Correlated Chained Gaussian Processes with Multiple Annotators

J. Gil-González, J. Giraldo, A. Álvarez-Meza, A. Orozco-Gutiérrez, and and M. Álvarez-López

***Abstract*—A dataset to train a supervised learning algorithm comprises features and their corresponding labels. The labeling process is usually carried out by an expert, which provides the ground truth/gold standard for each sample. However, in many real-world applications, instead of such a gold standard, we usually have access to annotations provided by crowds, which hold different and unknown expertise levels. Thus, Learning from crowds is a subject undergoing intense study, and its main aim is to face different machine learning paradigms in the presence of multiple annotators. Most state-of-the-art approaches reside on two key assumptions: i) the labeler's performance does not depend on the input feature space, and ii) independence among the annotators is imposed. Accordingly, we introduce the chained Gaussian processes–(CGPMA) model and the correlated chained Gaussian processes–(CCGPMA) model to deal with multi-labelers problems. Namely, CGPMA allows modeling each annotator's performance as a function of the input space. Similarly, CCGPMA is an extension of CGPMA to exploit correlations among the labelers' answers. Experimental results devoted to regression and classification show that our CCGPMA achieves suitable performances from inconsistent labelers, even if the gold standard is not available.**

***Index Terms*—Multiple annotators, Chained Gaussian Processes, Classification, Regression.**

I. Introduction

SUPERVISED learning requires the acquisition of high-quality training sets. The labeling process is performed by a domain expert, which is supposed to provide the absolute true label (termed gold standard/ground truth) (1). However, the experts are scarce; their time is expensive; the labeling task is tedious and time-consuming. Therefore, the gold standard's availability is hardly feasible for many real-world applications (2). An alternative is to distribute the labeling task to multiple heterogeneous annotators (also known as labelers or sources), where each labeler annotates part of the whole dataset by providing its version (possibly noisy) of the hidden ground truth (3). Those annotations can be collected in several ways. For instance, in (4), the sources are physicians who make a diagnosis about the presence of cancer, based on the analysis of medical images. Likewise, in (5), the labelers are algorithms that measure the QT interval in an electrocardiogram (ECG) signal. Besides, crowdsourcing platforms such as Amazon Mechanical Turk (AMT)[1] are becoming a valuable tool to obtain labels from multiple sources for large datasets in an efficient way (regarding labeling time). The attractiveness of these platforms lies in that at a low cost, it is possible to obtain suitable quality labels that, in some cases, can compete with those provided by experts (6).

Accordingly, in a multi-labeler setting, each input feature is matched with multiple annotations provided by different sources with unknown and diverse expertise. Thus, it is clear that such a problem cannot be faced with traditional supervised learning algorithms (7; 8). Taking the above into account, the question is how to perform predictions with information from multiple annotators? Such a question has been addressed by an area named *Learning from crowds*, which is undergoing intense study in the machine learning context (9). *Learning from crowds* offers two options to deal with multi-labeler problems data: *i)* to adapt the labels from multiple annotators or *ii)* to accommodate the supervised learning techniques (10).

The first approach is known in the literature as "Label aggregation" or "Truth inference". It comprises the computation of a single hard label per sample as an estimation of the unknown ground truth. These hard labels are then used to feed a standard supervised learning algorithm (11). The most straightforward approach is the so-called majority voting–(MV); it has been used in different multi-labeler problems due to its simplicity (12). However, MV assumes homogeneity in annotators' reliability, which is hardly feasible in real applications, e.g., experts vs. spammers. Furthermore, the consensus is profoundly impacted by incorrect labels and outliers (3). Conversely, more elaborated models have been considered to improve the estimation of the correct tag. For example, authors in (13) use an Expectation-Maximization (EM) algorithm to compute each labeler's reliability and infer the ground truth models. Besides, the work in (12) gives an estimation of the actual label by facing the problem of imbalanced labeling.

The second approach comprises the modification of supervised learning algorithms to jointly train the supervised learning algorithm and the annotators' behavior. It has been shown that such strategies lead to better performance compared to the ones belonging to Label aggregation. Namely, the features used to train the learning algorithm provides valuable information to puzzle out the ground truth (14). The most representative work in this area is the exposed in (4), which offers an EM-based framework to learn the parameters of a logistic regression classifier and model the annotators' behavior by computing their sensitivities and specificities. This work has inspired several models in the context of multi-labelers scenarios. For instance, for binary classification (15; 14), multi-

J. Gil-González and A. Orozco are with the Faculty of Engineering, Universidad Tecnológica de Pereira, Colombia, 660003, e-mail:jugil@utp.edu.co

J. Giraldo and M. Álvarez are with Department of Computer Science, University of Sheffield, UK.

A. Álvarez is with Universidad Nacional de Colombia sede Manizales, Colombia

[1]https://www.mturk.com/

class classification (11; 16), regression (9; 17), and sequence labeling (18). Furthermore, in the last years, some works have faced the multi-labeler problem using deep learning approaches (19; 20; 21). Chiefly, the idea of this kind of approach is to design an extra layer to code multiple annotators' information.

From the above, we can note an increasing interest in developing algorithms to deal with data from multiple labelers. Notwithstanding, as was analyzed in (22), there exist some problems that are not entirely solved; specifically, we are interested in the two following: *i)* to code the relationship between the input features and the labelers' behavior, and *ii)* to reveal the annotators' interdependencies. Regarding the first assumption, to model the annotators' behavior, it is necessary to learn some parameters related to their performance. Such parameters include the accuracy (23), the confusion matrix (16), the error variance (4), and the bias (17). In the literature, we commonly find that the parameters are modeled as fixed points (15) or as random variables (11), where it is considered that such parameters are homogeneous across the input data. The latter assumption is not correct since an expert makes decisions based not only on his/her expertise but also on the features observed from raw data (4). For the second assumption, it is widespread to consider independence among the annotators, aiming to reduce the complexity of the model (24), or based on the fact that it is plausible to guarantee that each labeler performs the annotation process individually (25). Notwithstanding, this assumption is not entirely true since there may exist correlations among the annotators (26). For example, if the sources are humans, the independence assumption is hardly feasible because knowledge is a social construction; then, people's decisions will be correlated since they share information, communicate with each other, or belong to a particular school of thought (27; 28). On the other hand, if we consider that the sources are algorithms, where some of them gather the same math principle, there likely exists a correlation among their labels (5). Accordingly, the relaxation of this restriction can improve the ground truth estimation (22).

Here, we propose a probabilistic framework based on Gaussian Processes–(GPs) to jointly build a prediction algorithm (regression and classification) and model the labelers' behavior as a function of the input features and taking into account annotators' interdependencies.
We perform an initial approach by applying the Chained GPs (CGP) (29) model to the prediction problem with multiple annotators. CGPs are a Multi-GPs framework, where the parameters of an arbitrary likelihood function are modeled with multiple independent GPs (one GP per parameter). From the multiple annotators' point of view, the likelihood's parameters are related to the labelers' behavior. Hence, by using CGPs for multiple annotators (CGPMA), we are modeling each annotator's performance as a function of the input features, which coincides with one of this work aims; however, due to CGP uses independent GPs, the modeling of dependencies (or correlation) among individual labelers is missing.
Conversely, aiming to model such labelers' interdependencies, we introduce an extension to the CGPs named Correlated Chained GP for multiple annotators–(CCGPMA), which induces correlations between multiple GPs that model the likelihood's parameters. To this end, we take as a basis the ideas from a Multi-output GP–(MOGP) regression (30), where each output is coded as a weighted sum of shared latent functions via a semi-parametric latent factor model (SLFM) (31). Nevertheless, conversely to MOGP, we do not have multiple outputs but multiple functions chained to the parameters of a given likelihood. Hence, to introduce correlations among the parameters' functions, we suppose them to be generated from an SLFM of Q latent functions, where each hidden function obeys a GP; thus, we are modeling the labelers' performance by taking into account interdependencies among them. The formulation of our CCGPs is based on the so-called inducing variables (32) in combination with stochastic variational inference (33), which make them scalable to large datasets. To the best of our knowledge, this is the first attempt to build a probabilistic approach that models both the interdependencies among the annotators and the relationship between the input features and the labelers' performance. Obtained results, using both simulated and real-world annotators, show how our methodology can deal with both regression and classification problems from multi-labelers data, outperforming state-of-the-art techniques.

The remainder is organized as follows. Section 2 exposes the related work and main contributions of the proposal. Section 3 describes the methods. Sections 4 and 5 present the experiments and discuss the results. Finally, Section 6 outlines the conclusions and future work.

II. Related work and main contributions

We note an increasing interest in developing algorithms to deal with data from multiple labelers. Notwithstanding, as was analyzed in (22), there exist some problems that are not entirely solved: *i)* to code the relationship between the input features and the labelers' behavior, and *ii)* to reveal the annotators' interdependencies.

First, aiming to model the annotators' behavior, it is necessary to learn some parameters related to their performance. Such parameters include the accuracy (23), the confusion matrix (16), the error variance (4), and the bias (17). In the literature, it is commonly founded that the parameters are modeled as fixed points (15) or as random variables (11), where it is considered that such parameters are homogeneous across the input data. The latter assumption is wrong since an expert makes decisions based not only on his/her expertise but also on the features observed from raw data (4). The first attempt to analyze the relationship between the annotators' parameters and the input features is the work in (26). The authors propose an approach for binary classification with multiple labelers, where the input data is represented by some clusters using a Gaussian Mixtures Model–(GMM). Then, they assumed that for each cluster, the annotators exhibit a particular performance measured in terms of sensitivity and specificity. However, this labelers' model is not adequate because they do not consider the information from multiple experts as an input for the GMM, which could generate some regions where the labelers' parameters may vary. In (34), the authors propose a binary

classification algorithm that employs two approaches to code the annotators' performance as a function of the input space: a Bernoulli and a Gaussian distribution. The parameters of these distributions are computing via a Logistic regression scheme. Nonetheless, they assume a linear dependence between the labeler expertise and the input space, which may not be appropriate in real-world scenarios. For example, if we consider online annotators assessing some documents, they may have different labeling accuracy. Such differences may rely on whether they are more familiar with some issues than other (8), which configures nonlinear behavior. Finally, (35) offers a GP-based regression scheme with multiple annotators. An additional GP is included to model the annotators' parameters as a nonlinear function of the input space. Nevertheless, such a model does not consider the uncertainty in the parameters (36).

Now, the independence among the annotators is commonly used to reduce the complexity of the model (24), or based on the fact that it is plausible to guarantee that each labeler performs the annotation process individually (25). Nevertheless, this is not entirely correct due to there may exist correlations among the annotators (26). For example, if the sources are humans, the independence assumption is hardly feasible because knowledge is a social construction; hence, people's decisions will be correlated when they share information, communicate with each other, or belong to a particular school of thought (27; 28). Accordingly, the relaxation of this restriction could be used to improve the ground truth estimation (22). To the best of our knowledge, only two works gather the dependencies among the labelers. First, authors in (5) expose an approach to deal with regression problems, where the labelers' behavior is addressed using a multivariate Gaussian distribution. Thus, the annotators' interdependencies are coded in the covariance matrix. In (37), authors propose a binary classification approach based on a weighted combination of classifiers, where each classifier models each annotator's response. In turn, the weights are estimated by using a kernel alignment-based approach considering dependencies among the labelers.

According to the related previously, we propose a framework based on GPs to face classification and regression settings with multiple annotators. Our proposal follows the line of the works in (15; 9; 14; 11; 38) in the sense that we are modeling the unknown ground truth trough a GP. However, while such approaches code the annotators' parameters as fixed points (15; 9) or as random variables (14; 11; 38); we model them as random processes (GPs) aiming to take into account dependencies between the input space and the labelers' behavior. On the other hand, our CGPMA and CCGPMA share some similarities (34; 35), due to these works assume dependencies between the input features and the labelers' performance. Besides, our CCGPMA is similar to approaches in (5; 37) because they assume the existence of dependencies among the annotators. Nonetheless, in contrast to these approaches, our CCGPMA is the only approach that includes such assumptions to code the annotators' behavior. Finally, We highlight that due to our approach is capable to model inconsistent annotators, it is more robust to outliers compared with such models that do not consider the relationship between the input features and the labelers' behavior. Namely, CCGPMA can estimate the annotators' performance for every region in the input space; meanwhile, the other approaches estimate such performance as an average of some parameters (17; 11; 14). Consequently, it is known that the average operator suffers under the presence of outliers (3). Table I summarizes the similarities and differences among our CGPMA, CCGPMA, and state-of-the-art approaches.

III. Methods

A. Supervised learning from multiple annotators

A supervised learning scenario involves the estimation of a function $g : \mathcal{X} \to \mathcal{Y}$ from a set $\{\boldsymbol{X} \in \mathcal{X}, \boldsymbol{y} \in \mathcal{Y}\}$, where $\boldsymbol{X} = \left\{\boldsymbol{x}_n \in \mathcal{X} \subseteq \mathbb{R}^P\right\}_{n=1}^{N}$ and $\boldsymbol{y} = \{y_n \in \mathcal{Y}\}_{n=1}^{N}$ are the input and output data, respectively. Here, depending on the nature of the output space $\mathcal{Y}$, it is possible to recognize different supervised learning settings: *i)* binary classification, holding $\mathcal{Y} \in \{-1, +1\}$; *ii)* multi-class classification, where $\mathcal{Y} \in \{1, \ldots, K\}$, being $K \in \mathbb{N}$ the number of classes, and *iii)* regression, where $\mathcal{Y} \subseteq \mathbb{R}$.

Commonly, each $\boldsymbol{x}_n$ is assigned to a single y_n, i.e., the ground truth. Still, in several real-world problems instead of the ground truth, we have multiple labels provided by $R \in \mathbb{N}$ annotators with different levels of expertise (4), where it is common to find that the r-th annotator only labels $N_r \leq N$ samples. Thereby, we build the set $\mathcal{D} = \{\boldsymbol{X}, \boldsymbol{Y} = \{y_n^r\}\}$, where $y_n^r \in \mathcal{Y}$ is the label given by labeler r to the sample n. Moreover, to code missing labels, we gather the set $R_n \subset \{1, \ldots, R\}$ holding the annotators that labelled the n-th instance ($r \in \{1, \ldots, R_n\}$). Given $\mathcal{D}$, A multi-labeler approach's main problems include i) the estimation of the unknown gold standard from the training set, ii) the annotators' performance assessment along with the input feature space, iii) the annotators' interdependencies coding, and iv) the model generalization against unseen data.

B. Chained Gaussian Processes for multiple annotators–(CGPMA)

To full-fill the issues above, we introduce a Gaussian Process–(GP)-based framework. First, we propose a chained Gaussian processes approach to model data from multiple annotators–(CGPMA), which uses various independent GPs to model the parameters of a given multi-labeler likelihood. Namely, such parameters are related to the unknown gold standard and the annotators' performance. Accordingly, our CGPMA estimates the unknown ground truth, assesses the labelers' performance as a function of the input features, and builds a supervised algorithm to make predictions on new data.

Let $\mathcal{D}$ be a multi-labeler annotator dataset; our CGPMA relies on estimating the following joint distribution (from multiple independent GPs priors):

$$\begin{aligned} p(\boldsymbol{Y}, \hat{\boldsymbol{f}}|\boldsymbol{X}) &= p(\boldsymbol{Y}|\boldsymbol{\theta}(\boldsymbol{x}))p(\hat{\boldsymbol{f}}|\boldsymbol{X}), \\ &= p(\boldsymbol{Y}|\boldsymbol{\theta}(\boldsymbol{x})) \prod_{j=1}^{J} \mathcal{N}(\boldsymbol{f}_j|\boldsymbol{0}, \boldsymbol{K}_{\boldsymbol{f}_j\boldsymbol{f}_j}), 5 \end{aligned} \quad (1)$$

where $\boldsymbol{\theta}(\boldsymbol{x}) = [\theta_1(\boldsymbol{x}), \ldots, \theta_J(\boldsymbol{x})] \in \mathbb{R}^J$ gathers the $J \in \mathbb{N}$ parameters representing the distribution over the label set

TABLE I
SURVEY OF RELEVANT SUPERVISED LEARNING MODELS DEVOTED TO MULTIPLE ANNOTATORS.

Source	Data type	Modeling the annotator's expertise	Expertise as a function of the input space	Modeling the annotators' interdependencies
Raykar et al., 2010 (4)	Regression-Binary-Categorical	✓	✗	✗
Zhang and Obradovic, 2011 (26)	Binary	✓	✓	✗
Xiao et al., 2013 (35)	Regression	✓	✓	✗
Yan et al., 2014 (34)	Binary	✓	✓	✗
Wang and Bi, 2016 (8)	Binary	✓	✓	✗
Rodrigues et al., 2017 (17)	Regression-Binary-Categorical	✓	✗	✗
Gil-Gonzalez et al., 2018 (37)	Binary	✓	✗	✓
Hua et al., 2018 (39)	Binary-Categorical	✓	✗	✗
Ruiz et al., 2019 (14)	Binary	✓	✗	✗
Morales- Alvarez et al., 2019 (11)	Binary	✓	✗	✗
Zhu et al., 2019 (5)	Regression	✓	✗	✓
Proposal-(CGPMA)	Regression-Binary-Categorical	✓	✓	✗
Proposal-(CCGPMA)	Regression-Binary-Categorical	✓	✓	✓

$\boldsymbol{Y}$ ($\boldsymbol{x} \in \boldsymbol{X}$). Each $\theta_j(\boldsymbol{x}) \in \mathcal{M}_j$, with $j \in \{1, \ldots, J\}$, holds a non-linear mapping from a GP-based prior $f_j(\boldsymbol{x})$, e.g., $\theta_j(\boldsymbol{x}) = h_j(f_j(\mathbf{x}))$, where $h_j : \mathbb{R} \rightarrow \mathcal{M}_j$ is a deterministic function that maps the GP output to the appropriate domain $\mathcal{M}_j$. Moreover, $\boldsymbol{f}_j = [f_j(\boldsymbol{x}_1), \ldots, f_j(\boldsymbol{x}_N)]^\top \in \mathbb{R}^N$ is a Latent Function (LF) vector that obey a GP prior, and $\hat{\boldsymbol{f}} = \left[\boldsymbol{f}_1^\top, \ldots, \boldsymbol{f}_J^\top\right]^\top \in \mathbb{R}^{NJ}$. $\boldsymbol{K}_{\boldsymbol{f}_j \boldsymbol{f}_j} \in \mathbb{R}^{N \times N}$ is the covariance matrix belonging to the j-th GP prior, which is computed based on the kernel function $\kappa_j : \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}$. Note that the CGPMA does not consider interdependencies among the labelers.

In turn, our CGPMA approximate inference framework is based on the inducing points-based method for sparse approximations of GPs (40). We introduce a set of $M \leq N$ "pseudo points" (known as inducing points) $\boldsymbol{u}_j = [f_j(\boldsymbol{z}_1^j), \ldots, f_j(\boldsymbol{z}_M^j)]^\top \in \mathbb{R}^M$ through additional evaluations of $f_j(\cdot)$ at unknown locations $\boldsymbol{Z}^j = [\boldsymbol{z}_1^j, \ldots, \boldsymbol{z}_M^j] \in \mathcal{X}^M$, which decreases the GP's computational complexity of $\mathcal{O}(N^3)$ to $\mathcal{O}(NM^2)$. Further, the following augmented GP prior arises:

$$p(\boldsymbol{f}_j, \boldsymbol{u}_j) = \mathcal{N}\left(\begin{bmatrix}\boldsymbol{f}_j \\ \boldsymbol{u}_j\end{bmatrix} \middle| \, \mathbf{0}, \begin{bmatrix}\boldsymbol{K}_{\boldsymbol{f}_j \boldsymbol{f}_j} & \boldsymbol{K}_{\boldsymbol{f}_j \boldsymbol{u}_j} \\ \boldsymbol{K}_{\boldsymbol{u}_j \boldsymbol{f}_j} & \boldsymbol{K}_{\boldsymbol{u}_j \boldsymbol{u}_j}\end{bmatrix}\right), \tag{2}$$

where $\boldsymbol{K}_{\boldsymbol{f}_j \boldsymbol{u}_j} \in \mathbb{R}^{N \times M}$ is the cross-covariance matrix formed by the kernel function $\kappa_j(\cdot, \cdot)$. Likewise, $\boldsymbol{K}_{\boldsymbol{u}_j \boldsymbol{u}_j} \in \mathbb{R}^{M \times M}$ is the inducing points-based covariance matrix. Then, the distribution of $\boldsymbol{f}_j$ conditioned to the inducing points $\boldsymbol{u}_j$ can be written as:

$$\begin{aligned} p(\boldsymbol{f}_j | \boldsymbol{u}_j) =& \mathcal{N}\left(\boldsymbol{f}_j | \boldsymbol{K}_{\boldsymbol{f}_j \boldsymbol{u}_j} \boldsymbol{K}_{\boldsymbol{u}_j \boldsymbol{u}_j}^{-1} \boldsymbol{u}_j, \boldsymbol{K}_{\boldsymbol{f}_j \boldsymbol{f}_j} - \cdots \right. \\ & \left. \cdots - \boldsymbol{K}_{\boldsymbol{f}_j \boldsymbol{u}_j} \boldsymbol{K}_{\boldsymbol{u}_j \boldsymbol{u}_j}^{-1} \boldsymbol{K}_{\boldsymbol{u}_j \boldsymbol{f}_j}\right), \end{aligned} \tag{3}$$

$$p(\boldsymbol{u}_j) = \mathcal{N}\left(\boldsymbol{u}_j | \mathbf{0}, \boldsymbol{K}_{\boldsymbol{u}_j \boldsymbol{u}_j}\right). \tag{4}$$

Solving the posterior from Eqs. (3) and (4) based on a non-Gaussian likelihood is no tractable analytically; therefore, a variational approximation is required. Let $p(\boldsymbol{Y}, \hat{\boldsymbol{f}}, \hat{\boldsymbol{u}})$ be a joint distribution with likelihood function $\prod_{n=1}^N p(\boldsymbol{y}_n | \boldsymbol{\theta}(\boldsymbol{x}_n))$ and an augmented prior:

$$p(\hat{\boldsymbol{f}}, \hat{\boldsymbol{u}}) = \prod_{j=1}^J p(\boldsymbol{f}_j | \boldsymbol{u}_j) p(\boldsymbol{u}_j), \tag{5}$$

where $\hat{\boldsymbol{u}} = \left[\boldsymbol{u}_1^\top, \ldots, \boldsymbol{u}_J^\top\right]^\top \in \mathbb{R}^{MJ}$ stores the inducing points. The actual posterior can be approximated by a parametrized variational $p(\hat{\boldsymbol{f}}, \hat{\boldsymbol{u}} | \boldsymbol{Y}) \approx q(\hat{\boldsymbol{f}}, \hat{\boldsymbol{u}})$, as:

$$q(\hat{\boldsymbol{f}}, \hat{\boldsymbol{u}}) = p(\hat{\boldsymbol{f}} | \hat{\boldsymbol{u}}) q(\hat{\boldsymbol{u}}) = \prod_{j=1}^J p(\boldsymbol{f}_j | \boldsymbol{u}_j) q(\boldsymbol{u}_j), \tag{6}$$

where $p(\boldsymbol{f}_j | \boldsymbol{u}_j)$ is defined in Eq. (3), and $q(\hat{\boldsymbol{u}})$ is the posterior approximation over the inducing variables:

$$q(\hat{\boldsymbol{u}}) = \prod_{j=1}^J q(\boldsymbol{u}_j) = \prod_{j=1}^J \mathcal{N}\left(\boldsymbol{u}_j | \boldsymbol{m}_j, \boldsymbol{V}_j\right). \tag{7}$$

Accordingly, the approximation for the posterior distribution comprises the estimation of the following variational parameters: the mean vectors $\boldsymbol{m}_j \in \mathbb{R}^M$ and the covariance matrices $\boldsymbol{V}_j \in \mathbb{R}^{M \times M}$. Such an assessment is carried out by maximizing an evidence lower bound–(ELBO). Thus, assuming that the instances $\boldsymbol{x}_n$ are independently sampled, the ELBO can be derived as:

$$\mathcal{L} = \mathbb{E}_{q(\hat{\boldsymbol{f}}, \hat{\boldsymbol{u}})}\left[\log\left(\frac{p(\boldsymbol{Y} | \hat{\boldsymbol{f}}) p(\hat{\boldsymbol{f}}, \hat{\boldsymbol{u}})}{q(\hat{\boldsymbol{f}}, \hat{\boldsymbol{u}})}\right)\right] \tag{8}$$

$$\begin{aligned} =& \sum_{n=1}^N \mathbb{E}_{\prod_{j=1}^J q(f_j(\boldsymbol{x}_n))}\left[\log\left(p\left(\boldsymbol{y}_n | \boldsymbol{\theta}(\boldsymbol{x}_n)\right)\right)\right] - \cdots \\ & \cdots - \sum_{j=1}^J \mathbb{D}_{KL}(q(\boldsymbol{u}_j) || p(\boldsymbol{u}_j)), \end{aligned} \tag{9}$$

where $\boldsymbol{y}_n \in \mathcal{Y}^{|R_n|}$ holds the labels given for the n-th instance, with $|\cdot|$ indicating the set cardinality. Besides, $\mathbb{D}_{KL}(\cdot || \cdot)$ is the Kullback-Leibler divergence and $q(\boldsymbol{f}_j)$ is defined as follows:

$$q(\boldsymbol{f}_j) = \int p(\boldsymbol{f}_j | \boldsymbol{u}_j) q(\boldsymbol{u}_j) d\boldsymbol{u}_j. \tag{10}$$

Solving Eq. (10), we have

$$\begin{aligned} q(\boldsymbol{f}_j) = & \mathcal{N}(\boldsymbol{f}_j | \boldsymbol{K}_{\boldsymbol{f}_j \boldsymbol{u}_j} \boldsymbol{K}_{\boldsymbol{u}_j \boldsymbol{u}_J}^{-1} \boldsymbol{m}_j, \boldsymbol{K}_{\boldsymbol{f}_j \boldsymbol{f}_j} + \cdots \\ & \cdots + \boldsymbol{K}_{\boldsymbol{f}_j \boldsymbol{u}_j} \boldsymbol{K}_{\boldsymbol{u}_j \boldsymbol{u}_j}^{-1} (\boldsymbol{V}_j - \boldsymbol{K}_{\boldsymbol{u}_j \boldsymbol{u}_j}) \boldsymbol{K}_{\boldsymbol{u}_j \boldsymbol{u}_j}^{-1} \boldsymbol{K}_{\boldsymbol{u}_j \boldsymbol{f}_j}). \end{aligned} \tag{11}$$

Yet, the computation of the variational expectations in Eq. (9) are intractable, due to for many likelihood functions, such expectations cannot be solved analytically (29; 41). Therefore, aiming to model different data types, i.e., classification and regression tasks, we need to find a generic alternative to solve

the integrals related to these expectations. In that sense, we use the Gaussian-Hermite quadratures approach as in (40; 29).

C. Correlated Chained Gaussian Processes fro multiple annotators-(CCGPMA)

The CGPMA introduced does not consider the annotators' dependencies because it models the likelihood's parameters using independent GPs. Conversely, we extend our CGPMA approach to modeling correlations between the GP latent functions, which are supposed to be generated from a semi-parametric latent factor model–(SLFM) (31), as follows:

$$f_j(\boldsymbol{x}_n) = \sum_{q=1}^{Q} w_{j,q}\mu_q(\mathbf{x}_n), \tag{12}$$

where $f_j : \mathcal{X} \to \mathbb{R}$ is a latent force function–(LF), $\mu_q(\cdot) \sim \mathcal{GP}(\mathbf{0}, \tilde{k}_q(\cdot,\cdot))$, $\tilde{k}_q : \mathcal{X}\times\mathcal{X} \to \mathbb{R}$ is a kernel function, and $w_{j,q} \in \mathbb{R}$ is a combination coefficient ($Q \in \mathbb{N}$). Here, each LF is chained to the likelihood's parameters, yielding:

$$p(\boldsymbol{Y}|\hat{\boldsymbol{f}}) = \prod_{n=1}^{N} p(\boldsymbol{y}_n|\boldsymbol{\theta}(\boldsymbol{x}_n)). \tag{13}$$

The augmented GP prior can be expressed as:

$$p(\hat{\boldsymbol{f}}, \hat{\boldsymbol{u}}) = \prod_{j=1}^{J} p(\boldsymbol{f}_j|\hat{\boldsymbol{u}})p(\hat{\boldsymbol{u}}), \tag{14}$$

where

$$p(\boldsymbol{f}_j|\hat{\boldsymbol{u}}) = \mathcal{N}\left(\boldsymbol{f}_j|\boldsymbol{K}_{\boldsymbol{f}_j\hat{\boldsymbol{u}}}\boldsymbol{K}_{\hat{\boldsymbol{u}}\hat{\boldsymbol{u}}}^{-1}\hat{\boldsymbol{u}}, \boldsymbol{K}_{\boldsymbol{f}_j\boldsymbol{f}_j} - \cdots \right. \\ \left. \cdots - \boldsymbol{K}_{\boldsymbol{f}_j\hat{\boldsymbol{u}}}\boldsymbol{K}_{\hat{\boldsymbol{u}}\hat{\boldsymbol{u}}}^{-1}\boldsymbol{K}_{\hat{\boldsymbol{u}}\boldsymbol{f}_j}\right), \tag{15}$$

$$p(\hat{\boldsymbol{u}}) = \mathcal{N}\left(\hat{\boldsymbol{u}}|\mathbf{0}, \boldsymbol{K}_{\hat{\boldsymbol{u}}\hat{\boldsymbol{u}}}\right), \tag{16}$$

where $\boldsymbol{K}_{\hat{\boldsymbol{u}}\hat{\boldsymbol{u}}} \in \mathbb{R}^{QM\times QM}$ is a block-diagonal matrix with blocks $\boldsymbol{K}_{\boldsymbol{u}_q\boldsymbol{u}_q} \in \mathbb{R}^{M\times M}$ (based on the kernel function $\hat{\kappa}_q : \mathcal{X}\times\mathcal{X} \to \mathbb{R}$). The vector $\hat{\boldsymbol{u}} = \left[\boldsymbol{u}_1^\top, \ldots, \boldsymbol{u}_Q^\top\right]^\top \in \mathbb{R}^{QM}$ is computed from the inducing points matrix $\boldsymbol{Z}^q = [\boldsymbol{z}_1^q, \ldots, \boldsymbol{z}_M^q] \in \mathcal{X}^M$, as follows: $\boldsymbol{u}_q = [\mu_q(\boldsymbol{z}_1^q), \ldots, \mu_q(\boldsymbol{z}_M^q)]^\top \in \mathbb{R}^M$. The covariance matrix $\boldsymbol{K}_{\boldsymbol{f}_j\boldsymbol{f}_j} \in \mathbb{R}^{N\times N}$ holds elements $\sum_{q=1}^{Q} w_{j,q}w_{j,q}\hat{\kappa}_q(\boldsymbol{x}_n, \boldsymbol{x}_{n'})$, with $\boldsymbol{x}_n, \boldsymbol{x}_{n'} \in \boldsymbol{X}$. Likewise, $\boldsymbol{K}_{\boldsymbol{f}_j\hat{\boldsymbol{u}}} = [\boldsymbol{K}_{\boldsymbol{f}_j\boldsymbol{u}_1}, \ldots, \boldsymbol{K}_{\boldsymbol{f}_j\boldsymbol{u}_Q}] \in \mathbb{R}^{N\times QM}$, where $\boldsymbol{K}_{\boldsymbol{f}_j\boldsymbol{u}_q} \in \mathbb{R}^{N\times M}$ gathers elements $w_{j,q}\hat{\kappa}_q(\boldsymbol{x}_n, \boldsymbol{z}_m^q)$, $m \in \{1, \ldots, M\}$. Again, to derive an ELBO, we define the following variational posterior:

$$q(\hat{\boldsymbol{f}}, \hat{\boldsymbol{u}}) = p(\hat{\boldsymbol{f}}|\hat{\boldsymbol{u}})q(\hat{\boldsymbol{u}}) = \prod_{j=1}^{J} p(\boldsymbol{f}_j|\hat{\boldsymbol{u}}) \prod_{q=1}^{Q} q(\boldsymbol{u}_q), \tag{17}$$

where $q(\boldsymbol{u}_q) = \mathcal{N}(\boldsymbol{u}_q|\boldsymbol{m}_q, \boldsymbol{V}_q)$ and $q(\hat{\boldsymbol{u}}) = \mathcal{N}(\hat{\boldsymbol{u}}|\boldsymbol{m}, \boldsymbol{V})$. Also, $\boldsymbol{m} = [\boldsymbol{m}_1^\top, \ldots, \boldsymbol{m}_Q^\top]^\top \in \mathbb{R}^{QM}$, $\boldsymbol{m}_q \in \mathbb{R}^M$, and $\boldsymbol{V} \in \mathbb{R}^{QM\times QM}$ is a diagonal block matrix computed from the covariance matrices $\boldsymbol{V}_q \in \mathbb{R}^{M\times M}$. Thus, a similar ELBO as in Eq. (9) can be derived, where the posterior $q(\boldsymbol{f}_j)$ is solved along $\hat{\boldsymbol{u}}$, yielding:

$$q(\boldsymbol{f}_j) = \mathcal{N}(\boldsymbol{f}_j|\boldsymbol{K}_{\boldsymbol{f}_j\boldsymbol{u}}\boldsymbol{K}_{\boldsymbol{u}\boldsymbol{u}}^{-1}\boldsymbol{m}, \boldsymbol{K}_{\boldsymbol{f}_j\boldsymbol{f}_j} + \cdots \\ \cdots + \boldsymbol{K}_{\boldsymbol{f}_j\boldsymbol{u}}\boldsymbol{K}_{\boldsymbol{u}\boldsymbol{u}}^{-1}(\mathbf{V} - \boldsymbol{K}_{\boldsymbol{u}\boldsymbol{u}})\boldsymbol{K}_{\boldsymbol{u}\boldsymbol{u}}^{-1}\boldsymbol{K}_{\boldsymbol{u}\boldsymbol{f}_j}). \tag{18}$$

Similar to CGPMA, the variational expectations are intractable for some likelihood functions. Again, the Gaussian-Hermite quadratures can be used.

Of note, CGPMA' inducing variables $\hat{\boldsymbol{u}}$ are not additional evaluations of $\hat{\boldsymbol{f}}$ but individual evaluations of each $\boldsymbol{u}_q(\cdot)$. Besides, CCGPMA conditions the distribution of each $\boldsymbol{f}_j$ to all Q latent functions μ_q; while in CGPMA $\boldsymbol{f}_j$ is only conditioned on an unique latent function $\boldsymbol{u}_j$ (independence assumption). Lastly, CGPMA is a particular case of CCGPMA when $Q = J$ and $w_{j,q} = 1$ if $j = q$, otherwise $w_{j,q} = 0$.

It is worth mentioning that the CGPMA and CCGPMA objective functions exhibit an ELBO that allows Stochastic Variational Inference–(SVI) (42). Hence, the optimization is solved through a *mini-batch*-based approach from noisy estimates of the global objective gradient, which allows leading with large scale datasets (40; 29; 41). The ELBO is optimized to estimate the variational parameters $\{\boldsymbol{m}_j, \boldsymbol{V}_j\}_{j=1}^{J}$ in CGPMA and $\{\boldsymbol{m}_q, \boldsymbol{V}_q\}_{q=1}^{Q}$ in CCGPMA. However, such ELBOs can be also used to infer the model's hyperparameters. Indeed, for CGPMA, we need to estimate the inducing points location and the kernel hyperparameters. Alike, for CCGPMA, we also require the combination factors associated with the LF.

D. CGPMA and CCGPMA applied to classification and regression tasks

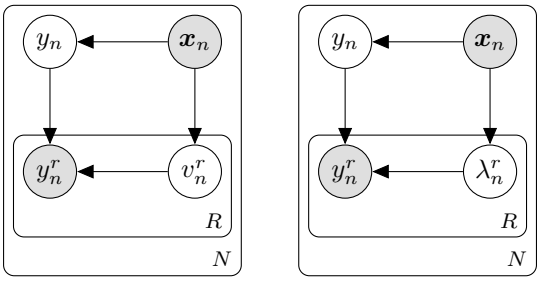
Our approach includes several multi-labeler likelihoods modulated by the latent functions $\boldsymbol{F}$ (41). We define two likelihoods, depending on the output domain, for concrete testing in this work: real-valued (regression) and categorical data (binary and multi-class classification). For real-valued outputs, e.g., $\mathcal{Y} \subset \mathbb{R}$, we follow the multi-annotator model used in (4; 9; 35; 17), where each output y_n^r is a corrupted version of the hidden ground truth y_n. Then, the likelihood function is given as:

$$p(\boldsymbol{Y}|\hat{\boldsymbol{f}}) = \prod_{n=1}^{N} \prod_{r\in R_n} \mathcal{N}\left(y_n^r|y_n, v_n^r\right), \tag{19}$$

where $v_n^r \in \mathbb{R}^+$ is the r-th annotator error-variance for the instance n. In turn, to model this likelihood function with CGPMA or CCGPMA, it is necessary to chain each likelihood's paramater to a latent function f_j. Thus, we require $J = R + 1$ LFs; one to model the hidden ground truth, such that $y_n = f_1(\boldsymbol{x}_n)$, and R LFs to model each of the error-variances $v_n^r = \exp(f_{l_r}(\boldsymbol{x}_n))$, with $r \in \{1, \ldots R\}$, and $l_r = r + 1 \in \{2, \ldots J\}$. Note that we use an exponential function to code f_{l_r} and v_n^r, aiming to guarantee $v_n^r > 0$ ($f_{l_r} \in \mathbb{R}$).

Next, to deal with binary and multi-class classification tasks, we use the model proposed in (23). Such a model introduces a binary variable $\lambda_n^r \in \{0, 1\}$, which indicates if the r-th labeler provides the correct label ($\lambda_n^r = 1$) or not ($\lambda_n^r = 0$); thus, $\boldsymbol{\lambda}^r$ codes the r-th labeler reliability. The likelihood function is defined as:

$$p(\boldsymbol{Y}|\hat{\boldsymbol{f}}) = \prod_{n=1}^{N} \prod_{r\in R_n} \left(\prod_{k=1}^{K} \zeta_{k,n}^{c_{k,n}^r}\right)^{\lambda_n^r} \left(\frac{1}{K}\right)^{1-\lambda_n^r}, \tag{20}$$



(a) Regression (b) Classification

Fig. 1. Graphical plates for regression and classification models (see Eq. (19) and Eq. (20)). Shaded nodes represent observed variables, and unshaded nodes indicate latent ones.

where $\zeta_{k,n}$ is the k-th component in $\boldsymbol{\zeta}_n \in \{1, \dots K\}^K$. The latter represents the unknown ground truth based on a $1-$of$-K$ encoding. Similarly, $c_{k,n}^r$ represents the r-th position in $\mathbf{c}_n^r \in \{1, \dots K\}^K$, which corresponds to the $1-$of$-K$ version of y_n^r. To model this likelihood function, we need $J = K + R$ LFs; K latent functions to model the hidden true label, yielding

$$\zeta_{k,n} = \text{Softmax}(f_k(\mathbf{x}_n)) = \frac{\exp(f_k(\mathbf{x}_n))}{\sum_{j=1}^K f_j(\mathbf{x}_n)}. \quad (21)$$

Besides, we need R LFs to model the reliability for each labeler; therefore, $\lambda_n^r = \sigma(f_{l_r}(\mathbf{x}_n))$ with $r \in \{1, \dots R\}$, and $l_r = K + r \in \{K+1, \dots J\}$. Here, $\sigma(f_{l_r}(\mathbf{x}_n))$ is the logistic sigmoid function: $\sigma(a) = 1/1 + \exp(-a)$, where $a \in \mathbb{R}$. The plate representations for classification and regression tasks are shown in Fig. 1. Note that the label generation y_n^r depends on the ground truth y_n. The annotators' parameters v_n^r or λ_n^r are related to their performance. Further, we remark that such annotators' parameters are modeled as independent functions of the input features using CGPMA, and as correlated functions through CCGPMA.[2]

IV. Experimental Set-up

A. Datasets and simulated/provided annotations

For both cases (regression and classification), we test our approaches CGPMA and CCGPMA using three types of datasets. Namely, *fully synthetic data*, *semi-synthetic data*, and *fully real datasets*.

1) Regression: First, we generate *fully synthetic data* as an one-dimensional regression problem, where the ground truth for the n-th sample correspond to $y_n = \sin(2\pi \mathbf{x}_n) \sin(6\pi \mathbf{x}_n)$, where the input matrix $\mathbf{X}$ is formed by randomly sampling 100 points in $[0, 1]$ from a uniform distribution. Besides, the test instances are obtained by extracting equally spaced samples from the interval $[0, 1]$.

Second, to control the label generation (14), we build *semi-synthetic data* from six datasets devoted regression from the well-known UCI repository. [3] The chosen datasets include: Auto MPG Data Set–(Auto), Bike Sharing Dataset Data Set–(Bike), Concrete Compressive Strength Data Set–(Concrete), The Boston Housing Dataset–(Housing),[4] Yacht Hydrodynamics Data Set–(Yacht), and Relative location of CT slices on axial axis Data Set–(CT).

TABLE II
DATASETS FOR REGRESSION.

	Name	Number of features	Number of instances
fully synthetic	synthetic	1	100
semi-synthetic	Auto	8	398
	Bike	13	17389
	Concrete	9	1030
	Housing	13	506
	Yacht	6	308
	CT	384	53500
fully real	Music	124	1000

Third, we evaluate our proposal on one *fully real dataset*, where both the input features and the annotations are captured from real-world problems. Namely, we use the music genre data, holding a collection of songs records labeled from one to ten depending on their music genre: classical, country, disco, hip-hop, jazz, rock, blues, reggae, pop, and metal. From this set, 700 samples were published randomly in the AMT platform to obtain labels from multiples sources, 2946 labels were obtained from 44 workers; however, we only take into account the annotators who labeled at least 20% of the available instances, thus, we use the information from $R = 7$ labelers. The feature extraction is performed by following the work by authors in (23), to obtain an input space with $P = 124$. Note that initially, this dataset configures a problem of multiclass classification (with 10 classes); however, in this experiment, we face it by using regression. Table II summarizes the tested datasets for the regression case.

As we pointed out previously, *fully synthetic* and *semi-synthetic* datasets do not hold real annotations. Thus, it is necessary to generate these labels synthetically as a version of the gold standard corrupted by Gaussian noise, i.e., $y_n^r = y_n + \epsilon_{r,n}$, where $\epsilon_{r,n} \sim \mathcal{N}(0, v_{r,n})$, being the r-th annotator error-variance for the sample n. Note that we are interested in to model such error-variance for the r-th annotator $\mathbf{v}_r$ as a function of the input features, $h_r(\cdot) : \mathcal{X} \to \mathbb{R}^+$, which is correlated with the variances of the other labelers. For doing so, the error variances are generated as follows:

- Define Q functions $u_q(\cdot)$, and the combination parameters $w_{r,q}$, $\forall r, n$.
- For each annotator r and sample n, compute $f_{r,n} = \sum_{q=1}^Q w_{r,q} u_q(x_n)$, where x_n is the n-th component of $\boldsymbol{x} \in \mathbb{R}$, which is an $1-$D representation of input features $\boldsymbol{X}$ by using the t-distributed Stochastic Neighbor Embedding approach (43).
- Finally, determine $v_{r,n} = \exp(f_{r,n})$.

2) Classification: Alike to regression, we test our CGPMA and CCGPMA approaches in three kinds of datasets. First, we generate *fully synthetic data* as one-dimensional multi-class classification problem (K=3). The input matrix $\mathbf{X}$ is

[2]The required equations for using CGPMA and CCGPMA with the classification and regression likelihoods are shown in appendix

[3]http://archive.ics.uci.edu/ml

[4]See https://www.cs.toronto.edu/~delve/data/boston/bostonDetail.html for housing

formed by randomly sampling $N = 100$ points in $[0, 1]$ from a uniform distribution. The true label for the n-th input sample $\mathbf{x}_n$ is generated by taking $\arg\max(t_{1,n}, t_{2,n}, t_{3,n})$, where each value $t_{1,n}, t_{2,n}, t_{3,n}$ is respectively generated by evaluating $\mathbf{x}_n$ in the following functions $\sin(2\pi\mathbf{x}_n))$, $-\sin(2\pi\mathbf{x}_n))$, and $-\sin(2\pi(\mathbf{x}_n + 0.25)) + 0.5$. Besides, the test instances are obtained by extracting 200 equally spaced samples from the interval $[0, 1]$.

Second, to control the label generation (14), we build *semi-synthetic data* from seven datasets devoted to binary and multi class-classification of the well-known UCI repository [5]. The chosen datasets include: Wisconsin Breast Cancer Database–(breast), BUPA liver disorders–(bupa), Johns Hopkins University Ionosphere database–(ionosphere), Pima Indians Diabetes Database–(pima), Tic-Tac-Toe Endgame database–(tic-tac-toe), Wine Data set–(Wine), and Image Segmentation Data Set–(Segmentation). Moreover, we use the publicly available bearing data collected by the Case Western Reserve University–(Western). The aim is to build a system to diagnose the status of an electric motor based on the information from two accelerometers. The feature extraction and selection stage was performed by following (44).

Third, we evaluate our proposal on two *fully real datasets*, where both the input features and the annotations are captured from real-world problems. Namely, we use a biosignal database, where the goal is to build a system to evaluate the presence/absence of voice pathologies. In particular, a subset ($N = 218$) of the Massachusetts Eye and Ear Infirmary Disordered Voice Database from the Kay Elemetrics company is utilized, which comprises voice records from healthy and different voice issues. Each record is parametrized by the Mel-frequency cepstral coefficients (MFCC) to obtain an input space with $P = 13$. A set of physicians assess the voice quality by following the GRBAS protocol that comprises the evaluation of five qualitative scales: Grade of dysphonia–(G), Roughness–(R), Breathiness–(B), Asthenia–(A), and Strain–(S). For each perceptual scale, the specialist assigns a tag ranging from 0 (healthy voice) to 3 (severe disease) (45). Accordingly, we face five multi-class classification problems (one per scale); however, we follow the procedure in (37), aiming to convert them into five binary classification problems due to the binary settings we have access to the ground truth (16). Further, the music genre data (used for the regression experiments) is also taken into account. We recall that such a dataset set configures a multi-class classification problem. Table III summarizes the tested datasets for the regression case.

Similar to the regression case, *fully synthetic* and *semi-synthetic* datasets do not hold real annotations. Therefore, it is necessary to simulation those labels as corrupted versions of the hidden ground truth. Such simulation is performed by taking into account that there exist dependencies among the annotators and that the labelers' performance is a function of the input features. Accordingly, we simulate these labels as follows

- Define Q functions $u_q(\cdot)$, and the combination parameters $w_{r,q}$, $\forall r, n$.
- For each annotator r and sample n, compute $f_{r,n} = \sum_{q=1}^{Q} w_{r,q} u_q(x_n)$, where x_n is the n-th component of $\boldsymbol{x} \in \mathbb{R}$, which is an $1-$D representation of input features $\boldsymbol{X}$ by using the t-distributed Stochastic Neighbor Embedding approach (43).
- Determine $\lambda_n^r = \sigma(f_{r,n})$.
- Finally, the label $y_n^r = \begin{cases} y_n, & \text{if } \lambda_n^r \geq 0.5 \\ \tilde{y}_n, & \text{if } \lambda_n^r < 0.5 \end{cases}$, where $\tilde{y}_n$ is the flipped version of the actual label y_n

[5]http://archive.ics.uci.edu/ml

TABLE III
DATASETS USED FOR CLASSIFICATION.

	Name	Number of features	Number of instances	Number of classes
fully synthetic	synthetic	1	100	3
semi-synthetic	Breast	9	683	2
	Bupa	6	345	2
	Ionosphere	34	351	2
	Pima	8	768	2
	Tic-tac-toe	9	958	2
	Western	7	3413	4
	Wine	13	178	3
	Segmentation	18	2310	7
fully real	Voice	13	218	2
	Music	124	1000	10

B. CGPMA and CCGPMA training

Overall, the Radial basis function–(RBF) kernel is preferred in pattern classification because of its universal approximating ability and mathematical tractability. Hence, all the kernel functions for both CGPMA and CCGPMA are fixed as

$$\kappa_j(\mathbf{x}_n, \mathbf{x}_{n'}) = \kappa_q(\mathbf{x}_n, \mathbf{x}_{n'}) = \theta_1 \exp\left(\frac{-\|\mathbf{x}_n - \mathbf{x}_{n'}\|_2^2}{2\boldsymbol{\theta}_2^2}\right), \tag{22}$$

where $\|\cdot\|_2$ is the L2 norm, $n, n' \in \{0, \ldots, N\}$, and $\theta_1 \in \mathbb{R}$ and $\boldsymbol{\theta}_2 \in (\mathbb{R}^+)^P$ ($\mathbb{R}^+$ stands for the positive real numbers) are the kernel hiperparameters. For concrete testing we fix $\theta_1 = 1$, while $\boldsymbol{\theta}_2$ is estimated by optimizing the corresponding ELBO for CGPMA and CCGPMA. We emphasize that all GP-based approaches in this work are based on this type of kernel.
Another relevant parameter in the training of our CGPMA and CCGPMA is the number of required LFs. For CGPMA, it is clear that Q have to be equal to the number of likelihood parameters J since each function $f_j(\cdot)$ is linked to a LF $u_q(\cdot)$; accordingly, for regression, we fix $Q = R+1$, and $Q = R+K$, for classification scenarios. On the other hand, for CCGPMA, each $f_j(\cdot)$ is built as a convex combination of LFs $u_q(\cdot)$ (see Eq. (12)), therefore, there is no any restriction about the number of Q. However, to make a fair comparison with CGPMA, we fix $Q = J$. Furthermore, for *fully synthetic datasets*, we use $M = 10$ inducing points per latent function, and for the remaining experiments, we test with $M = 40$, and $M = 80$. Finally, for all the experiments, we use stochastic inference with a mini-batch size of 100.

C. Method comparison and performance metrics

1) Regression: The quality assessment is carried out by estimating the regression performance as the coefficient of determination–(R^2). A cross-validation scheme is employed with 15 repetitions where 70% of the samples are utilized for

TABLE IV
A BRIEF OVERVIEW OF STATE-OF-THE-ART METHODS TESTED FOR REGRESSION TASKS. GPR: GAUSSIAN PROCESSES REGRESSION, LR: LOGISTIC REGRESSION, AV: AVERAGE, MA: MULTIPLE ANNOTATORS, DL: DEEP LEARNING, LFCR: LEARNING FROM CROWDS FOR REGRESSION.

Algorithm	Description
GPR-GOLD	A GPR using the real labels (upper bound).
GPR-Av	A GPR using the average of the labels as the ground truth.
MA-LFCR (4)	A LR model for MA where the labelers' parameters are supposed to be constant across the input space.
MA-GPR (15)	A multi-labeler GPR, which is as an extension of MA-LFCR.
MA-DL (20)	A Crowd Layer for DL, where the annotators' parameters are constant across the input space.

TABLE V
A BRIEF OVERVIEW OF THE STATE-OF-THE-ART METHODS TESTED. GPC: GAUSSIAN PROCESSES CLASSIFIER, LRC: LOGISTIC REGRESSION CLASSIFIER, MV: MAJORITY VOTING, MA: MULTIPLE ANNOTATORS, MAE: MODELLING ANNOTATORS EXPERTISE, LFC: LEARNING FROM CROWDS, DGRL: DISTINGUISHING GOOD FROM RANDOM LABELERS, KAAR: KERNEL ALIGNMENT-BASED ANNOTATOR RELEVANCE ANALYSIS.

Algorithm	Description
GPC-GOLD	A GPC using the real labels (upper bound).
GPC-MV	A GPC using the MV of the labels as the ground truth.
MA-LFC-C (4)	A LRC with constant parameters across the input space.
MA-DGRL (23)	A multi-labeler approach that considers as latent variables the annotator performance.
MA-GPC (15)	A multi-labeler GPC, which is as an extension of MA-LFC
MA-GPCV (11)	An extension of MA-GPC, by using variational inference and including priors over the labelers' parameters.
MA-DL (20)	A Crowd Layer for DL, where the annotators' parameters are constant across the input space
KAAR (37)	A kernel-based approach that employs a convex combination of classifiers and codes labelers dependencies.

training and the remaining 30% for testing (except for *fully synthetic dataset*, since it clearly defines the training and testing sets). Table IV displays the employed methods of the state-of-the-art for comparison purposes. From Table IV, we highlight that for the model MA-DL, the authors provided three different annotators' codification: MA-DL-B, where the bias for the annotators is measured; MA-DL-S, where the labelers' scale is computed; and measured; MA-DL-B+S, which is a version with both (20). Besides, it is worth clarifying that GPR-GOLD and GPR-Av are built from a sparse approximation based on inducing points (similar to the exposed in Section III-B) combined with stochastic variational inference; hence we use mini-batches with a size of 100. Further, for the *fully synthetic dataset* we fix $M = 10$, and we tested with $M = 40$, and $M = 80$ for the rest of experiments.

2) Classification: The classification performance is assessed as the Area Under the Curve–(AUC). Further, the AUC is extended for multi-class settings, as discussed by authors in (46). Similarly to the regression case, a cross-validation scheme is employed with 15 repetitions where 70% of the samples are utilized for training and the remaining 30% for testing (except for music dataset, since its clearly define the training and testing sets). Table V displays the employed methods of the state-of-the-art for comparison purposes.

V. RESULTS AND DISCUSSION

In this section, we expose the results regarding real-valued data (regression), and categorical data (classification), for the three types of cases that we have defined in previous section: *fully synthetic data*, *semi synthetic data*, and *fully real data*.

A. Regression

1) fully synthetic data: We perform a controlled experiment aiming to verify the capability of our CGPMA and CCGPMA to estimate the performance of inconsistent annotators as a function of the input space and taking into account their dependencies. For this first experiment, we use the *fully synthetic* dataset described in Section IV-A1. We simulate five labelers ($R = 5$) with different levels of expertise. To simulate the error-variances, we define $Q = 3$ functions $u_q(\cdot)$, which are given as

$$u_1(x) = 4.5\cos(2\pi x + 1.5\pi) - 3\sin(4.3\pi x + 0.3\pi) + \cdots \\ \cdots + 4\cos(7\pi x + 2.4\pi), \quad (23)$$

$$u_2(x) = 4.5\cos(1.5\pi x + 0.5\pi) + 5\sin(3\pi x + 1.5\pi) - \cdots \\ \cdots - 4.5\cos(8\pi x + 0.25\pi), \quad (24)$$

$$u_3(x) = 1, \quad (25)$$

where $x \in [0, 1]$. Besides, we define the following combination matrix $\mathbf{W} \in \mathbb{R}^{Q\times R}$, where

$$\mathbf{W} = \begin{bmatrix} -0.10 & 0.01 & -0.05 & 0.01 & -0.01 \\ 0.10 & -0.01 & 0.01 & -0.05 & 0.05 \\ -2.3 & -1.77 & 0.54 & 0.9 & 1.42 \end{bmatrix}, \quad (26)$$

holding elements $w_{r,q}$. Fig. 2 shows the predictive performance of all methods in this first experiment. The results show two clear groups: those based on GPs (GPR-Av, MA-GPR, CGPMA-R, and CCGPMA-R), which expose the best performance in terms of the R^2 score, and those based on other types of approaches (MA-LFCR, and MA-DL), whose performance is not satisfactory. The behavior of MA-LFCR is not unexpected since it only can deal with linear problems. Besides, concerning MA-DL and its three variations (S, B, and S+B), we note that this approach, in general terms, has the capability of modeling the non-linearities present in the regression problem; however, MA-DL exposes a significant low performance (even lower than the most naive approach, GPR-Av), which is a bit surprising due to the DL models have shown significant advances in the artificial intelligence context (47). Such an outcome is explained because these crowd layers provide a very simple codification of the annotators' performance to guarantee a low computational cost (38); therefore, MA-DL does not provide a proper codification of the annotators' behavior.
Among the GP-based methods, the proposed CCGPMA-R exposes the best result, followed closely by our CGPMA-R and MA-GPR. Notice that the performance of GPR-Av is also close to them, which indicates that this experiment seems not to be too challenging for the GPs-based approaches; nevertheless, notorious differences will be appreciated in the next experiments.

On the other hand, concerning the significant high performance of our CCGPMA-R (the best in terms of R^2 score), we hypothesize that such an outcome is a consequence that our approach offers a better representation of the labelers' behavior when compared with its competitors. To empirically support the

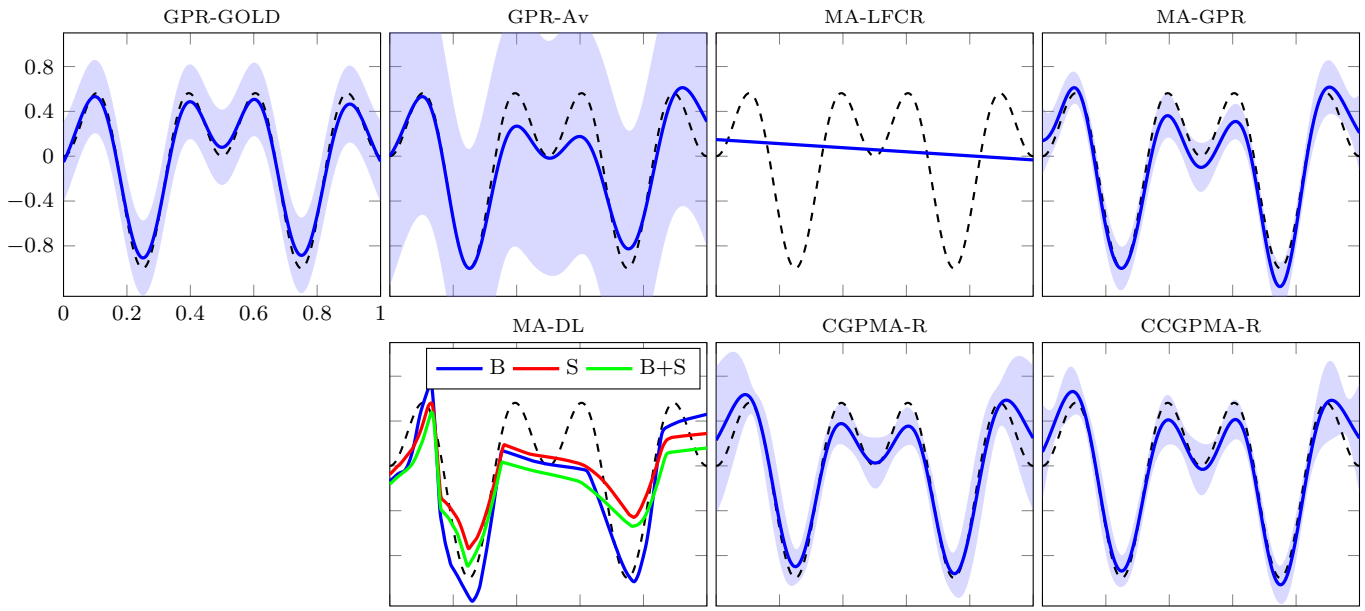


Fig. 2. Fully synthetic dataset results. We compare the prediction of our CCGPMA-R($R^2 = 0.9438$), and CCGPMA-R($R^2 = 0.9280$) with the theoretical upper bound GPR-GOLD($R^2 = 0.9843$) and lower bound GPR-Av($R^2 = 0.8718$), and state-of-the-art approaches, MA-LFCR($R^2 = -0.0245$), MA-GPR($R^2 = 0.9208$), MA-DL-B($R^2 = 0.7020$), MA-DL-S($R^2 = 0.6559$), MA-DL-B+S($R^2 = 0.5997$). Note that we provided the Gold Standard in dashed lines. The shaded region in GPR-Av, MA-GPR, CGPMA-R, and CCGPMA-R indicates the area enclosed by the mean plus or minus two standard deviations. We remark that there is no shaded region for MA-LFCR, and DLMA since these approaches do not provide information about the prediction uncertainty.

above hypothesis, Fig. 3 shows the estimated error-variances for this first experiment; here, we only take into account the models that include these parameters in their formulations. From Fig. 3, we note from the R^2 score and making a visual inspection that the approaches MA-LFCR and MA-GPR (see columns 2 and 3 in Fig. 3) offer the worst representation for the annotator's performance, which is expected due to such models do not take into account the relationship between the annotators' performance and the input space. Conversely, our CGPMA-R and CCGPMA-R (see columns 4 and 5 in Fig. 3) clearly outperforms the models named previously. This outcome is a consequence that our two approaches compute such error-variances as functions of the input features, allowing for a better codification of the labelers' behavior. Besides, by making a visual inspection and analyzing the R^2 scores, CCGPMA-R performs better that CGPMA-R; in principle, this is a bit unexpected since both approaches compute the labelers' parameters as non-linear functions; however, we highlight that contrary to CGPMA-R, CCGPMA-R models the annotators' interdependencies, which improves the modeling of such performances as was empirically demonstrated in (5). Finally, we remark that although our CCGPMA-R offers the best representation of the annotators' performance, the results for Annotators 2 and 3 (rows 2 and 3 in the fifth column of Fig. 3) seem to be unsatisfactory. Such an outcome is caused by the quasi-periodic behavior in the error-variances for those labelers, which cannot be captured by our approach because we are using a kernel RBF, as we pointed out in Section IV-B.

2) Semi-synthetic data results: as for the *fully synthetic data* (see Eqs. (23) to (26)). Table VI shows the results for this second experiment with *semi synthetic dataset*. On average, our CCGPMA-R exhibits the best generalization performance in terms of the R^2 score. Besides, regarding its GPs-based competitors (GPR-Av, MA-GPR, and CGPMA-R), we first note that the performance of our CGPMA-R is a bit lower than CCGPMA-R. The above is not an unexpected outcome since both approaches estimate the annotators' performances as functions of the input features, which fixes the process that we use to simulate the labels for this experiment (see **??**). Secondly, as expected, the intuitive lower bound GPR-Av exhibits a significantly worse prediction than our approaches. On the other hand, the behavior of MA-GPR is surprising due to it exhibits the most deficient prediction capability, even far worse than the supposed lower bound GPR-Av. The key to this abnormal behavior lies in the formulation of this approach; MA-GPR is based on a basic Gaussian process (i.e., without considering sparse approximations neither stochastic variational inference) that cannot handle large datasets. In fact, we note that MA-GPR exhibits the worst performances for datasets Bike, Concrete, and CT, which are conformed by a large number of samples (Respectively 17389, 1030, and 53500). Next, we analyze the results concerning the linear model MA-LFR; attained to the results, we note that this approach's prediction capacity is not satisfactory since its performance is far lower than our approaches; the above outcome suggests that there may exist a non-linear structure in most databases. However, we highlight a particular result for the dataset CT, where MA-LFCR exhibits the best performance defeating all its competitors based on non-linear models. From the above, we intuit that the CT dataset may have a linear structure. To confirm this supposition, we perform an additional experiment

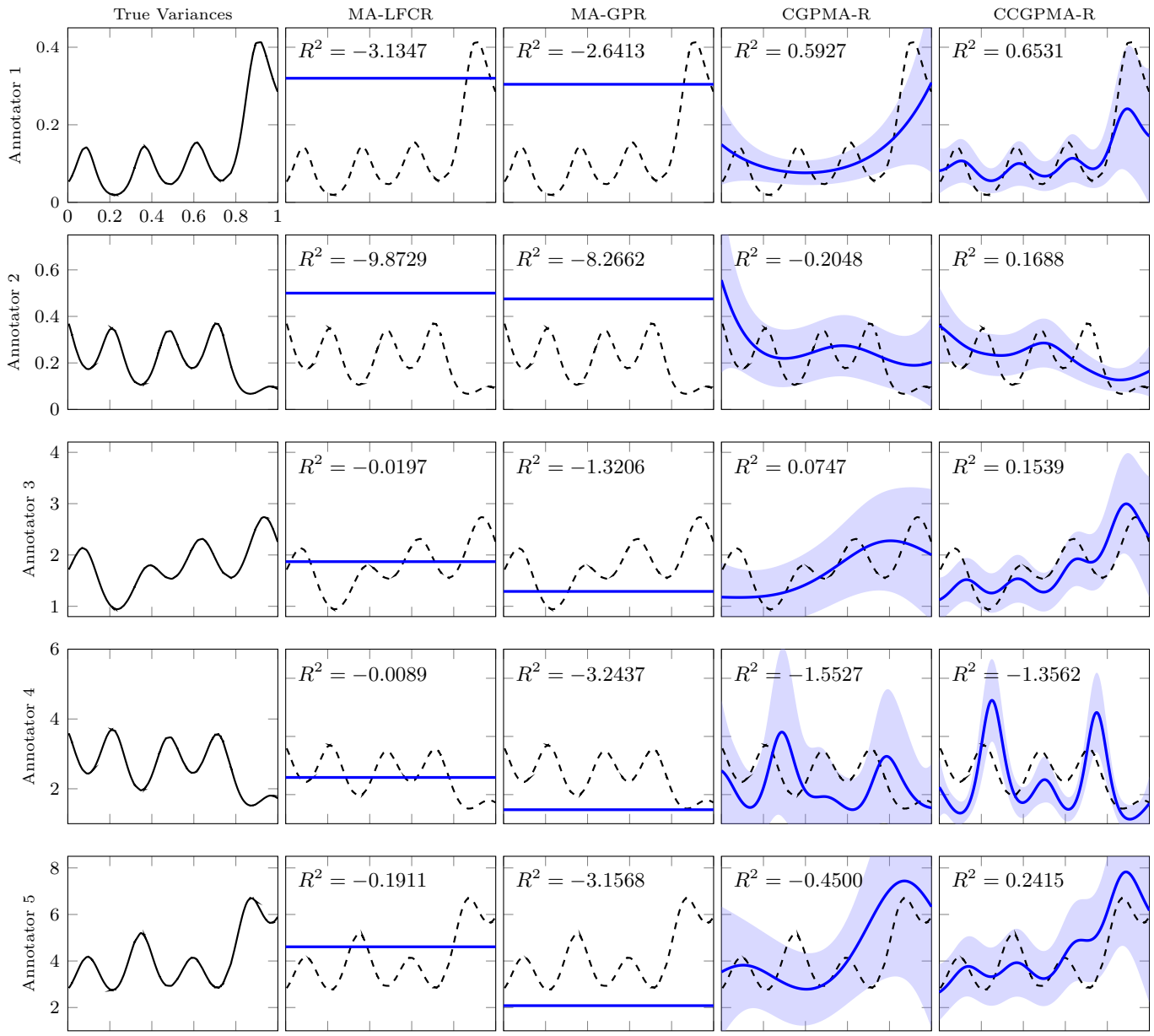


Fig. 3. Estimated values of error-variance for the five annotators in the *fully synthetic* experiment. In the first column, from top to bottom, we expose the error-variances $\mathbf{v}_r$ used to simulate the labels from each annotator. Furthermore, the subsequent columns from top to bottom present the estimation of such error-variances performed by state-of-the-art models that include these kinds of parameters in their formulation; moreover, the true error-variances are provided in dashed lines. The shaded region in CGPMA-R and CCGPMA-R indicates the area enclosed by the mean plus or minus two standard deviations. We remark that there is no shaded region for MA-LFCR, and MA-GPR since these approaches perform a fixed-point estimation for the annotators' parameters. Finally, we remark that the R^2 score between the true and estimated error variances are provided.

over CT by training a regression scheme based on LR with the actual labels (we follow the same scheme as for GPR-GOLD). We obtain an R^2 score equal to 0.8541 (on average), which is close to the results obtained by GPR-GOLD. Thus, we can elucidate that there exists a linear structure in the dataset CT. Hence, considering linear datasets, MA-LFR sets an attractive option. Finally, we analyze the results for the DL-based models. Similar to the experiments over *fully synthetic datasets*, we note a considerable low prediction capacity; in fact, they are even defeated by the linear model MA-LFR Again, we attribute this behavior to the fact that the CrowdLayer (used to manage the data from multiple annotators) does not offer a suitable codification of the labelers' behavior. Nevertheless, taking the above into account, we observe an unusual result in the dataset Bike, where the DL-based approaches offer the best performance, even defeating the supposed upper-bound GPR-GOLD. To explain that, it is necessary to analyze the meaning of the target variable in such a dataset. Attained to the

TABLE VI
REGRESSION RESULTS IN TERMS OF R^2 SCORE OVER *semi synthetic datasets*. BOLD: THE HIGHEST R^2 EXCLUDING THE UPPER BOUND GPR-GOLD.

Method	Auto	Bike	Concrete	Housing	Yacht	CT	Average
GPR-GOLD($M = 40$)	0.8604 ± 0.0271	0.5529 ± 0.0065	0.8037 ± 0.0254	0.8235 ± 0.0419	0.8354 ± 0.0412	0.8569 ± 0.0055	0.7888
GPR-GOLD($M = 80$)	0.8612 ± 0.0279	0.5603 ± 0.0063	0.8271 ± 0.0230	0.8275 ± 0.0399	0.8087 ± 0.0423	0.8648 ± 0.0047	0.7916
GPR-Av($M = 40$)	0.8425 ± 0.0286	0.5280 ± 0.0100	0.7589 ± 0.0279	0.7834 ± 0.0463	0.7588 ± 0.0498	0.8070 ± 0.0130	0.7464
GPR-Av($M = 80$)	0.8406 ± 0.0304	0.5397 ± 0.0085	0.7765 ± 0.0274	0.7903 ± 0.0451	0.7676 ± 0.0535	0.8167 ± 0.0089	0.7552
MA-LFCR	0.7973 ± 0.0218	0.3385 ± 0.0051	0.6064 ± 0.0384	0.7122 ± 0.0509	0.6403 ± 0.0186	$\mathbf{0.8400 \pm 0.0014}$	0.6558
MA-GPR	0.8456 ± 0.0281	0.4448 ± 0.0187	0.7769 ± 0.0367	0.7685 ± 0.0632	0.7842 ± 0.1027	0.0105 ± 0.0045	0.6051
MA-DL-B	0.7766 ± 0.0253	0.5854 ± 0.0107	0.2319 ± 0.0328	0.5317 ± 0.1005	0.2089 ± 0.0783	0.6903 ± 0.2689	0.5041
MA-DL-S	0.7761 ± 0.0279	$\mathbf{0.5828 \pm 0.0149}$	0.2363 ± 0.0252	0.5352 ± 0.0948	0.1822 ± 0.0985	0.9394 ± 0.0257	0.5420
MA-DL-B+S	0.7717 ± 0.0239	0.5816 ± 0.0181	0.2369 ± 0.0322	0.5330 ± 0.0850	0.1974 ± 0.0895	0.5517 ± 0.2316	0.4787
CGPMA-R($M = 40$)	0.8474 ± 0.0221	0.5464 ± 0.0069	0.8169 ± 0.0231	0.7946 ± 0.0498	0.7545 ± 0.1029	0.8236 ± 0.0132	0.7639
CGPMA-R($M = 80$)	0.7768 ± 0.0708	0.5560 ± 0.0074	0.8190 ± 0.0254	0.8058 ± 0.0493	0.8230 ± 0.0760	0.8371 ± 0.0104	0.7696
CCGPMA-R($M = 40$)	0.8563 ± 0.0247	0.5284 ± 0.0117	0.7976 ± 0.0270	0.7994 ± 0.0462	0.8436 ± 0.0507	0.8219 ± 0.0062	0.7745
CCGPMA-R($M = 80$)	$\mathbf{0.8578 \pm 0.0244}$	0.5467 ± 0.0069	$\mathbf{0.8220 \pm 0.0259}$	$\mathbf{0.8110 \pm 0.0453}$	$\mathbf{0.8476 \pm 0.0544}$	0.8252 ± 0.0083	**0.7850**

description of this dataset,[6] the target variables indicates the count of total rental bikes, including both casual and registered in a day. The above suggests that there may exist a quasi-periodic structure in the dataset, which cannot be captured by the GPR-GOLD since it uses a non-periodic kernel (it uses the RBF kernel). To support our suppositions, an additional experiment was performed over this dataset by training the model GPR-GOLD with the kernel defined as follows.

$$\kappa(\mathbf{x}_n, \mathbf{x}_{n'}) = \varphi \exp\left[-\frac{1}{2}\sum_{p=1}^{P}\left(\frac{\sin(\frac{\pi}{T_p}(x_{p,n} - x_{p,n'}))}{l_p}\right)^2\right], \tag{27}$$

where $\varphi \in \mathbb{R}$ is the variance parameter, $l_p \in (\mathbb{R}^+)$ is the length-scale parameter for the p-th dimension, and $T_p \in (\mathbb{R}^+)$ is the period for the p-th dimension. Therefore, we obtain an R^2 score equal to 0.5952 (on average), which is greater than the obtained by the DL-based approaches, indicating a quasi-periodic structure in the Bike dataset as we had supposed.

3) fully real data: Until now, we have empirically demonstrated that our approaches CGPMA and CCGPMA offer a better representation of the labelers' behavior. Nevertheless, the previous experiments configure a controlled scenario due to the labels were simulated; hence, these results could be biased by the simulation method. In this sense, the *fully real datasets* present the most challenging scenario, where both the input samples and the labels come from real-world applications. Table VII outlines the achieved performances. We remark that our CCGPMA-R with $M = 80$ obtains the best generalization performance in terms of R^2 score. Further, as theoretically expected, its performance lies between that of GPR-GOLD and GP-Av. Moreover, regarding the GPs-based competitors (MA-GPR and CGPMA-R), we note that similarly to previous experiments, our CGPMA-R is just a bit lower than CCGPMA-R, which is expected behavior due to both frameworks estimate the annotators' performances as functions of the input features. On the other hand, the behavior of MA-GPR is again surprising, due to it exhibits a prediction capability worse than GPR-Av. We remark that this outcome is due to over-fitting; in fact, the training R^2 score for MA-GPR is 0.4731, which is comparable with GPR-GOLD. The linear approach MA-LFCR exhibits the second-lowest performance and performs worse than the theoretical lower bound GP-Av, which indicates a non-linear structure in the Music dataset. Finally, analyzing the results from the deep learning approaches, we note that the variation MA-DL-B+S exhibit performance similar to our CGPMA-R; however, it is a bit lower than our CCGPMA-R. We highlight that despite the capacities of deep learning, our approach CCGPMA-R offers a better representation of annotators' behavior, unlike the deep learning approaches, which measure such performance using a single parameter.
On the other hand, we observe that all regression models exposed a lower generalization performance than previous results (see Table VI), which is a repercussion of solving a multi-class classification problem with regression models. Such an outcome is not uncommon due to similar results are exposed in (17; 20), where they used regression approaches to solve a multi-class classification problem.

[6]Such description can be found in https://archive.ics.uci.edu/ml/datasets/bike+sharing+dataset

TABLE VII
REGRESSION RESULTS IN TERMS OF R^2 SCORE OVER *fully real dataset*. BOLD: THE HIGHEST R^2 EXCLUDING THE UPPER BOUND GPR-GOLD.

Method	Music
GPR-GOLD($M = 40$)	0.4704
GPR-GOLD($M = 80$)	0.4889
GPR-Av($M = 40$)	0.2572
GPR-Av($M = 80$)	0.2744
MA-LFCR	0.1404
MA-GPR	0.0090
MA-DL-B	0.2339
MA-DL-S	0.2934
MA-DL-B+S	0.3519
CGPMA-R($M = 40$)	0.3345
CGPMA-R($M = 80$)	0.3531
CCGPMA-R($M = 40$)	0.3337
CCGPMA-R($M = 80$)	**0.3872**

B. Classification

1) Fully synthetic data: Alike the regression case, we first perform a controlled experiment to verify the capability of our CGPMA and CCGPMA with binary and multi-class classification with multiple annotators. For this first experiment, we use the *fully synthetic* dataset described in Section IV-A2. We simulate five labelers ($R = 5$) with different levels of expertise. To simulate the error-variances, we define $Q = 3$

functions $u_q(\cdot)$, which are given as

$$u_1(x) = 4.5\cos(2\pi x + 1.5\pi) - 3\sin(4.3\pi x + 0.3\pi), \quad (28)$$

$$u_2(x) = 4.5\cos(1.5\pi x + 0.5\pi) + 5\sin(3\pi x + 1.5\pi), \quad (29)$$

$$u_3(x) = 1, \quad (30)$$

where $x \in [0, 1]$. Besides, we define the combination matrix $\mathbf{W} \in \mathbb{R}^{Q \times R}$

$$\mathbf{W} = \begin{bmatrix} 0.4 & 0.7 & -0.5 & 0.0 & -0.7 \\ 0.4 & -1.0 & -0.1 & -0.8 & 1.0 \\ 3.1 & -1.8 & -0.6 & -1.2 & 1.0 \end{bmatrix}, \quad (31)$$

holding elements $w_{r,q}$. Fig. 4 exposes the predictive performance for all approaches for the *fully synthetic* data. First, we highlight that the predicted mean label value–(PMLV) for KAAR, MA-GPC, and MA-GPCV presents an unexpected shape compared with the ground truth; moreover, KAAR and MA-GPCV exhibit the worst AUC, even worsen than the intuitive lower bound GPC-MV. Such conduct is not unexpected because these approaches are designed to deal with binary labels (37; 15; 14), while this dataset configures a 3-class classification problem. To face such a problem, we use the *one-vs-all* scheme, which is not the most suitable alternative because, among other aspects, it leads to regions of input space that are ambiguously classified (36). On the other hand, concerning MA-DL methods and the linear approaches MA-LFC-C and MA-DGRL, we note an akin predictive AUC; however, the linear approaches exhibit PMLV less similar to the Ground truth, which is due to MA-LFC-C and MA-DGRL only can deal with linearly separable data. Next, we analyze the results of our CGPM-C and CCGPM-C. We remark that the predictive AUC of our methods is pretty close to the deep learning and linear models; however, unlike them, our CGPMA-C and CCGPMA-C show the most accurate PMLV compared with the absolute gold standard. In fact, CCGPMA-C behaves pretty similar to GPC-GOLD, which is the theoretical upper bound. Finally, from the GPC-MV, we do not identify notably differences with the rest of the approaches (excluding KAAR and MA-GPCV), indicating that this first experiment seems not to be difficult, notwithstanding, posteriors experiments will expose notorious differences.

From the above, we recognize that analyzing both the predictive AUC and the PMLV; our CCGPMA-C exhibits the best performance obtaining quite similar results compared with the intuitive upper bound (GPC-GOLD). Accordingly, we note CCGPMA-C proffers a more suitable representation of the labelers' behavior than its competitors because CCGPMA-C is the unique approach that models both the dependencies among the annotators and the relationship between the input features and the annotators' performance. To empirically support the above statement, Fig. 5 shows the estimated per-annotator reliability, where we only take into account models that include such types of parameters (MA-DGRL, CGPMA, and CCGPMA). From these results, we can note that based on a visual inspection and the accuracy score that the approach MA-DGRL (see column 2 in Fig. 5) do not offer a proper representation of the annotators' behavior, which is not unexpected due to such a model does not consider the relationship between the input features and the labelers' decisions. On the other hand, our approaches CGPMA-C and CCGPMA-C (columns 3 and 4 in Fig. 5) outperforms MA-DGRL, which is a direct repercussion of modeling the labelers' parameters as functions of the input features, which leads to a better representation of the labelers' behavior. We observe that CCGPMA-C exhibits the best performance in terms of accuracy; such an outcome is due to this method improves the quality of the annotators' model by considering correlations among the labelers (as was empirically established in (5; 37)).

2) Semi-synthetic data results: We recall that for this type of data we have features from real-world problems whilst the data from multiple annotators were simulated as for *fully synthetic data* (see Eqs. (28) to (31)). Table VIII shows the results concerning this second experiment, where we highlight that the annotators were simulated by considering correlations among the labelers' opinions and modeling dependencies between such opinions and the input features. From Table VIII, we can elucidate that, on average, our CCGPMA-C exhibits the best predictive AUC; moreover, we note that our CGPMA-C reaches the second-best performance in terms of AUC. The above is an expected result due to the formulation of CGPMA-C and CCGPMA-C is based on similar assumptions to those that were used to simulate the multiple annotators' labels. Regarding the GPs-based competitors (GPC-MV, MA-GPC, MA-GPCV, and KAAR), we note that they offer the worst predictive performance in terms of AUC. The performance of MA-GPCV is expected because it represents the most naive method to deal with multi-labelers scenarios, which can be considered the theoretical lower bound. Conversely, the results of MA-GPC, MA-GPCV, and KAAR are, in principle, surprising because they perform worsen than MA-GPCV. Nevertheless, we explain such an outcome in two regards. First, these approaches do not model the relationship between the input features and the annotators' performance, which does not fit how this experiment was conducted. Second, as we commented in the previous experiment, we use *one-vs-all* scheme aiming to adapt these models for multi-class problems, which could be problematic as we have demonstrated in Fig. 5; the above can be confirmed in the results for the multi-class datasets "Western" ($K = 4$), "Wine" ($K = 3$), and "Segmentation" ($K = 10$), where the predictive AUC is low compared with the remaining approaches. Then, analyzing the results from the DL-based approaches, we note a slightly better performance compared with the GPs-based methods (excluding our CGPMA-C and CCGPMA-C). We argue that such an outcome is caused by the fact that DL models can handle both binary and multi-class classification problems; however, these DL models perform considerable worsen that our proposals, which is due to crowd layers provide a very simple codification of the labelers' performance to guarantee a low computational cost (38). Finally, from the linear models, we first analyze the protruding performance from MA-DGRL, which defeats all its non-linear competitors, and obtains a performance similar to the DL-based approaches (non-linear methods). Initially, one can think that such an outcome is anomalous because a linear model is overcoming non-linear models; however, we argue that this is due to that the process of generating the labels from multiple annotators

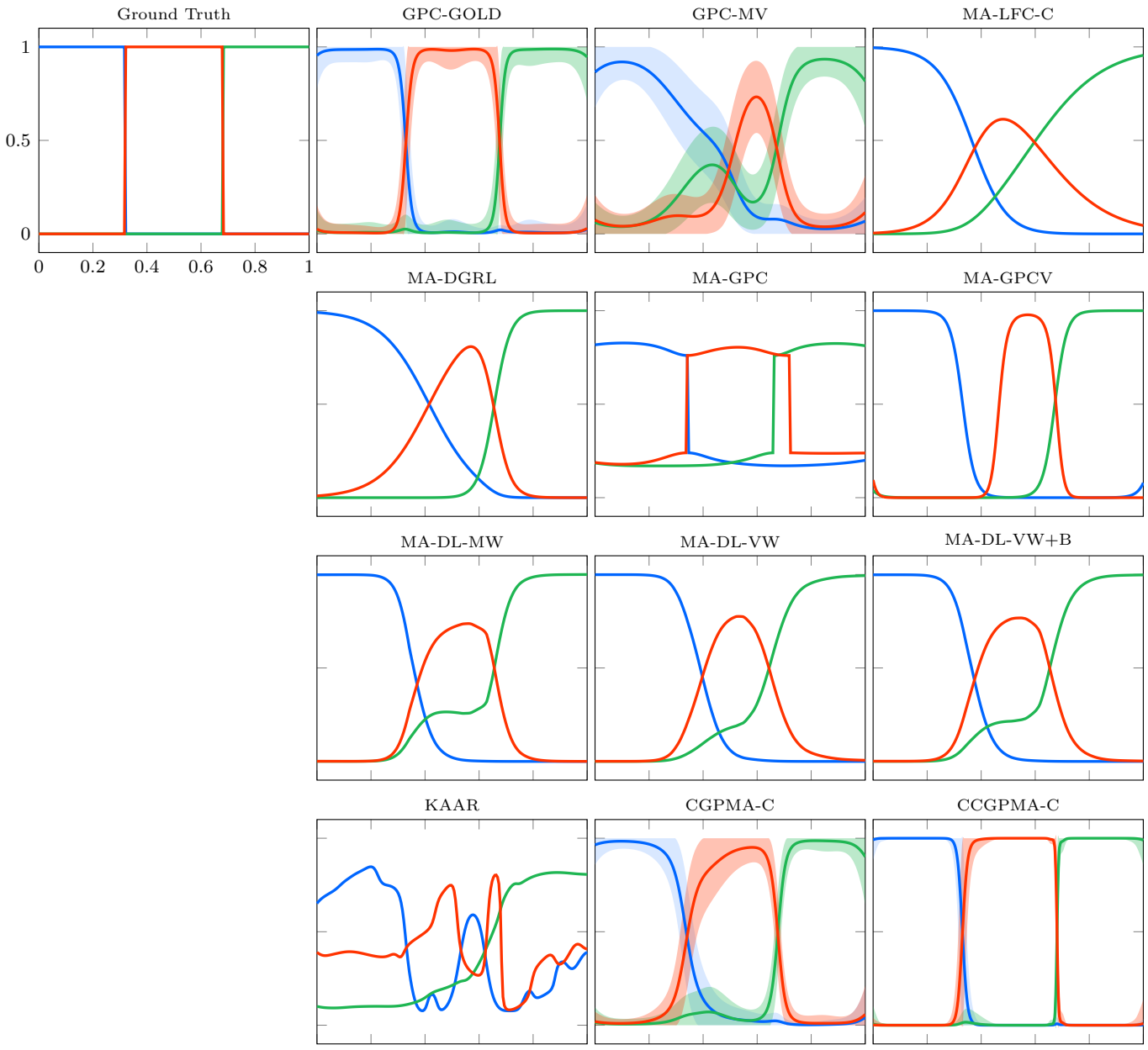


Fig. 4. Fully synthetic dataset results (blue=class 1, red=class 2, and green=class3). We compare the prediction of our CCGPMA-C($AUC = 1$), and CCGPMA-C($AUC = 0.9999$) with the theoretical upper bound GPC-GOLD($AUC = 1.0$) and lower bound GPC-MV($AUC = 0.9809$), and state-of-the-art approaches, MA-LFC-C($AUC = 0.9993$), MA-DGRL($AUC = 0.9999$), MA-GPC($AUC = 0.9977$), MA-GPCV($AUC = 0.9515$), MA-DL-MW($AUC = 0.9989$), MA-DL-VW($AUC = 0.9972$), MA-DL-VW+B($AUC = 0.9994$), KAAR(0.9099). Note that the shaded region in GPC-MV, CGPMA-C, and CCGPMA-C indicates the area enclosed by the mean plus or minus two standard deviations. We remark that there is no shaded region for the rest of approaches since they do not provide information about the prediction uncertainty.

(see Section IV-A2) is similar to the labelers' model followed by MA-DGRL. On the other hand, regarding MA-LFC-C, we note a performance similar to the DL-based methods. Still, it is considerably lower than our proposals, which is due to that MA-LFC-C formulation is based on the assumption that the annotators' behavior is homogeneous across the input space, which does not correspond to the labels simulation procedure.

3) fully real data: We have empirically demonstrated that our approach offers a better representation of the labelers' behavior. Notwithstanding, previous experiments represent a very controlled scenario due to the labels are simulated. Accordingly, these results could be biased by the simulation procedure. Thus, the fully real datasets configure the most challenging scenario, where both the input features and the labels from multiple labelers come from real-world applications. Table IX outlines the achieved predictive AUC. First, we observe that for the voice data, the scales G and R exhibit a similar performance for all considered approaches; in fact, we highlight that GPC-MV obtains a performance comparable with the upper bound GPC-GOLD. The latter can be explained in

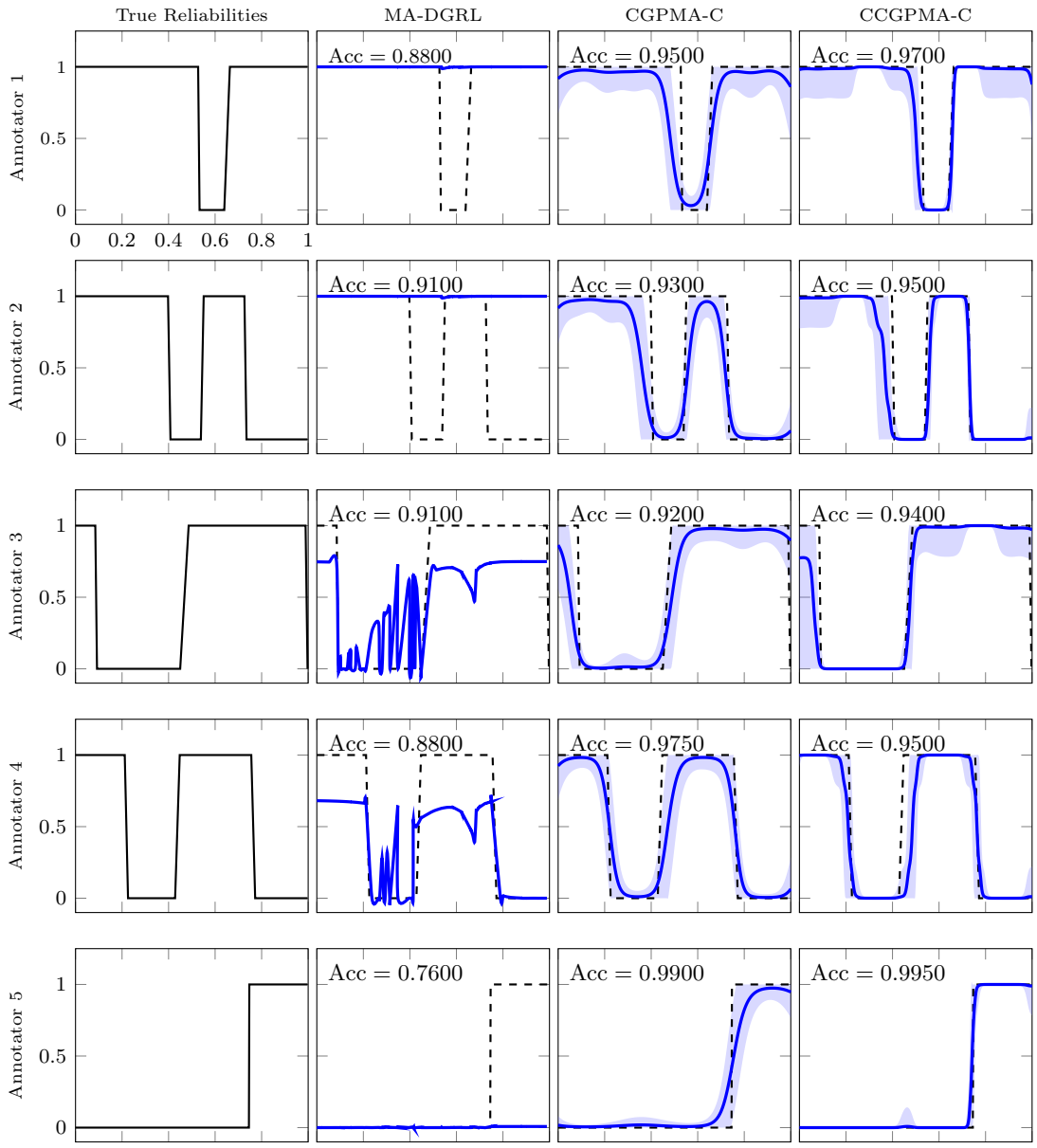


Fig. 5. Estimated reliabilities for the five annotators in the. In the first column, from top to bottom, we expose the true reliabilities λ_r used to simulate the labels from each annotator. The subsequent columns from top to bottom present the estimation of such reliabilities performed by state-of-the-art models that include these kinds of parameters in their formulation, where true values are provided in dashed lines. The shaded region in CGPMA-C and CCGPMA-C indicates the area enclosed by the mean plus or minus two standard deviations. Finally, we remark that the accuracy (Acc) is provided.

TABLE VIII
REGRESSION RESULTS IN TERMS OF AUC SCORE OVER *semi synthetic datasets*. BOLD: THE HIGHEST AUC EXCLUDING THE UPPER BOUND GPC-GOLD.

Method	Breast	Bupa	Ionosphere	Pima	TicTacToe	Western	Wine	Segmentation	Average
GPC-GOLD($M = 40$)	0.9907 ± 0.0045	0.6975 ± 0.0466	0.9490 ± 0.0235	0.8378 ± 0.0302	0.8429 ± 0.0334	0.9185 ± 0.0061	0.9987 ± 0.0015	0.9596 ± 0.0196	0.8993
GPC-GOLD($M = 80$)	0.9903 ± 0.0046	0.6997 ± 0.0483	0.9513 ± 0.0225	0.8374 ± 0.0297	0.8491 ± 0.0323	0.9250 ± 0.0057	0.9988 ± 0.0016	0.9781 ± 0.0041	0.9037
GPC-MV($M = 40$)	0.9897 ± 0.0045	0.5366 ± 0.0516	0.7566 ± 0.0572	0.5399 ± 0.0760	0.6620 ± 0.0357	0.8658 ± 0.0331	0.8179 ± 0.0212	0.9562 ± 0.0228	0.7656
GPC-MV($M = 80$)	0.9892 ± 0.0048	0.5698 ± 0.0529	0.7779 ± 0.0550	0.5302 ± 0.0674	0.6744 ± 0.0357	0.8446 ± 0.0089	0.8323 ± 0.0487	0.9749 ± 0.0047	0.7742
MA-LFC-C	0.8789 ± 0.0510	0.4593 ± 0.1444	0.7358 ± 0.0901	0.8119 ± 0.0313	0.6004 ± 0.0261	0.8400 ± 0.0211	0.9692 ± 0.0357	0.9892 ± 0.0031	0.7856
MA-DGRL	0.9757 ± 0.0189	0.5724 ± 0.0336	0.6453 ± 0.0721	0.8138 ± 0.0290	0.6129 ± 0.0230	0.8143 ± 0.0150	0.9795 ± 0.0221	0.9897 ± 0.0038	0.8005
MA-GPC	0.9811 ± 0.0116	0.5446 ± 0.0578	0.6631 ± 0.1474	0.5325 ± 0.1780	0.6079 ± 0.0995	0.8671 ± 0.0114	0.9417 ± 0.0262	0.9734 ± 0.0035	0.7639
MA-GPCV	0.8270 ± 0.0547	0.5567 ± 0.0683	0.6238 ± 0.0871	0.6217 ± 0.0590	0.6104 ± 0.1003	0.8451 ± 0.0147	0.9735 ± 0.0172	0.9924 ± 0.0027	0.7563
MA-DL-MW	0.9470 ± 0.0173	0.5237 ± 0.0568	0.7535 ± 0.0543	0.6178 ± 0.0267	0.6827 ± 0.0296	0.9092 ± 0.0056	0.9728 ± 0.0109	0.9950 ± 0.0017	0.8002
MA-DL-VW	0.9526 ± 0.0245	0.5327 ± 0.0618	0.6987 ± 0.0497	0.6063 ± 0.0336	0.6771 ± 0.0267	0.9173 ± 0.0067	0.9807 ± 0.0152	$\mathbf{0.9972 \pm 0.0011}$	0.7953
MA-DL-VW+B	0.9465 ± 0.0242	0.5281 ± 0.0631	0.7196 ± 0.0453	0.6123 ± 0.0378	0.6780 ± 0.0342	0.9164 ± 0.0085	0.9817 ± 0.0155	$\mathbf{0.9972 \pm 0.0009}$	0.7975
KAAR	0.8058 ± 0.0274	0.5920 ± 0.0663	0.7046 ± 0.0739	0.5802 ± 0.0406	0.6381 ± 0.0545	0.8588 ± 0.0120	0.9943 ± 0.0105	0.9217 ± 0.0190	0.7619
CGPMA-C($M = 40$)	0.9920 ± 0.0038	0.5537 ± 0.0630	0.8356 ± 0.1002	0.8201 ± 0.0314	0.7056 ± 0.0304	0.9178 ± 0.0066	0.9969 ± 0.0028	0.9679 ± 0.0065	0.8487
CGPMA-C($M = 80$)	0.9914 ± 0.0038	0.5945 ± 0.0642	0.8615 ± 0.0696	$\mathbf{0.8204 \pm 0.0318}$	0.7048 ± 0.0312	0.9185 ± 0.0057	$\mathbf{0.9986 \pm 0.0016}$	0.9406 ± 0.0061	0.8538
CCGPMA-C($M = 40$)	$\mathbf{0.9938 \pm 0.0027}$	0.5734 ± 0.0533	0.9021 ± 0.1079	0.7810 ± 0.0622	$\mathbf{0.7495 \pm 0.0539}$	0.9269 ± 0.0058	0.9952 ± 0.0040	0.9774 ± 0.0048	0.8624
CCGPMA-C($M = 80$)	0.9933 ± 0.0030	$\mathbf{0.6022 \pm 0.0487}$	$\mathbf{0.9023 \pm 0.1066}$	0.8045 ± 0.0510	0.7312 ± 0.0323	$\mathbf{0.9307 \pm 0.0049}$	0.9955 ± 0.0039	0.9774 ± 0.0045	**0.8671**

the sense that for these scales, the annotators exhibit a suitable performance (i.e., the provided labels are similar to the ground truth). On the other hand, a reduction in the predictive AUC is observed for scale B, which is a consequence of a diminution in the labelers' performance compared with scales G and R, as demonstrated in (16). We highlight that our approaches exhibit the best generalization performances for the three scales in the voice dataset. Remarkably, we note that CGPMA-C and CCGPMA-C do not suffer significant changes in the scale B, which is an outstanding outcome because it reflects that our approach offers a better representation of the labelers' behavior even if the labels' quality decreases.

Finally, we review the results from the Music dataset. We note that our CCGPMA-C reaches the best predictive AUC; in fact, we highlight CCGPMA-C is the unique approach with performance comparable with the intuitive upper bound. We elucidate that such outstanding performance is due to that our approach improves the annotators' representation since it models dependencies among the labelers and computes the parameters of such model (related to the annotators' performance) as a function of the input features. On the other hand, we note a considerably low performance for MA-GPC, even lower than their intuitive lower bound (GPC-MV). This behavior is not uncommon, given that it has been a constant in the experiments; we argue that this outcome is because the Music dataset configures a multi-class classification problem, and we use a one-vs-all scheme for all of the binary classification (including MA-GP). Hence, as we have explained, such a scheme is not the most proper for multi-class problems.

TABLE IX
FULLY REAL DATASETS RESULTS. BOLD: THE METHOD WITH THE HIGHEST PERFORMANCE EXCLUDING THE UPPER BOUND (TARGET) CLASSIFIER GPC-GOLD.

Method	Voice			Music	Average
	G	R	B		
GPC-GOLD($M = 40$)	0.9481	0.9481	0.9481	0.9358	0.9450
GPC-GOLD($M = 80$)	0.9484	0.9484	0.9484	0.9178	0.9407
GPC-MV($M = 40$)	0.8942	0.9373	0.8001	0.8871	0.8797
GPC-MV($M = 80$)	0.9301	0.9377	0.7962	0.8897	0.8884
MA-LFC-C	0.9122	0.9130	0.8406	0.8599	0.8814
MA-DGRL	0.9127	0.9164	0.8259	0.8832	0.8845
MA-GPC	0.8660	0.8597	0.4489	0.8253	0.7500
MA-GPCV	0.9283	0.9208	0.8835	0.8677	0.9001
MA-DL-MW	0.8957	0.8966	0.8123	0.8567	0.8653
MA-DL-VW	0.8942	0.8929	0.8092	0.9167	0.8782
MA-DL-VW+B	0.9030	0.8937	0.8218	0.8573	0.8689
KAAR	0.9109	0.9351	0.8969	0.8896	0.9081
CGPMA-C($M = 40$)	**0.9324**	0.9406	0.8696	0.9025	0.9113
CGPMA-C($M = 80$)	**0.9324**	0.9417	0.8708	0.8987	0.9109
CCGPMA-C($M = 40$)	0.9318	**0.9422**	**0.9002**	0.9446	**0.9297**
CCGPMA-C($M = 80$)	0.9243	0.9383	0.8907	**0.9456**	0.9247

VI. CONCLUSION

We have introduced two models to face with multi-labeler scenarios. The first, the Chained Gaussian processes model-CGPMA, allows modeling the each annotator's performance as a function of the input space. The second is the correlated chained Gaussian processes model-CCGPMA; here, we propose an extension of CGPMA introducing a semi-parametric latent factor model-(SLFM) to exploit correlations among the labelers' answers. We emphasize that to the best of our knowledge; CCGPMA is the first attempt to build a multi-labeler probabilistic framework capable of dealing with regression, binary classification, and multi-class classification, and that models the annotators' behavior by considering both dependencies among the labelers and computing the annotators' performance as a function of the input features. We tested our approach for regression and classification and using different scenarios regarding if the labels' were simulated (synthetic, semi-synthetic) or if they come from a real-world application (real-world dataset). Attained to the results, we remark that the CCGPMA can achieve robust predictive properties for both regression and classification scenarios. In fact, in most cases, our approach outperforms the models considered in this work.

As future work, we believe that our work could be extended by using convolution processes (48) instead of the SLFM aiming to obtain a better representation of the correlations among the labelers. On the other hand, our approach could be extended to deal with multi-output Gaussian processes with multiple annotators, namely, a multi-output scenario where for each output, we do not have access to an absolute ground truth but to a set of labels provided by multiple annotators.

APPENDIX A
PROOF OF THE FIRST ZONKLAR EQUATION

ACKNOWLEDGMENT

Under grants provided by the Minciencias project: "Desarrollo de un prototipo funcional para el monitoreo no intrusivo de vehículos usando data analitycs para innnovar en el proceso de mantenimiento basado en la condición en empresas de transporte público."-code 643885271399. Julian Gil-Gonzalez is funded by the program "Doctorados Nacionales - Convocatoria 785 de 2017".

REFERENCES

[1] J. Zhang, V. S. Sheng, and J. Wu, "Crowdsourced label aggregation using bilayer collaborative clustering," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 30, no. 10, pp. 3172–3185, 2019.

[2] Y. Liu, W. Zhang, Y. Yu *et al.*, "Truth inference with a deep clustering-based aggregation model," *IEEE Access*, vol. 8, pp. 16 662–16 675, 2020.

[3] Y. E. Kara, G. Genc, O. Aran, and L. Akarun, "Modeling annotator behaviors for crowd labeling," *Neurocomputing*, vol. 160, pp. 141–156, 2015.

[4] V. C. Raykar, S. Yu, L. H. Zhao, G. H. Valadez, C. Florin, L. Bogoni, and L. Moy, "Learning from crowds," *J. Speech Lang. Hear. Res.*, vol. 11, no. Apr, pp. 1297–1322, 2010.

[5] T. Zhu, M. A. Pimentel, G. D. Clifford, and D. A. Clifton, "Unsupervised bayesian inference to fuse biosignal sensory estimates for personalising care," *IEEE journal of biomedical and health informatics*, vol. 23, no. 1, p. 47, 2019.

[6] R. Snow, B. O'Connor, D. Jurafsky, and A. Y. Ng, "Cheap and fast—but is it good?: evaluating non-expert annotations for natural language tasks," in *EMNLP*. Association for Computational Linguistics, 2008, pp. 254–263.

[7] D. Tao, J. Cheng, Z. Yu, K. Yue, and L. Wang, "Domain-weighted majority voting for crowdsourcing," *IEEE transactions on neural networks and learning systems*, vol. 30, no. 1, pp. 163–174, 2018.

[8] X. Wang and J. Bi, "Bi-convex optimization to learn classifiers from multiple biomedical annotations," *IEEE/ACM transactions on computational biology and bioinformatics*, vol. 14, no. 3, pp. 564–575, 2016.

[9] P. Groot, A. Birlutiu, and T. Heskes, "Learning from multiple annotators with Gaussian processes," in *ICANN*. Springer, 2011, pp. 159–164.

[10] G. Rizos and B. W. Schuller, "Average jane, where art thou?–recent avenues in efficient machine learning under subjectivity uncertainty," in *International Conference on Information Processing and Management of Uncertainty in Knowledge-Based Systems*. Springer, 2020, pp. 42–55.

[11] P. Morales-Álvarez, P. Ruiz, S. Coughlin, R. Molina, and A. K. Katsaggelos, "Scalable variational gaussian processes for crowdsourcing: Glitch detection in ligo," *arXiv preprint arXiv:1911.01915*, 2019.

[12] J. Zhang, X. Wu, and V. S. Sheng, "Imbalanced multiple noisy labeling," *IEEE Transactions on Knowledge and Data Engineering*, vol. 27, no. 2, pp. 489–503, 2014.

[13] A. P. Dawid and A. M. Skene, "Maximum likelihood estimation of observer error-rates using the em algorithm," *Applied statistics*, pp. 20–28, 1979.

[14] P. Ruiz, P. Morales-Álvarez, R. Molina, and A. K. Katsaggelos, "Learning from crowds with variational gaussian processes," *Pattern Recognition*, vol. 88, pp. 298–311, 2019.

[15] F. Rodrigues, F. C. Pereira, and B. Ribeiro, "Gaussian process classification and active learning with multiple annotators." in *ICML*, 2014, pp. 433–441.

[16] J. Gil, M. Álvarez, and Á. Orozco, "Automatic assessment of voice quality in the context of multiple annotations," in *EMBC*. IEEE, 2015, pp. 6236–6239.

[17] F. Rodrigues, M. Lourenco, B. Ribeiro, and F. Pereira, "Learning supervised topic models for classification and regression from crowds," *IEEE transactions on PAMI*, 2017.

[18] F. Rodrigues, F. Pereira, and B. Ribeiro, "Sequence labeling with multiple annotators," *Machine learning*, vol. 95, no. 2, pp. 165–181, 2014.

[19] S. Albarqouni, C. Baur, F. Achilles, V. Belagiannis, S. Demirci, and N. Navab, "Aggnet: deep learning from crowds for mitosis detection in breast cancer histology images," *IEEE transactions on medical imaging*, vol. 35, no. 5, pp. 1313–1321, 2016.

[20] F. Rodrigues and F. C. Pereira, "Deep learning from crowds," in *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.

[21] M. Y. Guan, V. Gulshan, A. M. Dai, and G. E. Hinton, "Who said what: Modeling individual labelers improves classification," in *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.

[22] E. G. Rodrigo, J. A. Aledo, and J. A. Gámez, "Machine learning from crowds: A systematic review of its applications," *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, vol. 9, no. 2, p. e1288, 2019.

[23] F. Rodrigues, F. Pereira, and B. Ribeiro, "Learning from multiple annotators: Distinguishing good from random labelers," *Pattern Recognition Letters*, vol. 34, no. 12, pp. 1428–1436, 2013.

[24] M. Venanzi, J. Guiver, G. Kazai, P. Kohli, and M. Shokouhi, "Community-based bayesian aggregation models for crowdsourcing," in *Proceedings of the 23rd international conference on World wide web*. ACM, 2014, pp. 155–164.

[25] W. Tang, M. Yin, and C.-J. Ho, "Leveraging peer communication to enhance crowdsourcing," in *The World Wide Web Conference*. ACM, 2019, pp. 1794–1805.

[26] P. Zhang and Z. Obradovic, "Learning from inconsistent and unreliable annotators by a gaussian mixture model and bayesian information criterion," in *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*. Springer, 2011, pp. 553–568.

[27] J. Surowiecki, *The wisdom of crowds*. Anchor, 2005.

[28] U. Hahn, M. von Sydow, and C. Merdes, "How communication can make voters choose less well," *Topics in cognitive science*, 2018.

[29] A. D. Saul, J. Hensman, A. Vehtari, and N. D. Lawrence, "Chained gaussian processes," in *Artificial Intelligence and Statistics*, 2016, pp. 1431–1440.

[30] M. A. Álvarez, L. Rosasco, N. D. Lawrence *et al.*, "Kernels for vector-valued functions: A review," *Foundations and Trends® in Machine Learning*, vol. 4, no. 3, pp. 195–266, 2012.

[31] Y. Teh, M. Seeger, and M. Jordan, "Semiparametric latent factor models," in *AISTATS 2005-Proceedings of the 10th International Workshop on Artificial Intelligence and Statistics*, 2005.

[32] M. Álvarez, D. Luengo, M. Titsias, and N. D. Lawrence, "Efficient multioutput gaussian processes through variational inducing kernels," in *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, 2010, pp. 25–32.

[33] M. D. Hoffman, D. M. Blei, C. Wang, and J. Paisley, "Stochastic variational inference," *The Journal of Machine Learning Research*, vol. 14, no. 1, pp. 1303–1347, 2013.

[34] Y. Yan, R. Rosales, G. Fung, R. Subramanian, and J. Dy, "Learning from multiple annotators with varying expertise," *Machine learning*, vol. 95, no. 3, pp. 291–327, 2014.

[35] H. Xiao, H. Xiao, and C. Eckert, "Learning from multiple observers with unknown expertise," in *Pacific-Asia Conference on Knowledge Discovery and Data Mining*. Springer, 2013, pp. 595–606.

[36] C. M. Bishop, *Pattern recognition and machine learning*. springer, 2006.

[37] J. Gil-Gonzalez, A. Alvarez-Meza, and A. Orozco-Gutierrez, "Learning from multiple annotators using kernel alignment," *Pattern Recognition Letters*, vol. 116, pp. 150–156, 2018.

[38] P. Morales-Álvarez, P. Ruiz, R. Santos-Rodríguez, R. Molina, and A. K. Katsaggelos, "Scalable and efficient learning from crowds with gaussian processes," *Information Fusion*, vol. 52, pp. 110–127, 2019.

[39] G. Hua, C. Long, M. Yang, and Y. Gao, "Collaborative active visual recognition from crowds: A distributed ensemble approach," *IEEE transactions on pattern analysis and machine intelligence*, vol. 40, no. 3, pp. 582–594, 2018.

[40] J. Hensman, A. G. Matthews, and Z. Ghahramani, "Scalable variational gaussian process classification," *Proceedings of Machine Learning Research*, vol. 38, pp. 351–360, 2015.

[41] P. Moreno-Muñoz, A. Artés, and M. Alvarez, "Heterogeneous multi-output gaussian process prediction," in *Advances in neural information processing systems*, 2018, pp. 6711–6720.

[42] D. M. Blei, A. Kucukelbir, and J. D. McAuliffe, "Variational inference: A review for statisticians," *Journal of the American statistical Association*, vol. 112, no. 518, pp. 859–877, 2017.

[43] L. v. d. Maaten and G. Hinton, "Visualizing data using t-sne," *Journal of machine learning research*, vol. 9, no. Nov, pp. 2579–2605, 2008.

[44] J. A. Hernández-Muriel, J. B. Bermeo-Ulloa, M. Holguin-Londoño, A. M. Álvarez-Meza, and Á. A. Orozco-Gutiérrez, "Bearing health monitoring using relief-f-based feature relevance analysis and hmm," *Applied Sciences*, vol. 10, no. 15, p. 5170, 2020.

[45] J. D. Arias-Londoño, J. I. Godino-Llorente, J. Gutiérrez-Arriola, V. Osma-Ruiz, and N. Sáenz-Lechón, "Automatic grbas assessment using complexity measures and a multiclass gmm-based detector," *Models and Analysis of Vocal Emissions for Biomedical Applications*, pp. 111–114, 2011.

[46] T. Fawcett, "An introduction to roc analysis," *Pattern recognition letters*, vol. 27, no. 8, pp. 861–874, 2006.

[47] J. Schmidhuber, "Deep learning in neural networks: An overview," *Neural networks*, vol. 61, pp. 85–117, 2015.

[48] M. A. Álvarez and N. D. Lawrence, "Computationally efficient convolved multiple output gaussian processes," *The Journal of Machine Learning Research*, vol. 12, pp. 1459–1500, 2011.

PLACE PHOTO HERE

Michael Shell Biography text here.

John Doe Biography text here.

Jane Doe Biography text here.