Hypothesis Testing for Class-Conditional Noise Using Local Maximum Likelihood

Weisong Yang[1*], Rafael Poyiadzi[2*], Niall Twomey[1], Raul Santos-Rodriguez[1]
[1]University of Bristol
[2]GSK
{ws.yang, niall.twomey, enrsr}@bristol.ac.uk, rafaelpoyiadzi@gmail.com

Abstract

In supervised learning, automatically assessing the quality of the labels before any learning takes place remains an open research question. In certain particular cases, hypothesis testing procedures have been proposed to assess whether a given instance-label dataset is contaminated with class-conditional label noise, as opposed to uniform label noise. The existing theory builds on the asymptotic properties of the *Maximum Likelihood Estimate* for parametric logistic regression. However, the parametric assumptions on top of which these approaches are constructed are often too strong and unrealistic in practice. To alleviate this problem, in this paper we propose an alternative path by showing how similar procedures can be followed when the underlying model is a product of *Local Maximum Likelihood Estimation* that leads to more flexible nonparametric logistic regression models, which in turn are less susceptible to model misspecification. This different view allows for wider applicability of the tests by offering users access to a richer model class. Similarly to existing works, we assume we have access to anchor points which are provided by the users. We introduce the necessary ingredients for the adaptation of the hypothesis tests to the case of nonparametric logistic regression and empirically compare against the parametric approach presenting both synthetic and real-world case studies and discussing the advantages and limitations of the proposed approach.

Introduction

Data quality checks are an essential part of the machine learning process, as a dataset of poor quality would lead to less trustworthy inferences. There have been plenty of studies targeting different aspects of data quality, e.g. missing value imputation (Jarrett et al. 2022; Lall and Robinson 2022), anomaly detection (Xia et al. 2022; Hilal, Gadsden, and Yawney 2022; Schmidl, Wenig, and Papenbrock 2022), or class imbalance (Sauber-Cole and Khoshgoftaar 2022), etc.

In this work, we focus on assessing the annotation quality of the dataset. In many real-world applications, data annotation is inherently suboptimal, especially when data is harvested from the web (Li et al. 2017; Gong et al. 2014), or when the annotation is not carried out by domain experts, such as in the case of crowdsourcing.

Here, our work aims to provide machine learning practitioners with tools for assessing the quality of the labels of a binary supervised dataset of the form: $(\boldsymbol{X}, \boldsymbol{y}) = \{(\boldsymbol{x}_i, y_i)\}_{i=1}^N \in (\mathbb{R}^d \times \{-1, 1\})$. These tools are of the form of hypothesis tests and are meant to take place after the data collection and annotation steps are done. The tests provide information regarding the presence of *class-conditional label noise*, as opposed to *uniform label noise*. Class-conditional label noise (CCN) refers to the setting where the per-class noise rates are different, i.e. $\mathbb{P}(\tilde{Y} = -1 \mid Y = 1) = \alpha$ and $\mathbb{P}(\tilde{Y} = 1 \mid Y = -1) = \beta$, with $\alpha \neq \beta$. Uniform label noise (UN) refers to the setting where per-class noise rates are identical, i.e. $\mathbb{P}(\tilde{Y} = 1 \mid Y = -1) = \mathbb{P}(\tilde{Y} = -1 \mid Y = 1) = \tau$.

As other efforts in this area, including those that try to estimate the noise rates directly, a necessary ingredient for our tests is *anchor points*, i.e. instances ($\boldsymbol{x}$), with a known (or, approximately known) *posterior* ($\eta(\boldsymbol{x}) = \mathbb{P}(Y = 1 \mid X = x)$). These need to be provided by the user, probably sourced by a domain expert.

In Poyiadzi et al. (2022), the authors introduced a framework that follows from the asymptotic properties of Maximum Likelihood Estimation (MLE) of the logistic regression model and thus its correctness relies on whether the assumptions of the model are met. In practice, the linear assumptions behind the method mean that, unless there is an element of control over the data generating process, the blind application of the test can lead to suboptimal outcomes. In this paper, we take an alternative view by considering more flexible nonparametric models. Instead of relying on the asymptotics of MLE for a parametric model, we build our approach on top of a nonparametric estimation of the underlying regression function based on local likelihood models, such as local polynomial regression, which importantly is more flexible and robust against model misspecification.

In the empirical analysis, we perform both experiments on synthetic data and real-world data, focusing on understanding the benefits and limitations of the approach as compared to parametric baselines. We show that only if the practitioner knows the data generation process beforehand and can ensure that all the assumptions of the parametric test hold, the

*These authors contributed equally.

parametric test can be used. If any of the assumptions of the parametric test does not hold or if they are unknown at all, we should rely on the nonparametric test, which is more robust against model misspecification. We then present a first real-world case study on a smart home dataset. Smart home data can be a typical example that involves dealing with noisy labels, since smart home systems are expected to be scalable and affordable, and most likely non-specialists will be responsible for the data annotation. Consequently, assessing the annotation quality of new datasets will become necessary and important, as it might affect subsequent analyses and conclusions.

The contributions of this paper are summarized as follows:

- We extend the parametric form of the hypothesis test for class-conditional label noise proposed in Poyiadzi et al. (2022), and consider a nonparametric estimation of the underlying regression function based on local likelihood models.
- We thoroughly compare the strengths and weaknesses of these two methods respectively, and provide guidelines for machine learning practitioners to know which one to use given a dataset.
- The effectiveness of our hypothesis test is further illustrated by performing the test on a real-world dataset. We additionally discuss some practical considerations when designing such a test in real life.

Background

We present here a summary of the main results of local likelihood models that will be used to derive the tests. We first outline some classical results, before introducing kernel-based extensions. Later, we highlight the bias injected by these methods, as well as methods to mitigate these.

We refer the reader to Wasserman (2006, Chapter 5) and Loader (2006, Chapter 4) for a complete presentation. We consider a nonparametric version of logistic regression, with data $(\boldsymbol{X}, \boldsymbol{y}) = \{(\boldsymbol{x}_i, y_i)\}_{i=1}^N \in (\mathbb{R}^d \times \{-1, 1\})$. We want to assess whether the labels have been corrupted with class-conditional noise, as opposed to uniform noise. We assume:

$$Y_i \sim Bernoulli\left(r\left(x_i\right)\right)$$

$$\text{with } r(x_i) = \mathbb{P}(Y_i = 1 \mid X_i = x_i) = p_i$$

for a smooth function $r(x)$ with $0 \leq r(x) \leq 1$. Our likelihood function then is:

$$\mathcal{L}(r) = \prod_{i=1}^n r(x_i)^{Y_i}(1 - r(x_i))^{1-Y_i},$$

and by using: $\xi(x) = \text{logit}(r(x)) = \log \frac{r(x)}{1-r(x)}$, the log-likelihood function:

$$\ell(r) = \sum_{i=1}^n \ell(Y_i, \xi(x_i))$$

$$\text{with } \ell(y, \xi) = \log\left[\left(\frac{e^\xi}{1+e^\xi}\right)^y \left(\frac{1}{1+e^\xi}\right)^{1-y}\right]$$

To proceed with the estimation of the regression function $r(x)$ near x, we approximate the function at z near x with the local logistic function:

$$r(z) \approx g\left(\langle \beta, A_p(z - x)\rangle\right)$$

$$g(v) = \text{logit}^{-1}(v) = \frac{\exp(v)}{1 + \exp(v)}$$

where $A_p(\cdot)$ is the vector of polynomial basis functions of order p. For example, $\beta_0 + \beta_1 \cdot x$, for $x \in \mathbb{R}^1$. Borrowing the example from Loader (2006, See Eq. 2.9):

$$A_2\left([v_0,\ v_1]\right) \ = \ [1,\ v_0,\ v_1,\ \frac{1}{2}v_0^2,\ v_0 v_1,\ \frac{1}{2}v_1^2] \quad (1)$$

We can now define the local log-likelihood:

$$\ell_x(\beta) = \sum_{i=1}^n w_{i,h}(x)\ell\left(y_i, \langle \beta, A_p\left(x_i - x\right)\rangle\right) \quad (2)$$

where $w_{i,h}(x)$ is the weight, or the kernel, and is of the form: $w_{i,h}(x) = K\left(\frac{x - x_i}{h}\right)$, where h is called the bandwidth. We use the Gaussian kernel, where: $K(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}$. See Section 4.2 of Wasserman (2006) for more information on kernels.

Let $\hat{\beta} = \arg\max \ell_x(\beta)$. The nonparametric estimate of $r(x)$ is then: $\hat{r}(x) = \frac{\exp \hat{\beta}_0}{1 + \exp \hat{\beta}_0}$. In the case of a local-likelihood model, we no longer assume a parametric form, as in a (global) likelihood model, but rather fit the polynomial model locally.

We now discuss common issues and limitations of nonparametric regression models.

The Bias Problem A well-known problem of smoothing methods is a non-vanishing term that introduces bias into the asymptotic normal distribution of the estimand. The implication is that the confidence interval will not be centred around the true function $r(x)$, but rather around: $\bar{r}_n(x) = \mathbb{E}\left[\hat{r}_n(x)\right]$. A few remedies, include: (a) Do nothing, (b) Estimate the bias function: $\bar{r}_n(x) - r(x)$, (c) Undersmooth, and, (d) Bootstrap bias correction. In this work, we follow (a) for illustrative purposes. More information can be found in Section 5.7 of Wasserman (2006).

Testing for Linearity In this work, we introduce an extension of the tests in Poyiadzi et al. (2022) that replaces the use of a parametric model, with a nonparametric model. For a practitioner, it would be necessary to know whether a parametric (or linear) fit is suitable for their needs. One option is an F-test (Loader 2012) with the null: $\mathcal{H}_0 : r(x) = \beta_0 + \beta_1 x$, for β_0, β_1, against the alternative that $\mathcal{H}_0$ is false (Wasserman 2006).

Bandwidth Selection (Model Selection) This can be chosen by leave-one-out cross-validation (LOO-CV) as follows,

$$\text{LOO-CV} = \sum_{i=1}^n \ell\left(Y_i, \widehat{\xi}_{(-i)}\left(x_i\right)\right) \quad (3)$$

In our experiments, we use sub-sampled LOO-CV[1].

[1]There are also approximations of it as presented in Wasserman (2006, Chapter 5) and Loader (2006, Definition 4.4) with describes

Hypothesis Tests Based on Local Maximum Likelihood Estimation

In this section, we introduce the necessary ingredients for the proposed hypothesis tests. These hypothesis tests can be used to provide evidence against the *null hypothesis*: uniform label noise, and for the *alternative hypothesis*: class-conditional label noise. These tests require anchor points to be provided by the user. As we will now discuss, anchor-points are instances for which the true posterior is known to either be strictly, or approximately, $1/2$.

We first introduce anchor points and useful identities. We then present properties on the asymptotic distribution of solutions of local maximum likelihood models discussed earlier, and then show how these can be used to construct tests in the cases of having: (1) a single strict anchor-point, (2) multiple strict anchors ($\eta(x_i) = 1/2,\ \forall i \in [k],\ k > 1$, where k is the number of anchor points), (3) multiple relaxed anchors ($\eta(x_i) \approx 1/2,\ \forall i \in [k],\ k > 1$),

Anchor Points

In binary classification, we are interested in the posterior predictive distribution: $\eta(\boldsymbol{x}) = \mathbb{P}(Y = 1 \mid X = \mathbf{x})$. We will denote the posterior under label noise with: $\tilde{\eta}(\mathbf{x}) = \mathbb{P}(\tilde{Y} = 1 \mid X = \mathbf{x})$. Under the two settings: uniform noise $UN(\tau)$ and class-conditional noise $CCN(\alpha, \beta)$, we have[2]:

$$\tilde{\eta}(\boldsymbol{x}) = \begin{cases} (1-\alpha-\beta)\cdot\eta(\boldsymbol{x}) + \beta & \text{if (CCN)} \\ (1-2\tau)\cdot\eta(\boldsymbol{x}) + \tau & \text{if (UN)} \end{cases} \tag{4}$$

Anchor points are instances for which we are provided with their true posterior. In this work, we are interested in anchor points for which the true posterior is $1/2$. Under this setting, we have:

$$\eta(\boldsymbol{x}) = 1/2 \quad \rightarrow \quad \tilde{\eta}(\boldsymbol{x}) = \frac{1-\alpha+\beta}{2} \tag{5}$$

We will also develop theory around *relaxed* anchor-points (as opposed to *strict* anchor-points discussed above), for which $\eta(\boldsymbol{x}) \approx 1/2$. More specifically, $\eta(\boldsymbol{x}_i) = 1/2 + \epsilon_i$, for $\epsilon_i \sim \mathbb{U}([-\delta,\ \delta])$, with $0 \leq \delta \leq 0.50$, and ideally $0 \leq \delta \ll 1$ (respecting $0 \leq \eta(\boldsymbol{x}) \leq 1$).

Asymptotics

A consistent estimator of the variance of the estimator for a local-likelihood model has the form (Frölich 2006):

$$B = \sum_{i=1}^{n} w_i p_i (1-p_i) x_i x_i^\top \quad \& \quad C^3 = \sum_{i=1}^{n} (y_i - p_i)^2 x_i x_i^\top w_i^2$$

$$\hat{\mathbb{V}}[\hat{\beta}_x] = B^{-1} C B^{-1} \tag{6}$$

the Akaike Information Criterion (AIC) adjusted for local likelihood.

[2]See of Poyiadzi et al. (2022, Section 7.1) for full derivation.

[3]See Frölich (2006, Section 2.3) for a discussion on similar estimators found in the literature. As noted in said paper, all should result in similar approximations under a small bandwidth.

and then with the application of the Delta Method (See for example Van der Vaart (2000, Chapter 3)) we have the variance for the predictions:

$$\mathbb{V}[\hat{\eta}(x)] \approx \eta(x)^2 (1-\eta(x))^2 \cdot \hat{\mathbb{V}}[\hat{\beta}_x]_{0,0} = \frac{1}{16} v(x) \tag{7}$$

For the covariance we have:

$$B_{\boldsymbol{x}_j} = \sum_{i=1}^{n} w_{i,\boldsymbol{x}_j} p_{i,\boldsymbol{x}_j} (1 - p_{i,\boldsymbol{x}_j}) x_i x_i^\top$$

$$C_{\boldsymbol{x}_j} = \sum_{i=1}^{n} \left(y_i - p_{i,\boldsymbol{x}_j}\right) x_i w_{i,\boldsymbol{x}_j}$$

Note that we now make the independence on $\boldsymbol{x}$ explicit.

$$Cov[\hat{\beta}_{x_k}, \hat{\beta}_{x_j}] = B_{\boldsymbol{x}_k}^{-1} C_{\boldsymbol{x}_k} C_{\boldsymbol{x}_j}^\top B_{\boldsymbol{x}_j}^{-1} \tag{8}$$

$$cov(\hat{\eta}(x_j), \hat{\eta}(x_k)) = \frac{1}{16} Cov[\hat{\beta}_{x_k}, \hat{\beta}_{x_j}]_{0,0} = \frac{1}{16} cov(x_k, x_j) \tag{9}$$

Equations 7 and 8 can be obtained by considering expectations[4]of Loader (2006, Eq. 4.18).

A Nonparametric Hypothesis Test for Class-Conditional Label Noise

We now define our null hypothesis ($\mathcal{H}_0$) and (implicit) alternative hypothesis ($\mathcal{H}_1$) as follows:

$$\mathcal{H}_0 : \alpha = \beta \quad \& \quad \mathcal{H}_1 : \alpha \neq \beta \tag{10}$$

The next sub-sections outline the varied tests under strict, multiple, and relaxed conditions.

Single Strict Anchors Under the null hypothesis, we have the following for the estimated posterior of the anchor:

$$\mathcal{H}_0 : \hat{\eta}(\boldsymbol{x}) \sim \mathcal{N}\left(\frac{1}{2},\ \frac{1}{16} \cdot v(\boldsymbol{x})\right) \tag{11}$$

Multiple Strict Anchors Let us consider a set $\mathcal{A}_{1/2}^k = \{x \mid \eta(x) = 1/2\}$, with $|\mathcal{A}_{1/2}^k| = k$. Let $\hat{\eta}_i$ correspond to the ith instance in $\mathcal{A}_{1/2}^k$. Then for $\bar{\eta} = \frac{1}{k}\sum_{i=1}^{k} \hat{\eta}_i$ we have:

$$\mathbb{E}\bar{\eta} = \mathbb{E}[\frac{1}{k}\sum_{i=1}^{k} \hat{\eta}_i] = \frac{1}{k}\sum_{i=1}^{k} \mathbb{E}\hat{\eta}_i = \frac{1}{2}$$

$$\begin{aligned} \mathbb{V}\bar{\eta} &= \mathbb{V}[\frac{1}{k}\sum_{i=1}^{k} \hat{\eta}_i] \\ &= \frac{1}{k^2}\left(\sum_{i=1}^{k} \mathbb{V}\hat{\eta}_i + 2 \cdot \sum_{i=1, j>i}^{k} Cov(\hat{\eta}_i, \hat{\eta}_j)\right) \\ &= \frac{1}{16k^2}\left(\sum_{i=1}^{k} v(x_i) + 2 \cdot \sum_{i=1, j>i}^{k} cov(x_k, x_j)\right) \\ &= \frac{1}{16k^2} var(x_{[1:k]}) \end{aligned}$$

[4]In the form of $\mathbb{E}\left[(X - \mathbb{E}X)^2\right]$ for Eq.7 and in the form of $\mathbb{E}\left[(X - \mathbb{E}X)(Y - \mathbb{E}Y)\right]$ for Eq.8.

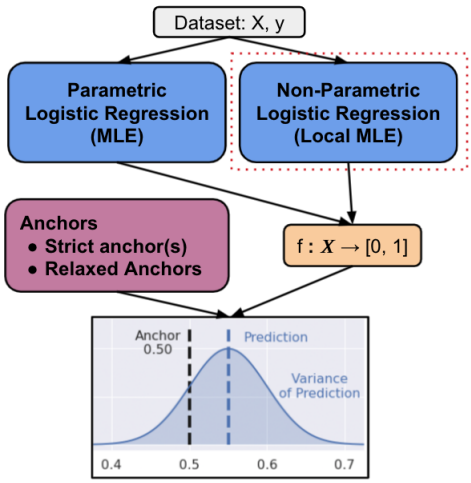

Figure 1: A comparison between the parametric test proposed by Poyiadzi et al. (2022) and ours, highlighting that our proposed method offers a richer model class.

Multiple Relaxed Anchors Providing anchor points with the property that $\eta(x) = 1/2$ might be challenging, but this strict requirement can be relaxed such that, i.e. $\eta(x) \approx 1/2$. Following Poyiadzi et al. (2022), we model this with: $\eta(x) = 1/2 + \epsilon$, where $\epsilon \sim \mathbb{U}([-\delta, +\delta])$ $(0 \leq \eta(x) \leq 1)$, with $0 \leq \delta \leq 0.50$ denoting the degree of relaxation. We now have two sources of stochasticity: from the model (denoted with the subscript of β), and from the fact that anchors are not perfect (denoted with a subscript of A). We then have:

For one (relaxed) anchor point:

$$\mathbb{E}_\beta \eta = \frac{1}{2} + \epsilon, \quad \text{and} \quad \mathbb{E}_A \mathbb{E}_\beta \eta = \frac{1}{2}$$

and, with the use of the *law of total variance*:

$$\begin{aligned}\mathbb{V}\eta &= \mathbb{E}[\mathbb{V}[\eta|\epsilon]] + \mathbb{V}[\mathbb{E}[\eta|\epsilon]] \\ &= \mathbb{E}\left[\left(\frac{1}{16} - \frac{\epsilon^2}{2}\right) \cdot v(x)\right] + \mathbb{V}[\frac{1}{2} + \epsilon] \\ &= \left(\frac{1}{16} - \frac{\delta^2}{6}\right) \cdot v(x) + \frac{\delta^2}{6}\end{aligned}$$

and, then for the case of multiple relaxed anchors:

$$\begin{aligned}\mathbb{V}\bar{\eta} &= \mathbb{E}\left[\left(\frac{1}{16} - \frac{\epsilon^2}{2}\right) \cdot v(x)\right] + \mathbb{V}[\frac{1}{2} + \frac{1}{k}\sum_{i=1}^{k} \epsilon_i] \\ &= \left(\frac{1}{k^2}\right)\left(\frac{1}{16} - \frac{\delta^2}{6}\right) \cdot var(x_{[1:k]}) + \frac{\delta^2}{6k} \\ &\approx \frac{1}{16k^2} var(x_{[1:k]}) \qquad (12)\end{aligned}$$

where the approximation at the last line follows from the assumption that δ is small.

Topics	Discussion
Flexibility	Nonparametric models are a richer model class than parametric models, and are more robust to model misspecification. This comes at a cost of slower convergence (Wasserman 2006).
Model Fitting	Bandwidth selection be more computationally demanding in the nonparametric case, although approximations exist.
Bias Problem	Smoothing methods are in general biased. See Section Background for discussion, references, and remedies

Table 1: Differences between parametric and nonparametric approaches.

Parametric vs. Nonparametric Approaches

As seen by the schematic diagram in Figure 1, in this work we propose a nonparametric alternative to the framework proposed in Poyiadzi et al. (2022) under nonparametric logistic regression models, as opposed to only parametric regression models. Therefore, the comparison of the two works can be discussed on the basis of parametric versus nonparametric generalised linear models (Wasserman 2006). In Table 1 we summarise the main differences between both approaches.

Related Work

Being investigated extensively in the literature, learning with noisy labels (Song et al. 2022) has recently had two dominant research directions: (a) direct estimation of noise (Cheng et al. 2022; Zhu, Wang, and Liu 2022; Liu, Cheng, and Zhang 2023), and (b) data separation (Key, Ragg, and Boukai 2023; Karim et al. 2022).

Direct noise estimation often involves the notion of the noise transition matrix $T(x)$, which reveals the transition probability of clean labels to noisy labels given an instance, i.e. its (i, j) entry $T_{ij}(x)$ is defined as $\mathbb{P}(\tilde{Y} = j \mid Y = i, X = \mathbf{x})$. The transition matrix usually needs to be estimated from the whole noisy dataset in a supervised way and typically requires a large number of training samples, and then it can be used to correct the training loss. However, $T(x)$, in reality, is hard to estimate (Cheng et al. 2022; Li et al. 2022), especially in scenarios with a high noise rate.

Data separation methods, on the other hand, work on the identification of noisy or clean examples. Typically, this line of research focuses on first filtering out the noisy examples from the clean ones, e.g. using the small-loss trick (Yu et al. 2019). Clean labels are then used to train the model under a supervised setting. In the meantime, noisy labels can be treated as unlabelled data and different semi-supervised learning techniques can be applied (Li, Socher, and Hoi 2020; Chen et al. 2023; Han et al. 2018; Yu et al. 2019). Ad-

ditionally, some data separation techniques are more data-centric as opposed to learning-centric (Zhu, Dong, and Liu 2022; Wei et al. 2022). A recent study (Key, Ragg, and Boukai 2023) proposes a nonparametric hypothesis test to filter out mislabelled instances from correctly labelled ones. The proposed hypothesis test is based on the distance between instances, with the primary assumption being that distances between instances within the same class are stochastically smaller than distances between instances from different classes. To detect mislabelled instances, this procedure is applied to all points in each class, and can be iterated across all classes. One possible limitation of this approach may be the case when there is a large number of mislabelled instances, which could bias the estimates. In addition, how to draw an accurate line between noisy examples and examples with large losses has always been a challenge (Wei et al. 2022; Pleiss et al. 2020)

Interestingly, by looking at these two lines of research from a causal perspective, the authors in Yao et al. (2023) investigate the influence of the data generation process (X causes Y or Y causes X) on the algorithm design, and introduce an intuitive method for the causal structure discovery.

By contrast, the nonparametric hypothesis tests we propose in this work build upon both categories but fundamentally differ on their purpose, as our tests are designed to be employed before any learning takes place – they are a tool to assess the validity of the annotations of a given dataset (similarly to Poyiadzi et al. (2022)). However, different to existing tests, our tests are nonparametric and are based on the solid asymptotic properties of local maximum likelihood estimation.

Experiments

The empirical analysis is designed to provide a clear understanding of the advantages and disadvantages of nonparametric vs parametric tests in order to provide practitioners with intuition for their application in practice[5]. To achieve this, we first explore controlled synthetic datasets and then move to a real-world scenario.

Synthetic Datasets

Symmetric XOR Dataset Firstly, we consider a synthetic XOR dataset with Gaussian distributions centred at: $[2, 2], [-2, -2], [-2, 2]$ and $[2, -2]$ with scale 1 (first two correspond to one class, and the latter two to another). Anchor points can be generated by first considering the Bayes Classifier (since we know the data generating distribution), and then optimising the input (the instance) such that the output is 0.50. In the case of wanting relaxed anchor points, we do it with rejection sampling. In both cases, anchor points are restricted within $[-4, 4]$ for both features. We vary the following parameters:

1. $N \in [200,\ 500,\ 1000]$: the training sample size.
2. $(\alpha,\ \beta)$: we let $\alpha = 0.0$, and $\beta = 0.10$. Please see the appendix of the extended version of this paper (Yang et al. 2023) for additional experiments on more combinations of $(\alpha,\ \beta)$.
3. $k \in [1,\ 2,\ 4,\ 8,\ 16]$: the number of anchor points.
4. $\delta \in [0,\ 0.05,\ 0.10,\ 0.20]$: how relaxed the anchor points are: $\eta(x) \in [0.50 - \delta,\ 0.50 + \delta]$.

[5]Our code is available on GitHub. https://github.com/weisongyang/htest

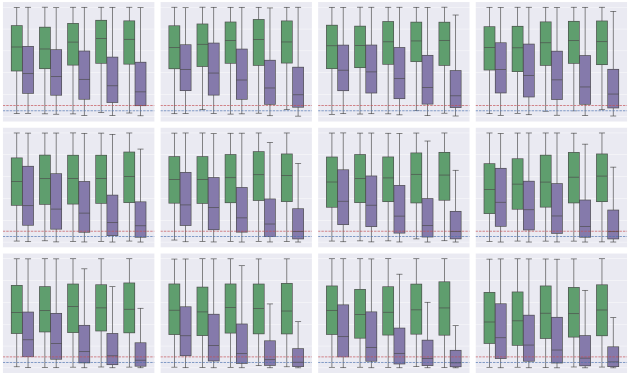

Figure 2: Nonparametric test on symmetric XOR data. Box-plots with $\alpha = 0.0$ and $\beta = 0.10$ (purple) against box-plots with $\alpha = 0.0$ and $\beta = 0.0$ (green).

For each combination of N and $(\alpha,\ \beta)$ we perform 100 runs from 100 new dataset draws. Then for each model, and for each combination of k and δ we do 10 draws.

The sub-figures in Figures 2, 3, 4 and 5 are related as follows: moving from top to bottom, from the first row to the last row, we increase N: $200,\ 500,\ 1000$. Moving left to right, in each column, we relax the anchors more: Strict, $\delta = 0.05, \delta = 0.10, \delta = 0.20$. Within each sub-figure, on the y-axis we have the p values, and on the x-axis, we vary the number of anchor points: $1,\ 2,\ 4,\ 8,\ 16$. Then, for every set of anchor points, we have two box-plots, the purple corresponds to the underlying model being trained on noisy data, and the green corresponds to the underlying models being trained on clean data. For example, the sub-figure at the bottom right of figure 2 refers to training with $N = 1000$ data points, anchor points being relaxed with $\delta = 0.20$. In all subplots, we indicate 0.10 with a red dashed line and 0.05 with a blue dashed line, which serves as the rejection threshold for the null hypothesis.

From Figure 2, we observe that: (1) as the sample size (N) increases, (2) as the number of anchor points increases, p values correctly decrease in value implying evidence against the null hypothesis (uniform label noise). We also observe that (per row) moving from right to left, i.e. from the most relaxed to strict anchor points, we again see better performance. In the appendix of Yang et al. (2023), we show that as the difference between α and β increases, p values correctly decrease as well.

In addition, we implement the parametric test proposed in Poyiadzi et al. (2022) and also run it on the symmetric XOR data. From Figure 3, we can observe that as the number of training samples and the number of anchor points increase, green boxes go up to 1 and purple boxes decrease to 0, which suggests that the parametric test also works on the symmet-

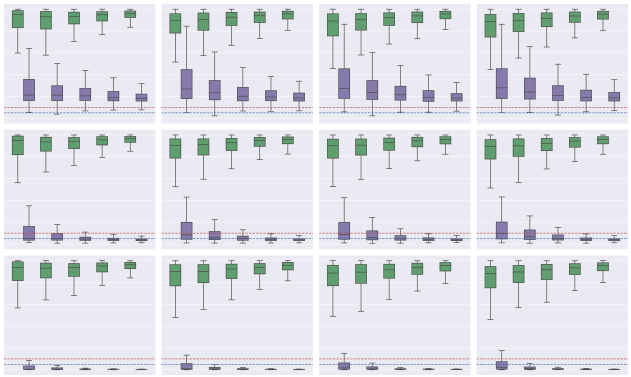

Figure 3: Parametric test on symmetric XOR data. Box-plots with $\alpha = 0.0$ and $\beta = 0.10$ (purple) against box-plots with $\alpha = 0.0$ and $\beta = 0.0$ (green).

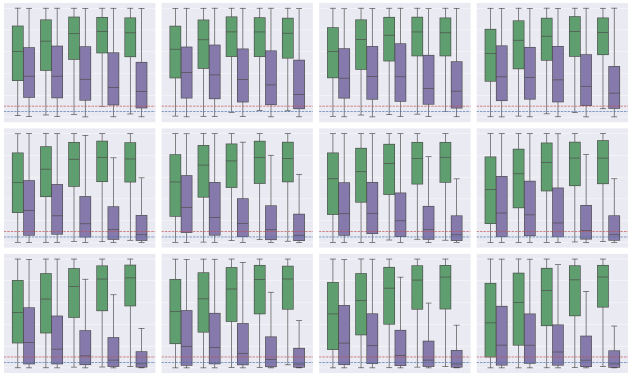

Figure 4: Nonparametric test on asymmetric XOR data. Box-plots with $\alpha = 0.0$ and $\beta = 0.10$ (purple) against box-plots with $\alpha = 0.0$ and $\beta = 0.0$ (green).

ric XOR even if its model assumptions are not met. This can be explained by the symmetry and overlap of the symmetric XOR. Thus, we extend the experiment to an asymmetric XOR.

Asymmetric XOR Dataset The asymmetric XOR dataset also is consisted of also four Gaussians, which are centred at $[4, 4], [-2, -2], [-1, 1]$ and $[1, -1]$. Apart from this, the rest of the experiment settings are the same as in the symmetric XOR's case.

For the nonparametric test, from Figure 4, we can make the same observations as in Figure 2, i.e. p values will decrease as the training sample size and the number of anchor points increase. Please see the appendix of Yang et al. (2023) for the results of more combinations of $(\alpha,\ \beta)$.

As for the results of the parametric test on the asymmetric XOR, however, we can now notice from Figure 5 that both green and purple boxes will go down as the training sample size and the number of anchor points increase. This verifies our claim that blind application of the parametric test can lead to suboptimal outcomes if any of the model assumptions doesn't hold.

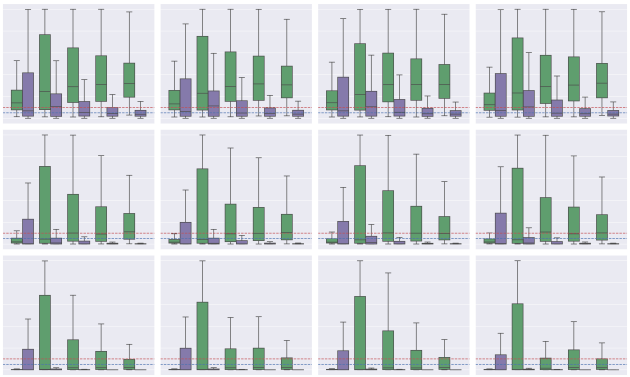

Figure 5: Parametric test on asymmetric XOR data. Box-plots with $\alpha = 0.0$ and $\beta = 0.10$ (purple) against box-plots with $\alpha = 0.0$ and $\beta = 0.0$ (green).

Limitations One open question when working with local likelihood problems is the empty-neighbourhood problem, which is more apparent in the low-data setting, and when the instance we are fitting for is rather far from the data. In the future, we would like to better understand remedies for this for our case. Overall the experiments carried out, we have encountered an issue of approximately $1/10000$.

Real-World Dataset

Dataset Description For the experiment on real-world data, we consider data routinely collected in smart homes. There are numerous initiatives that work on digital health technologies to monitor the health status of people living in smart homes (Brumitt et al. 2000; Cook 2006; Zhu et al. 2015; Woznowski et al. 2017; Twomey et al. 2016). With our rapidly growing aging population and relatively few clinicians, there is a pressing need to automatically understand the state and progression of chronic diseases. The emergence of state-of-the-art sensing platforms offers unprecedented opportunities for indirect and automatic evaluation of disease state through the lens of behavioural monitoring. One of the end goals of the research and development of these smart home systems will be to make them affordable and scalable, so that they could be installed by non-specialists and even general older adults. Naturally, the need to evaluate the quality of data annotation upon installation will emerge. In these settings, a particularly important task is that of localisation which is helpful for quantifying behaviours in the home environment (D'Souza, Ros, and Karunanithi 2012; Otsason et al. 2005; Klingbeil and Wark 2008). Our real-world data is collected from a similar smart home project (Yang et al. 2022). For the localisation task, we have access to room-level annotations. The experiments we carry out in this section serve as a case study to show how hypothesis tests can be used in real life and the practical considerations needed.

Experiment Settings There are seven gateways installed at different locations in our example house, i.e. Living room 1, Kitchen 1, Hall 1, Dining room 1 (two gateways installed),

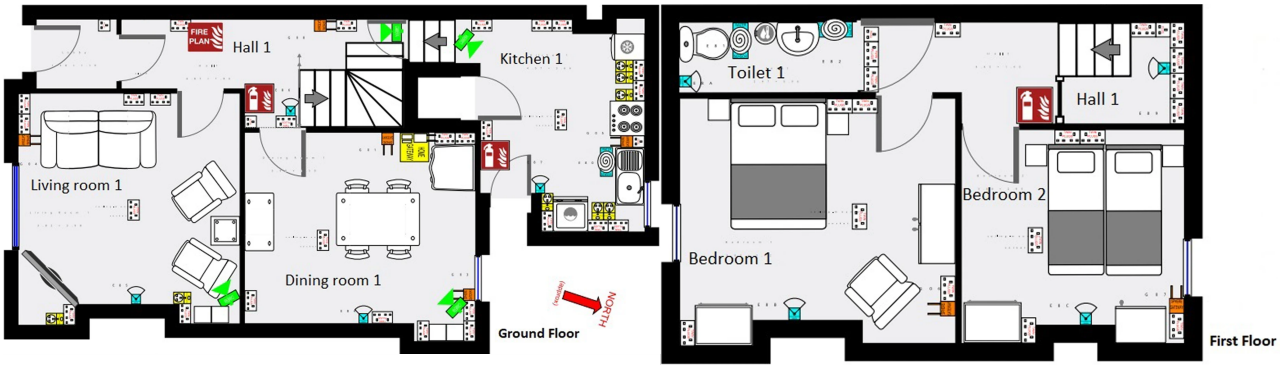

Figure 6: An example house with two floors.

Bedroom 1, and Bedroom 2. Figure 6 is its floorplan. These gateways simultaneously record accelerometer data and the Received Signal Strength Intensity (RSSI), which is the basis for our localisation efforts. To be consistent with the binary classification setting in this paper, we group our room-level labels by floor so that we have two classes in this two-floor house. Location annotations were collected by the technician upon installation, who carried all the wearables in a bag and walked through every room. The real-time location of the technician was recorded with an annotation app while the RSSI data is simultaneously recorded. For this example house, the annotation walkaround lasted less than 30 minutes. For input, we extract basic features from windowed RSSI data (5 seconds length with an overlap of 2.50 seconds following Twomey et al. (2018)). There are two wearables for this house, which gives us roughly 1200 data points. We randomly sample 1000 data points to train our model in each run, and we perform 100 runs in total.

We use RSSI readings received by the gateways when the technician was on the stairs as *anchor points*. The intuition is that the stairs are in the middle of two floors, and RSSI readings on the stairs should be close to the decision boundary of the binary classification task. In other words, the true posteriors of RSSI readings on stairs should be close to 0.5. We vary the number of anchor points $k \in [1,\ 2,\ 4,\ 8]$. Importantly, the notion of stairs also requires extra care when designing the experiment. Depending on the height of the ceiling, the width of the stairs, and on which floor the gateway labelled 'stairs' is installed, the posterior of steps at the bottom of the stairs and at the top of the stairs can be very different, which may introduce bias if we group stairs into one of the classes and include them in the training data. Accordingly, we consider two settings in the end: (a) include RSSI readings on stairs in the training process and also use them as anchor points, and (b) exclude RSSI readings on stairs from training, and only use them as anchor points.

Results In Figure 7, the plot on the left corresponds to scenario (a), while scenario (b) is on the right. We do not expect there to be label noise in our walkaround dataset. Yet we observe that if we include RSSI readings received on stairs in training, p values will decrease as the number of anchor points increases, and finally go below the cut-off point. But

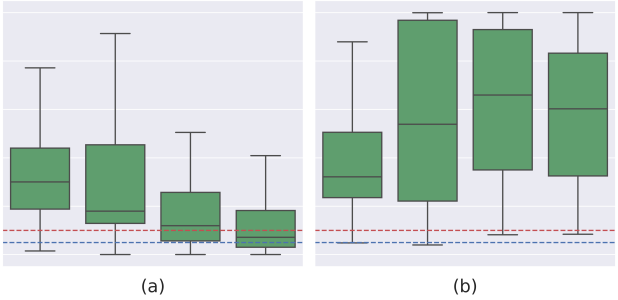

Figure 7: Nonparametric test on the smart home dataset.

if we exclude stairs from training, p values remain above the cut-off points, which says there is no class-conditional label noise in our walkaround dataset. This validates our assumption that grouping stairs into one of the floors might introduce noise in the labels. The stairs in our example house are wide and high enough, so that they will cause bias in training under setting (a), and finally lead us to wrong conclusions.

Conclusions

In this work we showed how (nonparametric) hypothesis testing procedures on class-dependent label noise can be used when the underlying model is a product of local maximum likelihood estimation procedure. Going beyond the state-of-the-art, local Maximum Likelihood Estimation results in the creation of more adaptable nonparametric logistic regression models, thereby reducing vulnerability to model misspecification. This extends the potential applications to more occasions where the underlying data generation is unknown. We further compare empirically both parametric and nonparametric options, focusing on discussing the advantages and limitations of the nonparametric approach. Finally, we presented a first real-world case study, demonstrating the steps of applying such tests in practice with a smart home dataset, which shows the effectiveness and practical considerations of the proposed method.

In future work, we want to explore bootstrap bias correction as well as consider bootstrap on its own as a means of obtaining variance for predictions of anchors.

Acknowledgements

This work was funded by the UKRI Turing AI Fellowship EP/V024817/1 and the MRC CUBOID project (UK MRC Momentum grant MC/PC/16029).

References

Brumitt, B.; Meyers, B.; Krumm, J.; Kern, A.; and Shafer, S. 2000. Easyliving: Technologies for intelligent environments. In *International Symposium on Handheld and Ubiquitous Computing*, 12–29. Springer.

Chen, M.; Cheng, H.; Du, Y.; Xu, M.; Jiang, W.; and Wang, C. 2023. Two wrongs don't make a right: Combating confirmation bias in learning with label noise. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, 14765–14773.

Cheng, D.; Liu, T.; Ning, Y.; Wang, N.; Han, B.; Niu, G.; Gao, X.; and Sugiyama, M. 2022. Instance-dependent label-noise learning with manifold-regularized transition matrix estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 16630–16639.

Cook, D. J. 2006. Health monitoring and assistance to support aging in place. *J. UCS*, 12(1): 15–29.

D'Souza, M.; Ros, M.; and Karunanithi, M. 2012. An indoor localisation and motion monitoring system to determine behavioural activity in dementia afflicted patients in aged care.

Frölich, M. 2006. Non-parametric regression for binary dependent variables. *The Econometrics Journal*, 9(3): 511–540.

Gong, Y.; Ke, Q.; Isard, M.; and Lazebnik, S. 2014. A multi-view embedding space for modeling internet images, tags, and their semantics. *International journal of computer vision*, 106: 210–233.

Han, B.; Yao, Q.; Yu, X.; Niu, G.; Xu, M.; Hu, W.; Tsang, I.; and Sugiyama, M. 2018. Co-teaching: Robust training of deep neural networks with extremely noisy labels. *Advances in neural information processing systems*, 31.

Hilal, W.; Gadsden, S. A.; and Yawney, J. 2022. Financial fraud: a review of anomaly detection techniques and recent advances. *Expert systems With applications*, 193: 116429.

Jarrett, D.; Cebere, B. C.; Liu, T.; Curth, A.; and van der Schaar, M. 2022. Hyperimpute: Generalized iterative imputation with automatic model selection. In *International Conference on Machine Learning*, 9916–9937. PMLR.

Karim, N.; Rizve, M. N.; Rahnavard, N.; Mian, A.; and Shah, M. 2022. Unicon: Combating label noise through uniform selection and contrastive learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 9676–9686.

Key, M. C.; Ragg, S.; and Boukai, B. 2023. A statistical testing procedure for validating class labels. *Journal of Applied Statistics*, 50(8): 1725–1749.

Klingbeil, L.; and Wark, T. 2008. A wireless sensor network for real-time indoor localisation and motion monitoring. In *2008 International Conference on Information Processing in Sensor Networks (ipsn 2008)*, 39–50. IEEE.

Lall, R.; and Robinson, T. 2022. The MIDAS touch: accurate and scalable missing-data imputation with deep learning. *Political Analysis*, 30(2): 179–196.

Li, J.; Socher, R.; and Hoi, S. C. 2020. Dividemix: Learning with noisy labels as semi-supervised learning. *arXiv preprint arXiv:2002.07394*.

Li, S.-Y.; Shi, Y.; Huang, S.-J.; and Chen, S. 2022. Improving deep label noise learning with dual active label correction. *Machine Learning*, 1–22.

Li, W.; Wang, L.; Li, W.; Agustsson, E.; and Van Gool, L. 2017. Webvision database: Visual learning and understanding from web data. *arXiv preprint arXiv:1708.02862*.

Liu, Y.; Cheng, H.; and Zhang, K. 2023. Identifiability of label noise transition matrix. In *International Conference on Machine Learning*, 21475–21496. PMLR.

Loader, C. 2006. *Local regression and likelihood*. Springer Science & Business Media.

Loader, C. 2012. Smoothing: local regression techniques. *Handbook of computational statistics: concepts and methods*, 571–596.

Otsason, V.; Varshavsky, A.; LaMarca, A.; and De Lara, E. 2005. Accurate GSM indoor localization. In *International conference on ubiquitous computing*, 141–158. Springer.

Pleiss, G.; Zhang, T.; Elenberg, E.; and Weinberger, K. Q. 2020. Identifying mislabeled data using the area under the margin ranking. *Advances in Neural Information Processing Systems*, 33: 17044–17056.

Poyiadzi, R.; Yang, W.; Twomey, N.; and Santos-Rodriguez, R. 2022. Statistical hypothesis testing for class-conditional label noise. In *European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases*.

Sauber-Cole, R.; and Khoshgoftaar, T. M. 2022. The use of generative adversarial networks to alleviate class imbalance in tabular data: a survey. *Journal of Big Data*, 9(1): 98.

Schmidl, S.; Wenig, P.; and Papenbrock, T. 2022. Anomaly detection in time series: a comprehensive evaluation. *Proceedings of the VLDB Endowment*, 15(9): 1779–1797.

Song, H.; Kim, M.; Park, D.; Shin, Y.; and Lee, J.-G. 2022. Learning From Noisy Labels With Deep Neural Networks: A Survey. *IEEE Transactions on Neural Networks and Learning Systems*, 1–19.

Twomey, N.; Diethe, T.; Fafoutis, X.; Elsts, A.; McConville, R.; Flach, P.; and Craddock, I. 2018. A comprehensive study of activity recognition using accelerometers. In *Informatics*, volume 5, 27. Multidisciplinary Digital Publishing Institute.

Twomey, N.; Diethe, T.; Kull, M.; Song, H.; Camplani, M.; Hannuna, S.; Fafoutis, X.; Zhu, N.; Woznowski, P.; Flach, P.; et al. 2016. The SPHERE challenge: Activity recognition with multimodal sensor data. *arXiv preprint arXiv:1603.00797*.

Van der Vaart, A. W. 2000. *Asymptotic statistics*, volume 3. Cambridge university press.

Wasserman, L. 2006. *All of nonparametric statistics*. Springer Science & Business Media.

Wei, Q.; Sun, H.; Lu, X.; and Yin, Y. 2022. Self-filtering: A noise-aware sample selection for label noise with confidence penalization. In *European Conference on Computer Vision*, 516–532. Springer.

Woznowski, P.; Burrows, A.; Diethe, T.; Fafoutis, X.; Hall, J.; Hannuna, S.; Camplani, M.; Twomey, N.; Kozlowski, M.; Tan, B.; et al. 2017. SPHERE: A sensor platform for healthcare in a residential environment. *Designing, Developing, and Facilitating Smart Cities: Urban Design to IoT Solutions*, 315–333.

Xia, X.; Pan, X.; Li, N.; He, X.; Ma, L.; Zhang, X.; and Ding, N. 2022. GAN-based anomaly detection: A review. *Neurocomputing*, 493: 497–535.

Yang, W.; Poyiadzi, R.; Ben-Shlomo, Y.; Craddock, I.; Coulthard, L.; Santos-Rodriguez, R.; Selwood, J.; and Twomey, N. 2022. *Detecting and Monitoring Behavioural Patterns in Individuals with Cognitive Disorders in the Home Environment with Partial Annotations*, 25–52. Cham: Springer International Publishing. ISBN 978-3-030-91181-2.

Yang, W.; Poyiadzi, R.; Twomey, N.; and Rodriguez, R. S. 2023. Hypothesis Testing for Class-Conditional Noise Using Local Maximum Likelihood. arXiv:2312.10238.

Yao, Y.; Gong, M.; Du, Y.; Yu, J.; Han, B.; Zhang, K.; and Liu, T. 2023. Which is Better for Learning with Noisy Labels: The Semi-supervised Method or Modeling Label Noise?

Yu, X.; Han, B.; Yao, J.; Niu, G.; Tsang, I.; and Sugiyama, M. 2019. How does disagreement help generalization against label corruption? In *International Conference on Machine Learning*, 7164–7173. PMLR.

Zhu, N.; Diethe, T.; Camplani, M.; Tao, L.; Burrows, A.; Twomey, N.; Kaleshi, D.; Mirmehdi, M.; Flach, P.; and Craddock, I. 2015. Bridging e-health and the internet of things: The sphere project. *IEEE Intelligent Systems*, 30(4): 39–46.

Zhu, Z.; Dong, Z.; and Liu, Y. 2022. Detecting corrupted labels without training a model to predict. In *International conference on machine learning*, 27412–27427. PMLR.

Zhu, Z.; Wang, J.; and Liu, Y. 2022. Beyond images: Label noise transition matrix estimation for tasks with lower-quality features. In *International Conference on Machine Learning*, 27633–27653. PMLR.