - \sum_{i=1}^{n} \left\{y_{i2} \log\{\mathcal{S}(g_2(\mathcal{X}_i))\} - (1 - y_{i2}) \log(1 - \mathcal{S}(g_2(\mathcal{X}_i)))\right\},\end{aligned}$$

where $\mathcal{S}$ is the sigmoid function. Then $\hat{\mathcal{G}}$ is the nonparametric maximum likelihood under the distribution

$$y_1 \perp y_2 | \mathcal{X},\ y_1 \sim N(g_1(\mathcal{X}), 1),\ y_2 \sim Bernoulli(g_2(\mathcal{X})).$$

If we view the fitted CNN $\hat{\mathcal{G}}$ as the MLE of $\mathcal{G}$, the above propositions indicate that the use of empirical losses acquiescence the conditionally independent model assumption between (y_1, y_2), given the covariate $\mathcal{X}$. Similarly, one can show that the uncertainty loss (Kendall et al., 2018) is also an MLE under the conditional independence assumption.

A.3 Proof of Theorem 4.1

Notations. For a vector $\boldsymbol{v} \in \mathbb{R}^p$, denote by $|\boldsymbol{v}| = \sum_{i=1}^{p} I(|v_i| > 0)$ the cardinality of $\boldsymbol{v}$, and by $||\boldsymbol{v}||_2$ the L_2 norm of $\boldsymbol{v}$. For a matrix $X \in \mathbb{R}^{n\times K}$, let X_p be the pth column of X and X_{-p} be the remaining colums, for $p = 1, \ldots, K$. Let n be the data size. Let $\boldsymbol{Y} = (y_1, \ldots, y_p)$ be the p-dimensional responses.

Recall the multi-response nonlinear tensor regression model (1)

$$E(\boldsymbol{Y}|\mathcal{X})^T = \mathcal{G}(\mathcal{X}) := (g_1(\mathcal{X}), \ldots, g_p(\mathcal{X}))^T,$$

where the nonlinear function $\mathcal{G}$ is a CNN with p-dimensional outputs, $\mathcal{X} \in \mathbb{R}^{a\times b\times c}$ is the tensor covariate. Specifically, in the bivariate regression-regression (R-R) case, we have

$$\begin{pmatrix} y_1 \\ y_2 \end{pmatrix} = \begin{pmatrix} g_1(\mathcal{X}) \\ g_2(\mathcal{X}) \end{pmatrix} + \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \end{pmatrix}, \ (\epsilon_1, \epsilon_2) \sim N(\mathbf{0}, \Sigma).$$

Let $\mathcal{F}_{cnn}$ be the space of all functions expressed through a CNN. The following propositions are trivial results given by the log-likelihood.

Linear model of fully-connected layer with identity activation

Suppose the a depth of CNN $\mathcal{G}$ is L. Let $H : \mathbb{R}^K \to \mathbb{R}^2$ be the Lth hidden layer of $\mathcal{G}$, i.e. the last F-C layer. Let $D : \mathbb{R}^{a\times b\times c} \to \mathbb{R}^K$ be the stacking of $(1, 2, \ldots, L-1)$ hidden layers. Then the the R-R task with a CNN is rewritten as the following linear model

$$\underset{2\times 1}{\begin{pmatrix} y_1 \\ y_2 \end{pmatrix}} = \underset{2\times K}{\begin{pmatrix} \boldsymbol{w}_1^T \\ \boldsymbol{w}_2^T \end{pmatrix}} D(\mathcal{X}) + \begin{pmatrix} b_1 \\ b_2 \end{pmatrix} + \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \end{pmatrix}, \tag{2}$$

where $D(\mathcal{X}) \in \mathbb{R}^K$ denotes the K feature maps and $\boldsymbol{w}_j \in \mathbb{R}^K$ and $b_j \in \mathbb{R}$ are the weights and bias of the jth output neuron of layer H respectively, for $j = 1, 2$.

We denote by $\mathcal{D}(\mathcal{X}) = (D(\mathcal{X}_1), \dots, D(\mathcal{X}_n))^T \in \mathbb{R}^{n \times K}$ the "design matrix" of feature maps of n tensor observations $\mathcal{X}_i$, for $i = 1, \dots, n$.

We first make the following assumption.

Assumption 1 (Uncovered feature maps)**.** *Let* $\mathbf{1}_{\boldsymbol{w}_j} = \{k : w_{jk} \neq 0\}$ *be the index set of non-zero entries in weights* $\boldsymbol{w}_j$*. Assume that* $\mathbf{1}_{\boldsymbol{w}_1} \setminus \mathbf{1}_{\boldsymbol{w}_2} \neq \emptyset$ *and* $\mathbf{1}_{\boldsymbol{w}_2} \setminus \mathbf{1}_{\boldsymbol{w}_2} \neq \emptyset$*.*

Model (2) is, however, unidentified. To identify the model, we further require the following assumption.

Without loss of generality, we assume that $\mathbf{1}_{\boldsymbol{w}_1} = \{1, 2, \dots, |\boldsymbol{w}_1|\}$ and that $\mathbf{1}_{\boldsymbol{w}_2} = \{K - |\boldsymbol{w}_2| + 1, K - |\boldsymbol{w}_2| + 2, \dots, K\}$. Otherwise, one can simply reorder the elements in $\boldsymbol{w}_j$ to obtain that. Therefore, we have

$$\boldsymbol{w}_1 = (w_{11}, w_{12}, \dots, w_{1|\boldsymbol{w}_1|}, \underbrace{0, \dots, 0}_{K-|\boldsymbol{w}_1|})^T,$$

$$\boldsymbol{w}_2 = (\underbrace{0, \dots, 0}_{K-|\boldsymbol{w}_2|}, w_{21}, w_{22}, \dots, w_{2|\boldsymbol{w}_2|})^T.$$

Under Assumption 1, we can rewrite linear model (2) as following

$$\begin{pmatrix} y_1 \\ y_2 \end{pmatrix} = \underbrace{\begin{pmatrix} D_1(\mathcal{X}) & \mathbf{0}_{2K-|\boldsymbol{w}_1|} \\ \mathbf{0}_{2K-\boldsymbol{w}_1} & D_2(\mathcal{X}) \end{pmatrix}}_{2\times 2K} \underbrace{\begin{pmatrix} \boldsymbol{w}_1 \\ \boldsymbol{w}_2 \end{pmatrix}}_{2K \times 1} + \begin{pmatrix} b_1 \\ b_2 \end{pmatrix} + \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \end{pmatrix}, \tag{3}$$

where $D_j(\mathcal{X})^T = D(\mathcal{X})_{\mathbf{1}_{\boldsymbol{w}_j}} \in \mathbb{R}^{|\boldsymbol{w}_j|}$ denotes the row vector of activated feature maps of $D(\mathcal{X})$ corresponding to non-zero $\boldsymbol{w}_j$, for $j = 1, 2$. Accordingly, model (3) is a seemingly unrelated regression (SUR) model (Zellner, 1962).

It is trivial to show that, under the empirical loss, the fitted weights $\hat{\boldsymbol{w}}_j^{emp}$ are the ordinary least square (OLS) estimator under SUR model (3), denoted as $\hat{\boldsymbol{w}}_j^{ols}$. Meanwhile, with Gaussian assumption on the model error $(\epsilon_{i1}, \epsilon_{i2})$ in SUR (3), the generalized least square (GLS) estimator of $\boldsymbol{w}_j$ is also the MLE of $\boldsymbol{w}_j$ (Kmenta and Gilbert, 1968). Recall that the copula-likelihood loss is the minus log-likelihood. We denote the fitted weights $\hat{\boldsymbol{w}}_j^{cop}$ under the copula-likelihood loss as $\hat{\boldsymbol{w}}_j^{gls}$.

Let $\mathcal{D}_j(\mathcal{X}) \in \mathbb{R}^{n\times|\boldsymbol{w}_j|}$ be the design matrix of activated feature maps of n observations for the jth output. Let R^2_{jk} be the R^2 of the following linear regression associating the columns in $\mathcal{D}_j(\mathcal{X})$

$$\mathcal{D}_j(\mathcal{X})_k = \mathcal{D}_j(\mathcal{X})_{-k}\boldsymbol{\beta}_k + \boldsymbol{r}_{jk},\ j = 1,\ k = 1,\dots,K,\ \boldsymbol{\beta} \in \mathbb{R}^{K-1},\ \boldsymbol{r}_{jk} \in \mathbb{R}^n.$$

Define LR_{1k} as the following linear regression that associates the kth column of $\mathcal{D}_1(\mathcal{X})$ with $\mathcal{D}_2(\mathcal{X})$

$$\mathcal{D}_1(\mathcal{X})_k = \mathcal{D}_2(\mathcal{X})\boldsymbol{\gamma}_{1k} + \boldsymbol{v}_{1k},\ k = 1,\dots,|\boldsymbol{w}_1|,\ \boldsymbol{\gamma}_{1k} \in \mathbb{R}^{|\boldsymbol{w}_2|},\ \boldsymbol{v}_{1k} \in \mathbb{R}^n.$$

Similarly, define LR_{2k} as the linear regression that associates the kth column of $\mathcal{D}_2(\mathcal{X})$ with $\mathcal{D}_1(\mathcal{X})$.

$$\mathcal{D}_2(\mathcal{X})_k = \mathcal{D}_1(\mathcal{X})\boldsymbol{\gamma}_{2p} + \boldsymbol{v}_{2k},\ k = 1,\dots,|\boldsymbol{w}_2|, \boldsymbol{\gamma}_{2p} \in \mathbb{R}^{|\boldsymbol{w}_j|},\ \boldsymbol{v}_{2k} \in \mathbb{R}^n.$$

Let $\boldsymbol{e}_{jk} = (e_{1jk},\dots,e_{njk}) \in \mathbb{R}^n$ be the residual vector of regression LR_{jk}, for $j = 1, 2$, $k = 1,\dots,|\boldsymbol{w}_j|$. Define $E^{(j)} = (\boldsymbol{e}_{j1},\dots,\boldsymbol{e}_{jn}) \in \mathbb{R}^{n\times|\boldsymbol{w}_j|}$ as the matrix of residuals of all regression of LP_{jk}. Define the $2n \times |\boldsymbol{w}_j|$ matrix

$$V^{(j)} = \begin{pmatrix} \mathcal{D}_j(\mathcal{X}) \\ \rho E^{(j)} \end{pmatrix}.$$

Define LR^*_{1k} as the linear regression associating the columns of $V^{(1)}$

$$V^{(1)}_k = V^{(1)}_{-k}\boldsymbol{\eta}_{1k} + \boldsymbol{u}_{1k},\ k = 1,\dots,|\boldsymbol{w}_1|,\ \boldsymbol{\eta}_{1k} \in \mathbb{R}^{|\boldsymbol{w}_1|-1},\ \boldsymbol{u}_{1k} \in \mathbb{R}^{2n}.$$

Similarly, define LR^*_{2k} as the regression associating the columns of $V^{(2)}$

$$V^{(2)}_k = V^{(2)}_{-k}\boldsymbol{\eta}_{2k} + \boldsymbol{u}_{2k},\ k = 1,\dots,|\boldsymbol{w}_2|,\ \boldsymbol{\eta}_{2k} \in \mathbb{R}^{|\boldsymbol{w}_1|-1},\ \boldsymbol{u}_{2k} \in \mathbb{R}^{2n}.$$

Define R^{*2}_{jk} as the R^2 of the linear regression LR^*_{jk}.

Assumption 2 (Consistency)**.** *We assume that* $||\boldsymbol{w}_j||_2 = 1$ *for* $j = 1, 2$. *With this constraint, the fitted layers* $\hat{D}$ *satisfies*

$$\sup_{\mathcal{X}} ||D(\mathcal{X}) - \hat{D}(\mathcal{X})||_2 = O_p(n^{-\alpha})$$

for some $\alpha > 0$*.*

To avoid mathematically too complicated technical proofs, we make the following two assumptions.

Assumption 3 (Asymptotic normality)**.** *Assume that for some $\xi > 0$,*

$$n^{-\xi/2}\{D(\mathcal{X}) - \hat{D}(\mathcal{X})\} \to N(\mathbf{0}, \Lambda)$$

where $\Lambda = \mathrm{diag}(\lambda_1^2, \dots, \lambda_K^2)$.

Assumption 4. *The correlation matrix Γ and the marginal variance (σ_1^2, σ_2^2) are known.*

Proposition A.3. *Under Assumptions 1 to 4, we have:*

i), both $\hat{\boldsymbol{w}}_j^{ols}$ and $\hat{\boldsymbol{w}}_j^{gls}$ are unbiased and consistent to the true $\boldsymbol{w}_{j0}$;

ii), for $k = 1, \dots, K$,

$$\begin{aligned} var(\hat{w}_{jk}^{ols}) &= \frac{\sigma_j^2 + O(n^{-\xi})}{||\mathcal{D}_j(\mathcal{X})_k||_2^2(1 - R_{jk}^2)}, \\ var(\hat{w}_{jk}^{gls}) &= \frac{\sigma_j^2 + O(n^{-\xi})}{(||\mathcal{D}_j(\mathcal{X})_k||_2^2 + \rho^2||E_k^{(j)}||_2^2)(1 - R_{jk}^{*2})}, \end{aligned}$$

for $j = 1, 2$.

Proof. We prove the two assertions separately.

Proof of assertion i). To prove assertion i), we take two steps. In the first step, we start by assuming the design matrices $\mathcal{D}_j(\mathcal{X})$ is known. Then assertion i) is trivial.

Next, we replace $\mathcal{D}_j(\mathcal{X})$ by $\hat{\mathcal{D}}_j(\mathcal{X})$. We have the regression

$$\begin{pmatrix} y_1 \\ y_2 \end{pmatrix} = \begin{pmatrix} D_1(\mathcal{X}) & \mathbf{0}_{2K-|\boldsymbol{w}_1|} \\ \mathbf{0}_{2K-|\boldsymbol{w}_2|} & D_2(\mathcal{X}) \end{pmatrix} \begin{pmatrix} \boldsymbol{w}_1 \\ \boldsymbol{w}_2 \end{pmatrix} + \begin{pmatrix} \hat{D}_1(\mathcal{X}) - D_1(\mathcal{X}) & \mathbf{0}_{2K-|\boldsymbol{w}_1|} \\ \mathbf{0}_{2K-|\boldsymbol{w}_2|} & \hat{D}_2(\mathcal{X}) - D_2(\mathcal{X}) \end{pmatrix} \begin{pmatrix} \boldsymbol{w}_1 \\ \boldsymbol{w}_2 \end{pmatrix}$$
$$+ \begin{pmatrix} b_1 \\ b_2 \end{pmatrix} + \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \end{pmatrix}$$

This can be viewed as an SUR model with noisy covariates. By Assumption 2, since $||\boldsymbol{w}_{j0}||_2 = 1$, we have

$$|\{D_j(\mathcal{X}) - \hat{D}_j(\mathcal{X})\}\boldsymbol{w}_{j0}| = o_p(n^{-\alpha}).$$

Hence, we obtain that

$$\begin{pmatrix} y_1 \\ y_2 \end{pmatrix} = \begin{pmatrix} D_1(\mathcal{X}) & \mathbf{0}_{2K-|\boldsymbol{w}_1|} \\ \mathbf{0}_{2K-|\boldsymbol{w}_2|} & D_2(\mathcal{X}) \end{pmatrix} \begin{pmatrix} \boldsymbol{w}_1 \\ \boldsymbol{w}_2 \end{pmatrix} + \begin{pmatrix} b_1 \\ b_2 \end{pmatrix} + \begin{pmatrix} \epsilon_1 + o_p(n^{-\alpha}) \\ \epsilon_2 + o_p(n^{-\alpha}), \end{pmatrix} \tag{4}$$

which yields the consistency in assertion *i*).

By taking expectation with respect to $\boldsymbol{w}_j$, we have the following mean regression

$$E\begin{pmatrix} y_1 \\ y_2 \end{pmatrix} = \begin{pmatrix} D_1(\mathcal{X}) & \mathbf{0}_{2K-|\boldsymbol{w}_1|} \\ \mathbf{0}_{2K-|\boldsymbol{w}_1|} & D_2(\mathcal{X}) \end{pmatrix} \begin{pmatrix} E\{\hat{\boldsymbol{w}}_1\} \\ E\{\hat{\boldsymbol{w}}_2\} \end{pmatrix} + \begin{pmatrix} b1 + E_{\hat{D}_1}(\{\hat{D}_1(\mathcal{X}) - D_1(\mathcal{X})\}\boldsymbol{w}_{10}) \\ b2 + E_{\hat{D}_2}(\{\hat{D}_2(\mathcal{X}) - D_2(\mathcal{X})\}\boldsymbol{w}_{20}) \end{pmatrix}. \tag{5}$$

Taking the OLS or GLS estimators to (5) we find that the unbiasedness still holds. That is, we remove the bias caused by $\hat{D}_j$ to the bias parameters b_j in the jth output neuron.

Proof of assertion *ii*). We also take two steps to prove assertion *ii*). We first assume $\mathcal{D}(\mathcal{X})$ is known. According to Equations (10.20) and (10.21) in Baltagi (2011), we have

$$\text{var}(\hat{w}_{jp}^{ols}) = \frac{\sigma_j^2}{\sum_{i=1}^n ||\mathcal{D}_j(\mathcal{X})_k||_2^2(1-R_{jk}^2)},$$
$$\text{var}(\hat{w}_{jk}^{gls}) = \frac{\sigma_j^2}{||\mathcal{D}_j(\mathcal{X})_k||_2^2 + \rho^2||V_k^{(j)}||_2^2\}(1-R_{jk}^{*2})}.$$

Then we replace $D(\mathcal{X})$ by $\hat{D}(\mathcal{X})$. Based on (4), we have

$$\begin{pmatrix} y_1 \\ y_2 \end{pmatrix} = \begin{pmatrix} D_1(\mathcal{X}) & \mathbf{0}_{2K-|\boldsymbol{w}_1|} \\ \mathbf{0}_{2K-|\boldsymbol{w}_2|} & D_2(\mathcal{X}) \end{pmatrix} \begin{pmatrix} \boldsymbol{w}_1 \\ \boldsymbol{w}_2 \end{pmatrix} + \begin{pmatrix} b_1 \\ b_2 \end{pmatrix} + \begin{pmatrix} Z_1 \\ Z_2 \end{pmatrix},$$

where

$$Z_j = \epsilon_j + \boldsymbol{w}_j^T\{D(\mathcal{X}) - \hat{D}(\mathcal{X})\}.$$

Assumption 3 yields that

$$Z_j \sim N(0, \sigma_j^2 + n^{-\xi}\boldsymbol{w}_j^T\Lambda\boldsymbol{w}_j).$$

Therefore,

$$\text{Var}(Z_j) \leq \sigma_j^2 + \max_{p=1,\ldots,K}(\lambda_p)n^{-\xi},$$

which completes the proof. □

By definition, R_{jk}^2 reflects the multicollinearity among the columns of $\mathcal{D}_j(\mathcal{X})$; R_{1k}^{*2} incorporates the similarity between $\mathcal{D}_1(\mathcal{X})_k$ and $\mathcal{D}_2(\mathcal{X})$; R_{2k}^{*2} incorporates the similarity between $\mathcal{D}_2(\mathcal{X})_k$ and $\mathcal{D}_1(\mathcal{X})$. Therefore, if the gap between $\mathcal{D}_1(\mathcal{X})$ and $\mathcal{D}_2(\mathcal{X})$ is large, it is more likely that $R_{jk}^2 \geq R_{jk}^{*2}$. In an extreme case where $\mathcal{D}_1(\mathcal{X}) \perp \mathcal{D}_2(\mathcal{X})$, we surely have $R_{jk}^2 = R_{jk}^{*2}$. That enables us to proof Theorem 4.1 in the manuscript.

Proof of Theorem 4.1

Proof. By the definition of R_{jk}^{*2}, based on Baltagi (2011, Problem 7), if the columns of $\mathcal{D}_1(\mathcal{X})$ are orthogonal to the columns of $\mathcal{D}_2(\mathcal{X})$, then $R_{jk}^{*2} = R_{jk}^2$ for all p. And consequently, $Var(\hat{w}_{jk}^{gls}) < Var(\hat{w}_{jk}^{ols})$ for all $k = 1, \ldots, K$. Thus, it suffices to show that

$$Pr\{\mathcal{D}_1(\mathcal{X}) \perp \mathcal{D}_2(\mathcal{X})\} \rightarrow 1.$$

This is true due to the well-known fact of almost orthogonality of independent vectors in high-dimensional spaces (Vershynin, 2018, Remark 3.2.5).

□

B Comparison with the uncertainty loss

In this section, we first compare CeCNN with the uncertainty loss (Kendall et al., 2018) on our UWF dataset in subsection B.1. The brief introduction to the loss in both R-R and R-C tasks is deferred to subsection B.2.

B.1 Comparison

We choose the ResNet (He et al., 2016) as the backbone CNN for both losses, and set the baseline as the ResNet equipped with the empirical loss. The results of 10 rounds 5-fold cross validation in regression-classification (R-C) and regression-regression (R-R) tasks are presented in Figures 1 and 2, respectively. From the two figures, we find that CeCNN significantly and robustly improves the baseline in both R-C and R-R tasks. In terms of the uncertainty loss, the results show that it is suitable for the R-R task, while fails to improve the baseline in the R-C task. In the R-R task, the uncertainty loss is comparable with CeCNN in AL prediction and performs slightly better in SE prediction. We conjecture the reason is that the scale parameters σ_j play a more important role in the bivariate Gaussian density, compared with the correlation coefficient ρ. Since the uncertainty loss takes a simpler form than the copula-likelihood loss, it is expected to enjoy a smaller optimization error than the copula-likelihood loss of the CeCNN. This explains why the uncertainty loss works for the R-R task.

However, in the R-C task, the performance of the uncertainty loss is unsatisfactory. The uncertainty loss only has higher classification accuracy than baseline, but suffers from a larger prediction error in the regression target. Meanwhile, the high classification accuracy under the uncertainty loss sacrifices the AUC in classification, indicating that it does not strike a balance between sensitivity and specificity. We conjecture that the uncertainty parameter σ_2 in (Kendall et al., 2018, Eq. (10)) may not have a statistical interpretation, incurring poor learning on σ_2 from the data. As we show in Theorem 2.1, the Gaussian score of the binary response y_2 is marginally standard normal, indicating that the variance is always 1. In summary, we feel that the uncertainty may not be suitable for the binary responses.

B.2 The uncertainty loss

We briefly introduce the uncertainty loss (Kendall et al., 2018) in both R-R and R-C tasks, respectively.

Uncertainty to weigh in the R-R task Let $\mathbf{W}$ be the weights in a backbone CNN; let $\mathbf{f}^{\mathbf{W}}$ denote the CNN with weights $\mathbf{W}$. In the R-R tasks, for continuous responses (y_1, y_2),

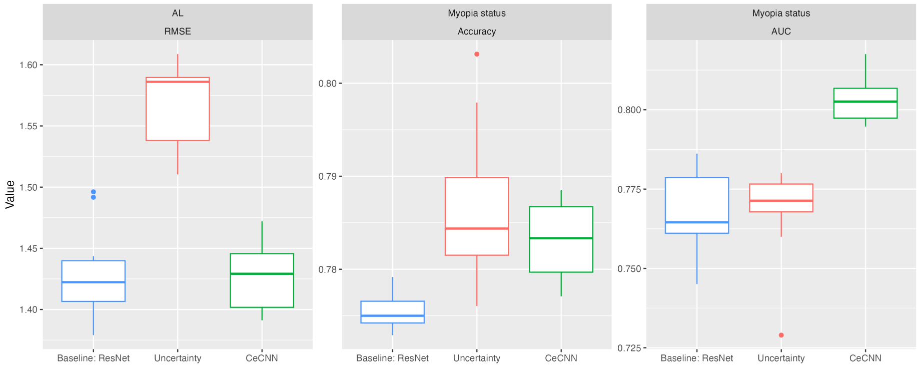

Figure 1: Box plots of MSE, Accuracy and AUC in 10 rounds of 5-fold validation of the R-C tasks for ResNet as the backbone model. “ResNet”: the baseline, ResNet with the empirical loss; “Uncertainty”: ResNet with uncertainty loss; “Copula”: CeCNN with ResNet backbone CNN.

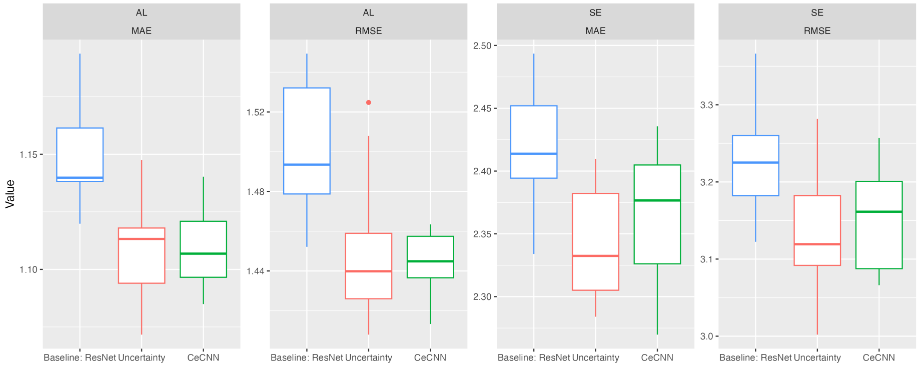

Figure 2: Box plots of RMSE and MAE in 10 rounds of 5-fold validation of the R-R tasks for ResNet as the backbone model.“ResNet”: the baseline, ResNet with the empirical loss; “Uncertainty”: the backbone CNN with uncertainty loss; “Copula”: CeCNN with ResNet backbone CNN.

the loss is given by

$$\begin{aligned}\mathcal{L}(\mathbf{W},\sigma_1,\sigma_2) &= -\log p(y_1,y_2|\mathbf{f}^{\mathbf{W}}(\mathcal{X}))\\ &\propto \frac{1}{2\sigma_1^2}||y_1-\mathbf{f}^{\mathbf{W}}(\mathcal{X})||^2+\frac{1}{2\sigma_2^2}||y_2-\mathbf{f}^{\mathbf{W}}(\mathcal{X})||^2+\log\sigma_1\sigma_2\\ &= \frac{1}{2\sigma_1^2}\mathcal{L}_1(\mathbf{W})+\frac{1}{2\sigma_2^2}\mathcal{L}_2(\mathbf{W})+\log\sigma_1\sigma_2 \qquad (6)\end{aligned}$$

where $\mathcal{L}_1(\mathbf{W}) = ||y_1-\mathbf{f}^{\mathbf{W}}(\mathcal{X})||^2$ for the loss of the first output variable, and similarly for $\mathcal{L}_2(\mathbf{W})$. Obviously, the uncertainty parameter σ_j represents the conditional standard deviation of the jth output y_j given covariate $\mathcal{X}$.

Uncertainty to weigh in the R-C task In the R-C task, the uncertainty loss is given by

$$\begin{aligned}
\mathcal{L}(\mathbf{W}, \sigma_1, \sigma_2) &= -\log p(y_1, y_2 = c|\mathbf{f}^{\mathbf{W}}(\mathcal{X})) \\
&= -\log \mathcal{N}(y_1; \mathbf{f}^{\mathbf{W}}(\mathcal{X}), \sigma_1^2) \cdot \text{Softmax}(y_2 = c; \mathbf{f}^{\mathbf{W}}(\mathcal{X}), \sigma_2) \\
&= \frac{1}{2\sigma_1^2}||y_1 - \mathbf{f}^{\mathbf{W}}(\mathcal{X})||^2 + \log \sigma_1 - \log p(y_2 = c|\mathbf{f}^{\mathbf{W}}(\mathcal{X}), \sigma_2) \\
&\approx \frac{1}{2\sigma_1^2}\mathcal{L}_1(\mathbf{W}) + \frac{1}{\sigma_2^2}\mathcal{L}_2(\mathbf{W}) + \log \sigma_1 + \log \sigma_2, \qquad (7)
\end{aligned}$$

where $\mathcal{L}_1(\mathbf{W}) = ||y_1 - \mathbf{f}^{\mathbf{W}}(\mathcal{X})||^2$ is the Euclidean loss of $\mathbf{y}_1$, $\mathcal{L}_2(\mathbf{W}) = -\log \text{Softmax}(y_2, \mathbf{f}^{\mathbf{W}}(\mathcal{X}))$ for the cross entropy loss of $\mathbf{y}_2$.

C Comparison with iteratively updating the copula parameter

In this section, we compare the CeCNN with iteratively updating the copula parameter. Since this is closely related to the alternating minimization (Jain and Tewari, 2015), we call it alternating minimization (AM) for abbreviation hereafter. We summarize the AM algorithm for the CeCNN as follows.

We present the comparisons between the CeCNN with the AM updating algorithm in the R-C and R-R tasks in Figure 3. We clearly find that the CeCNN outperforms the AM algorithm in all metrics for both the R-C and the R-R tasks.

Algorithm 1 Alternating minimization for the CeCNN

```
Set $T$, the total number of epochs to update; set initial copula parameters $\Gamma_0$ and $\sigma_0$.
for $t = 1, \dots, T$ do
    Update the backbone CNN by optimizing the copula-likelihood loss parameterized by $\Gamma_{t-1}$ and $\sigma_{t-1}$ (in the R-C and the R-R tasks respectively).
    Update $\Gamma_t$ and $\sigma_t$ by the residuals and Gaussian scores computed from the updated CNN.
end for
```

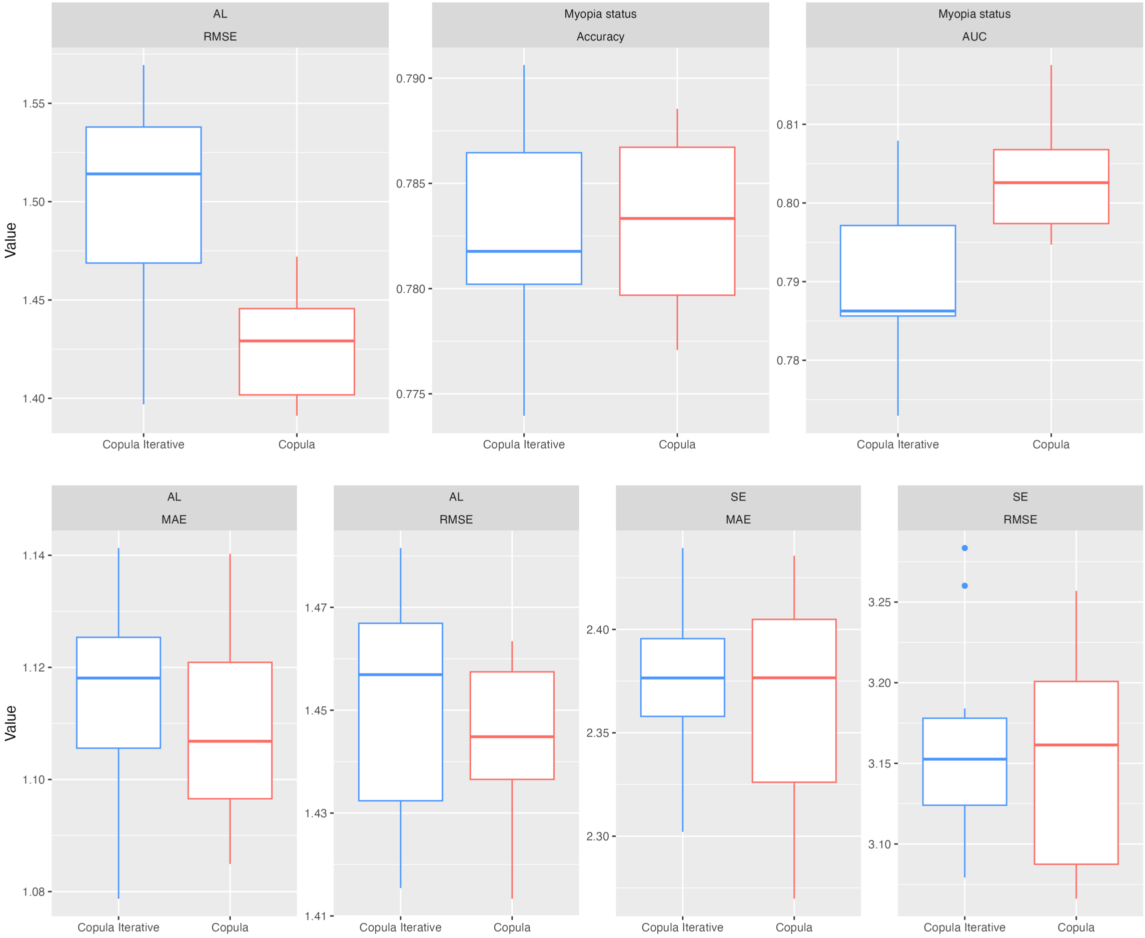

Figure 3: Comparison between the CeCNN and the AM algorithm. Top: the R-C task; bottom: the R-R task.

D Loss trace with regularization

In this section, we report the behaviors of the backbone CNN with different candidates of the tuning parameter λ for regularization. We present the traces of training/validation/testing losses in Figures 4, 5, and 6, with three different candidates.

We find that on our UWF dataset, with a small λ, the overfitting issue is still serious. Otherwise, with a large λ, the predictive performance on the test set is worse than that of no regularization, as shown in Table 1. Therefore, we do not consider regularization in our application to our UWF dataset.

Table 1: Comparison of Regularization Effects on RMSE and MAE of AL and SE

	Average RMSE		Average MAE	
	AL	SE	AL	SE
$\lambda = 0$ (No regularization)	1.401	3.017	1.077	2.293
$\lambda = 0.01$ (Minor regularization)	1.529	3.313	1.172	2.520
$\lambda = 1$ (Strong regularization)	1.998	4.46	1.545	3.389

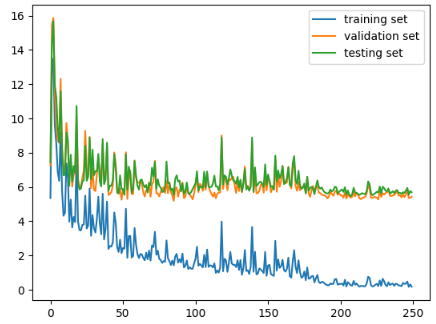

Figure 4: Loss trace of $\lambda = 0$ (no regularization).

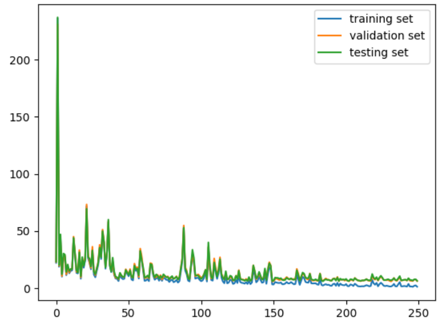

Figure 5: Loss trace of $\lambda = 0.01$ (weak regularization).

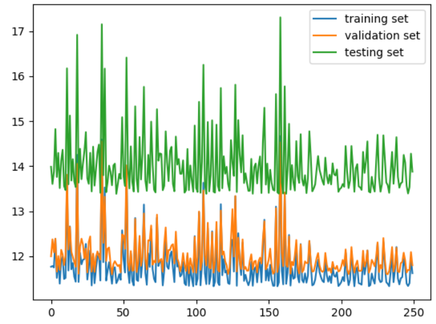

Figure 6: Loss trace of $\lambda = 1$ (strong regularization).

References

Baltagi, B. H. (2011). Seemingly unrelated regressions. In *Econometrics*. Springer. 7, 8

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. 8

Jain, P. and Tewari, A. (2015). Alternating minimization for regression problems with vector-valued outputs. *Advances in Neural Information Processing Systems*, 28. 11

Kendall, A., Gal, Y., and Cipolla, R. (2018). Multi-task learning using uncertainty to weigh losses for scene geometry and semantics. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7482–7491. 3, 8, 9

Kmenta, J. and Gilbert, R. F. (1968). Small sample properties of alternative estimators of seemingly unrelated regressions. *Journal of the American Statistical Association*, 63(324):1180–1200. 4

Vershynin, R. (2018). *High-dimensional probability: An introduction with applications in data science*, volume 47. Cambridge university press. 8

Zellner, A. (1962). An efficient method of estimating seemingly unrelated regressions and tests for aggregation bias. *Journal of the American statistical Association*, 57(298):348–368. 4