Learning from Multiple Observers with Unknown Expertise

Han Xiao, Huang Xiao, and Claudia Eckert

Institute of Informatics
Technische Universität München, Germany
{xiaoh,xiaohu,claudia.eckert}@in.tum.de

Abstract. Internet has emerged as a powerful technology for collecting labeled data from a large number of users around the world at very low cost. Consequently, each instance is often associated with a handful of labels, precluding any assessment of an individual user's quality. We present a probabilistic model for regression when there are multiple yet some unreliable observers providing continuous responses. Our approach simultaneously learns the regression function and the expertise of each observer that allow us to predict the ground truth and observers' responses on the new data. Experimental results on both synthetic and real-world data sets indicate that the proposed method has clear advantages over "taking the average" baseline and some state-of-art models.

1 Introduction

With the recent advent of social web services, the data can now be shared and processed by a large number of users. As a consequence, researchers are faced with data sets that are labeled by multiple users. For example, Wikipedia provides a feedback tool to engage readers in the assessment of article quality based on four criteria, i.e. "trustworthy", "objective", "complete" and "well-written". The Amazon Mechanical Turk is an online system that allows the requesters to hire users from all over the world to perform crowdsourcing tasks. Galaxy Zoo is a website where visitors label astronomical images. While providing large amounts of cheap labeled data in a short time, these platforms usually have little quality control over users. Thus, the response of each user can vary widely, and in some cases may even be adversarial. A natural question to ask is how to integrate opinions from multiple users for obtaining an objective opinion. The commonly used "majority vote" and "take the average" heuristics completely ignore the individual expertise and may fail in the settings with non-Gaussian or adversarial noise. This casts a challenge of *learning from multiple sources* for the machine learning and data mining researchers [2].

Despite these web applications, one can find this problem in wide range of domains. Recently, *sensor networks* have been deployed for the scientific monitoring of remote and hostile environments. For example, researchers deployed a 16-node sensor network on a tree to study its elevation under different weather fronts [9]. Each node samples climate data at regular time intervals and the statistics are collected. Using sensor data in this manner presents many novel challenges, such as fusing noisy readings from several sensors, detecting faulty and aging sensors. Importantly, it is necessary to use the

J. Pei et al. (Eds.): PAKDD 2013, Part I, LNAI 7818, pp. 595–606, 2013.

trends and correlations observed in previous data to predict the value of environmental parameters into the future, or to predict the reading of a sensor that is temporarily unavailable (e.g. due to network outages). However, these tasks may have to be performed with only limited knowledge of the location, reliability, and accuracy of each sensor.

In this work, the labeler (including user, annotator and sensor) mentioned above is referred to as the *observer*. Given an *instance*, the label (e.g. annotation, reading) provided by an observer is called the *response*. Unlike the conventional supervised learning scenario, in our setting each instance is associated with a set of responses, yet the *ground truth* is unknown as some responses may be subjective or come from unreliable observers. We concentrate on the regression problem with continuous responses from multiple observers. Specifically, our method provides a principled way to answer the following questions:

1. How to learn a regression function to predict the ground truth precluding the prior knowledge of observers?
2. How to estimate the expertise of each observer without knowing the ground truth?

2 Related Work and Novel Contributions

There is a number of studies dealing with the setting involving multiple labelers, yet most of them focus on the classification problem. Early work such as [3,4,8] focus on estimating the error rates of observers. In the machine learning community, the problem of estimating the ground truth from multiple noisy labels is addressed in [7]. Instead of estimating the ground truth and learning the classifier separately, recent interest has shifted towards on learning classifiers directly from such data. Authors of [2] provide a general theory of selecting the most informative samples from each source for model training. Later, a probabilistic framework is presented by [5,6] to address the classification, regression and ordinal regression problem with multiple annotators. The framework is based on a simple assumption that the expertise of each annotator does not depend on the given data. This assumption is infringed in [10,13] and later is extended to the active learning scenario [12]. There are some other related work that focus on different settings [1,11].

The above studies paid little attention to the regression problem under multiple observers, which is the main core of this paper. Moreover, our work differs from the related work in various aspects. First, we employ a less-parametric method, i.e. the *Gaussian process* (GP), to model the observers and the regression function. This allows us to associate the observer's expertise with both ground truth and input instance. Moreover, our model is presented in an extensible probabilistic framework. The missing data and prior knowledge can be straightforwardly incorporated into the model.

The rest of this paper is organized as follows. Section 3 formulates the problem and introduces a probabilistic framework. The framework consists of two parts. The regression model is introduced in Section 3.2. A linear and a non-linear observer model is proposed in Section 3.3 and Section 3.4, respectively. Section 4 reports the experimental results on both synthetic and real-world data sets. Conclusions are drawn in Section 5.

3 Probabilistic Formulation

Denote the *instance space* $\mathcal{X} \subseteq \mathbb{R}^L$ and the *response space* $\mathcal{Y} \subseteq \mathbb{R}^D$ and the *ground truth space* $\mathcal{Z} \subseteq \mathbb{R}^D$. Given N instances $\mathbf{x}_1, \ldots, \mathbf{x}_N$ where $\mathbf{x}_n \in \mathcal{X}$, denote the *objective ground truth* for $\mathbf{x}_n$ as $\mathbf{z}_n \in \mathcal{Z}$. In our setting, the ground truth is unknown. Instead, we have multiple responses $\mathbf{y}_{n,1}, \ldots, \mathbf{y}_{n,M} \in \mathcal{Y}$ for $\mathbf{x}_n$ provided by M different observers. For compactness, the $N \times L$ matrix of instance $x_{n,l}$ is represented as $\mathbf{X} := [\mathbf{x}_1, \ldots, \mathbf{x}_N]^\top$. The $N \times M \times D$ tensor of observers' responses $y_{n,m,d}$ is denoted by $\mathbf{Y} := [\mathbf{y}_{1,1}, \ldots, \mathbf{y}_{1,M}; \ldots; \mathbf{y}_{N,1}, \ldots, \mathbf{y}_{N,M}]$. The $N \times D$ matrix of ground truth $z_{n,d}$ is denoted by $\mathbf{Z} := [\mathbf{z}_1, \ldots, \mathbf{z}_N]^\top$.

Given the training data $\mathbf{X}$ and $\mathbf{Y}$, our goal is threefold. First, it is of interest to get an estimate of the unknown ground truth $\mathbf{Z}$. The second goal is to learn a regression function $f : \mathcal{X} \to \mathcal{Z}$ which generalizes well on unseen instances. Finally, for each observer we want to model its *expertise* as a function of the input instance and the ground truth, i.e. $g : \mathcal{X} \times \mathcal{Z} \to \mathcal{Y}$.

3.1 Probabilistic Framework

To formulate this problem from the probabilistic perspective, we consider the training data $\mathbf{X}$ and $\mathbf{Y}$ as random variables. The ground truth $\mathbf{Z}$ is unknown and hence is a latent variable. In general, the observed response $\mathbf{Y}$ depends both on the unknown ground truth and the instance. That is, observers may exhibit varying levels of expertise on different instances. On Wikipedia the assumption is particularly true for the novice readers, whereas the rating from an expert reader is consistent across different types of articles. Figure 1 illustrates the conditional dependence between $\mathbf{X}, \mathbf{Y}$ and $\mathbf{Z}$ with a graphical model. As a consequence, the joint conditional distribution can be expressed as

$$\begin{aligned} p(\mathbf{Y}, \mathbf{Z}, \mathbf{X}) &= p(\mathbf{Z} \,|\, \mathbf{X}) p(\mathbf{Y} \,|\, \mathbf{Z}, \mathbf{X}) p(\mathbf{X}) \\ &\propto \prod_{n=1}^{N} \prod_{d=1}^{D} p(z_{n,d} \,|\, \mathbf{x}_n) \prod_{m=1}^{M} p(y_{n,m,d} \,|\, \mathbf{x}_n, z_{n,d}), \end{aligned} \tag{1}$$

where the term $p(\mathbf{X})$ is dropped as we are more interested in the other two conditional distributions. There are two underlying assumptions in this model. First, each dimension of the ground truth is independent, but is not identically distributed. Second, all observers respond independently.

Note that the first term in (1) indicates the probabilistic dependence between the ground truth and the input instance, whereas the second term characterizes the observers' expertise. Previous work have explored different parametric methods to model these two conditional distributions [10,13,5,12,6]. A distinguishing factor in this paper is that, we employ the Gaussian process as the backbone to construct the model. Specifically, the generative process of $\mathbf{Y}$ can be interpreted as follows

$$z_{n,d} = f_d(\mathbf{x}_n) + \epsilon_n, \tag{2}$$

$$y_{n,m,d} = g_{m,d}(\mathbf{x}_n, z_{n,d}) + \xi_{m,d}, \tag{3}$$

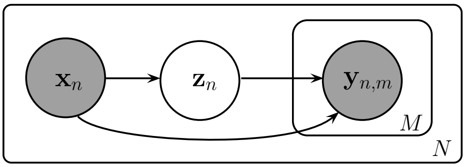

Fig. 1. Graphical model of instances $\mathbf{X}$, unknown ground truth $\mathbf{Z}$ and responses $\mathbf{Y}$ from M different observers. Only the shaded variables are observed.

where ϵ and ξ is independent identically distributed Gaussian noise, respectively. Note that the choice of $\{f_d\}$ and $\{g_{m,d}\}$ characterizes the regression function and the observers, respectively. In particular, an ideal observer would have $g_{m,d}(z_{n,d}) = z_{n,d}$ on every d. Therefore, our goal can be understood as searching $\{f_d\}$ and $\{g_{m,d}\}$ given the training data. Intuitively, if two instances are close to each other in $\mathcal{X}$, then their corresponding ground truth should be close in $\mathcal{Z}$ through the mapping of $\{f_d\}$, which in turn restricts the searching space of $\{g_{m,d}\}$ when $\mathbf{Y}$ is known.

3.2 Regression Model

We first concentrate on Eq. (2) and represent functions $\{f_d\}$ by the Gaussian process with some non-linear kernel. Specifically, the conditional distribution of the ground truth given the training instances is assumed to be

$$p(\mathbf{Z}\,|\,\mathbf{X}) = \prod_{d=1}^{D} \mathcal{N}\left(\mathbf{z}_{:,d}\,|\,\mathbf{0}, \mathbf{K}_d\right), \tag{4}$$

where the d^{th} dimension of the ground truth is denoted as $\mathbf{z}_{:,d}$. We introduce a $N \times N$ kernel matrix $\mathbf{K}_d$ that depends on $\mathbf{X}$, where each element is given by the value of a composite covariance function $k_d : \mathcal{X} \times \mathcal{X} \to \mathbb{R}_{0+}$, made up of several contributions as follows

$$k_d(\mathbf{x}_i, \mathbf{x}_j) := \kappa_{1,d}^2 \exp\left(-\frac{\kappa_{2,d}^2}{2}\|\mathbf{x}_i - \mathbf{x}_j\|^2\right) + \kappa_{3,d}^2 + \kappa_{4,d}^2 \mathbf{x}_i^\top \mathbf{x}_j + \kappa_{5,d}^2 \delta(\mathbf{x}_i, \mathbf{x}_j). \tag{5}$$

The noise term ϵ in Eq. (2) is folded into the Kronecker delta function $\delta(\mathbf{x}_i, \mathbf{x}_j)$. The covariance function involves an exponential of a quadratic term, with the addition of a constant bias, a linear and a noise terms. For each dimension, the parameters need to be learned from the data are $\kappa_{1,d}, \ldots, \kappa_{5,d}$.

3.3 Linear Observer Model

To model the observer's expertise, we now concentrate on (3) and assume that $\{g_{m,d}\}$ is a linear mapping from $\mathcal{Z}$ to $\mathcal{Y}$, which does not depend on the instance at all.

Denote $\mathbf{y}_{:,m,d}$ the d^{th} dimension response of all training instances provided by the m^{th} observer. The second conditional distribution in (1) is assumed to be

$$p(\mathbf{Y} \mid \mathbf{Z}, \mathbf{X}) = p(\mathbf{Y} \mid \mathbf{Z}) = \prod_{m=1}^{M} \prod_{d=1}^{D} \mathcal{N}\left(\mathbf{y}_{:,m,d} \,\middle|\, w_{m,d}\mathbf{z}_{:,d} + \mu_{m,d}\mathbf{1}, \sigma_{m,d}^2\mathbf{I}\right), \tag{6}$$

where $\mathbf{1}$ is an all-ones vector with length N and $\mathbf{I}$ is a $N \times N$ identity matrix. Each observer is characterized by $3 \times D$ parameters, i.e. $w_{m,d}, \mu_{m,d}, \sigma_{m,d} \in \mathbb{R}$.

Parameter Estimation. Now we can combine Eq. (6) with Eq. (4) and estimate the set of all parameters, i.e. $\boldsymbol{\Theta} := \{\{\kappa_{1,d}, \ldots, \kappa_{5,d}\}, \{w_{m,d}\}, \{\mu_{m,d}\}, \{\sigma_{m,d}\}\}$, by maximizing the likelihood function $p(\mathbf{Y} \mid \mathbf{X}, \boldsymbol{\Theta})$. In the linear observer model, the latent variable $\mathbf{Z}$ can be marginalized out, which yields

$$p(\mathbf{Y} \mid \mathbf{X}, \boldsymbol{\Theta}) = \prod_{m=1}^{M} \prod_{d=1}^{D} \mathcal{N}\left(\mu_{m,d}\mathbf{1}, w_{m,d}^2\mathbf{K}_d + \sigma_{m,d}^2\mathbf{I}\right).$$

The maximum likelihood estimator of $\mu_{m,d}$ is given by $\widetilde{\mu}_{m,d} = \frac{1}{N}\sum_{n=1}^{N} y_{n,m,d}$. We hereinafter use the short-hand $\overline{\mathbf{y}}_{:,m,d} := \mathbf{y}_{:,m,d} - \widetilde{\mu}_{m,d}\mathbf{1}$. As a consequence, the log-likelihood function is given by

$$\begin{aligned} F^{\text{LOB}} &:= \log p(\mathbf{Y} \mid \mathbf{X}, \boldsymbol{\Theta}) = \sum_{m=1}^{M}\sum_{d=1}^{D} \log p(\mathbf{y}_{:,m,d} \mid \mathbf{X}, \boldsymbol{\Theta}) \\ &= \sum_{m=1}^{M}\sum_{d=1}^{D} -\frac{N}{2}\log(2\pi) - \frac{1}{2}\log|\mathbf{C}| - \frac{1}{2}\operatorname{tr}\left(\overline{\mathbf{y}}_{:,m,d}^{\top}\mathbf{C}^{-1}\overline{\mathbf{y}}_{:,m,d}\right), \end{aligned} \tag{7}$$

where $\mathbf{C} := w_{m,d}^2\mathbf{K}_d + \sigma_{m,d}^2\mathbf{I}$. To find the parameters by maximizing Eq. (7), we take the partial derivatives of F^{LOB} with respect to the parameters and obtain

$$\frac{\partial F^{\text{LOB}}}{\partial w_{m,d}} = w_{m,d}\operatorname{tr}\left(\mathbf{B}\mathbf{C}^{-1}\mathbf{K}_d\right), \tag{8}$$

$$\frac{\partial F^{\text{LOB}}}{\partial \sigma_{m,d}} = \sigma_{m,d}\operatorname{tr}\left(\mathbf{B}\mathbf{C}^{-1}\right), \tag{9}$$

$$\frac{\partial F^{\text{LOB}}}{\partial \kappa_{i,d}} = \sum_{m=1}^{M} \frac{1}{2} w_{m,d}^2 \operatorname{tr}\left(\mathbf{B}\mathbf{C}^{-1}\frac{\partial \mathbf{K}_d}{\partial \kappa_{i,d}}\right), \tag{10}$$

where $\mathbf{B} := \mathbf{C}^{-1}\overline{\mathbf{y}}_{:,m,d}\overline{\mathbf{y}}_{:,m,d}^{\top} - \mathbf{I}$ and $\frac{\partial \mathbf{K}_d}{\partial \kappa_{i,d}}$ is a matrix of element-wise partial derivatives of Eq. (5) with respect to $\kappa_{1,d}, \ldots, \kappa_{5,d}$. As there exists no closed-form solution, we resort to L-BFGS quasi-Newton method to maximize F^{LOB}. Essentially, in each iteration the gradients are computed by Eqs. (8) to (10) and the parameters are updated accordingly.

Estimate of Ground Truth. Note that the ground truth $\mathbf{Z}$ is marginalized out from Eq. (7) and still remains unknown. To estimate the ground truth of all training instances, we need to find the posterior of $\mathbf{Z}$, i.e. $p(\mathbf{Z} \,|\, \mathbf{Y}, \mathbf{X}) = p(\mathbf{Y} \,|\, \mathbf{Z}, \mathbf{X}) p(\mathbf{Z} \,|\, \mathbf{X}) / p(\mathbf{Y} \,|\, \mathbf{X})$. By using the property of Gaussian distribution, one can show that the posterior of $\mathbf{z}_{:,d}$ follows $\mathcal{N}(\mathbf{u}, \mathbf{V})$, where

$$\mathbf{u} = \mathbf{V} \left(\sum_{m=1}^{M} \frac{w_{m,d}}{\sigma_{m,d}^2} \mathbf{y}_{:,m,d} \right), \quad \mathbf{V} = \left(\sum_{m=1}^{M} \frac{w_{m,d}^2}{\sigma_{m,d}^2} \mathbf{I} + \mathbf{K}_d^{-1} \right)^{-1}. \tag{11}$$

The above computation is repeated D times on every dimension to obtain the estimate of ground truth $\widetilde{\mathbf{Z}}$.

Prediction on New Instance. Given a new instance $\mathbf{x}_*$, we are interested in predicting the ground truth $\mathbf{z}_*$ by using the learned regression function. This can be derived from the joint distribution

$$\begin{bmatrix} \widetilde{\mathbf{z}}_{:,d} \\ z_{*,d} \end{bmatrix} \sim \mathcal{N} \left(\mathbf{0}, \begin{bmatrix} \mathbf{K}_d & \mathbf{k}_*^\top \\ \mathbf{k}_* & k_d(\mathbf{x}_*, \mathbf{x}_*) \end{bmatrix} \right), \tag{12}$$

where $\mathbf{k}_* := [k_d(\mathbf{x}_*, \mathbf{x}_1), \ldots, k_d(\mathbf{x}_*, \mathbf{x}_N)]$. It turns out that $p(z_{*,f} \,|\, \mathbf{X}, \widetilde{\mathbf{z}}_{:,d}, \mathbf{x}_*)$ follows a Gaussian distribution. Hence, the best estimate for the ground truth is

$$\widetilde{z}_{*,d} = \mathbf{k}_* \mathbf{K}_d^{-1} \widetilde{\mathbf{z}}_{:,d}, \tag{13}$$

and the uncertainty is captured in its variance

$$\mathrm{var}(\widetilde{z}_{*,d}) = k_d(\mathbf{x}_*, \mathbf{x}_*) - \mathbf{k}_* \mathbf{K}_d^{-1} \mathbf{k}_*^\top. \tag{14}$$

As a consequence, the response from an observer can be also predicted by

$$\widetilde{y}_{*,m,d} = (1 + \widetilde{w}_{m,d}) \widetilde{z}_{*,d} + \widetilde{\mu}_{m,d}, \tag{15}$$

with variance $\widetilde{\sigma}_{m,d}$.

Priors on Parameters. Note that $w_{m,d}$ is an important indicator of the observer's expertise. On the one hand, a genuine observer would have $w_{m,d}$ close to 1, whereas an adversarial observer gives $w_{m,d}$ close to -1. On the other hand, we encourage $w_{m,d}$ to be a small value unless supported by the data. Without any knowledge on observers, we can only expect that $w_{m,d}$ takes value either around 1 or -1, which inspires the following penalty function

$$\mathrm{penalty}(w_{m,d}) := \begin{cases} \eta(w_{m,d} - 1)^2 & \text{if } w_{m,d} > 1; \\ 0 & \text{if } -1 \leq w_{m,d} \leq 1; \\ \eta(w_{m,d} + 1)^2 & \text{if } w_{m,d} < -1, \end{cases} \tag{16}$$

where η controls the value of penalty as shown in Fig. 2 (see "general"). When $w_{m,d}$ takes value between $[-1, 1]$, there is no penalty and the gradient is given by Eq. (8) directly. When $|w_{m,d}| > 1$ we penalize $w_{m,d}$ and keep it from being too large. This allows our model to search a reasonable solution for $w_{m,d}$ without over-fitting on the training data.

In the case that observers are highly reliable, the learned $w_{m,d}$ should be close to 1 and $\mu_{m,d}, \sigma_{m,d}$ close to 0. One can add a Laplacian prior for observers' parameters, which leads to an L_1 regularization. The penalty term induced by the Laplacian prior for $w_{m,d}$ is $-(\frac{1}{2}\log\lambda + \sqrt{\frac{2}{\lambda}}|w_{m,d} - 1|)$, where a smaller value of λ suggests that the observer is more reliable. The maximization of F^{LOB} can be carried out by computing the sub-gradient of $w_{m,d}, \mu_{m,d}$ and $\sigma_{m,d}$, respectively.

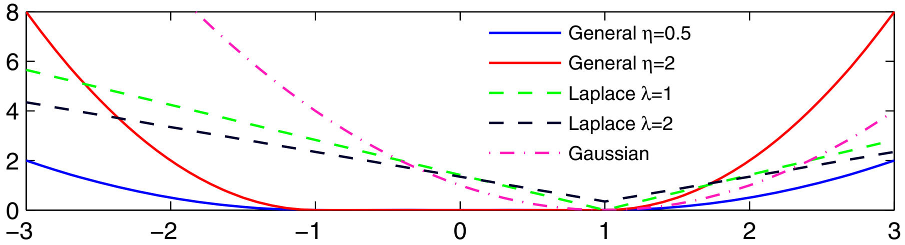

Fig. 2. Penalty functions of $w_{m,d}$ induced by different prior models. The "general" penalty function corresponds to Eq. (16). Similar penalty functions can be added to $\mu_{m,d}$ and $\sigma_{m,d}$ as well.

The relationship between observers can be incorporated into the model as well. For example, the demographic information of users or the geographic location of sensors can be represented as a $M \times M$ proximity matrix $\mathbf{P}$. In particular, we expect two observers have similar parameters if they are highly correlated in $\mathbf{P}$. Assuming $\mathbf{P}$ is a positive definite matrix, we can set the prior distribution of $\mathbf{w}_{:,d}$ set as $\mathcal{N}(\mathbf{w}_{:,d} \mid \mathbf{1}, \mathbf{P})$. As a consequence, we add a penalty term $-\sum_{d=1}^{D} \mathrm{tr}(\mathbf{w}_{:,d}^{\top}\mathbf{P}\mathbf{w}_{:,d})$ to Eq. (6). The gradient of $w_{m,d}$ is computed by Eq. (8) with an additional term $-2\mathbf{P}_{m,:}\mathbf{w}_{:,d}$. Figure 2 illustrates different penalty functions of $w_{m,d}$.

Missing Responses. The model can be extended to handle the training data with missing responses. First of all, we partition the responses $\mathbf{Y} = (\mathbf{Y}^o, \mathbf{Y}^u)$, where $\mathbf{Y}^o$ represents the observed part and $\mathbf{Y}^u$ is the missing part of the responses. Consequently, the latent variables in our model consists of $\mathbf{Z}$ and $\mathbf{Y}^u$. The *expectation maximization* (EM) algorithm can be developed for estimating the model parameters. In the E-step, we fix the model parameter $\boldsymbol{\Theta}$ and compute the sufficient statistics of $\widetilde{\mathbf{Z}}$ by Eq. (11) and then update $\widetilde{\mathbf{Y}}^u$ by its prediction using Eq. (15). In the M-step, we use L-BFGS to maximize $\log p(\widetilde{\mathbf{Y}}, \widetilde{\mathbf{Z}} \mid \mathbf{X}, \boldsymbol{\Theta})$ and update $\boldsymbol{\Theta}$. The two steps are repeated until the likelihood reaches a local maximum.

3.4 Non-linear Observer Model

The assumptions behind the linear observer model may not be appropriate in some scenarios. For instance, if the thermistor is being used to measure the temperature of the environment, due to the self-heating effect the electrical heating may introduce a significant error, which is known as a nonlinear function of the actual environment temperature. Moreover, the observers' responses may depend on the input instance. With these considerations in mind, we propose a more sophisticated model which assumes that $\{g_{m,d}\}$ is a nonlinear mapping from $\mathcal{X} \times \mathcal{Z}$ to $\mathcal{Y}$. By representing $\{g_{m,d}\}$ as the Gaussian process, the second conditional distribution in (1) has the form of

$$p(\mathbf{Y} \,|\, \mathbf{Z}, \mathbf{X}) = \prod_{m=1}^{M} \prod_{d=1}^{D} \mathcal{N}\left(\mathbf{y}_{:,m,d} \,|\, \mathbf{0}, \mathbf{S}_{m,d}\right), \tag{17}$$

where $\mathbf{Y}$ is connected with $\mathbf{X}$ and $\mathbf{Z}$ by a $N \times N$ kernel matrix $\mathbf{S}_{m,d}$. The $(i,j)^{\text{th}}$ element in $\mathbf{S}_{m,d}$ is given by

$$\begin{aligned} s_{m,d}\left(\{\mathbf{z}_i, \mathbf{x}_i\}, \{\mathbf{z}_j, \mathbf{x}_j\}\right) := \phi_{m,1,d}^2 \exp\left[-\frac{\phi_{m,2,d}^2}{2}(z_{i,d} - z_{j,d})^2\right] + \phi_{m,3,d}^2 \\ + \phi_{m,4,d}^2 z_{i,d} z_{j,d} + \phi_{m,5,d}^2 \delta(z_{i,d}, z_{j,d}) \\ + \phi_{m,6,d}^2 \exp\left[-\frac{1}{2}\sum_{l=1}^{L} \eta_{m,l,d}^2 (x_{i,l} - x_{j,l})^2\right], \end{aligned} \tag{18}$$

where $x_{i,l}$ is the l^{th} dimension of the instance $\mathbf{x}_i$. This covariance function has a similar form as Eq. (5), but with the addition of an *automatic relevance determination* kernel on $\mathbf{X}$. By incorporating a separate parameter $\eta_{m,l,d}$ for each input dimension l, we can optimize these parameters to infer the relative importance of different dimensions of an instance from the data. One can see that, as $\eta_{m,l,d}$ becomes small, the response $y_{n,m,d}$ becomes relatively insensitive to $x_{n,l}$. This allows us to detect the dimensions of $\mathcal{X}$ that substantially affect the observer's response.

Parameter Estimation. The observer model in Eq. (17) can be combined with Eq. (4) to form our new model,

$$p(\mathbf{Y} \,|\, \mathbf{X}, \boldsymbol{\Theta}) = \int p(\mathbf{Y} \,|\, \mathbf{Z}, \mathbf{X}, \boldsymbol{\Theta}) p(\mathbf{Z} \,|\, \mathbf{X}, \boldsymbol{\Theta}) \mathrm{d}\mathbf{Z},$$

where $\boldsymbol{\Theta} := \{\{\kappa_{1,d}, \ldots, \kappa_{5,d}\}, \{\phi_{m,1,d}, \ldots, \phi_{m,6,d}\}, \{\eta_{m,l,d}\}\}$ is the set of model parameters to be inferred from the data. Unfortunately, such marginalization of $\mathbf{Z}$ intractable as the latent variable $\mathbf{z}$ appears nonlinear in the kernel matrix. Instead, we seek a *maximum a posterior* (MAP) solution by maximizing

$$\log p(\mathbf{Z}, \boldsymbol{\Theta} \,|\, \mathbf{Y}, \mathbf{X}) = \log p(\mathbf{Y} \,|\, \mathbf{Z}, \mathbf{X}, \boldsymbol{\Theta}) + \log p(\mathbf{Z} \,|\, \mathbf{X}, \boldsymbol{\Theta}) + \text{constant}, \tag{19}$$

with respect to $\mathbf{Z}$ and $\boldsymbol{\Theta}$. Substituting Eq. (17) and Eq. (4) into Eq. (19) gives

$$\begin{aligned} F^{\text{NLOB}} := \log p(\mathbf{Z}, \boldsymbol{\Theta} \,|\, \mathbf{Y}, \mathbf{X}) = -\frac{1}{2}\sum_{d=1}^{D}\sum_{m=1}^{M}\Big(\ln|\mathbf{S}_{m,d}| + \operatorname{tr}(\mathbf{S}_{m,d}^{-1}\mathbf{y}_{:,m,d}\mathbf{y}_{:,m,d}^{\top})\Big) \\ -\frac{1}{2}\sum_{d=1}^{D}\left(\ln|\mathbf{K}_d| + \operatorname{tr}(\mathbf{K}_d^{-1}\mathbf{z}_{:,d}\mathbf{z}_{:,d}^{\top})\right) + \text{constant}. \end{aligned} \tag{20}$$

The partial derivative of F^{NLOB} with respect to the latent variable is given by

$$\frac{\partial F^{\text{NLOB}}}{\partial \mathbf{z}_{:,d}} = \operatorname{tr}\left(\left(\mathbf{S}^{-1}_{m,d}\mathbf{y}^{\top}_{:,m,d}\mathbf{y}_{:,m,d}\mathbf{S}^{-1}_{m,d} - \mathbf{S}^{-1}_{m,d}\right)\frac{\partial \mathbf{S}_{m,d}}{\partial \mathbf{z}_{:,d}}\right) - \mathbf{K}_d^{-1}\mathbf{z}_{:,d}. \tag{21}$$

The gradients with respect to the parameters of kernel matrix can be likewise derived as in the linear observer model. Finally, these gradients are used in the L-BFGS algorithm for maximizing F^{NLOB}.

When the algorithm converges, the estimate of ground truth is directly given by the stationary point of F^{NLOB}. Predicting the response of a new instance can be carried out in the same way as in Eq. (11). Moreover, the estimation of the m^{th} observer's response is given by

$$\widetilde{y}_{*,m,d} = \mathbf{s}_*\mathbf{S}^{-1}_{m,d}\widetilde{\mathbf{y}}_{:,m,d},$$

where $\mathbf{s}_* := [s_{m,d}(\widetilde{\mathbf{z}}_*, \widetilde{\mathbf{z}}_1, \mathbf{x}_*, \mathbf{x}_1), \ldots, s_{m,d}(\widetilde{\mathbf{z}}_*, \widetilde{\mathbf{z}}_N, \mathbf{x}_*, \mathbf{x}_N)]$.

Initialization. Note that seeking the MAP solution of $\mathbf{Z}$ and $\boldsymbol{\Theta}$ simultaneously may lead to a bad local optimum. Specifically, the model may stuck in a solution where $\{f_d\}$ is too trivial (e.g. close to a constant) and $\{g_{m,d}\}$ is too complicated (e.g. highly non-linear), which contradicts our intuition. To mitigate this problem, we first fit the training data with the linear observer model. The idea is to find an initial approximation of $\{f_d\}$ by restricting $\{g_{m,d}\}$ as linear. Then, we take $\widetilde{\mathbf{Z}}$ estimated by the linear observer model as the initialization of the ground truth, and train the nonlinear observer model to further refine $\{f_d\}$ and $\{g_{m,d}\}$.

4 Experimental Results

To evaluate the performance of our algorithm on predicting the ground truth and the observers' responses, we set up two experiments[1]. First, the effectiveness of our models is demonstrated on the synthetic data. The second experiment is conducted on the real-world data. In both experiments, the ground truth is known and observers' responses are simulated by mapping the ground truth with some random nonlinear functions. As a consequence, the performance can be evaluated straightforwardly. Two metrics are considered here, i.e. the mean absolute normalized error (MANE) and the Pearson correlation coefficient (PCC). In MANE, we first rescale the actual value and its predicted value into $[0, 1]$ respectively, and then measure the mean absolute error. MANE value close to 0 and PCC value close to 1 indicate that the algorithm performs well. In particular, the expected MANE of a random predictor is 0.5.

The proposed linear observer model (`LOB`) and nonlinear observer model (`NLOB`) are compared with several baselines. We first refer `SVR` and `GPR` as the Support Vector Regression and Gaussian Process Regression trained with the ground truth, respectively. Then we combine responses from multiple observers by taking the average and then using it for training, which we denote as `SVR-AVG` and `GPR-AVG`, respectively.

[1] For reproducing the experimental results, our MATLAB implementation is available at `http://home.in.tum.de/~xiaoh`.

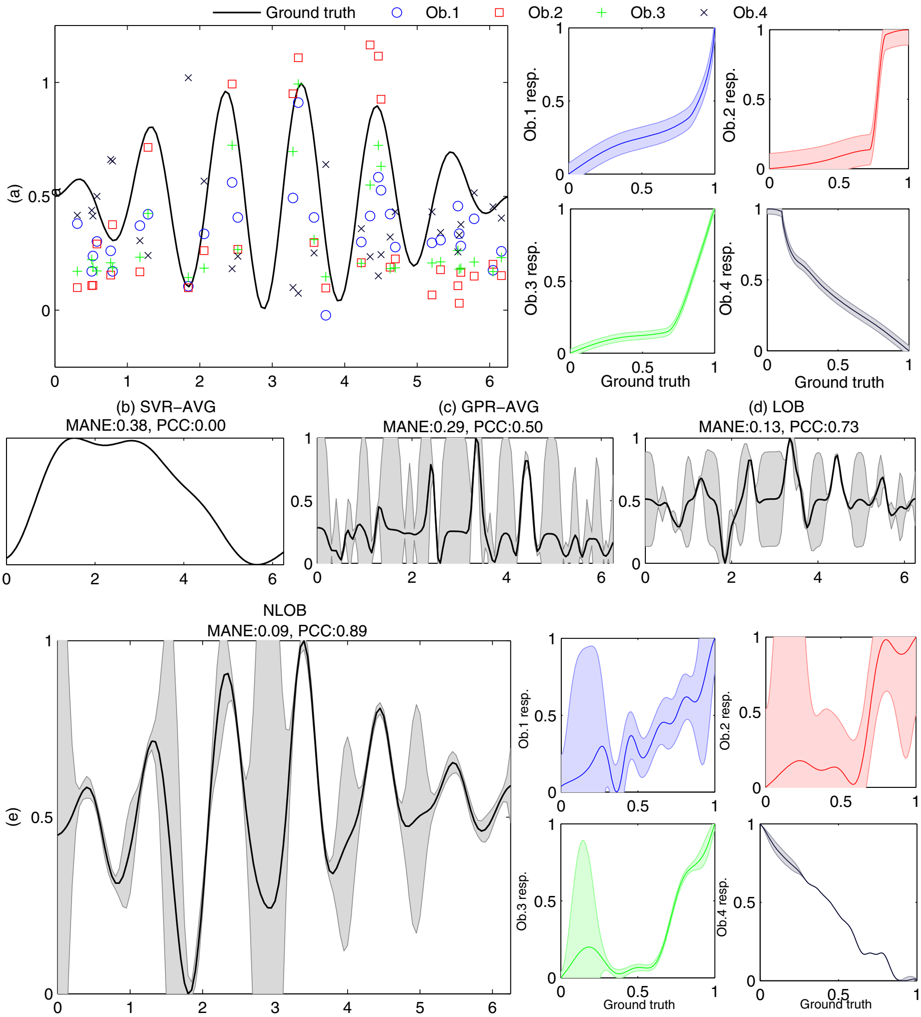

Fig. 3. (a) Synthetic data generated for the experiment. Responses from observers are represented by markers with different colors. The right panel illustrates randomly generated $\{g_m\}$ used for simulating four observers. Shaded area represents the pointwise variance. Note that the 4^{th} observer is *adversarial*, as his response tends to be the *opposite* of the ground truth. **(b, c, d)** Predicted ground truth on the test set by applying `SVR-AVG`, `GPR-AVG` and `LOB`, respectively. **(e)** Predicted ground truth and learned observer functions given by `NLOB`.

For a fair comparison, the covariance function of $\mathbf{x}$ in `GPR` and `GPR-AVG` has the same composite form as in Eq. (5). In addition to these non-parametric methods, `Raykar` refers to the model in which both $p(\mathbf{Z} \mid \mathbf{X})$ and $p(\mathbf{Y} \mid \mathbf{Z})$ are Gaussian in the spirit of [6].

4.1 Synthetic Examples

To create one-dimensional synthetic data (i.e. $L := 1$ and $D := 1$), we set $f(x) := \sin(6x)\sin(\frac{x}{2})$. The training instances $\mathbf{X}$ are generated by randomly sampling 30 points in $[0, 2\pi]$ from the uniform distribution. The test instances are obtained using a discretization of $[0, 2\pi]$ with equal space of 0.05, which results in 126 points. Four simulated observers are obtained by setting the corresponding $\{g_m\}$ as a random nonlinear monotonic function. For a training instance x, the m^{th} observer provides its response by $g_m(f(x))$ plus some Gaussian noise. An illustration of our synthetic data is depicted in Fig. 3(a). Figure 3(b, c, d, e) shows the results given by the baselines and our method. Not surprisingly, taking the average of observers' responses is not an effective solution. In contrast, our LOB and NLOB models outperform baseline methods significantly, which yield lower MANE and higher PCC. Moreover, the observers' functions learned by NLOB are very close to those predefined $\{g_m\}$ in Fig. 3(a).

4.2 On Real-World Data

We download four real-world data sets from UCI Machine Learning Repository, namely AUTO, COMMUNITY, CONCRETE and WINE. On each data set, we randomly select 500 instances and generate 20 observers in the same manner as in Section 4.1. The number of adversarial observers is fixed to 6. The experiment is conducted with 10-fold cross-validation. The prediction result of the ground truth and observers' responses is summarized in Table 1. It is notable that the proposed LOB and NLOB significantly outperform SVR/GPR-AVG and Raykar on inferring the ground truth. In general, additional improvements are observed when NLOB is used. Comparing it with the SVR/GPR column, one can see that the regression function learned by NLOB is almost as good as the one trained using the ground truth. We remark that the promising performance of NLOB is achieved by merely learning from a set of observers without any prior knowledge of their expertise and the ground truth. Furthermore, LOB and NLOB also show encouraging performance on predicting responses of observers, which can be proved useful in many applications such as the recommendation system.

Table 1. Prediction of the ground truth and observers' responses. In each cell, the upper value is MANE, while PCC is at the bottom. For the ground truth and the average baselines we only report the best performance, where a superscript $^{\text{S}}$ denotes that the performance is achieved by SVR or SVR-AVG; for GPR and GPR-AVG we use the superscript $^{\text{G}}$. The best model on each data set is highlighted by bold font. Note that only LOB and NLOB can predict observers' responses.

Data set	Ground truth					Observers' responses	
	SVR/GPR	SVR/GPR-AVG	Raykar	LOB	NLOB	LOB	NLOB
AUTO	$0.19 \pm 0.05^{\text{G}}$ $0.84 \pm 0.07^{\text{G}}$	$0.21 \pm 0.07^{\text{G}}$ $0.63 \pm 0.43^{\text{G}}$	0.25 ± 0.08 0.50 ± 0.22	0.26 ± 0.05 $\mathbf{0.84 \pm 0.05}$	$\mathbf{0.20 \pm 0.04}$ 0.82 ± 0.08	0.26 ± 0.04 $\mathbf{0.75 \pm 0.05}$	$\mathbf{0.25 \pm 0.09}$ 0.70 ± 0.11
COMMUNITY	$0.15 \pm 0.03^{\text{G}}$ $0.80 \pm 0.08^{\text{G}}$	$0.27 \pm 0.08^{\text{S}}$ $0.44 \pm 0.38^{\text{S}}$	0.22 ± 0.10 0.70 ± 0.13	0.17 ± 0.03 0.76 ± 0.04	$\mathbf{0.16 \pm 0.03}$ $\mathbf{0.77 \pm 0.04}$	0.26 ± 0.04 $\mathbf{0.62 \pm 0.09}$	$\mathbf{0.25 \pm 0.09}$ 0.55 ± 0.15
CONCRETE	$0.15 \pm 0.02^{\text{G}}$ $0.76 \pm 0.08^{\text{G}}$	$0.22 \pm 0.08^{\text{G}}$ $0.60 \pm 0.46^{\text{G}}$	0.20 ± 0.08 0.66 ± 0.21	0.18 ± 0.07 0.78 ± 0.11	$\mathbf{0.17 \pm 0.06}$ $\mathbf{0.79 \pm 0.09}$	0.26 ± 0.04 0.66 ± 0.18	$\mathbf{0.15 \pm 0.06}$ $\mathbf{0.72 \pm 0.15}$
WINE	$0.20 \pm 0.06^{\text{G}}$ $0.67 \pm 0.12^{\text{G}}$	$0.30 \pm 0.05^{\text{S}}$ $0.52 \pm 0.30^{\text{G}}$	0.29 ± 0.06 0.38 ± 0.19	0.27 ± 0.09 0.58 ± 0.20	$\mathbf{0.25 \pm 0.07}$ $\mathbf{0.61 \pm 0.17}$	0.32 ± 0.07 0.47 ± 0.18	$\mathbf{0.24 \pm 0.07}$ $\mathbf{0.48 \pm 0.15}$

5 Conclusion

This paper investigates the regression problem under multiple observers providing responses that are not absolutely accurate. The problem involves learning a regression function and observers' expertise from such data without any prior information of the observers. Based on the Gaussian process, we propose a probabilistic framework and develop two models. Our approach provides an estimate of the ground truth and also predicts the responses of each observer given new instances. Experiments show that the proposed method outperforms several baselines and leads to a performance close to the model trained with the ground truth.

There are many opportunities for future research. One possible direction is to extend our model with *multiple kernel learning*. The idea is to let the algorithm pick or composite different covariance functions instead of fixing the combination in advance. As a consequence, the algorithm may learn complex fits for the observers by selecting multiple kernels in a data-dependent way. Moreover, it would be highly beneficial to design *active sampling* methods for selecting which instance and whose response should be learned next.

References

1. Chen, S., Zhang, J., Chen, G., Zhang, C.: What if the irresponsible teachers are dominating? In: Proc. 24th AAAI (2010)
2. Crammer, K., Kearns, M., Wortman, J.: Learning from multiple sources. JMLR 9, 1757–1774 (2008)
3. Dawid, A., Skene, A.: Maximum likelihood estimation of observer error-rates using the em algorithm. Applied Statistics, 20–28 (1979)
4. Hui, S., Walter, S.: Estimating the error rates of diagnostic tests. Biometrics, 167–171 (1980)
5. Raykar, V., Yu, S., Zhao, L., Jerebko, A., Florin, C., Valadez, G., Bogoni, L., Moy, L.: Supervised learning from multiple experts: Whom to trust when everyone lies a bit. In: Proc. 26th ICML, pp. 889–896. ACM (2009)
6. Raykar, V., Yu, S., Zhao, L., Valadez, G., Florin, C., Bogoni, L., Moy, L.: Learning from crowds. JMLR 11, 1297–1322 (2010)
7. Smyth, P., Fayyad, U., Burl, M., Perona, P., Baldi, P.: Inferring ground truth from subjective labelling of venus images. In: Proc. 9th NIPS, pp. 1085–1092 (1995)
8. Spiegelhalter, D., Stovin, P.: An analysis of repeated biopsies following cardiac transplantation. Statistics in Medicine 2(1), 33–40 (1983)
9. Tubaishat, M., Madria, S.: Sensor networks: an overview. IEEE Potentials 22(2), 20–23 (2003)
10. Whitehill, J., Ruvolo, P., Wu, T., Bergsma, J., Movellan, J.: Whose vote should count more: Optimal integration of labels from labelers of unknown expertise. In: Proc. 23rd NIPS, vol. 22, pp. 2035–2043 (2009)
11. Wu, O., Hu, W., Gao, J.: Learning to rank under multiple annotators. In: Proc. 22nd IJCAI (2011)
12. Yan, Y., Rosales, R., Fung, G., Dy, J.: Active learning from crowds. In: Proc. 28th ICML (2011)
13. Yan, Y., Rosales, R., Fung, G., Schmidt, M., Hermosillo, G., Bogoni, L., Moy, L., Dy, J., Malvern, P.: Modeling annotator expertise: Learning when everybody knows a bit of something. In: Proc. AISTATS (2010)