



Deposited via The University of Sheffield.

White Rose Research Online URL for this paper:

<https://eprints.whiterose.ac.uk/id/eprint/179506/>

Version: Accepted Version

Article:

Gil-Gonzalez, J., Giraldo, J.-J., Alvarez-Meza, A.M. et al. (2021) Correlated chained Gaussian processes for datasets with multiple annotators. *IEEE Transactions on Neural Networks and Learning Systems*, 34 (8). pp. 4514-4528. ISSN: 2162-237X

<https://doi.org/10.1109/tnnls.2021.3116943>

© 2021 IEEE. Personal use of this material is permitted. Permission from IEEE must be obtained for all other users, including reprinting/ republishing this material for advertising or promotional purposes, creating new collective works for resale or redistribution to servers or lists, or reuse of any copyrighted components of this work in other works. Reproduced in accordance with the publisher's self-archiving policy.

Reuse

Items deposited in White Rose Research Online are protected by copyright, with all rights reserved unless indicated otherwise. They may be downloaded and/or printed for private study, or other acts as permitted by national copyright laws. The publisher or other rights holders may allow further reproduction and re-use of the full text version. This is indicated by the licence information on the White Rose Research Online record for the item.

Takedown

If you consider content in White Rose Research Online to be in breach of UK law, please notify us by emailing eprints@whiterose.ac.uk including the URL of the record and the reason for the withdrawal request.

Correlated Chained Gaussian Processes for Datasets with Multiple Annotators

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

Abstract—The labeling process within a supervised learning task is usually carried out by an expert, which provides the ground truth (gold standard) for each sample. However, in many real-world applications, we typically have access to annotations provided by crowds holding different and unknown expertise levels. Learning from crowds intends to configure machine learning paradigms in the presence of multi-labelers, residing on two key assumptions: the labeler’s performance does not depend on the input space, and independence among the annotators is imposed. Here, we propose the correlated chained Gaussian processes from multiple annotators–(CCGPMA) approach, which models each annotator’s performance as a function of the input space and exploits the correlations among experts. Experimental results associated with classification and regression tasks show that our CCGPMA performs better modeling of the labelers’ behaviour, indicating that it consistently outperforms other state-of-the-art learning from crowds approaches.

Index Terms—Multiple annotators, Correlated Chained Gaussian Processes, Variational inference, Semi-parametric latent factor model.

I. INTRODUCTION

SUPERVISED learning requires that a domain expert labels the instances to build the gold standard (ground truth) (1). Yet, experts are scarce, or their time is expensive, not mentioning that the labeling task is tedious and time-consuming (2). As an alternative, the labeling is distributed through multiple heterogeneous annotators, who annotate part of the whole dataset by providing their version of the hidden ground truth (3). Recently, crowdsourcing platforms, i.e., Amazon Mechanical Turk– (AMT)¹, have been introduced to capture labels from multiple sources on large datasets efficiently. The attractiveness of these platforms lies in that, at a low cost, it is possible to obtain suitable quality labels. Indeed, in some cases, such a labeling process can compete with those provided by experts (4). However, in such multi-labeler scenario, each instance is matched with multiple annotations provided by different sources with unknown and diverse expertise, being difficult to apply traditional supervised learning algorithms (5). In this sense, *learning from crowds* has been introduced as a general framework from two main perspectives: to fit the labels from multiple annotators or to adapt the supervised learning algorithms (6).

The first approach is known in the literature as “label aggregation” or “truth inference”, comprising the computation of a single hard label per sample as an estimation of the ground truth. The hard labels are then used to feed a standard supervised learning algorithm (7). The straightforward method is the so-called majority voting–(MV), and it has been used in different multi-labeler problems due to its simplicity (8). Still, 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 through the well-known Expectation-Maximization–(EM) framework and by facing the imbalanced labeling issue (9; 8).

The second approach jointly trains the supervised learning algorithm and models the annotators’ behavior. It has been shown that such strategies lead to better performance compared to the ones belonging to label aggregation. Thus, the features used to train the learning algorithm provide valuable information to puzzle out the ground truth (10). The most representative work in this area is exposed in (11), 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. In fact, such a technique has inspired several models in the context of multi-labeler scenarios, including binary classification (12; 10), multi-class discrimination (7; 13), regression (14; 15), and sequence labeling (16). Furthermore, some works have addressed the multi-labeler problem using deep learning approaches typically including an extra layer that codes the annotators’ information (17; 18; 19).

Two main issues are still unsolved in the context of learning from crowds (20): we need to code the relationships between the input features and the labelers’ performance while revealing relevant annotators’ interdependencies. In general, the annotators’ behavior is parametrized through a homogeneous constraint across the input samples. 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 (11). Besides, it is widespread to consider independence in the annotators’ labels, aiming to reduce the complexity of the model (21), or based on the fact that it is plausible to guarantee that each labeler performs the annotation process individually (22). However, this assumption is not true since there may exist correlations among the annotators (23). 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 because they share information or belong to a particular school of

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

J. Giraldo and M. A. Álvarez are with the University of Sheffield, UK. email: {jgiraldogutierrez1,mauricio.alvarez}@sheffield.ac.uk

A. Álvarez is with the Universidad Nacional de Colombia sede Manizales, 170001, Colombia. email: amalvarezme@unal.edu.co

¹<https://www.mturk.com/>

thought (24; 25). Now, if we consider that the sources are algorithms, where some of them gather the same math principles there likely exists a correlation in their labels (26).

In this work, we propose a probabilistic model, named the correlated chained Gaussian Processes for multiple annotators (CCGPMA), to jointly build a prediction algorithm applicable to classification and regression tasks. CCGPMA is based on the chained GPs model—(CGP) (27), which is a Multi-GPs framework where the parameters of an arbitrary likelihood function are modeled with multiple independent GPs (one GP prior per parameter). Unlike CGP, we consider that multiple correlated GPs model the likelihood’s parameters. For doing so we take as a basis the ideas from a Multi-output GP—(MOGP) regression (28), where each output is coded as a weighted sum of shared latent functions via a semi-parametric latent factor model—(SLFM) (29). In contrast to the MOGP, we do not have multiple outputs but multiple functions chained to the given likelihood parameters. From the multiple annotators’ point of view, the likelihood parameters are related to the labelers’ behavior; thereby, CCGPMA models the labelers’ behavior as a function of the input features while also taking into account annotators’ interdependencies. Moreover, our proposal is based on the so-called inducing variables framework (30), in combination with stochastic variational inference (31). To the best of our knowledge, this is the first attempt to build a probabilistic approach to model the labelers’ behavior as a function of the input features while also considering annotators’ interdependencies. Achieved results, using both simulated and real-world data, show how our method can deal with both regression and classification problems from multi-labelers data.

The remainder is organized as follows. Section 2 exposes the related work and the 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

Most of the learning from crowds-based methods aim to model the annotators’ behavior based on the accuracy (32), the confusion matrix (13), the error variance (11), and the bias (15). Concerning this, the expert parameters are modeled as fixed points (12), or as random variables, where it is considered that such parameters are homogeneous across the input data (7).

The first attempt to analyze the relationship between the annotators’ parameters and the input features is the work in (23). The authors propose an approach for binary classification with multiple labelers, where the input data is represented by a defined cluster using a Gaussian Mixture Model—(GMM). The approach assumes that the annotators exhibit a particular performance measured in terms of sensitivity and specificity for each group. However, the model does not consider the information from multiple experts as an input for the GMM, yielding variations in the labelers’ parameters. Similarly, in (33), the authors propose a binary classification algorithm that employs two probability models to code the annotators’ performance as a function of the input space, namely a Bernoulli and a Gaussian distribution. The parameters of these

distributions are computed via Logistic regression. Nonetheless, a linear dependence between the labeler expertise and the input space is assumed, which may not be appropriate because of the data structure’s nonlinearities. 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 specific topics related to studied documents (34). Authors in (35) offer a GP-based regression with multiple annotators. An additional GP models the annotators’ parameters as a nonlinear function of the input space. Yet, the inference is carried out based on maximum a posteriori (MAP), without including the uncertainty of the posterior distribution.

On the other hand, it has been shown that the relaxation of the annotators’ independence restriction can improve the ground truth estimation (23; 20). To the best of our knowledge, only two works address such an issue. First, the authors in (26) describe an approach to deal with regression problems, where the labelers’ behavior is modeled using a multivariate Gaussian distribution. Thus, the annotators’ interdependencies are coded in the covariance matrix. Further, in (36), the authors propose a binary classification method based on a weighted combination of classifiers. In turn, the weights are estimated by using a kernel alignment-based algorithm considering dependencies among the labelers.

Here, we propose a GPs-based framework to face classification and regression settings with multiple annotators. Our proposal follows the line of the works in (12; 14; 10; 7; 37) in the sense that we are modeling the unknown ground truth through a GP prior. However, while such approaches code the annotators’ parameters as fixed points (12; 14); or as random variables (10; 7; 37); we model them as random processes to take into account dependencies between the input space and the labelers’ behavior. Besides, our CCGPMA shares some similarities with the works in (33; 35), because we aim to model the dependencies between the input features and the labelers’ performance. Our method is also similar to the works in (26; 36), because they assume dependencies in the annotators’ labels. In contrast, CCGPMA is the only one that includes both assumptions to code the annotators’ behavior. Of note, we highlight that our proposal codes inconsistent annotations, being robust against outliers. Namely, CCGPMA can estimate the annotators’ performance for every region in the input space; meanwhile, state-of-the-art techniques assess it based on a conventional averaging (15; 7; 10). Table I summarizes the key insights of our CCGPMA and state-of-the-art approaches.

III. METHODS

A. Chained Gaussian processes

Let us consider an input-output dataset $\mathcal{D} = \{\mathbf{X} \in \mathcal{X}, \mathbf{y} \in \mathcal{Y}\}$, where $\mathbf{X} = \{\mathbf{x}_n \in \mathcal{X} \subseteq \mathbb{R}^P\}_{n=1}^N$ and $\mathbf{y} = \{y_n \in \mathcal{Y}\}_{n=1}^N$. In turn, let a GP be a collection of random variables $f(\mathbf{x})$ indexed by the input samples $\mathbf{x} \in \mathcal{X}$ holding a joint multivariate Gaussian distribution (39). A GP is defined by its mean $m(\mathbf{x}) = \mathbb{E}[f(\mathbf{x})]$ (we consider $m(\mathbf{x}) = 0$) and covariance function $\kappa_f(\mathbf{x}, \mathbf{x}') = \mathbb{E}[(f(\mathbf{x}) - m(\mathbf{x}))(f(\mathbf{x}') - m(\mathbf{x}'))]$, where $\kappa_f: \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}$ is a given kernel function and $\mathbf{x}' \in \mathcal{X}$, yielding:

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

Source	Data type	Type of model	Modeling the annotator's expertise	Expertise as a function of the input space	Modeling the annotators' inter-dependencies
<i>Raykar et al., 2010</i> (11)	Regression-Binary-Categorical	Probabilistic	✓	✗	✗
<i>Zhang and Obradovic, 2011</i> (23)	Binary	Probabilistic	✓	✓	✗
<i>Xiao et al., 2013</i> (35)	Regression	Probabilistic	✓	✓	✗
<i>Yan et al., 2014</i> (33)	Binary	Probabilistic	✓	✓	✗
<i>Wang and Bi, 2016</i> (34)	Binary	Deterministic	✓	✓	✗
<i>Rodrigues et al., 2017</i> (15)	Regression-Binary-Categorical	Probabilistic	✓	✗	✗
<i>Gil-Gonzalez et al., 2018</i> (36)	Binary	Deterministic	✓	✗	✓
<i>Hua et al., 2018</i> (38)	Binary-Categorical	Deterministic	✓	✗	✗
<i>Ruiz et al., 2019</i> (10)	Binary	Probabilistic	✓	✗	✗
<i>Morales- Alvarez et al., 2019</i> (7)	Binary	Probabilistic	✓	✗	✗
<i>Zhu et al., 2019</i> (26)	Regression	Probabilistic	✓	✗	✓
Proposal-(CCGPMA)	Regression-Binary-Categorical	Probabilistic	✓	✓	✓

234 $\mathbb{R}^{M \times P}$, which decreases the GP's computational complexity to
 235 $\mathcal{O}(NM^2)$. Further, the following augmented GP prior arises:
 (1)
$$f(\mathbf{x}) \sim \mathcal{GP}(0, \kappa_f(\mathbf{x}, \mathbf{x}')).$$

203 If we consider the finite set of inputs in \mathbf{X} , then
 204 $\mathbf{f} = [f(\mathbf{x}_1), \dots, f(\mathbf{x}_N)]^\top \in \mathbb{R}^N$ is drawn for a multivariate
 205 Gaussian distribution $\mathbf{f} \sim \mathcal{N}(\mathbf{f}|\mathbf{0}, \mathbf{K}_{\mathbf{f}\mathbf{f}})$, where $\mathbf{K}_{\mathbf{f}\mathbf{f}} \in \mathbb{R}^{N \times N}$
 206 is the covariance matrix formed by the evaluation of $\kappa_f(\cdot, \cdot)$
 207 over the input set \mathbf{X} .
 208 Accordingly, using GPs for modeling the input-output data
 209 collection \mathcal{D} consists of constructing a joint distribution
 210 between a given likelihood function and one or multiple GP-
 211 based priors. To code each likelihood parameter as a random
 212 process, we employ the so-called chained GP-(CGP) that
 213 attaches such parameters to multiple independent GP priors,
 214 as follows (27):

$$p(\mathbf{y}, \hat{\mathbf{f}}|\mathbf{X}) = \prod_{n=1}^N p(y_n|\theta_1(\mathbf{x}_n), \dots, \theta_J(\mathbf{x}_n)) \times \dots \quad 241$$

$$\dots \times \prod_{j=1}^J \mathcal{N}(\mathbf{f}_j|\mathbf{0}, \mathbf{K}_{\mathbf{f}_j\mathbf{f}_j}), \quad (2) \quad 242$$

215 where each $\{\theta_j(\mathbf{x}) \in \mathcal{M}_j\}_{j=1}^J$ represents the likelihood's param-
 216 eters, being $J \in \mathbb{N}$ the number of parameters to repre-
 217 sent the likelihood. Besides, each $\theta_j(\mathbf{x})$ holds a non-linear
 218 mapping from a GP prior, e.g., $\theta_j(\mathbf{x}) = h_j(f_j(\mathbf{x}))$, where
 219 $h_j: \mathbb{R} \rightarrow \mathcal{M}_j$ is a deterministic function that maps each latent
 220 function-(LF) $f_j(\mathbf{x})$, to the appropriate domain \mathcal{M}_j . Moreover,
 221 $\mathbf{f}_j = [f_j(\mathbf{x}_1), \dots, f_j(\mathbf{x}_N)]^\top \in \mathbb{R}^N$ is a LF vector that follows
 222 a GP prior, and $\hat{\mathbf{f}} = [\hat{\mathbf{f}}_1, \dots, \hat{\mathbf{f}}_J]^\top \in \mathbb{R}^{N \times J}$. $\mathbf{K}_{\mathbf{f}_j\mathbf{f}_j} \in \mathbb{R}^{N \times N}$
 223 is the covariance matrix belonging to the j -th GP prior, which is
 224 computed based on the kernel function $\kappa_j: \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}$. The
 225 non-parametric formulation of a GP introduces computational
 226 loads through the inference process. For instance, considering
 227 that the dataset \mathcal{D} configures a regression problem, a GP
 228 modeling involves a computational complexity of $\mathcal{O}(N^3)$
 229 to invert the matrix $\mathbf{K}_{\mathbf{f}_j\mathbf{f}_j}$ (39). A common approach to
 230 reduce such computational complexity is to augment the
 231 GP prior with a set of $M \ll N$ inducing variables (40)
 232 $\mathbf{u}_j = [f_j(\mathbf{z}_1^j), \dots, f_j(\mathbf{z}_M^j)]^\top \in \mathbb{R}^M$ through additional eval-
 233 uations of $f_j(\cdot)$ at unknown locations $\mathbf{Z}_j = [\mathbf{z}_1^j, \dots, \mathbf{z}_M^j]$

234 $\mathbb{R}^{M \times P}$, which decreases the GP's computational complexity to
 235 $\mathcal{O}(NM^2)$. Further, the following augmented GP prior arises:

$$p(\mathbf{f}_j, \mathbf{u}_j) = \mathcal{N} \left(\begin{bmatrix} \mathbf{f}_j \\ \mathbf{u}_j \end{bmatrix} \middle| \mathbf{0}, \begin{bmatrix} \mathbf{K}_{\mathbf{f}_j\mathbf{f}_j} & \mathbf{K}_{\mathbf{f}_j\mathbf{u}_j} \\ \mathbf{K}_{\mathbf{u}_j\mathbf{f}_j} & \mathbf{K}_{\mathbf{u}_j\mathbf{u}_j} \end{bmatrix} \right), \quad (3)$$

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

$$p(\mathbf{f}_j|\mathbf{u}_j) = \mathcal{N} \left(\mathbf{f}_j | \mathbf{K}_{\mathbf{f}_j\mathbf{u}_j} \mathbf{K}_{\mathbf{u}_j\mathbf{u}_j}^{-1} \mathbf{u}_j, \mathbf{K}_{\mathbf{f}_j\mathbf{f}_j} - \dots \right) \quad (4)$$

$$\dots - \mathbf{K}_{\mathbf{f}_j\mathbf{u}_j} \mathbf{K}_{\mathbf{u}_j\mathbf{u}_j}^{-1} \mathbf{K}_{\mathbf{u}_j\mathbf{f}_j} \Big),$$

$$p(\mathbf{u}_j) = \mathcal{N}(\mathbf{u}_j | \mathbf{0}, \mathbf{K}_{\mathbf{u}_j\mathbf{u}_j}). \quad (5)$$

In most cases Eqs. (4) and (5) are non-conjugate to the
 likelihood, finding the posterior distribution $p(\mathbf{f}, \mathbf{u}|\mathbf{y})$ is not
 tractable analytically; therefore, we resort to a deterministic
 approximation of the posterior distribution using variational
 inference. Hence, the actual posterior can be approximated by
 a parametrized variational distribution $p(\hat{\mathbf{f}}, \mathbf{u}|\mathbf{y}) \approx q(\hat{\mathbf{f}}, \mathbf{u})$, as:

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

where $\mathbf{u} = [\mathbf{u}_1^\top, \dots, \mathbf{u}_J^\top]^\top \in \mathbb{R}^{MJ}$; moreover, $p(\mathbf{f}_j|\mathbf{u}_j)$ is
 defined in Eq. (4), and $q(\mathbf{u})$ is the posterior approximation
 over the inducing variables:

$$q(\mathbf{u}) = \prod_{j=1}^J q(\mathbf{u}_j) = \prod_{j=1}^J \mathcal{N}(\mathbf{u}_j | \mathbf{m}_j, \mathbf{V}_j). \quad (7)$$

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

$$\begin{aligned} \mathcal{L} = & \sum_{n=1}^N \mathbb{E}_{q(\mathbf{f}_1), \dots, q(\mathbf{f}_J)} [\log p(y_n | \theta_{1,n}, \dots, \theta_{J,n}) - \dots \\ & \dots - \sum_{j=1}^J \mathbb{D}_{KL}(q(\mathbf{u}_j) || p(\mathbf{u}_j)), \end{aligned} \quad (8)$$

where $\mathbb{D}_{KL}(\cdot || \cdot)$ is the Kullback-Leibler divergence and $q(\mathbf{f}_j)$ is defined as follows:

$$q(\mathbf{f}_j) = \int p(\mathbf{f}_j | \mathbf{u}_j) q(\mathbf{u}_j) d\mathbf{u}_j. \quad (9)$$

258

B. Correlated chained Gaussian processes

From Section III-A, we note that the CGP model assumes independence between priors, thereby lacking a correlation structure between GPs. As mentioned before, we consider that the annotators are correlated. We will enable this aspect of the model by assuming dependencies among the latent parameters of the chained GP. In particular, we introduce the correlated chained GPs (CCGP) to model correlations between the GP latent functions, which are supposed to be generated from a semi-parametric latent factor model (SLFM) (29):

$$f_j(\mathbf{x}_n) = \sum_{q=1}^Q w_{j,q} \mu_q(\mathbf{x}_n), \quad (10)$$

where $f_j : \mathcal{X} \rightarrow \mathbb{R}$ is an LF, $\mu_q(\cdot) \sim \mathcal{GP}(0, k_q(\cdot, \cdot))$ with $k_q : \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}$ being 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 to extend the joint distribution in Eq. (2) as follows:

$$p(\mathbf{y}, \hat{\mathbf{f}}, \mathbf{u} | \mathbf{X}) = p(\mathbf{y} | \boldsymbol{\theta}) \prod_{j=1}^J p(\mathbf{f}_j | \mathbf{u}) p(\mathbf{u}), \quad (11)$$

where $\boldsymbol{\theta} = [\boldsymbol{\theta}_1, \dots, \boldsymbol{\theta}_J]^\top \in \mathbb{R}^{NJ}$ holds the model's parameters and $\boldsymbol{\theta}_j = [\theta_j(\mathbf{x}_1), \dots, \theta_j(\mathbf{x}_N)]^\top \in \mathbb{R}^N$ relates the j -th parameter with the input space. Our CCGP employs the inducing variables-based method for sparse approximations of GPs (40). For each $\mu_q(\cdot)$, we introduce a set of $M \leq N$ "pseudo variables" $\mathbf{u}_q = [\mu_q(\mathbf{z}_1^q), \dots, \mu_q(\mathbf{z}_M^q)]^\top \in \mathbb{R}^M$ through evaluations of $\mu_q(\cdot)$ at unknown locations $\mathbf{Z}_q = [\mathbf{z}_1^q, \dots, \mathbf{z}_M^q] \in \mathbb{R}^{M \times P^2}$. Note that $\mathbf{u} = [\mathbf{u}_1^\top, \dots, \mathbf{u}_Q^\top]^\top \in \mathbb{R}^{QM}$, yielding:

$$p(\mathbf{f}_j | \mathbf{u}) = \mathcal{N}(\mathbf{f}_j | \mathbf{K}_{\mathbf{f}_j \mathbf{u}} \mathbf{K}_{\mathbf{u} \mathbf{u}}^{-1} \mathbf{u}, \mathbf{K}_{\mathbf{f}_j \mathbf{f}_j} - \dots - \mathbf{K}_{\mathbf{f}_j \mathbf{u}} \mathbf{K}_{\mathbf{u} \mathbf{u}}^{-1} \mathbf{K}_{\mathbf{u} \mathbf{f}_j}), \quad (12)$$

$$p(\mathbf{u}) = \mathcal{N}(\mathbf{u} | \mathbf{0}, \mathbf{K}_{\mathbf{u} \mathbf{u}}) = \prod_{q=1}^Q \mathcal{N}(\mathbf{u}_q | \mathbf{0}, \mathbf{K}_{\mathbf{u}_q \mathbf{u}_q}), \quad (13)$$

where $\mathbf{K}_{\mathbf{u} \mathbf{u}} \in \mathbb{R}^{QM \times QM}$ is a block-diagonal matrix with blocks $\mathbf{K}_{\mathbf{u}_q \mathbf{u}_q} \in \mathbb{R}^{M \times M}$, based on the kernel function $\kappa_q(\cdot, \cdot)$. The covariance matrix $\mathbf{K}_{\mathbf{f}_j \mathbf{f}_j} \in \mathbb{R}^{N \times N}$ holds

elements $\sum_{q=1}^Q w_{j,q} w_{j',q} \kappa_q(\mathbf{x}_n, \mathbf{x}_{n'})$, with $\mathbf{x}_n, \mathbf{x}_{n'} \in \mathcal{X}$. Likewise, $\mathbf{K}_{\mathbf{f}_j \mathbf{u}} = [\mathbf{K}_{\mathbf{f}_j \mathbf{u}_1}, \dots, \mathbf{K}_{\mathbf{f}_j \mathbf{u}_Q}] \in \mathbb{R}^{N \times QM}$, where $\mathbf{K}_{\mathbf{f}_j \mathbf{u}_q} \in \mathbb{R}^{N \times M}$ gathers elements $w_{j,q} \kappa_q(\mathbf{x}_n, \mathbf{z}_m^q)$, $m \in \{1, \dots, M\}$. Alike CGP, in most cases, the CCGP posterior distribution $p(\hat{\mathbf{f}}, \mathbf{u} | \mathbf{y})$ has not an analytical solution, so the actual posterior can be approximated by a parametrized variational distribution $p(\hat{\mathbf{f}}, \mathbf{u} | \mathbf{y}) \approx q(\hat{\mathbf{f}}, \mathbf{u})$, as:

$$q(\hat{\mathbf{f}}, \mathbf{u}) = p(\hat{\mathbf{f}} | \mathbf{u}) q(\mathbf{u}) = \prod_{j=1}^J p(\mathbf{f}_j | \mathbf{u}) \prod_{q=1}^Q q(\mathbf{u}_q), \quad (14)$$

where $p(\mathbf{f}_j | \mathbf{u})$ is given by Eq. (12), $q(\mathbf{u}_q) = \mathcal{N}(\mathbf{u}_q | \mathbf{m}_q, \mathbf{V}_q)$, and $q(\mathbf{u}) = \mathcal{N}(\mathbf{u} | \mathbf{m}, \mathbf{V})$. Also, $\mathbf{m}_q \in \mathbb{R}^M$, and $\mathbf{V}_q \in \mathbb{R}^{M \times M}$ are respectively the mean and covariance of variational distribution $q(\mathbf{u}_q)$; similarly, $\mathbf{m} = [\mathbf{m}_1^\top, \dots, \mathbf{m}_Q^\top]^\top \in \mathbb{R}^{QM}$, and $\mathbf{V} \in \mathbb{R}^{QM \times QM}$ is a block-diagonal matrix with blocks given by the covariance matrices \mathbf{V}_q . We remark that the variational approximation given by Eq. (14) is not uncommon, and it has been used in several GPs models, including (27; 41). The approximation for the posterior distribution comprises the computation of the following variational parameters: the mean vectors $\{\mathbf{m}_q\}_{q=1}^Q$ and the covariance matrices $\{\mathbf{V}_q\}_{q=1}^Q$. Such an estimation is carried out by maximizing an evidence lower bound (ELBO), which is given as:

$$\begin{aligned} \mathcal{L} = & \sum_{n=1}^N \mathbb{E}_{q(\mathbf{f}_1), \dots, q(\mathbf{f}_J)} [\log p(y_n | \theta_{1,n}, \dots, \theta_{J,n}) - \dots \\ & \dots - \sum_{q=1}^Q \mathbb{D}_{KL}(q(\mathbf{u}_q) || p(\mathbf{u}_q)), \end{aligned} \quad (15)$$

where $\theta_{j,n} = \theta_j(\mathbf{x}_n)$, with $j \in \{1, \dots, J\}$, and $\mathbb{D}_{KL}(\cdot || \cdot)$ is the Kullback-Leibler divergence and $q(\mathbf{f}_j)$ is defined as follows:

$$\begin{aligned} q(\mathbf{f}_j) = & \mathcal{N}(\mathbf{f}_j | \mathbf{K}_{\mathbf{f}_j \mathbf{u}} \mathbf{K}_{\mathbf{u} \mathbf{u}}^{-1} \mathbf{m}, \mathbf{K}_{\mathbf{f}_j \mathbf{f}_j} + \dots \\ & \dots + \mathbf{K}_{\mathbf{f}_j \mathbf{u}} \mathbf{K}_{\mathbf{u} \mathbf{u}}^{-1} (\mathbf{V} - \mathbf{K}_{\mathbf{u} \mathbf{u}}) \mathbf{K}_{\mathbf{u} \mathbf{f}_j}). \end{aligned} \quad (16)$$

Yet, in presence of non-Gaussian likelihoods, the computation of the variational expectations (VEs) in Eq. (15) cannot be solved analytically (27; 41). Hence, 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; 27). We remark such ELBO is used to infer the model's hyperparameters such as the inducing points, the kernel hyperparameters, and the combination factors $w_{j,q}$ Eq. (10). It is worth mentioning that the CCGPs 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 dealing with large scale datasets (40; 27; 41). Finally, we notice that the computational complexity for our CCGP is similar to the model in (41). Accordingly, it is dominated by the inversion of $\mathbf{K}_{\mathbf{u} \mathbf{u}}$ with $\mathcal{O}(QM^3)$ and products like $\mathbf{K}_{\hat{\mathbf{f}} \mathbf{u}}$ with $\mathcal{O}(JNQM^2)$.

C. Correlated chained GP for multiple annotators-CCGPMA

Let us consider that a predefined panel of $R \in \mathbb{N}$ annotators (with different and unknown levels of expertise) label a given dataset of N instances. It is common to find that the each annotator r only labels $|N_r| \leq N$ samples, being $|N_r|$ the cardinality of the set $N_r \subseteq \{1, \dots, N\}$ that contains the indexes of samples labeled by the r -th annotator. Besides, we define the set $R_n \subseteq \{1, \dots, R\}$ holding the indexes of annotators that labeled the n -th instance. The input-output set is coupled within a multiple annotators scenario as $\mathcal{D} = \{\mathbf{X}, \mathbf{Y} = \{y_n^r\}_{n \in N, r \in R_n}\}$, where $y_n^r \in \mathcal{Y}$ is the output given by labeler r to the sample n ; accordingly, our main aims are: *i*) to code each labeler's performance as a function of the input space and taking into account inter-annotator dependencies, and *ii*) to predict the true output $y_* \in \mathcal{Y}$ of a new instance $\mathbf{x}_* \in \mathbb{R}^P$. We highlight that to achieve such objectives, no extra information about the annotators' behaviour is provided (e.g., extra labels or information about her/his experience).

1) *Classification*: To model categorical data from multiple annotators with K classes ($\mathcal{Y} = \{1, \dots, K\}$) using our CCGPMA, we use the framework proposed in (32), which introduces a binary variable $\lambda_n^r \in \{0, 1\}$ representing the r -th labeler's reliability as a function of each sample \mathbf{x}_n . If $\lambda_n^r = 1$, the r -th annotator is supposed to provide the actual label, yielding to a categorical distribution. Conversely, $\lambda_n^r = 0$ indicates that the r -th annotator gives an incorrect output, which is modeled by a uniform distribution. Therefore, the likelihood function is given as:

$$p(\mathbf{Y}|\boldsymbol{\theta}) = \prod_{n=1}^N \prod_{r \in R_n} \left(\prod_{k=1}^K \zeta_{k,n}^{\delta(y_n^r, k)} \right)^{\lambda_n^r} \left(\frac{1}{K} \right)^{(1-\lambda_n^r)}, \quad (17)$$

where $\delta(y_n^r, k) = 1$, if $y_n^r = k$, otherwise $\delta(y_n^r, k) = 0$. Besides, $\zeta_{k,n} = p(y_n^r = k | \lambda_n^r = 1)$ is an estimation of the unknown ground truth. Accordingly, $J = K + R$ LFs are required within our CCGPMA approach, aiming to model the likelihood parameters $\boldsymbol{\theta}$. In particular, K LFs are used to model $\zeta_{k,n}$ based on a softmax function ι as:

$$\zeta_{k,n} = \iota(f_k(\mathbf{x}_n)) = \frac{\exp(f_k(\mathbf{x}_n))}{\sum_{j=1}^K \exp(f_j(\mathbf{x}_n))}. \quad (18)$$

Besides, R LFs are utilized to compute each λ_n^r from a step function; therefore, $\lambda_n^r = 1$ if $f_{l_r}(\mathbf{x}_n) \geq 0$, otherwise, $\lambda_n^r = 0$ ($r \in \{1, \dots, R\}$). $l_r = K + r \in \{K + 1, \dots, J\}$ indexes the r -th annotator' LF. Of note, we approximate the step function through the well-known sigmoid function ς to avoid discontinuities and favor the CCGPMA implementation. Unlike to CCGP, we use variational inference to approximate the posterior distribution of our CCGPMA. In consequence, the actual posterior $p(\mathbf{f}, \mathbf{u}|\mathbf{Y})$ is approximated following Eq. (14). Besides, we can derive a CCGPMA ELBO, yielding:

$$\begin{aligned} \mathcal{L} = & \sum_{n=1}^N \sum_{r \in R_n} \mathbb{E}_{q(\mathbf{f}_1), \dots, q(\mathbf{f}_J)} [\log p(y_n^r | \theta_{1,n}, \dots, \theta_{J,n})] - \dots \\ & \dots - \sum_{q=1}^Q \mathbb{D}_{KL}(q(\mathbf{u}_q) || p(\mathbf{u}_q)), \end{aligned} \quad (19)$$

where for the classification case, we have

$$p(y_n^r | \theta_{1,n}, \dots, \theta_{J,n}) = \left(\prod_{k=1}^K \zeta_{k,n}^{\delta(y_n^r, k)} \right)^{\lambda_n^r} \left(\frac{1}{K} \right)^{(1-\lambda_n^r)}. \quad (20)$$

Finally, given a new sample \mathbf{x}_* , we are interested in the mean and variance for predictive distributions related to the ground truth $\zeta_{k,*} = p(y_* = k | \mathbf{x}_*, \mathbf{f}, \mathbf{u})$, and the labelers' reliabilities λ_*^r . Accordingly, for $\zeta_{k,*}$ we obtain

$$\mathbb{E}[\zeta_{k,*}] \approx \int \iota(f_k(\mathbf{x}_*)) q(\mathbf{f}_*) d\mathbf{f}_*, \quad (21)$$

where $q(\mathbf{f}_*) = \int p(\mathbf{f}_* | \mathbf{u}) q(\mathbf{u}) d\mathbf{u}$. Similarly, for the predictive variance of $\zeta_{k,*}$, we use the expression $\text{Var}[\zeta_{k,*}] = \mathbb{E}[\zeta_{k,*}^2] - \mathbb{E}[\zeta_{k,*}]^2$; hence, we need to compute $\mathbb{E}[\zeta_{k,*}^2]$ as

$$\mathbb{E}[\zeta_{k,*}^2] \approx \int \iota(f_k(\mathbf{x}_*))^2 q(\mathbf{f}_*) d\mathbf{f}_*. \quad (22)$$

On the other hand, regarding the predictive mean and variance for λ_*^r , we have

$$\mathbb{E}[\lambda_*^r] = \int \varsigma(f_{l_r}(\mathbf{x}_*)) q(\mathbf{f}_*) d\mathbf{f}_*. \quad (23)$$

For the variance of λ_*^r , we use the expression $\text{Var}[\lambda_*^r] = \mathbb{E}[(\lambda_*^r)^2] - \mathbb{E}[\lambda_*^r]^2$; hence, we need to compute

$$\mathbb{E}[(\lambda_*^r)^2] = \int \varsigma(f_{l_r}(\mathbf{x}_*))^2 q(\mathbf{f}_*) d\mathbf{f}_*. \quad (24)$$

In this case, integrals in Eqs. (21) to (24) have not closed solution; hence, we approximate them using the Gaussian-Hermite quadrature.

2) *Regression*: For real-valued outputs, e.g., $\mathcal{Y} \subset \mathbb{R}$, we follow the multi-annotator model used in (11; 14; 35; 15), where each output y_n^r is considered to be a corrupted version of the hidden ground truth y_n . Then:

$$p(\mathbf{Y}|\boldsymbol{\theta}) = \prod_{n=1}^N \prod_{r \in R_n} \mathcal{N}(y_n^r | y_n, v_n^r), \quad (25)$$

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 CCGPMA, it is necessary to chain each likelihood's parameter to a latent function f_j . Thus, we require $J = R + 1$ LFs; one to model the hidden ground truth, such that $y_n = f_1(\mathbf{x}_n)$, and R LFs to model each error-variance $v_n^r = \exp(f_{l_r}(\mathbf{x}_n))$, with $r \in \{1, \dots, R\}$, and $l_r = r + 1 \in \{2, \dots, J\}$. Note that we use an exponential function to map from f_{l_r} to v_n^r , aiming

396 to guarantee $v_n^r > 0$ ($f_{l_r} \in \mathbb{R}$). Similar to the classification
 397 problem, the ELBO in regression settings is given by Eq. (19),
 398 where $p(y_n^r | \theta_{1,n}, \dots, \theta_{J,n}) = \mathcal{N}(y_n^r | y_n, v_n^r)$.

399 Now, given a new sample \mathbf{x}_* , we are interested in the
 400 mean and variances for predictive distributions concerning the
 401 ground truth y_* , and the labelers' error-variances v_*^r . First, for
 402 y_* we have that since $\mathbf{y} = \mathbf{f}_1$, the posterior distribution for y_*
 403 corresponds to $q(f_{1*})$, yielding:

$$\mathbb{E}[y_*] = \mu_{1,*} \tag{26}$$

$$\text{Var}[y_*] = s_{1,*}, \tag{27}$$

404 where $\mu_{1,*}$, and $s_{1,*}$ are respectively the mean and variance of
 405 $q(f_{1*})$. Then, for v_*^r , we note that due to $\mathbf{v}_r = \exp(\mathbf{f}_{l_r})$, the
 406 posterior distribution for v_*^r follows a log-normal distribution
 407 with parameters $\mu_{l_r,*}$ and $s_{l_r,*}$, which respectively correspond
 408 to the mean and variance of $q(f_{l_r,*})$. In this sense, the mean
 409 and variance of v_*^r are given as:

$$\mathbb{E}[v_*^r] = \exp\left(\mu_{l_r,*} + \frac{s_{l_r,*}}{2}\right). \tag{28}$$

$$\text{Var}[v_*^r] = \exp(2\mu_{l_r,*} + s_{l_r,*}) (\exp(s_{l_r,*}) - 1). \tag{29}$$

411 IV. EXPERIMENTAL SET-UP

412 In this section, we describe the experiments' configurations
 413 to validate our CCGPMA concerning multiple annotators
 414 classification and regression tasks.

415 A. Classification

416 1) *Datasets and simulated/provided annotations:* We test
 417 our approach using three types of datasets: *fully synthetic data*,
 418 *semi-synthetic data*, and *fully real datasets*.

419 First, we generate *fully synthetic data* as one-dimensional
 420 ($P=1$) multi-class classification problem ($K=3$). The input
 421 feature matrix \mathbf{X} is built by randomly sampling $N=1000$
 422 points from an uniform distribution within the interval $[0, 1]$.
 423 The true label for the n -th sample is generated by taking
 424 the $\arg \max_i \{t_{n,i} : i \in \{1, 2, 3\}\}$, where $t_{n,1} = \sin(2\pi x_n)$,
 425 $t_{n,2} = -\sin(2\pi x_n)$, and $t_{n,3} = -\sin(2\pi(x_n + 0.25)) + 0.5$.
 426 Besides, the test instances are obtained by extracting 200
 427 equally spaced samples from the interval $[0, 1]$.

428 Second, to control the label generation, we build *semi-*
 429 *synthetic data* from seven datasets of the UCI repository
 430 focused on binary and multi class-classification: Wisconsin
 431 sin Breast Cancer Database–(breast), BUPA liver disorders
 432 (bupa), Johns Hopkins University Ionosphere database
 433 (ionosphere), Pima Indians Diabetes Database–(pima), Tic-
 434 Tac-Toe Endgame database–(tic-tac-toe), **Occupancy Detection**
 435 **Data Set–(Occupancy)**, **Skin Segmentation Data Set–(Skin)**,
 436 **Wine Data set–(Wine)**, and **Image Segmentation Data Set–**
 437 **(Segmentation)**. Also, we test the publicly available bearing data
 438 collected by the Case Western Reserve University–(Western).
 439 The aim is to build a system to diagnose an electric motor's

TABLE II
TESTED DATASETS.

	Name	Number of features	Number of instances	Number of classes	
<i>fully synthetic</i>	synthetic	1	100	3	
	Breast	9	683	2	
	Bupa	6	345	2	
<i>semi-synthetic</i>	Ionosphere	34	351	2	
	Pima	8	768	2	
	Tic-tac-toe	9	958	2	
	Occupancy	7	20560	2	
	Skin	4	245057	2	
	Western	7	3413	4	
	Wine	13	178	3	
	Segmentation	18	2310	7	
	<i>fully real</i>	Voice	13	218	2
		Music	124	1000	10

status based on two accelerometers. The feature extraction was performed as in (43).

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 bio-signal 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 signal 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) (44). Accordingly, we face five multi-class classification problems (one per scale). We follow the procedure in (36) to rewrite five binary classification tasks preserving the available ground truth (13). Further, we use the music genre data³, 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 annotations from 44 workers). Yet, we only consider the annotators who labeled at least 20% of the instances; thus, we use the information from $R=7$ labelers. The feature extraction is performed by following the work by authors in (32), to obtain an input space with $P=124$. Table II summarizes the tested datasets for the classification case.

Note that the *fully synthetic* and the *semi-synthetic* datasets do not hold real annotations. Therefore, it is necessary to simulate those labels as corrupted versions of the hidden ground truth. Here, the simulations are performed by assuming: i) dependencies among annotators, and ii) the labelers' performance is modeled as a function of the input features. In turn, an SLFM-based approach (termed **SLFM-C**) is used to build the labels, as follows:

- Define Q deterministic functions $\hat{\mu}_q : \mathcal{X} \rightarrow \mathbb{R}$, and their combination parameters $\hat{w}_{l_r,q} \in \mathbb{R}, \forall r \in R, n \in N$.

³<http://fprodigues.com/publications/learning-from-multiple-annotators-distinguishing-good-from-random-labelers/>

²<http://archive.ics.uci.edu/ml>

TABLE III

A BRIEF OVERVIEW OF THE STATE-OF-THE-ART METHODS TESTED.

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 (11)	A LRC with constant parameters across the input space.
MA-DGRL (32)	A multi-labeler approach that considers as latent variables the annotator performance.
MA-GPC (12)	A multi-labeler GPC, which is as an extension of MA-LFC.
MA-GPCV (7)	An extension of MA-GPC that includes variational inference and priors over the labelers' parameters.
MA-DL (18)	A Crowd Layer for DL, where the annotators' parameters are constant across the input space.
KAAR (36)	A kernel-based approach that employs a convex combination of classifiers and codes labelers dependencies.
CGPMA-C	A particular case of our CCGPMA for classification, where $Q = J$, and we fix $w_{j,q} = 1$, if $j = q$, otherwise $w_{j,q} = 0$.

- Compute $\hat{f}_{l,r,n} = \sum_{q=1}^Q \hat{w}_{l,r,q} \hat{\mu}_q(\hat{x}_n)$, where $\hat{x}_n \in \mathbb{R}$ is the n -th component of $\hat{\mathbf{x}} \in \mathbb{R}^N$, being $\hat{\mathbf{x}}$ the 1–D representation of the input features in \mathbf{X} by using the well-known t -distributed Stochastic Neighbor Embedding approach (45).
- Calculate $\hat{\lambda}_n^r = \varsigma(\hat{f}_{l,r,n})$, where $\varsigma(\cdot) \in [0, 1]$ is the sigmoid function.
- Finally, find the r -th label as $y_n^r = \begin{cases} y_n, & \text{if } \lambda_n^r \geq 0.5 \\ \tilde{y}_n, & \text{if } \lambda_n^r < 0.5 \end{cases}$, where \tilde{y}_n is a flipped version of the actual label y_n .

2) *Method comparison and performance metrics:* 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). We use a cross-validation scheme with 15 repetitions where 70% of the samples are utilized for training and the remaining 30% for testing (except for the music dataset training and testing sets are clearly defined). Table III displays the employed methods of the state-of-the-art for comparison purposes. The abbreviations are fixed as: Gaussian Processes classifier (GPC), logistic regression classifier (LRC), majority voting (MV), multiple annotators (MA), Modelling annotators expertise (MAE), Learning from crowds (LFC), Distinguishing good from random labelers (DGRL), kernel alignment-based annotator relevance analysis (KAAR).

B. Regression

1) *Datasets and simulated/provided annotations:* We test our approach using three types of datasets: fully synthetic data, semi-synthetic data, and fully real datasets. First, We generate *fully synthetic data* as an one-dimensional regression problem, where the ground truth for the n -th sample corresponds to $y_n = \sin(2\pi x_n) \sin(6\pi x_n)$, where the input matrix \mathbf{X} is formed by randomly sampling 100 points within the range $[0, 1]$ from an uniform distribution. The test instances are obtained by extracting equally spaced samples from the interval $[0, 1]$. Second, to control the label generation (10), we build *semi-synthetic data* from six datasets related to regression tasks from the well-known UCI repository. We selected the following datasets: Auto MPG Data Set–(Auto), Bike Sharing Dataset Data Set–(Bike), Concrete Compressive Strength Data Set–(Concrete), The Boston Housing Dataset–(Housing),⁴ Yacht

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

TABLE IV

DATASETS FOR REGRESSION.

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

TABLE V

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 (11)	A LR model for MA where the labelers' parameters are supposed to be constant across the input space.
MA-GPR (12)	A multi-labeler GPR, which is as an extension of MA-LFCR.
MA-DL (18)	A Crowd Layer for DL, where the annotators' parameters are constant across the input space.
CGPMA-R	A particular case of our CCGPMA for regression, where $Q = J$, and $w_{j,q} = 1$ if $j = q$, otherwise $w_{j,q} = 0$.

Hydrodynamics Data Set–(Yacht), and Relative location of CT slices on axial axis Data Set–(CT). Third, we evaluate our proposal on one *fully real dataset*. In particular, we use the Music dataset introduced in Section IV-A1. Notice that the music dataset configures a 10-class classification problem; however, in this experiment, we are using our CCGPMA with a likelihood function designed for real-valued labels Eq. (25). Such practice is not uncommon in machine learning, and it is usually known as “Least-square classification” (39). Table IV 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., $y_n^r = y_n + \epsilon_n^r$, where $\epsilon_n^r \sim \mathcal{N}(0, v_n^r)$, being v_n^r the r -th annotator error-variance for the sample n . Note that we are interested in modeling such an error-variance for the r -th annotator as a function of the input features, which is correlated with the other labelers' variances. In turn, an SLFM-based approach (termed SLFM-R) is used to build the labels, as follows:

- Define Q functions $\hat{\mu}_q : \mathcal{X} \rightarrow \mathbb{R}$, and the combination parameters $\hat{w}_{l,r,q} \in \mathbb{R}$, $\forall r, q$.
- Compute $\hat{f}_{l,r,n} = \sum_{q=1}^Q \hat{w}_{l,r,q} \hat{\mu}_q(\hat{x}_n)$, where \hat{x}_n is the n -th component of $\hat{\mathbf{x}} \in \mathbb{R}$, which is a 1–D representation of input features \mathbf{X} by using the t -distributed Stochastic Neighbor Embedding approach (45).
- Finally, determine $\hat{v}_n^r = \exp(\hat{f}_{l,r,n})$.

2) *Method comparison and performance metrics:* 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 training and the remaining

30% for testing (except for *fully synthetic dataset*, since it clearly defines the training and testing sets). Table V displays the employed methods of the state-of-the-art for comparison purposes. From Table V, 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 (18).

$$\hat{\mu}_1(x) = 4.5 \cos(2\pi x + 1.5\pi) - 3 \sin(4.3\pi x + 0.3\pi), \quad (31)$$

$$\hat{\mu}_2(x) = 4.5 \cos(1.5\pi x + 0.5\pi) + 5 \sin(3\pi x + 1.5\pi), \quad (32)$$

$$\hat{\mu}_3(x) = 1, \quad (33)$$

where $x \in [0, 1]$. Besides, the combination weights are gathered within the following combination matrix $\hat{W} \in \mathbb{R}^{Q \times R}$:

$$\hat{W} = \begin{bmatrix} 0.4 & 0.7 & -0.5 & 0.0 & -0.7 \\ 0.4 & -1.0 & -0.1 & -0.8 & 1.0 \\ 3.1 & -1.8 & -0.6 & -1.2 & 1.0 \end{bmatrix}, \quad (34)$$

C. CCGPMA training

Overall, the Radial basis function-(RBF) kernel is preferred in both classification and regression tasks because of its universal approximating ability and mathematical tractability. Hence, for all GP-based approaches, the kernel functions are fixed as:

$$\kappa(\mathbf{x}_n, \mathbf{x}_{n'}) = \phi_1 \exp\left(\frac{-\|\mathbf{x}_n - \mathbf{x}_{n'}\|_2^2}{2\phi_2^2}\right), \quad (30)$$

where $\|\cdot\|_2^2$ stands for the L2 norm, $n, n' \in \{1, 2, \dots, N\}$ and $\phi_1, \phi_2 \in \mathbb{R}^+$ are the kernel hyper-parameters. For concrete testing, we fix $\phi_1 = 1$, while ϕ_2 is estimated by optimizing the corresponding ELBO (as exposed in Eq. (19)). Moreover, for CGPMA, since each LF $f_j(\cdot)$ is linked to $u_q(\cdot)$, we fix $Q = R + K$, and $Q = R + 1$ for classification and regression respectively. On the other hand, for CCGPMA, each $f_j(\cdot)$ is built as a convex combination of $\mu_q(\cdot)$ (see Eq. (10)); therefore, there is no restriction concerning Q . However, to make a fair comparison with CGPMA, we also fix $Q = R + K$ (classification), and $Q = R + 1$ (regression) in CCGPMA. For the *fully synthetic datasets*, we use $M = 10$ inducing points per latent function, and for the remaining experiments, we test with $M = 40$, and $M = 80$. For all the experiments, we use the ADADELTA included in the climin library with a mini-batch size of 100 samples to perform SVI. However, for small datasets ($N \leq 500$), we employ mini-batches with a size equal to the number of samples in the training set. Finally, for all experiments related to our CCGPMA, the variational parameters' initialization is carried out as follows: the variational mean is set $\mathbf{m}_q = \mathbf{0}, \forall q \in \{1, \dots, Q\}$, where $\mathbf{0} \in \mathbb{R}^M$ is an all-zeros vector; the variational covariances $\mathbf{V}_q = \mathbf{I}, \forall q \in \{1, \dots, Q\}$ are fixed as the identity matrix $\mathbf{I} \in \mathbb{R}^{M \times M}$. The CCGPMA's Python code is publicly available.⁵

V. RESULTS AND DISCUSSION

A. Classification

1) *Fully synthetic data results.*: We first perform a controlled experiment to test the CCGPMA capability when dealing with binary and multi-class classification. We use the *fully synthetic* dataset described in Section IV-A1. Besides, five labelers ($R = 5$) are simulated with different levels of expertise. To simulate the error-variances, we define $Q = 3$ $\hat{\mu}_q(\cdot)$ functions, yielding

holding elements $\hat{w}_{l,r,q}$. For visual inspection purposes, Fig. 1 shows the predictive label's probability-(PLP), $p(y_* = k | \mathbf{x}_*)$, and the AUC for all studied approaches regarding the *fully synthetic* data. Notice that for methods MA-GPC, MA-GPCV, and KAAR, we use the *one-vs-all* scheme to face this experiment (such methods were defined only for binary classification settings). Accordingly, for those models, the PLP corresponds to scores rather than probabilities. Besides, regarding the PLP of our CGPMA and CCGPMA, we provide the mean and variance for the predictive distribution $\zeta_{k,*} = p(y_* = k | \mathbf{x}_*, \hat{\mathbf{f}}, \mathbf{u})$, which are computed based on Eqs. (21) and (22). As seen in Fig. 1, KAAR, MA-GPC, and MA-GPCV presents a different shape than the ground truth; moreover, KAAR and MA-GPCV exhibit the worst AUC, even worse than the intuitive lower bound GPC-MV. We explain such conduct in the sense that these approaches are designed to deal with binary labels (36; 12; 10). To face such a problem, we use the *one-vs-all* scheme; still, it can lead to ambiguously classified regions (47). We note an akin predictive AUC concerning MA-DL methods and the linear approaches MA-LFC-C and MA-DGRL. Nonetheless, the linear techniques exhibit a PLP less similar to the Ground truth, which is due to MA-LFC-C and MA-DGRL only can deal with linearly separable data. Further, we analyze the results of our CGPMA-C and its particular enhancement CCGPMA-C. We remark that our methods' predictive AUC is pretty close to deep learning and linear models. Unlike them, our CGPMA-C and CCGPMA-C show the most accurate PLP compared with the absolute gold standard. CCGPMA-C behaves quite similarly to GPC-GOLD, which is the theoretical upper bound. Finally, from the GPC-MV, we do not identify notable differences with the rest of the approaches (excluding KAAR and MA-GPCV).

From the above, we recognize that analyzing both the predictive AUC and the PLP, our CCGPMA-C exhibits the best performance obtaining similar results compared with the intuitive upper bound (GPC-GOLD). Accordingly, CCGPMA-C proffers a more suitable representation of the labelers' behavior than its competitors. Indeed, CCGPMA-C codes both the annotators' dependencies and the relationship between the input features and the annotators' performance. To empirically support the above statement, Fig. 2 shows the estimated per-annotator reliability, where we only take into account models that include such types of parameters (MA-DGRL, CGPMA-C, and CCGPMA-C). As seen, MA-DGRL (see column 2 in Fig. 2) does not offer a proper representation of the annotators' behavior. CGPMA-C and CCGPMA-C (columns 3 and 4 in Fig. 2) outperforms MA-DGRL, which is a direct repercussion

⁵<https://github.com/juliangilg/CCGPMA>

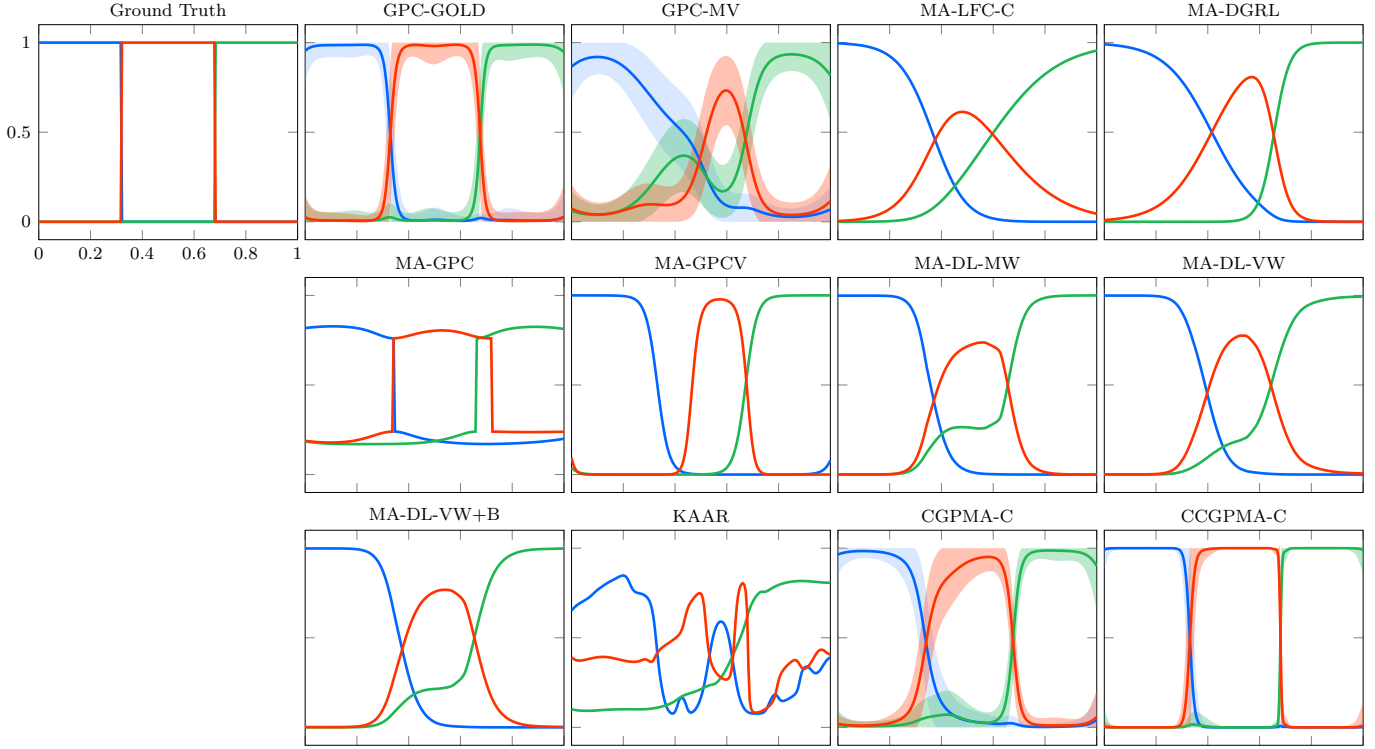


Fig. 1. Fully synthetic dataset results. The PLP is shown, comparing the prediction of our CCGPMA-C ($AUC = 1$) and CCGPMA-C ($AUC = 0.9999$) against: the theoretical upper bound GPC-GOLD ($AUC = 1.0$), the lower bound GPC-MV ($AUC = 0.9809$), and the state-of-the-art approaches MA-LFC-C ($AUC = 0.9993$), 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 \pm two standard deviations. There is no shaded region for approaches lacking prediction uncertainty.

TABLE VI
AUC(%) CLASSIFICATION RESULTS FOR THE SEMI SYNTHETIC DATASETS. BOLD: THE HIGHEST AUC EXCLUDING THE UPPER BOUND (GPC-GOLD).

Method	Breast	Bupa	Ionosphere	Pima	TicTacToe	Occupancy	Skin	Western	Wine	Segmentation	Average
GPC-GOLD($M = 40$)	99.07 \pm 0.45	69.75 \pm 4.66	94.90 \pm 2.35	83.78 \pm 3.02	84.29 \pm 3.34	99.56 \pm 0.06	99.97 \pm 0.01	91.85 \pm 0.61	99.87 \pm 0.15	95.96 \pm 1.96	91.90
GPC-GOLD($M = 80$)	99.03 \pm 0.46	69.97 \pm 4.83	95.13 \pm 2.25	83.74 \pm 2.97	84.91 \pm 3.23	99.56 \pm 0.06	99.97 \pm 0.01	92.50 \pm 0.57	99.88 \pm 0.16	97.81 \pm 0.41	92.25
GPC-MV($M = 40$)	98.97 \pm 0.45	53.66 \pm 5.16	75.66 \pm 5.72	53.99 \pm 7.60	66.20 \pm 3.57	75.85 \pm 19.16	84.58 \pm 0.90	86.58 \pm 3.31	81.79 \pm 2.12	95.62 \pm 2.28	77.29
GPC-MV($M = 80$)	98.92 \pm 0.48	56.98 \pm 5.29	77.79 \pm 5.50	53.02 \pm 6.74	67.44 \pm 3.57	63.12 \pm 19.68	84.20 \pm 0.80	84.46 \pm 0.89	83.23 \pm 3.87	97.49 \pm 0.47	76.66
MA-LFC-C	87.89 \pm 5.10	45.93 \pm 14.44	73.58 \pm 9.01	81.19 \pm 3.13	60.04 \pm 2.61	89.42 \pm 0.79	94.40 \pm 0.08	84.00 \pm 2.11	96.92 \pm 3.57	98.92 \pm 0.31	81.23
MA-DGRL	97.57 \pm 1.89	57.24 \pm 3.36	64.53 \pm 7.21	81.38 \pm 2.90	61.29 \pm 2.30	49.71 \pm 1.05	93.79 \pm 1.07	81.43 \pm 1.50	97.95 \pm 2.21	98.97 \pm 0.38	78.39
MA-GPC	98.11 \pm 1.16	54.46 \pm 5.78	66.31 \pm 14.74	53.25 \pm 17.80	60.79 \pm 9.95	92.57 \pm 7.96	80.89 \pm 0.60	86.71 \pm 1.14	94.17 \pm 2.62	97.34 \pm 0.35	78.46
MA-GPCV	82.70 \pm 5.47	55.67 \pm 6.83	62.38 \pm 8.71	62.17 \pm 5.90	61.04 \pm 10.03	60.22 \pm 2.66	76.29 \pm 3.74	84.51 \pm 1.47	97.35 \pm 1.72	99.24 \pm 0.27	74.16
MA-DL-MW	94.70 \pm 1.73	52.37 \pm 5.68	75.35 \pm 5.43	61.78 \pm 2.67	68.27 \pm 2.96	64.09 \pm 2.26	86.36 \pm 0.57	90.92 \pm 0.56	97.28 \pm 1.09	99.50 \pm 0.17	79.06
MA-DL-VW	95.26 \pm 2.45	53.27 \pm 6.18	69.87 \pm 4.97	60.63 \pm 3.36	67.71 \pm 2.67	68.40 \pm 3.45	86.56 \pm 0.68	91.73 \pm 0.67	98.07 \pm 1.52	99.72 \pm 0.11	79.12
MA-DL-VW+B	94.65 \pm 2.42	52.81 \pm 6.31	71.96 \pm 4.53	61.23 \pm 3.78	67.80 \pm 3.42	67.82 \pm 3.86	86.68 \pm 0.67	91.64 \pm 0.85	98.17 \pm 1.55	99.72 \pm 0.09	79.25
KAAR	80.58 \pm 2.74	59.20 \pm 6.63	70.46 \pm 7.39	58.02 \pm 4.06	63.81 \pm 5.45	69.16 \pm 2.06	51.58 \pm 4.74	85.88 \pm 1.20	99.43 \pm 1.05	92.17 \pm 1.90	73.03
CGPMA-C($M = 40$)	99.20 \pm 0.38	57.13 \pm 4.68	83.56 \pm 10.02	82.01 \pm 3.14	70.56 \pm 3.04	82.20 \pm 2.73	92.62 \pm 1.20	91.78 \pm 0.66	99.82 \pm 0.18	96.79 \pm 0.65	85.56
CGPMA-C($M = 80$)	99.14 \pm 0.38	56.96 \pm 4.74	86.15 \pm 6.96	82.04 \pm 3.18	70.48 \pm 3.12	99.08 \pm 0.26	90.46 \pm 1.64	91.85 \pm 0.57	99.84 \pm 0.12	94.06 \pm 0.61	87.01
CCGPMA-C($M = 40$)	99.38 \pm 0.27	60.22 \pm 5.06	87.84 \pm 6.72	78.10 \pm 6.22	74.95 \pm 5.39	91.98 \pm 2.00	85.70 \pm 2.66	93.09 \pm 0.51	99.44 \pm 0.33	97.67 \pm 0.53	86.84
CCGPMA-C($M = 80$)	99.33 \pm 0.30	59.19 \pm 5.65	90.55 \pm 6.29	80.45 \pm 5.10	73.12 \pm 3.23	97.75 \pm 2.00	89.42 \pm 2.20	93.15 \pm 0.50	99.43 \pm 0.33	97.58 \pm 0.43	88.00

of modeling the labelers' parameters as functions of the input features. 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 their decisions (26; 36).

2) *Semi-synthetic data results.*: It is worth mentioning that the Semi-synthetic experiments are a common practice in the *learning from crowds* area (10; 36; 7), where the input features comes from real-world datasets whilst the labels from multiple annotators are simulated following the fully synthetic data set-up (see Eqs. (31) to (34)). Table VI shows the results concerning this second experiment. On average, CCGPMA-C accomplishes the best predictive AUC; moreover,

we note that CGPMA-C reaches the second-best performance. Furthermore, the GPs-based competitors achieve competitive results (GPC-MV, MA-GPC, MA-GPCV, and KAAR). On the other hand, the GPC-MV method obtains a significantly lower performance than our CCGPMA-C, which is explained because GPC-MV is the most naive approach since it considers that the whole annotators exhibit the same performance. Conversely, analyzing the results from MA-GPC, MA-GPCV, and KAAR, we note that they perform worse than GPC-MV. We explain such an outcome in two ways. First, these approaches do not model the relationship between the input features and the annotators' performance. Second, as exposed in a previous experiment MA-GPC, MA-GPCV, and KAAR use a *one-*

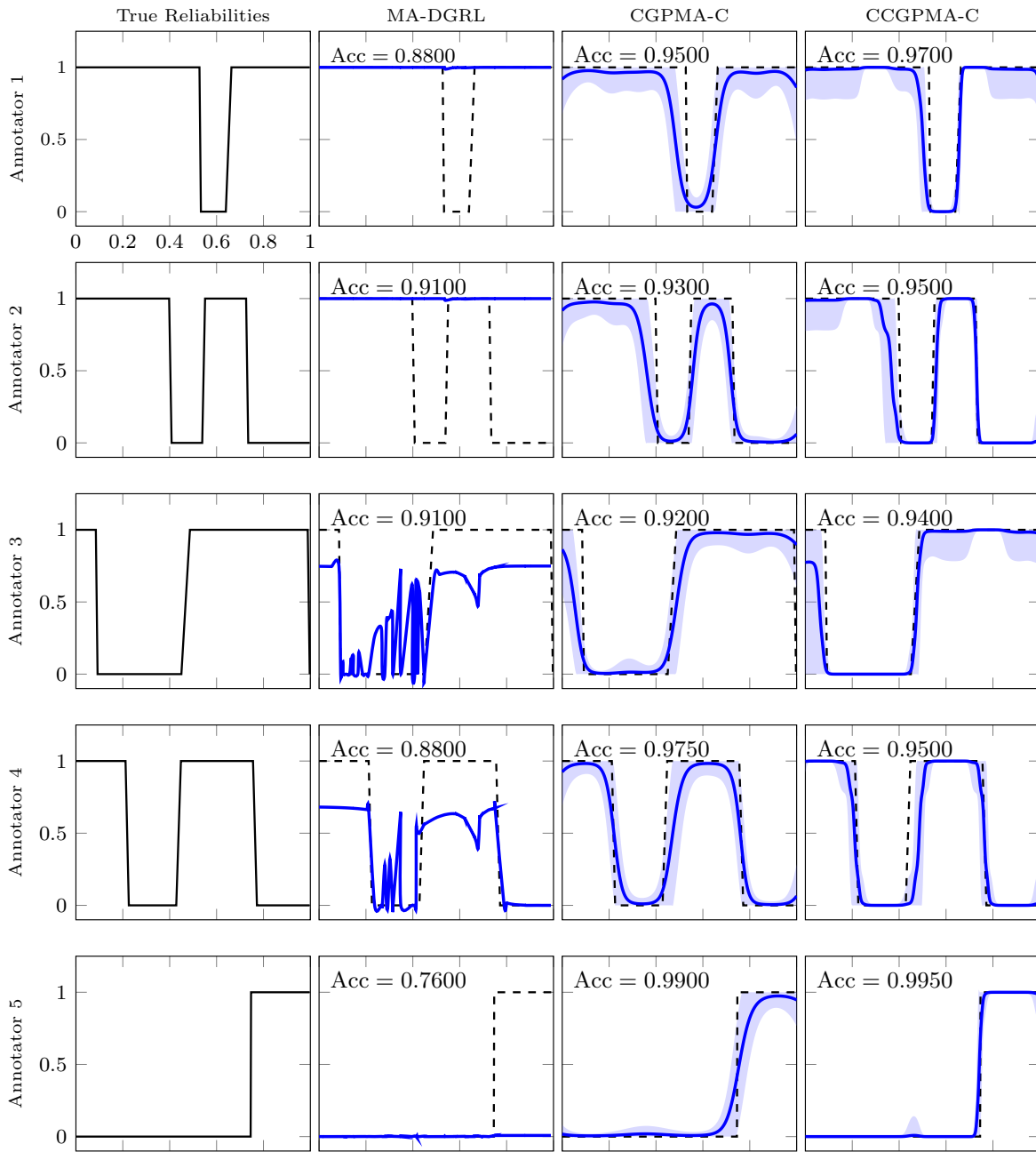


Fig. 2. Fully synthetic data reliability results. From top to bottom, the first column exposes the true reliabilities (λ_r). The subsequent columns present the estimation of the reliabilities performed by state-of-the-art models, where the correct values are provided in dashed lines. The shaded region in CGPMA-C and CCGPMA-C indicates the area enclosed by the mean \pm two standard deviations. Also, the accuracy (Acc) is provided.

673 *vs-all* to deal with multi-class problems, which can lead to
 674 ambiguously classified regions (47). The latter can be confirmed
 675 in the results for the multi-class dataset “Western” ($K = 4$)
 676 where the predictive AUC for such approaches are the lowest
 677 Then, analyzing the results from the DL-based strategies
 678 we note a slightly better performance compared with the
 679 GPs-based methods (excluding CGPMA-C and CCGPMA-
 680 C). However, the DL-based performs considerably worse than
 681 our proposal because the CrowdLayer provides straightforward
 682 codification of the labelers’ performance to guarantee a low
 683 computational cost (37). Finally, from the linear models, we

first analyze the outstanding performance from MA-DGRL,
 which defeats all its non-linear competitors. In particular, the
 simulated labels (see Section IV-A1) follows the MA-DGRL
 model, favoring its performance. Though MA-LFC-C achieves
 competitive performance compared to the DL-based methods,
 it is considerably lower than our proposal. In fact, the MA-
 LFC-C formulation assumes that the annotators’ behavior is
 homogeneous across the input space, which does not correspond
 to the labels simulation procedure.

3) *Fully real data results.*: We test the *fully real datasets*,
 which configure the most challenging scenario. The input

695 features and the labels from multiple experts come from real-
 696 world applications. Table VII outlines the achieved AUC. First,
 697 we observe that for the voice data, G and R scales exhibit a
 698 similar AUC for all considered approaches; in fact, GPC-MV
 699 obtains a result comparable with the upper bound GPC-GOLD.
 700 The latter can be explained in the sense that the annotators
 701 exhibit a suitable performance for these scales, i.e., the provided
 702 labels are similar to the ground truth. On the other hand, \mathcal{B}
 703 reduction in the predictive AUC is observed for scale B, which
 704 is a consequence of diminishing the labelers' performance
 705 compared with scales G and R, as demonstrated in (13). Our
 706 approaches exhibit the best generalization performances for
 707 the three scales in the voice dataset. Remarkably, CGPMA-
 708 C and CCGPMA-C do not suffer significant changes in the
 709 scale B, which is an outstanding outcome because it reflects
 710 that our method offers a better representation of the labelers'
 711 behavior against low-quality annotations. Finally, we review
 712 the AUC for the Music dataset. Achieved results show a low
 713 performance for the MA-GPC, even lower than their intuitive
 714 lower bound (GPC-MV). Notably, our CCGPMA-C reaches
 715 the best predictive AUC, being comparable with the intuitive
 716 upper bound.

743
744
745
746
747
748
749
750
751
752
753
754
755
756
757
758
759
760
761
762
763
764
765
766
767
768
769
770
771
772
773
774
775
776
777
778
779
780
781
782
783
784
785
786
787
788
789
790
791
792
793
794
795
796
797
798
799
800
801
802
803
804
805
806
807
808
809
810
811
812
813
814
815
816
817
818
819
820
821
822
823
824
825
826
827
828
829
830
831
832
833
834
835
836
837
838
839
840
841
842
843
844
845
846
847
848
849
850
851
852
853
854
855
856
857
858
859
860
861
862
863
864
865
866
867
868
869
870
871
872
873
874
875
876
877
878
879
880
881
882
883
884
885
886
887
888
889
890
891
892
893
894
895
896
897
898
899
900
901
902
903
904
905
906
907
908
909
910
911
912
913
914
915
916
917
918
919
920
921
922
923
924
925
926
927
928
929
930
931
932
933
934
935
936
937
938
939
940
941
942
943
944
945
946
947
948
949
950
951
952
953
954
955
956
957
958
959
960
961
962
963
964
965
966
967
968
969
970
971
972
973
974
975
976
977
978
979
980
981
982
983
984
985
986
987
988
989
990
991
992
993
994
995
996
997
998
999
1000

Method	G	Voice R	B	Music	Average
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-LFCR-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

B. Regression

1) *Fully synthetic data results* : 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-B1. We simulate five labelers ($R = 5$) with different levels of expertise. To simulate the error-variances, we define $Q = 3$ functions $\hat{\mu}_q(\cdot)$, which are given as

$$\hat{\mu}_1(x) = 4.5 \cos(2\pi x + 1.5\pi) - 3 \sin(4.3\pi x + 0.3\pi) + \dots + 4 \cos(7\pi x + 2.4\pi), \quad (35)$$

$$\hat{\mu}_2(x) = 4.5 \cos(1.5\pi x + 0.5\pi) + 5 \sin(3\pi x + 1.5\pi) - \dots - 4.5 \cos(8\pi x + 0.25\pi), \quad (36)$$

$$\hat{\mu}_3(x) = 1, \quad (37)$$

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

$$\hat{W} = \begin{bmatrix} -0.10 & 0.01 & -0.05 & 0.01 & -0.01 \\ 0.10 & -0.01 & 0.01 & -0.05 & 0.05 \\ -2.3 & -1.77 & 0.54 & 0.9 & 1.42 \end{bmatrix}, \quad (38)$$

holding elements $w_{l,r,q}$.

Fig. 3 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 low 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 can deal with non-linear dynamics. However, MA-DL reaches a significantly low performance (even lower than the most naive approach, GPR-Av). To explain such an outcome, we remark that MA-DL comprises the introduction of an additional layer, the ‘‘CrowdLayer’’, which allows the training of neural networks directly from the noisy labels of multiple annotators (18). Yet, such a CrowdLayer provides a very simple codification of the annotators’ performance to guarantee a low computational cost (37); therefore, MA-DL does not provide a proper codification of the annotators’ behavior. On the other hand, among the GP-based methods, the proposed CCGPMA-R achieves the best performance in terms of R^2 , followed closely by CGPMA-R and MA-GPR.

Besides, concerning the high performance of our CCGPMA-R (the best in terms of R^2 score), we hypothesize that such an outcome is a consequence of our method offers a better representation of the labelers’ behavior when compared with its competitors. To empirically support the above hypothesis, Fig. 4 shows the estimated error-variances for this first experiment; here, we only take into account the models that include these parameters in their formulations. As seen in Fig. 4, MA-LFCR and MA-GPR offer the worst representation for the annotators’ performance, which is due to such methods do not take into account the relationship between the annotators and the input space. Conversely, CGPMA-R and CCGPMA-R outperform the models named previously. This outcome is a consequence that such two approaches compute the error-variance as a function 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 than CGPMA-R because the former codes properly the annotators’ interdependencies (26). Finally, we remark that although our CCGPMA-R achieves the best representation of the annotators’ performance, Annotator 4 exhibits a lower performance in terms of R^2 score compared with the other labelers. Such an outcome is caused by the quasi-periodic behavior in the error-variances for those labelers, which cannot be captured because we are using an RBF-based kernel.

2) *Results over semi-synthetic data*: Table VIII shows the results of the semi synthetic datasets. On average, our CCGPMA-R exhibits the best generalization performance in

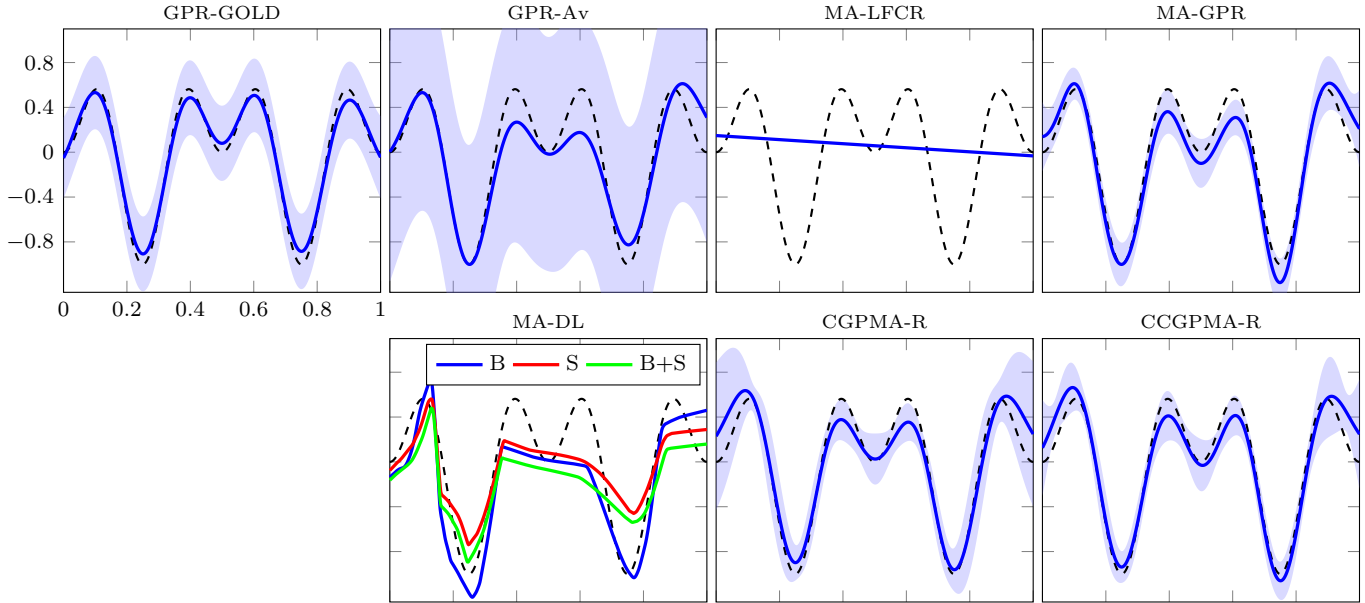


Fig. 3. Fully synthetic dataset results. We compare the prediction of our CCGPMA-R ($R^2 = 0.9438$), and CGPMA-R ($R^2 = 0.9280$) with the theoretical upper bound GPR-GOLD ($R^2 = 0.9843$) and lower bound GPR-Av ($R^2 = 0.8718$), and state-of-the-art approaches, MA-LFCR ($R^2 = -0.0245$), MA-GPR ($R^2 = 0.9208$), 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 they do not provide information about the prediction uncertainty.

780 terms of the R^2 score. On the other hand, regarding its GP-
 781 based competitors (GPR-Av, MA-GPR, and CGPMA-R), we
 782 first note that the performance of CGPMA-R exhibits a similar
 783 (but lower) performance than CCGPMA-R. The above is
 784 consequence of that conversely to CGPMA-R, our CCGPMA-
 785 R models the annotators' interdependencies. Secondly, the
 786 intuitive lower bound GPR-Av exhibits a significantly worse
 787 prediction than our approaches. We remark on MA-GPR's
 788 behavior, which is lowest compared with its GPs-based com-
 789 petitors, even far worse than the supposed lower bound GPR-
 790 Av. The key to this abnormal outcome lies in the formulation
 791 of this approach; MA-GPR models the annotators' behavior
 792 by assuming that their performance does not depend on the
 793 input features and considering that the labelers make their
 794 decisions independently, which does fit the process that we
 795 use to simulate the labels.

796 Next, we analyze the results concerning the linear model

MA-LFR; attained to the results, we note that this approach's
 prediction capacity is far lower than ours. 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 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 GPR-GOLD results. Thus, we
 can elucidate that there exists a linear structure in the dataset
 CT. 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

TABLE VIII

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 \pm 0.0271	0.5529 \pm 0.0065	0.8037 \pm 0.0254	0.8235 \pm 0.0419	0.8354 \pm 0.0412	0.8569 \pm 0.0055	0.7888
GPR-GOLD ($M = 80$)	0.8612 \pm 0.0279	0.5603 \pm 0.0063	0.8271 \pm 0.0230	0.8275 \pm 0.0399	0.8240 \pm 0.0339	0.8648 \pm 0.0047	0.7942
GPR-Av ($M = 40$)	0.8425 \pm 0.0286	0.5280 \pm 0.0100	0.7589 \pm 0.0279	0.7834 \pm 0.0463	0.7588 \pm 0.0498	0.8070 \pm 0.0130	0.7464
GPR-Av ($M = 80$)	0.8406 \pm 0.0304	0.5397 \pm 0.0085	0.7765 \pm 0.0274	0.7903 \pm 0.0451	0.7676 \pm 0.0535	0.8167 \pm 0.0089	0.7552
MA-LFCR	0.7973 \pm 0.0218	0.3385 \pm 0.0051	0.6064 \pm 0.0384	0.7122 \pm 0.0509	0.6403 \pm 0.0186	0.8400 \pm 0.0014	0.6558
MA-GPR	0.8456 \pm 0.0281	0.4448 \pm 0.0187	0.7769 \pm 0.0367	0.7685 \pm 0.0632	0.7842 \pm 0.1027	0.0105 \pm 0.0045	0.6051
MA-DL-B	0.7766 \pm 0.0253	0.5854 \pm 0.0107	0.2319 \pm 0.0328	0.5317 \pm 0.1005	0.2089 \pm 0.0783	0.6903 \pm 0.2689	0.5041
MA-DL-S	0.7761 \pm 0.0279	0.5828 \pm 0.0149	0.2363 \pm 0.0252	0.5352 \pm 0.0948	0.1822 \pm 0.0985	0.8418 \pm 0.2288	0.5257
MA-DL-B+S	0.7717 \pm 0.0239	0.5816 \pm 0.0181	0.2369 \pm 0.0322	0.5330 \pm 0.0850	0.1974 \pm 0.0895	0.5517 \pm 0.2316	0.4787
CGPMA-R ($M = 40$)	0.8476 \pm 0.0229	0.5464 \pm 0.0069	0.8169 \pm 0.0231	0.7244 \pm 0.2973	0.8049 \pm 0.0482	0.8236 \pm 0.0132	0.7606
CGPMA-R ($M = 80$)	0.8342 \pm 0.0217	0.5560 \pm 0.0074	0.8190 \pm 0.0254	0.7259 \pm 0.3018	0.7928 \pm 0.0884	0.8371 \pm 0.0104	0.7608
CCGPMA-R ($M = 40$)	0.8558 \pm 0.0248	0.5284 \pm 0.0117	0.7976 \pm 0.0270	0.8169 \pm 0.0468	0.8409 \pm 0.0548	0.8219 \pm 0.0062	0.7769
CCGPMA-R ($M = 80$)	0.8534 \pm 0.0243	0.5467 \pm 0.0069	0.8220 \pm 0.0259	0.8215 \pm 0.0466	0.8691 \pm 0.0473	0.8252 \pm 0.0083	0.7897

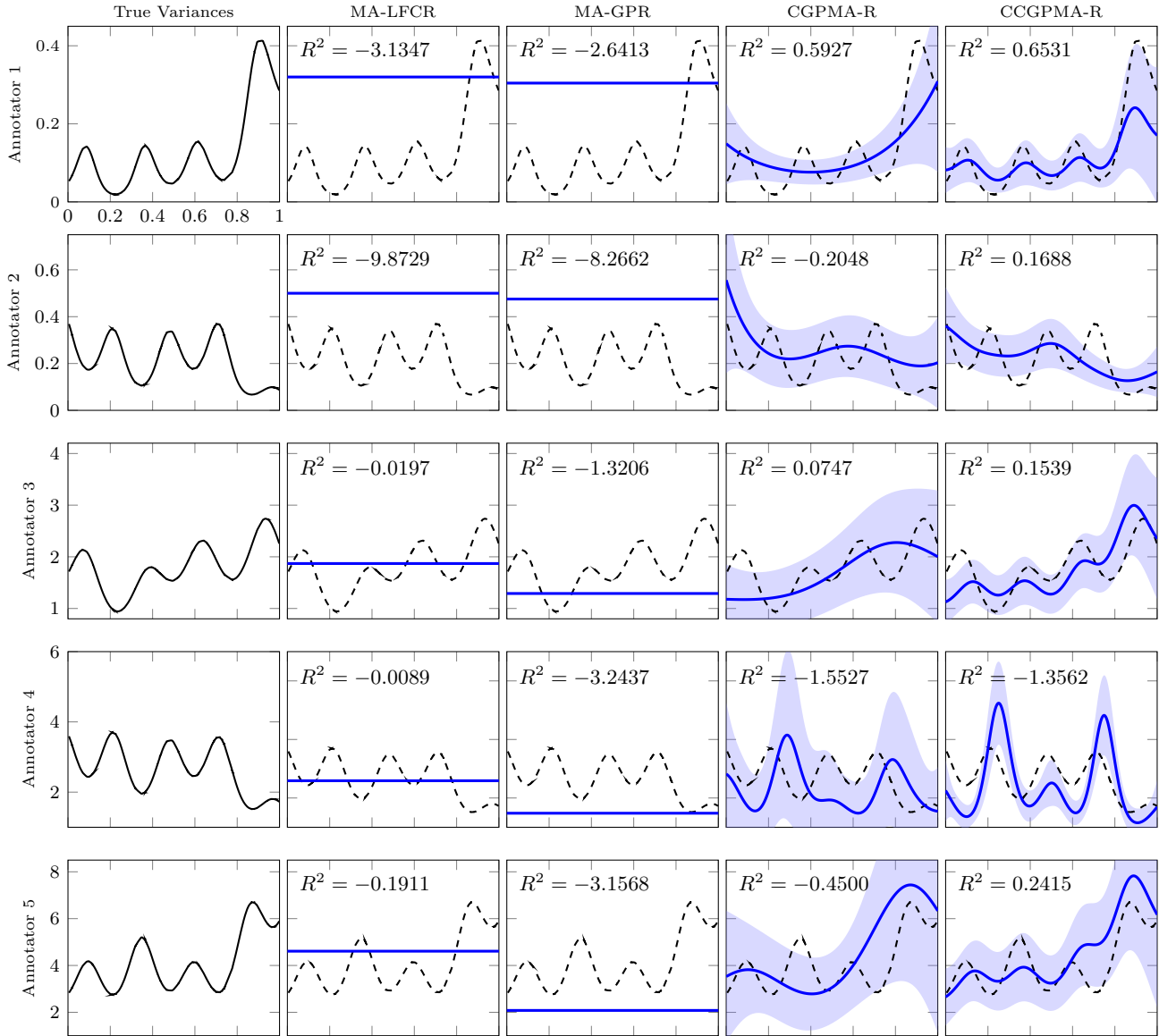


Fig. 4. Estimated values of error-variance for the five annotators in the *fully synthetic* experiment. In the first column, from top to bottom, we expose the error-variances used to simulate the labels from each annotator. Furthermore, the subsequent columns from top to bottom present the estimation 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.

814 this behavior to the fact that the CrowdLayer (used to manage
 815 the data from multiple annotators) does not offer a suitable
 816 codification of the labelers’ behavior. Nevertheless, taking the
 817 above into account, we observe a remarkable result in the Bike
 818 dataset. The DL-based approaches offer the best performance,
 819 even defeating the supposed upper-bound GPR-GOLD. To
 820 explain that, it is necessary to analyze the meaning of the
 821 target variable in such a dataset. Regarding the description of
 822 this dataset,⁶ the target variables indicate the count of total
 823 rental bikes, including both casual and registered in a day. The
 824 above suggests that there may exist a quasi-periodic structure
 825 in the dataset, which the GPR-GOLD cannot capture since

uses a non-periodic kernel (RBF). To support our suppositions,
 an additional experiment was performed over this dataset by
 training the model GPR-GOLD with the following kernel:

$$\kappa(\mathbf{x}_n, \mathbf{x}_{n'}) = \varphi \exp \left[-\frac{1}{2} \sum_{p=1}^P \left(\frac{\sin \left(\frac{\pi(x_{p,n} - x_{p,n'})}{T_p} \right)}{l_p} \right)^2 \right], \quad (39)$$

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

⁶<https://archive.ics.uci.edu/ml/datasets/bike+sharing+dataset>

the obtained by the DL-based approaches, indicating a quasi-periodic structure in the Bike dataset as we had supposed.

3) *Fully real data results:* Finally, we use the *fully real datasets*, which present the most challenging scenario, where both the input samples and the labels come from real-world applications. Table IX outlines the achieved performances. We

TABLE IX

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

remark that our CCGPMA-R with $M = 80$ obtains the best generalization performance in terms of the R^2 score. Further, as theoretically expected, its performance lies between that of GPR-GOLD and GP-Av. Moreover, regarding the GP-based competitors (MA-GPR and CGPMA-R), we note that our CGPMA-R is just a bit lower than CCGPMA-R. On the other hand, MA-GPR exhibits the worst prediction capability with a R^2 close to zero. We suppose the above is a symptom of overfitting, which can be confirmed because the training R^2 score for MA-GPR is 0.4731, comparable with GPR-GOLD. Conversely, 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 exhibits a similar performance compared with our CGPMA-R; however, it is slightly lower than our CCGPMA-R. We highlight that despite deep learning capacities, our approach CCGPMA-R offers a better representation of annotators' behavior, unlike the deep learning techniques, which measure such performance using a single parameter.

Also, we observe that all regression models presented a lower generalization performance than previous results (see Table V in the paper) over the same dataset. The above is a repercussion of solving a multi-class classification problem with regression models. Such an outcome is not uncommon, and it can be founded in works (18; 15).

VI. CONCLUSION

This paper introduces a novel Gaussian Process-based approach to deal with Multiple Annotators scenarios, termed Correlated Chain Gaussian Process for Multiple Annotators (CCGPMA). Our method is built as an extension of the chained GP (27), introducing a semi-parametric latent factor model (SLFM) to exploit correlations between the GP latent functions that model the parameters of a given likelihood function. To the

best of our knowledge, CCGPMA is the first attempt to build a probabilistic framework that codes the annotators' expertise as a function of the input data and exploits the correlations among the labelers' answers. Besides, we highlight that our approach can be used with different likelihood, which allows us to deal with both categorical data (classification) and real-valued (regression). We tested our approach for classification tasks using different scenarios concerning the provided annotations: synthetic, semi-synthetic, real-world experts. According to the results, we remark that our CCGPMA can achieve robust predictive properties for the studied datasets, outperforming state-of-the-art methods.

As future work, CCGPMA can be extended by using convolution processes (48) instead of the SLFM, aiming to obtain a better representation of the correlations among the labelers. Also, our approach can be extended for multi-task learning in the context of multiple annotators (49). Finally, we note that the performance of our approach heavily depend on kernel selection (see Section V-B2); accordingly, it would be interesting to automatically perform such kernel selection (50) as an input block of our framework.

ACKNOWLEDGMENT

Under grants provided by the Minciencias project: "Desarrollo de un prototipo funcional para el monitoreo no intrusivo de vehículos usando data analytics para innovar en el proceso de mantenimiento basado en la condición en empresas de transporte público."-code 643885271399. J. Gil is funded by the program "Doctorados Nacionales - Convocatoria 785 de 2017". MAA has been financed by the EPSRC Research Projects EP/R034303/1 and EP/T00343X/1. MAA has also been supported by the Rosetrees Trust (ref: A2501). A.M. Alvarez is financed by the project "Prototipo de interfaz cerebro-computador multimodal para la detección de patrones relevantes relacionados con trastornos de impulsividad" (Universidad Nacional de Colombia - code 50835).

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. 16662–16675, 2020.
- [3] Y. Kara, G. Genc, O. Aran, and L. Akarun, "Modeling annotator behaviors for crowd labeling," *Neurocomputing*, vol. 160, pp. 141–156, 2015.
- [4] R. Snow, B. O'Connor, D. Jurafsky, and A. Ng, "Cheap and fast—but is it good?: evaluating non-expert annotations for natural language tasks," in *EMNLP*. ACL, 2008, pp. 254–263.
- [5] 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.
- [6] 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.
- [7] 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.
- [8] 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.
- [9] A. Dawid and A. Skene, "Maximum likelihood estimation of observer error-rates using the EM algorithm," *Appl. Stat.*, pp. 20–28, 1979.

- [10] 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.
- [11] 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. 101, no. Apr, pp. 1297–1322, 2010.
- [12] F. Rodrigues, F. C. Pereira, and B. Ribeiro, "Gaussian process classification and active learning with multiple annotators," in *ICML*, 2014, pp. 433–441.
- [13] 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.
- [14] P. Groot, A. Birlutiu, and T. Heskes, "Learning from multiple annotators with Gaussian processes," in *ICANN*. Springer, 2011, pp. 159–164.
- [15] F. Rodrigues, M. Lourenco, B. Ribeiro, and F. Pereira, "Learning supervised topic models for classification and regression from crowds," *IEEE transactions on PAMI*, 2017.
- [16] F. Rodrigues, F. Pereira, and B. Ribeiro, "Sequence labeling with multiple annotators," *Machine learning*, vol. 95, no. 2, pp. 165–181, 2014.
- [17] S. Albarqouni, C. Baur, F. Achilles, V. Belagiannis, S. Demirci, and N. Navab, "Aggnet: deep learning from crowds for mitosis detection of breast cancer histology images," *IEEE transactions on medical imaging*, vol. 35, no. 5, pp. 1313–1321, 2016.
- [18] F. Rodrigues and F. C. Pereira, "Deep learning from crowds," in *Thirty-Sixth AAAI Conference on Artificial Intelligence*, 2018.
- [19] M. Y. Guan, V. Gulshan, A. M. Dai, and G. E. Hinton, "Who said what? Modeling individual labelers improves classification," in *Thirty-Sixth AAAI Conference on Artificial Intelligence*, 2018.
- [20] E. Rodrigo, J. Aledo, and J. 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.
- [21] 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.
- [22] 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.
- [23] 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.
- [24] J. Surowiecki, *The wisdom of crowds*. Anchor, 2005.
- [25] U. Hahn, M. von Sydow, and C. Merdes, "How communication make voters choose less well," *Topics in cognitive science*, 2018.
- [26] T. Zhu, M. A. Pimentel, G. D. Clifford, and D. A. Clifton, "Unsupervised bayesian inference to fuse biosignal sensory estimates for personalised care," *IEEE J. BIOMED. HEALTH*, vol. 23, no. 1, p. 47, 2019.
- [27] A. Saul, J. Hensman, A. Vehtari, and N. Lawrence, "Chained Gaussian processes," in *Artificial Intelligence and Statistics*, 2016, pp. 1431–1440.
- [28] 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.
- [29] 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.
- [30] 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.
- [31] 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.
- [32] 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.
- [33] 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.
- [34] 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.
- [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] 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.
- [37] 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.
- [38] 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.
- [39] C. E. Rasmussen and C. K. Williams, *Gaussian processes for machine learning*. MIT press Cambridge, 2006, vol. 1.
- [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. Álvarez, "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] 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.
- [44] J. Arias, J. Godino, J. Gutiérrez, V. Osuma, and N. Sáenz, "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.
- [45] 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.
- [46] T. Fawcett, "An introduction to ROC analysis," *Pattern recognition letters*, vol. 27, no. 8, pp. 861–874, 2006.
- [47] C. Bishop, *Pattern recognition and machine learning*. springer, 2006.
- [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.
- [49] M. A. Alvarez, L. Rosasco, and N. D. Lawrence, "Kernels for vector-valued functions: A review," *arXiv preprint arXiv:1106.6251*, 2011.
- [50] A. B. Abdessalem, N. Dervilis, D. J. Wagg, and K. Worden, "Automatic kernel selection for gaussian processes regression with approximate bayesian computation and sequential monte carlo," *Frontiers in Built Environment*, vol. 3, p. 52, 2017.

J. Gil-Gonzalez received his undergraduate degree in electronic engineering (2014) from the Universidad Tecnológica de Pereira, Colombia. His M.Sc. in electrical engineering (2016) from the same university. Currently, he is PhD student from the same university. His research interests include probabilistic models for machine learning, learning from crowds, and Bayesian inference.

Juan-José Giraldo Received a degree in Electronics Engineering (B. Eng.) with Honours, from Universidad del Quindío, Colombia in 2009, a master degree in Electrical Engineering (M. Eng.) from Universidad Tecnológica de Pereira, Colombia in 2015. Currently, Mr. Giraldo is a Ph.D student in Comp. Science at the University of Sheffield, UK

A.M. Álvarez-Meza received his undergraduate degree in electronic engineering (2009), his M.Sc. degree in engineering (2011), and his Ph.D. in automatics from the Universidad Nacional de Colombia. He is a Professor in the Department of Electrical, Electronic and Computation Engineering at the Universidad Nacional de Colombia - Manizales. His research interests include machine learning and signal processing.

A. Orozco-Gutierrez received his undergraduate degree in electrical engineering (1985) and his M.Sc. degree in electrical engineering (2004) from Universidad Tecnológica de Pereira, and his Ph.D. in bioengineering (2009) from Universidad Politecnica de Valencia (Spain). He received his undergraduate degree in law (1996) from Universidad Libre de Colombia. He is a Professor in the Department of Electrical Engineering at the Universidad Tecnológica de Pereira. His research interests include bioengineering.

M. A. Álvarez received the BEng degree in electronics engineering from the Universidad Nacional de Colombia (2004), the M.Sc. degree in electrical engineering from the Universidad Tecnológica de Pereira, Colombia (2006), and the PhD degree in computer science from The University of Manchester, UK (2011). Currently, he is a Lecturer of Machine Learning in the Department of Computer Science, University of Sheffield, United Kingdom. His research interests include probabilistic models, kernel methods, and stochastic processes.