Learning Supervised Topic Models for Classification and Regression from Crowds

Filipe Rodrigues, Mariana Lourenço, Bernardete Ribeiro, *Senior Member, IEEE,*
and Francisco C. Pereira, *Member, IEEE*

Abstract—The growing need to analyze large collections of documents has led to great developments in topic modeling. Since documents are frequently associated with other related variables, such as labels or ratings, much interest has been placed on supervised topic models. However, the nature of most annotation tasks, prone to ambiguity and noise, often with high volumes of documents, deem learning under a single-annotator assumption unrealistic or unpractical for most real-world applications. In this article, we propose two supervised topic models, one for classification and another for regression problems, which account for the heterogeneity and biases among different annotators that are encountered in practice when learning from crowds. We develop an efficient stochastic variational inference algorithm that is able to scale to very large datasets, and we empirically demonstrate the advantages of the proposed model over state-of-the-art approaches.

Index Terms—Topic models, crowdsoucing, multiple annotators, supervised learning

✦

1 INTRODUCTION

TOPIC models, such as latent Dirichlet allocation (LDA), allow us to analyze large collections of documents by revealing their underlying themes, or topics, and how each document exhibits them [1]. Therefore, it is not surprising that topic models have become a standard tool in data analysis, with many applications that go even beyond their original purpose of modeling textual data, such as analyzing images [2], [3], videos [4], survey data [5] or social networks data [6].

Since documents are frequently associated with other variables such as labels, tags or ratings, much interest has been placed on supervised topic models [7], which allow the use of that extra information to "guide" the topics discovery. By jointly learning the topics distributions and a classification or regression model, supervised topic models have been shown to outperform the separate use of their unsupervised analogues together with an external regression/classification algorithm [3], [8].

Supervised topics models are then state-of-the-art approaches for predicting target variables associated with complex high-dimensional data, such as documents or images. Unfortunately, the size of modern datasets makes the use of a single annotator unrealistic and unpractical for the majority of the real-world applications that involve some form of human labeling. For instance, the popular Reuters-21578 benchmark corpus was categorized by a group of personnel from Reuters Ltd and Carnegie Group, Inc. Similarly, the LabelMe[1] project asks volunteers to annotate images from a large collection using an online tool. Hence, it is seldom the case where a single oracle labels an entire collection.

Furthermore, the Web, through its social nature, also exploits the wisdom of crowds to annotate large collections of documents and images. By categorizing texts, tagging images or rating products and places, Web users are generating large volumes of labeled content. However, when learning supervised models from crowds, the quality of labels can vary significantly due to task subjectivity and differences in annotator reliability (or bias) [9], [10]. If we consider a sentiment analysis task, it becomes clear that the subjectiveness of the exercise is prone to generate considerably distinct labels from different annotators. Similarly, online product reviews are known to vary considerably depending on the personal biases and volatility of the reviewer's opinions. It is therefore essential to account for these issues when learning from this increasingly common type of data. Hence, the interest of researchers on building models that take the reliabilities of different annotators into consideration and mitigate the effect of their biases has spiked during the last few years (e.g., [11], [12]).

The increasing popularity of crowdsourcing platforms like Amazon Mechanical Turk (AMT) has further contributed to the recent advances in learning from crowds. This kind of platforms offers a fast, scalable and inexpensive solution for labeling large amounts of data. However, their heterogeneous nature in terms of contributors makes their

---

- *F. Rodrigues is with the Technical University of Denmark (DTU), Bygning 115, Kgs. Lyngby 2800, Denmark. E-mail: fmpr@dei.uc.pt.*
- *M. Lourenço and B. Ribeiro are with the CISUC, Department of Informatics Engineering, University of Coimbra, Coimbra 3000-213, Portugal. E-mail: mrlouren@student.dei.uc.pt, bribeiro@dei.uc.pt.*
- *F.C. Pereira is with the Massachusetts Institute of Technology (MIT), 77 Massachusetts Avenue, Cambridge, MA 02139, and Technical University of Denmark (DTU), Bygning 115, Kgs. Lyngby 2800, Denmark. E-mail: camara@transport.dtu.dk.*

1. http://labelme.csail.mit.edu

straightforward application prone to many sorts of labeling noise and bias. Hence, a careless use of crowdsourced data as training data risks generating flawed models.

In this article, we propose a fully generative supervised topic model that is able to account for the different reliabilities of multiple annotators and correct their biases. The proposed model is then capable of jointly modeling the words in documents as arising from a mixture of topics, the latent true target variables as a result of the empirical distribution over topics of the documents, and the labels of the multiple annotators as noisy versions of that latent ground truth. We propose two different models, one for classification [13] and another for regression problems, thus covering a very wide range of possible practical applications, as we empirically demonstrate. Since the majority of the tasks for which multiple annotators are used generally involve complex data such as text, images and video, by developing a multi-annotator supervised topic model we are contributing with a powerful tool for learning predictive models of complex high-dimensional data from crowds.

Given that the increasing sizes of modern datasets can pose a problem for obtaining human labels as well as for Bayesian inference, we propose an efficient stochastic variational inference algorithm [14] that is able to scale to very large datasets. We empirically show, using both simulated and real multiple-annotator labels obtained from AMT for popular text and image collections, that the proposed models are able to outperform other state-of-the-art approaches in both classification and regression tasks. We further show the computational and predictive advantages of the stochastic variational inference algorithm over its batch counterpart.

2 STATE OF THE ART

2.1 Supervised Topic Models

Latent Dirichlet allocation soon proved to be a powerful tool for modeling documents [1] and images [2] by extracting their underlying topics, where topics are probability distributions across words, and each document is characterized by a probability distribution across topics. However, the need to model the relationship between documents and labels quickly gave rise to many supervised variants of LDA. One of the first notable works was that of supervised LDA (sLDA) [7]. By extending LDA through the inclusion of a response variable that is linearly dependent on the mean topic-assignments of the words in a document, sLDA is able to jointly model the documents and their responses, in order to find latent topics that will best predict the response variables for future unlabeled documents. Although initially developed for general continuous response variables, sLDA was later extended to classification problems [3], by modeling the relationship between topic-assignments and labels with a softmax function as in logistic regression.

From a classification perspective, there are several ways in which document classes can be included in LDA. The most natural one in this setting is probably the sLDA approach, since the classes are directly dependent on the empirical topic mixture distributions. This approach is coherent with the generative perspective of LDA but, nevertheless, several discriminative alternatives also exist. For example, DiscLDA [15] introduces a class-dependent linear transformation on the topic mixture proportions of each document, such that the

per-word topic assignments are drawn from linearly transformed mixture proportions. The class-specific transformation matrices are then able to reposition the topic mixture proportions so that documents with the same class labels have similar topics mixture proportions. The transformation matrices can be estimated by maximizing the conditional likelihood of response variables as the authors propose [15].

An alternative way of including classes in LDA for supervision is the one proposed in the Labeled-LDA model [16]. Labeled-LDA is a variant of LDA that incorporates supervision by constraining the topic model to assign to a document only topics that correspond to its label set. While this allows for multiple labels per document, it is restrictive in the sense that the number of topics needs to be the same as the number of possible labels.

From a regression perspective, other than sLDA, the most relevant approaches are the Dirichlet-multimonial regression [17] and the inverse regression topic models [18]. The Dirichlet-multimonial regression (DMR) topic model [17] includes a log-linear prior on the document's mixture proportions that is a function of a set of arbitrary features, such as author, date, publication venue or references in scientific articles. The inferred Dirichlet-multinomial distribution can then be used to make predictions about the values of theses features. The inverse regression topic model (IRTM) [18] is a mixed-membership extension of the multinomial inverse regression (MNIR) model proposed in [19] that exploits the topical structure of text corpora to improve its predictions and facilitate exploratory data analysis. However, this results in a rather complex and inefficient inference procedure. Furthermore, making predictions in the IRTM is not trivial. For example, MAP estimates of targets will be in a different scale than the original document's metadata. Hence, the authors propose the use of a linear model to regress metadata values onto their MAP predictions.

The approaches discussed so far rely on likelihood-based estimation procedures. The work in [8] contrasts with these approaches by proposing MedLDA, a supervised topic model that utilizes the max-margin principle for estimation. Despite its margin-based advantages, MedLDA looses the probabilistic interpretation of the document classes given the topic mixture distributions. On the contrary, in this article we propose a fully generative probabilistic model of the answers of multiple annotators and of the words of documents arising from a mixture of topics.

2.2 Learning from Multiple Annotators

Learning from multiple annotators is an increasingly important research topic. Since the early work of Dawid and Skeene [20], who attempted to obtain point estimates of the error rates of patients given repeated but conflicting responses to various medical questions, many approaches have been proposed. These usually rely on latent variable models. For example, in [21] the authors propose a model to estimate the ground truth from the labels of multiple experts, which is then used to train a classifier.

While earlier works usually focused on estimating the ground truth and the error rates of different annotators, recent works are more focused on the problem of learning classifiers using multiple-annotator data. This idea was explored by Raykar et al. [22], who proposed an approach for jointly

learning the levels of expertise of different annotators and the parameters of a logistic regression classifier, by modeling the ground truth labels as latent variables. This work was later extended in [12] by considering the dependencies of the annotators' labels on the instances they are labeling, and also in [23] through the use of Gaussian process classifiers. The model proposed in this article for classification problems shares the same intuition with this line of work and models the true labels as latent variables. However, it differs significantly by using a fully Bayesian approach for estimating the reliabilities and biases of the different annotators. Furthermore, it considers the problems of learning a low-dimensional representation of the input data (through topic modeling) and modeling the answers of multiple annotators jointly, providing an efficient stochastic variational inference algorithm.

Despite the considerable amount of approaches for learning classifiers from the noisy answers of multiple annotators, for continuous response variables this problem has been approached in a much smaller extent. For example, Groot et al. [24] address this problem in the context of Gaussian processes. In their work, the authors assign a different variance to the likelihood of the data points provided by the different annotators, thereby allowing them to have different noise levels, which can be estimated by maximizing the marginal likelihood of the data. Similarly, the authors in [22] propose an extension of their own classification approach to regression problems by assigning different variances to the Gaussian noise models of the different annotators. In this article, we take this idea one step further by also considering a per-annotator bias parameter, which gives the proposed model the ability to overcome certain personal tendencies in the annotators labeling styles that are quite common, for example, in product ratings and document reviews. Furthermore, we empirically validate the proposed model using real multi-annotator data obtained from Amazon Mechanical Turk. This contrasts with the previously mentioned works, which rely only on simulated annotators.

3 CLASSIFICATION MODEL

In this section, we develop a multi-annotator supervised topic model for classification problems. The model for regression settings will be presented in Section 4. We start by deriving a (*batch*) variational inference algorithm for approximating the posterior distribution over the latent variables and an algorithm to estimate the model parameters. We then develop a stochastic variational inference algorithm that gives the model the capability of handling large collections of documents. Finally, we show how to use the learned model to classify new documents.

3.1 Proposed Model

Let $\mathcal{D} = \{\mathbf{w}^d, \mathbf{y}^d\}_{d=1}^D$ be an annotated corpus of size D, where each document $\mathbf{w}^d$ is given a set of labels $\mathbf{y}^d = \{y_r^d\}_{r=1}^{R_d}$ from R_d distinct annotators. We can take advantage of the inherent topical structure of documents and model their words as arising from a mixture of topics, each being defined as a distribution over the words in a vocabulary, as in LDA. In LDA, the nth word, w_n^d, in a document d is provided a discrete topic-assignment z_n^d, which is drawn from the documents' distribution over topics θ^d. This allows us to build lower-dimensional
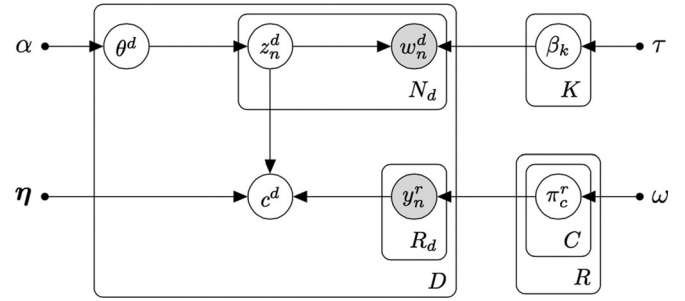

Fig. 1. Graphical representation of the proposed model for classification.

representations of documents, which we can explore to build classification models by assigning coefficients $\boldsymbol{\eta}$ to the mean topic-assignment of the words in the document, $\bar{z}^d$, and applying a softmax function in order to obtain a distribution over classes. Alternatively, one could consider more flexible models such as Gaussian processes, however that would considerably increase the complexity of inference.

Unfortunately, a direct mapping between document classes and the labels provided by the different annotators in a multiple-annotator setting would correspond to assuming that they are all equally reliable, an assumption that is violated in practice, as previous works clearly demonstrate (e.g., [9], [10]). Hence, we assume the existence of a latent ground truth class, and model the labels from the different annotators using a noise model that states that, given a true class c, each annotator r provides the label l with some probability $\pi_{c,l}^r$. Hence, by modeling the matrix $\boldsymbol{\pi}^r$ we are in fact modeling a per-annotator (normalized) confusion matrix, which allows us to account for their different levels of expertise and correct their potential biases.

The generative process of the proposed model for classification problems can then be summarized as follows:

1) For each annotator r
   a) For each class c
      i) Draw reliability parameter $\pi_c^r|\omega \sim \text{Dir}(\omega)$
2) For each topic k
   a) Draw topic distribution $\beta_k|\tau \sim \text{Dir}(\tau)$
3) For each document d
   a) Draw topic proportions $\theta^d|\alpha \sim \text{Dir}(\alpha)$
   b) For the nth word
      i) Draw topic assignment $z_n^d|\theta^d \sim \text{Mult}(\theta^d)$
      ii) Draw word $w_n^d|z_n^d, \boldsymbol{\beta} \sim \text{Mult}(\beta_{z_n^d})$
   c) Draw latent (true) class $c^d|\mathbf{z}^d, \boldsymbol{\eta} \sim \text{Softmax}(\bar{z}^d, \boldsymbol{\eta})$
   d) For each annotator $r \in R_d$
      i) Draw annotator's label $y^{d,r}|c^d, \boldsymbol{\pi}^r \sim \text{Mult}(\pi_{c^d}^r)$

where R_d denotes the set of annotators that labeled the dth document, $\bar{z}^d = \frac{1}{N_d}\sum_{n=1}^{N_d} z_n^d$, and the softmax is given by

$$p(c^d|\bar{z}^d, \boldsymbol{\eta}) = \frac{\exp(\eta_c^T \bar{z}^d)}{\sum_{l=1}^C \exp(\eta_l^T \bar{z}^d)}.$$

Fig. 1 shows a graphical model representation of the proposed model, where K denotes the number of topics, C is the number of classes, R is the total number of annotators and N_d is the number of words in the document d. Shaded nodes are used to distinguish latent variable from the observed ones and small solid circles are used to denote model parameters.

TABLE 1
Correspondence Between Variational Parameters
and the Original Parameters

Original parameter	π_c^r	β_k	θ^d	c^d	z_n^d
Variational parameter	ξ_c^r	ζ_k	γ^d	λ^d	ϕ_n^d

Notice that we included a Dirichlet prior over the topics β_k to produce a smooth posterior and control sparsity. Similarly, instead of computing maximum likelihood or MAP estimates for the annotators reliability parameters π_c^r, we place a Dirichlet prior over these variables and perform approximate Bayesian inference. This contrasts with previous works on learning classification models from crowds [22], [25].

3.2 Approximate Inference

Given a dataset $\mathcal{D}$, the goal of inference is to compute the posterior distribution of the per-document topic proportions θ^d, the per-word topic assignments z_n^d, the per-topic distribution over words β_k, the per-document latent true class c^d, and the per-annotator confusion parameters π^r. As with LDA, computing the exact posterior distribution of the latent variables is computationally intractable. Hence, we employ mean-field variational inference to perform approximate Bayesian inference.

Variational inference methods seek to minimize the KL divergence between the variational and the true posterior distribution. We assume a fully-factorized (mean-field) variational distribution of the form

$$q(\boldsymbol{\theta}, \mathbf{z}_{1:D}, \mathbf{c}, \boldsymbol{\beta}, \boldsymbol{\pi}_{1:R}) = \left(\prod_{r=1}^{R} \prod_{c=1}^{C} q(\pi_c^r|\xi_c^r) \right) \left(\prod_{i=1}^{K} q(\beta_i|\zeta_i) \right)$$
$$\times \prod_{d=1}^{D} q(\theta^d|\gamma^d) \, q(c^d|\lambda^d) \prod_{n=1}^{N_d} q(z_n^d|\phi_n^d),$$

where $\xi_{1:R}$, ζ, γ, λ and $\phi_{1:D}$ are variational parameters. Table 1 shows the correspondence between variational parameters and the original parameters.

Let $\Theta = \{\alpha, \tau, \omega, \eta\}$ denote the model parameters. Following [26], the KL minimization can be equivalently formulated as maximizing the following lower bound on the log marginal likelihood

$$\log p(\mathbf{w}_{1:D}, \mathbf{y}_{1:D}|\Theta)$$
$$= \log \int \sum_{\mathbf{z},\mathbf{c}} q(\boldsymbol{\theta}, \mathbf{z}_{1:D}, \mathbf{c}, \boldsymbol{\beta}, \boldsymbol{\pi}_{1:R})$$
$$\times \frac{p(\boldsymbol{\theta}, \mathbf{z}_{1:D}, \mathbf{c}, \mathbf{w}_{1:D}, \mathbf{y}_{1:D}, \boldsymbol{\beta}, \boldsymbol{\pi}_{1:R}|\Theta)}{q(\boldsymbol{\theta}, \mathbf{z}_{1:D}, \mathbf{c}, \boldsymbol{\beta}, \boldsymbol{\pi}_{1:R})} \, d\boldsymbol{\theta} \, d\boldsymbol{\beta} \, d\boldsymbol{\pi}_{1:R} \quad (1)$$
$$\geq \mathbb{E}_q[\log p(\boldsymbol{\theta}, \mathbf{z}_{1:D}, \mathbf{c}, \mathbf{w}_{1:D}, \mathbf{y}_{1:D}, \boldsymbol{\beta}, \boldsymbol{\pi}_{1:R}|\Theta)]$$
$$- \mathbb{E}_q[\log q(\boldsymbol{\theta}, \mathbf{z}_{1:D}, \mathbf{c}, \boldsymbol{\beta}, \boldsymbol{\pi}_{1:R})]$$
$$= \mathcal{L}(\boldsymbol{\gamma}, \boldsymbol{\phi}_{1:D}, \lambda, \zeta, \xi_{1:R}|\Theta),$$

which we maximize using coordinate ascent.

Optimizing $\mathcal{L}$ w.r.t. γ and ζ gives the same coordinate ascent updates as in LDA [1]

$$\gamma_i^d = \alpha + \sum_{n=1}^{N_d} \phi_{n,i}^d, \quad (2)$$

$$\zeta_{i,j} = \tau + \sum_{d=1}^{D} \sum_{n=1}^{N_d} w_{n,j}^d \phi_{n,i}^d. \quad (3)$$

The variational Dirichlet parameters ξ can be optimized by collecting only the terms in $\mathcal{L}$ that contain ξ

$$\mathcal{L}_{[\xi]} = \sum_{r=1}^{R} \sum_{c=1}^{C} \sum_{l=1}^{C} \mathbb{E}_q[\log \pi_{c,l}^r] \left(\omega + \sum_{d=1}^{D_r} \lambda_c^d y_l^{d,r} - \xi_{c,l}^r \right)$$
$$- \sum_{r=1}^{R} \sum_{c=1}^{C} \log \Gamma \left(\sum_{t=1}^{C} \xi_{c,t}^r \right) + \sum_{r=1}^{R} \sum_{c=1}^{C} \sum_{l=1}^{C} \log \Gamma(\xi_{c,l}^r),$$

where D_r denotes the documents labeled by the rth annotator, $\mathbb{E}_q[\log \pi_{c,l}^r] = \Psi(\xi_{c,l}^r) - \Psi(\sum_{t=1}^{C} \xi_{c,t}^r)$, and $\Gamma(\cdot)$ and $\Psi(\cdot)$ are the gamma and digamma functions, respectively. Taking derivatives of $\mathcal{L}_{[\xi]}$ w.r.t. ξ and setting them to zero, yields the following update

$$\xi_{c,l}^r = \omega + \sum_{d=1}^{D_r} \lambda_c^d y_l^{d,r}. \quad (4)$$

Similarly, the coordinate ascent updates for the documents distribution over classes λ can be found by considering the terms in $\mathcal{L}$ that contain λ

$$\mathcal{L}_{[\lambda]} = \sum_{d=1}^{D} \sum_{l=1}^{C} \lambda_l^d \eta_l^T \bar{\phi}^d - \sum_{l=1}^{C} \lambda_l^d \log \lambda_l^d$$
$$+ \sum_{d=1}^{D} \sum_{r=1}^{R_d} \sum_{l=1}^{C} \sum_{c=1}^{C} \lambda_l^d y_c^{d,r} \mathbb{E}_q[\log \pi_{l,c}^r],$$

where $\bar{\phi}^d = \frac{1}{N_d} \sum_{n=1}^{N_d} \phi_n^d$. Adding the necessary Lagrange multipliers to ensure that $\sum_{l=1}^{C} \lambda_l^d = 1$ and setting the derivatives w.r.t. λ_l^d to zero gives the following update

$$\lambda_l^d \propto \exp \left(\eta_l^T \bar{\phi}^d + \sum_{r=1}^{R_d} \sum_{c=1}^{C} y_c^{d,r} \mathbb{E}_q[\log \pi_{l,c}^r] \right). \quad (5)$$

Observe how the variational distribution over the true classes results from a combination between the dot product of the inferred mean topic assignment $\bar{\phi}^d$ with the coefficients η and the labels $\mathbf{y}$ from the multiple annotators "weighted" by their expected log probability $\mathbb{E}_q[\log \pi_{l,c}^r]$.

The main difficulty of applying standard variational inference methods to the proposed model is the non-conjugacy between the distribution of the mean topic-assignment $\bar{z}^d$ and the softmax. Namely, in the expectation

$$\mathbb{E}_q[\log p(c^d|\bar{z}^d, \eta)] = \mathbb{E}_q[\eta_{c^d}^T \bar{z}^d] - \mathbb{E}_q \left[\log \sum_{l=1}^{C} \exp(\eta_l^T \bar{z}^d) \right],$$

the second term is intractable to compute. We can make progress by applying Jensen's inequality to bound it as follows

$$-\mathbb{E}_q \left[\log \sum_{l=1}^{C} \exp(\eta_l^T \bar{z}^d) \right] \geq - \log \sum_{l=1}^{C} \mathbb{E}_q[\exp(\eta_l^T \bar{z}^d)]$$
$$= -\log \sum_{l=1}^{C} \prod_{j=1}^{N_d} (\phi_j^d)^T \exp \left(\eta_l \frac{1}{N_d} \right)$$
$$= -\log (a^T \phi_n^d),$$

where $a \triangleq \sum_{l=1}^{C} \exp(\frac{\eta_l}{N_d}) \prod_{j=1, j \neq n}^{N_d} (\phi_j^d)^T \exp(\frac{\eta_l}{N_d})$, which is constant w.r.t. ϕ_n^d. This local variational bound can be made tight by noticing that $\log(x) \leqslant \epsilon^{-1}x + \log(\epsilon) - 1, \forall x > 0$, $\epsilon > 0$, where equality holds if and only if $x = \epsilon$. Hence, given the current parameter estimates $(\phi_n^d)^{old}$, if we set $x = a^T \phi_n^d$ and $\epsilon = a^T (\phi_n^d)^{old}$ then, for an individual parameter ϕ_n^d, we have that

$$-\mathbb{E}_q \left[\log \sum_{l=1}^{C} \exp(\eta_l^T \bar{z}^d) \right]$$
$$\geqslant -(a^T (\phi_n^d)^{old})^{-1}(a^T \phi_n^d) - \log(a^T (\phi_n^d)^{old}) + 1.$$

Using this local bound to approximate the expectation of the log-sum-exp term, and taking derivatives of the evidence lower bound w.r.t. ϕ_n with the constraint that $\sum_{i=1}^{K} \phi_{n,i}^d = 1$, yields the following fix-point update

$$\phi_{n,i}^d \propto \exp\left(\Psi(\gamma_i) + \sum_{j=1}^{V} w_{n,j}^d \left(\Psi(\zeta_{i,j}) - \Psi\left(\sum_{k=1}^{V} \zeta_{i,k} \right) \right) \right.$$
$$\left. + \frac{\sum_{l=1}^{C} \lambda_l^d \eta_{l,i}}{N_d} - (a^T (\phi_n^d)^{old})^{-1} a_i \right). \quad (6)$$

where V denotes the size of the vocabulary. Notice how the per-word variational distribution over topics ϕ depends on the variational distribution over the true class label λ.

The variational inference algorithm iterates between Eqs. (2), (3), (4), (5), and (6) until the evidence lower bound, Eq. (1), converges. Additional details are provided as supplementary material, which can be found on the Computer Society Digital Library at http://doi.ieeecomputersociety.org/10.1109/TPAMI.2017.2648786.[2]

3.3 Parameter Estimation

The model parameters are $\Theta = \{\alpha, \tau, \omega, \eta\}$. The parameters $\{\alpha, \tau, \omega\}$ of the Dirichlet priors can be regarded as hyperparameters of the proposed model. As with many works on topic models (e.g., [3], [27]), we assume hyper-parameters to be fixed, since they can be effectively selected by grid-search procedures which are able to explore well the parameter space without suffering from local optima. Our focus is then on estimating the coefficients η using a variational EM algorithm. Therefore, in the E-step we use the variational inference algorithm from Section 3.2 to estimate the posterior distribution of the latent variables, and in the M-step we find maximum likelihood estimates of η by maximizing the evidence lower bound $\mathcal{L}$. Unfortunately, taking derivatives of $\mathcal{L}$ w.r.t. η does not yield a closed-form solution. Hence, we use a numerical method, namely L-BFGS [28], to find an optimum. The objective function and gradients are given by

$$\mathcal{L}_{[\eta]} = \sum_{d=1}^{D} \left(\sum_{l=1}^{C} \lambda_l^d \eta_l^T \bar{\phi}^d - \log \sum_{l=1}^{C} b_l^d \right)$$

$$\nabla_{\eta_{l,i}} = \sum_{d=1}^{D} \left(\lambda_{l,i}^d \bar{\phi}_i^d - \frac{b_l^d}{\sum_{t=1}^{C} b_t^d} \sum_{n=1}^{N_d} \frac{\frac{1}{N_d} \phi_{n,i}^d \exp(\frac{1}{N_d} \eta_{l,i})}{\sum_{j=1}^{K} \phi_{n,j}^d \exp(\frac{1}{N_d} \eta_{l,j})} \right),$$

2. Supplementary material available at: http://www.fprodrigues.com/maslda-supp-mat.pdf

where, for convenience, we defined the following variable: $b_l^d \triangleq \prod_{n=1}^{N_d} (\sum_{i=1}^{K} \phi_{n,i}^d \exp(\frac{1}{N_d} \eta_{l,i}))$.

3.4 Stochastic Variational Inference

In Section 3.2, we proposed a batch coordinate ascent algorithm for doing variational inference in the proposed model. This algorithm iterates between analyzing every document in the corpus to infer the local hidden structure, and estimating the global hidden variables. However, this can be inefficient for large datasets, since it requires a full pass through the data at each iteration before updating the global variables. In this section, we develop a stochastic variational inference algorithm [14], which follows noisy estimates of the gradients of the evidence lower bound $\mathcal{L}$.

Based on the theory of stochastic optimization [29], we can find unbiased estimates of the gradients by subsampling a document (or a mini-batch of documents) from the corpus, and using it to compute the gradients as if that document was observed D times. Hence, given an uniformly sampled document d, we use the current posterior distributions of the global latent variables, β and $\pi_{1:R}$, and the current coefficient estimates η, to compute the posterior distribution over the local hidden variables θ^d, $\mathbf{z}^d$ and c^d using Eqs. (2), (6), and (5) respectively. These posteriors are then used to update the global variational parameters, ζ and $\xi_{1:R}$ by taking a step of size ρ_t in the direction of the noisy estimates of the natural gradients.

Algorithm 1 describes a stochastic variational inference algorithm for the proposed model. Given an appropriate schedule for the learning rates $\{\rho_t\}$, such that $\sum_t \rho_t$ and $\sum_t \rho_t^2 < \infty$, the stochastic optimization algorithm is guaranteed to converge to a local maximum of the evidence lower bound [29].

Algorithm 1. Stochastic Variational Inference for the Proposed Classification Model

---

1: Initialize $\gamma^{(0)}, \phi_{1:D}^{(0)}, \lambda^{(0)}, \zeta^{(0)}, \xi_{1:R}^{(0)}, t = 0$
2: **repeat**
3:     Set t = t + 1
4:     Sample a document $\mathbf{w}^d$ uniformly from the corpus
5:     **repeat**
6:         Compute ϕ_n^d using Eq. (6), for $n \in \{1..N_d\}$
7:         Compute γ^d using Eq. (2)
8:         Compute λ^d using Eq. (5)
9:     **until** local parameters ϕ_n^d, γ^d and λ^d converge
10:     Compute step-size $\rho_t = (t + delay)^{-\kappa}$
11:     Update topics variational parameters

$$\zeta_{i,j}^{(t)} = (1 - \rho_t) \zeta_{i,j}^{(t-1)} + \rho_t \left(\tau + D \sum_{n=1}^{N_d} w_{n,j}^d \phi_{n,i}^d \right)$$

12:     Update annotators confusion parameters

$$\xi_{c,l}^{r\,(t)} = (1 - \rho_t) \xi_{c,l}^{r\,(t-1)} + \rho_t (\omega + D \lambda_c^d y_l^{d,r})$$

13: **until** global convergence criterion is met

3.5 Document Classification

In order to make predictions for a new (unlabeled) document d, we start by computing the approximate posterior distribution over the latent variables θ^d and $\mathbf{z}^d$. This can be
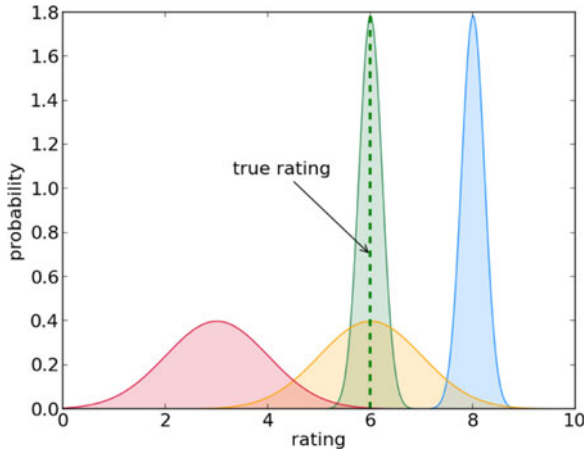
Fig. 2. Example of four different annotators (represented by different colours) with different biases and precisions.

Fig. 3. Graphical representation of the proposed model for regression.

achieved by dropping the terms that involve y, c and π from the model's joint distribution (since, at prediction time, the multi-annotator labels are no longer observed) and averaging over the estimated topics distributions. Letting the topics distribution over words inferred during training be $q(\boldsymbol{\beta}|\boldsymbol{\zeta})$, the joint distribution for a single document is now simply given by

$$p(\theta^d, \mathbf{z}^d) = \int q(\boldsymbol{\beta}|\boldsymbol{\zeta})\, p(\theta^d|\boldsymbol{\alpha}) \prod_{n=1}^{N_d} p(z_n^d|\theta^d)\, p(w_n^d|z_n^d, \boldsymbol{\beta})\, d\boldsymbol{\beta}.$$

Deriving a mean-field variational inference algorithm for computing the posterior over $q(\theta^d, \mathbf{z}^d) = q(\theta^d|\gamma^d) \prod_{n=1}^{N_d} q(z_n^d|\phi_n^d)$ results in the same fixed-point updates as in LDA [1] for γ_i^d (Eq. (2)) and $\phi_{n,i}^d$

$$\phi_{n,i}^d \propto \exp\left(\Psi(\gamma_i) + \sum_{j=1}^{V} w_{n,j}^d \left(\Psi(\zeta_{i,j}) - \Psi\left(\sum_{k=1}^{V} \zeta_{i,k} \right) \right) \right). \quad (7)$$

Using the inferred posteriors and the coefficients $\boldsymbol{\eta}$ estimated during training, we can make predictions as follows:

$$c_*^d = \arg\max_c \eta_c^T \bar{\phi}^d. \quad (8)$$

This is equivalent to making predictions in the classification version of sLDA [3].

4 REGRESSION MODEL

In this section, we develop a variant of the model proposed in Section 3 for regression problems. We shall start by describing the proposed model with a special focus on the how to handle multiple annotators with different biases and reliabilities when the target variables are continuous variables. Next, we present a variational inference algorithm, highlighting the differences to the classification version. Finally, we show how to optimize the model parameters.

4.1 Proposed Model

For developing a multi-annotator supervised topic model for regression, we shall follow a similar intuition as the one we considered for classification. Namely, we shall assume that, for a given document d, each annotator provides a noisy version, $y^{d,r} \in \mathbb{R}$, of the true (continuous) target variable, which
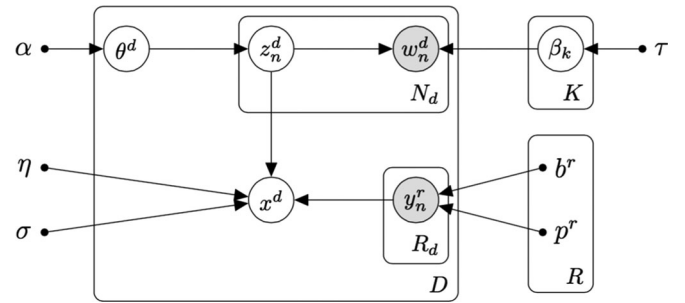
we denote by $x^d \in \mathbb{R}$. This can be, for example, the true rating of a product or the true sentiment of a document. Assuming that each annotator r has its own personal bias b^r and precision p^r (inverse variance), and assuming a Gaussian noise model for the annotators' answers, we have that

$$y^{d,r} \sim \mathcal{N}(y^{d,r}|x^d + b^r, 1/p^r). \quad (9)$$

This approach is therefore more powerful than previous works [22], [24], where a single precision parameter was used to model the annotators' expertise. Fig. 2 illustrates this intuition for four annotators, represented by different colors. The "green annotator" is the best one, since he is right on the target and his answers vary very little (low bias, high precision). The "yellow annotator" has a low bias, but his answers are very uncertain, as they can vary a lot. Contrarily, the "blue annotator" is very precise, but consistently over-estimates the true target (high bias, high precision). Finally, the "red annotator" corresponds to the worst kind of annotator (with high bias and low precision).

Having specified a model for annotators answers given the true targets, the only thing left is to do is to specify a model of the latent true targets x^d given the empirical topic mixture distributions $\bar{z}^d$. For this, we shall keep things simple and assume a linear model as in sLDA [7]. The generative process of the proposed model for continuous target variables can then be summarized as follows:

1) For each annotator r
   a) For each class c
      i) Draw reliability parameter $\pi_c^r|\omega \sim \mathrm{Dir}(\omega)$
2) For each topic k
   a) Draw topic distribution $\beta_k|\tau \sim \mathrm{Dir}(\tau)$
3) For each document d
   a) Draw topic proportions $\theta^d|\alpha \sim \mathrm{Dir}(\alpha)$
   b) For the nth word
      i) Draw topic assignment $z_n^d|\theta^d \sim \mathrm{Mult}(\theta^d)$
      ii) Draw word $w_n^d|z_n^d, \boldsymbol{\beta} \sim \mathrm{Mult}(\beta_{z_n^d})$
   c) Draw latent (true) target $x^d|\mathbf{z}^d, \eta, \sigma \sim \mathcal{N}(\eta^T \bar{z}^d, \sigma^2)$
   d) For each annotator $r \in R_d$
      i) Draw answer $y^{d,r}|x^d, b^r, p^r \sim \mathcal{N}(x^d + b^r, 1/p^r)$

Fig. 3 shows a graphical representation of the proposed model.

4.2 Approximate Inference

The goal of inference is to compute the posterior distribution of the per-document topic proportions θ^d, the per-word topic assignments z_n^d, the per-topic distribution over words β_k and the per-document latent true targets x^d. As we did for the classification model, we shall

develop a variational inference algorithm using coordinate ascent. The lower-bound on the log marginal likelihood is now given by

$$\mathcal{L}(\boldsymbol{\gamma}, \boldsymbol{\phi}_{1:D}, \mathbf{m}, \mathbf{v}, \boldsymbol{\zeta}|\Theta) = \mathbb{E}_q[\log p(\boldsymbol{\theta}, \mathbf{z}_{1:D}, \boldsymbol{x}, \mathbf{w}_{1:D}, \mathbf{y}_{1:D}, \boldsymbol{\beta}|\Theta)]$$
$$- \mathbb{E}_q[\log q(\boldsymbol{\theta}, \mathbf{z}_{1:D}, \boldsymbol{x}, \boldsymbol{\beta})], \quad (10)$$

where $\Theta = \{\alpha, \tau, \eta, \sigma, \mathbf{b}, \mathbf{p}\}$ are the model parameters. We assume a fully-factorized (mean-field) variational distribution q of the form

$$q(\boldsymbol{\theta}, \mathbf{z}_{1:D}, \boldsymbol{c}, \boldsymbol{\beta}) = \left(\prod_{i=1}^{K} q(\beta_i|\zeta_i) \right) \prod_{d=1}^{D} q(x^d|m^d, v^d) \, q(\theta^d|\gamma^d)$$
$$\times \prod_{n=1}^{N_d} q(z_n^d|\phi_n^d). \quad (11)$$

where $\boldsymbol{\zeta}$, $\mathbf{m}$, $\mathbf{v}$, $\boldsymbol{\gamma}$ and $\boldsymbol{\phi}_{1:D}$ are the variational parameters. Notice the new Gaussian term, $q(x^d|m^d, v^d)$, corresponding to the approximate posterior distribution of the unobserved true targets.

Optimizing the variational objective $\mathcal{L}$ w.r.t. $\boldsymbol{\gamma}$ and $\boldsymbol{\zeta}$ yields the same updates from Eqs. (2) and (3). Optimizing w.r.t. $\boldsymbol{\phi}$ gives a similar update to the one in sLDA [7]

$$\phi_{n,i}^d \propto \exp\left(\Psi(\gamma_i) + \sum_{j=1}^{V} w_{n,j}^d \left(\Psi(\zeta_{i,j}) - \Psi\left(\sum_{k=1}^{V} \zeta_{i,k} \right) \right) \right.$$
$$\left. + \frac{m^d}{N^d \sigma^2} \eta - \frac{2(\eta^T \phi_{-n}^d)\eta + (\eta \circ \eta)}{2(N^d)^2 \sigma^2} \right), \quad (12)$$

where we defined $\phi_{-n}^d \triangleq \sum_{m \neq n} \phi_m^d$. Notice how this update differs only from the one in [7] by replacing the true target variable by its expected value under the variational distribution, which is given by $\mathbb{E}_q[x^d] = m^d$.

The only variables left for doing inference on are then the latent true targets $\mathbf{x} = \{x^d\}_{d=1}^{D}$. The variational distribution of x^d is governed by two parameters: a mean m^d and a variance v^d. Collecting all the terms in $\mathcal{L}$ that contain m gives

$$\mathcal{L}_{[m]} = -\sum_{d=1}^{D} \sum_{r=1}^{R_d} \frac{p^r}{2} \left((m^d)^2 + 2m^d b^r - 2y^{d,r} m^d \right)$$
$$- \sum_{d=1}^{D} \frac{1}{2\sigma^2} \left((m^d)^2 - 2m^d(\eta^T \bar{\phi}^d) \right). \quad (13)$$

Taking derivatives of $\mathcal{L}_{[m]}$ and setting them to zero gives the following update for m^d

$$m^d = \frac{\sigma^{-2}(\eta^T \bar{\phi}^d) + \sum_{r=1}^{R_d} p^r(y^{d,r} - b^r)}{\sigma^{-2} + \sum_{r=1}^{R} p^r}. \quad (14)$$

Notice how the value of m^d is a weighted average of what the linear regression model on the empirical topic mixture believes the true target should be, and the bias-corrected answers of the different annotators weighted by their individual precisions.

As for m, we can optimize $\mathcal{L}$ w.r.t. v by collecting all terms that contain v

$$\mathcal{L}_{[v]} = \sum_{d=1}^{D} \left(\frac{1}{2} \log(v^d) - \sum_{r=1}^{R_d} \frac{p^r v^d}{2} - \frac{v^d}{2\sigma^2} \right), \quad (15)$$

and taking derivatives, yielding the update

$$v^d = \sigma^2 + \sum_{r=1}^{R_d} \frac{1}{p^r}. \quad (16)$$

4.3 Parameter Estimation

The parameters of the proposed regression model are $\Theta = \{\alpha, \tau, \eta, \sigma, \mathbf{b}, \mathbf{p}\}$. As we did for the classification model, we shall assume the Dirichlet parameters, α and τ, to be fixed. Similarly, we shall assume that the variance of the true targets, σ^2, to be constant. The only parameters left to estimate are then the regression coefficients η and the annotators biases, $\mathbf{b} = \{b^r\}_{r=1}^{R}$, and precisions, $\mathbf{p} = \{p^r\}_{r=1}^{R}$, which we estimate using variational Bayesian EM.

Since the latent true targets are now linear functions of the documents' empirical topic mixtures (i.e., there is no softmax function), we can find a closed form solution for the regression coefficients η. Taking derivatives of $\mathcal{L}$ w.r.t. η and setting them to zero, gives the following solution for η

$$\eta^T = \sum_{d=1}^{D} \mathbb{E}_q[\bar{z}^d(\bar{z}^d)^T]^{-1} (\bar{\phi}^d)^T m^d, \quad (17)$$

where

$$\mathbb{E}_q[\bar{z}^d(\bar{z}^d)^T] = \frac{1}{(N^d)^2} \left(\sum_{n=1}^{N^d} \sum_{m \neq n}^{N^d} \phi_n^d(\phi_m^d)^T + \sum_{n=1}^{N^d} diag(\phi_n^d) \right).$$

We can find maximum likelihood estimates for the annotator biases b^r by optimizing the lower bound on the marginal likelihood. The terms in $\mathcal{L}$ that involve b are

$$\mathcal{L}_{[b]} = \sum_{d=1}^{D} \sum_{r=1}^{R_d} \frac{p^r}{2} \left(2y^{d,r} b^r - 2m^d b^r - (b^r)^2 \right). \quad (18)$$

Taking derivatives w.r.t. b^r gives the following estimate for the bias of the rth annotator

$$b^r = \frac{1}{D_r} \sum_{d=1}^{D_r} \left(y^{d,r} - m^d \right). \quad (19)$$

Similarly, we can find maximum likelihood estimates for the precisions p^r of the different annotators by considering the terms in $\mathcal{L}$ that contain p

$$\mathcal{L}_{[p]} = \sum_{d=1}^{D} \sum_{r=1}^{R_d} \left(\frac{1}{2} \log(p^r) - \frac{p^r v^d}{2} - \frac{p^r}{2}(y^{d,r} - m^d - b^r)^2 \right). \quad (20)$$

The maximum likelihood estimate for the precision (inverse variance) of the rth annotator is then given by

$$p^r = \left(\frac{1}{D_r} \sum_{d=1}^{D_r} \left(v^d + (y^{d,r} - m^d - b^r)^2 \right) \right)^{-1}. \quad (21)$$

Given a set of fitted parameters, it is then straightforward to make predictions for new documents: it is just necessary

to infer the (approximate) posterior distribution over the word-topic assignments z_n^d for all the words using the coordinates ascent updates of standard LDA (Eqs. (2) and (7)), and then use the mean topic assignments $\bar{\phi}^d$ to make predictions $x_*^d = \eta^T \bar{\phi}^d$.

4.4 Stochastic Variational Inference

As we did for the classification model from Section 3, we can envision developing a stochastic variational inference for the proposed regression model. In this case, the only "global" latent variables are the per-topic distributions over words β_k. As for the "local" latent variables, instead of a single variable λ^d, we now have two variables per-document: m^d and v^d. The stochastic variational inference can then be summarized as shown in Algorithm 2. For added efficiency, one can also perform stochastic updates of the annotators biases b^r and precisions p^r, by taking a step in the direction of the gradient of the noisy evidence lower bound scaled by the step-size ρ_t.

Algorithm 2. Stochastic Variational Inference for the Proposed Regression Model

---

1: Initialize $\boldsymbol{\gamma}^{(0)}, \boldsymbol{\phi}_{1:D}^{(0)}, \mathbf{m}^{(0)}, \mathbf{v}^{(0)}, \boldsymbol{\zeta}^{(0)}, \boldsymbol{\xi}_{1:R}^{(0)}, t = 0$
2: **repeat**
3:   Set t = t + 1
4:   Sample a document $\mathbf{w}^d$ uniformly from the corpus
5:   **repeat**
6:     Compute ϕ_n^d using Eq. (12), for $n \in \{1..N_d\}$
7:     Compute γ^d using Eq. (2)
8:     Compute m^d using Eq. (14)
9:     Compute v^d using Eq. (16)
10:   **until** local parameters ϕ_n^d, γ^d and λ^d converge
11:   Compute step-size $\rho_t = (t + delay)^{-\kappa}$
12:   Update topics variational parameters

$$\zeta_{i,j}^{(t)} = (1 - \rho_t)\zeta_{i,j}^{(t-1)} + \rho_t \left(\tau + D \sum_{n=1}^{N_d} w_{n,j}^d \phi_{n,i}^d \right)$$

13: **until** global convergence criterion is met

---

5 EXPERIMENTS

In this section, the proposed multi-annotator supervised LDA models for classification and regression (MA-sLDAc and MA-sLDAr, respectively) are validated using both simulated annotators on popular corpora and using real multiple-annotator labels obtained from Amazon Mechanical Turk.[3] Namely, we shall consider the following real-world problems: classifying posts and news stories; classifying images according to their content; predicting number of stars that a given user gave to a restaurant based on the review; predicting movie ratings using the text of the reviews.

5.1 Classification

5.1.1 Simulated Annotators

In order to first validate the proposed model for classification problems in a slightly more controlled environment, the well-known 20-Newsgroups benchmark corpus [30] was used by

---

3. Source code and datasets used are available at: http://www. fprodrigues.com/software/

simulating multiple annotators with different levels of expertise. The 20-Newsgroups consists of twenty thousand messages taken from twenty newsgroups, and is divided in six super-classes, which are, in turn, partitioned in several sub-classes. For this first set of experiments, only the four most populated super-classes were used: "computers", "science", "politics" and "recreative". The preprocessing of the documents consisted of stemming and stop-words removal. After that, 75 percent of the documents were randomly selected for training and the remaining 25 percent for testing.

The different annotators were simulated by sampling their answers from a multinomial distribution, where the parameters are given by the lines of the annotators' confusion matrices. Hence, for each annotator r, we start by pre-defining a confusion matrix $\boldsymbol{\pi}^r$ with elements $\pi_{c,l}^r$, which correspond to the probability that the annotators' answer is l given that the true label is c, $p(y^r = l|c)$. Then, the answers are sampled i.i.d. from $y^r \sim \text{Mult}(y^r|\pi_{c,l}^r)$. This procedure was used to simulate five different annotators with the following accuracies: 0.737, 0.468, 0.284, 0.278, 0.260. In this experiment, no repeated labelling was used. Hence, each annotator only labels roughly one-fifth of the data. When compared to the ground truth, the simulated answers revealed an accuracy of 0.405. See Table 2 for an overview of the details of the classification datasets used.

Both the *batch* and the stochastic variational inference (*svi*) versions of the proposed model (MA-sLDAc) are compared with the following baselines:

- *LDA + LogReg (mv)*: This baseline corresponds to applying unsupervised LDA to the data, and learning a logistic regression classifier on the inferred topics distributions of the documents. The labels from the different annotators were aggregated using majority voting (mv). Notice that, when there is a single annotator label per instance, majority voting is equivalent to using that label for training. This is the case of the 20-Newsgroups' simulated annotators, but the same does not apply for the experiments in Section 5.1.2.

- *LDA + Raykar*: For this baseline, the model of [22] was applied using the documents' topic distributions inferred by LDA as features.

- *LDA + Rodrigues*: This baseline is similar to the previous one, but uses the model of [10] instead.

- *Blei 2003 (mv)*: The idea of this baseline is to replicate a popular state-of-the-art approach for document classification. Hence, the approach of [1] was used. It consists of applying LDA to extract the documents' topics distributions, which are then used to train a SVM. Similarly to the previous approach, the labels from the different annotators were aggregated using majority voting (mv).

- *sLDA (mv)*: This corresponds to using the classification version of sLDA [3] with the labels obtained by performing majority voting (mv) on the annotators' answers.

For all the experiments the hyper-parameters α, τ and ω were set using a simple grid search in the collection $\{0.01, 0.1, 1.0, 10.0\}$. The same approach was used to optimize the hyper-parameters of the all the baselines. For the

TABLE 2
Overall Statistics of the Classification Datasets Used in the Experiments

Dataset	Num. classes	Train/test sizes	Annotators source	Num. answers per instance ($\pm$ stddev.)	Mean annotators accuracy ($\pm$ stddev.)	Maj. vot. accuracy
20 Newsgroups	4	11,536/3,846	Simulated	1.000 ± 0.000	0.405 ± 0.182	0.405
Reuters-21578	8	1,800/5,216	Mech. Turk	3.007 ± 1.019	0.568 ± 0.262	0.710
LabelMe	8	1,000/1,688	Mech. Turk	2.547 ± 0.576	0.692 ± 0.181	0.769

svi algorithm, different mini-batch sizes and forgetting rates κ were tested. For the 20-Newsgroup dataset, the best results were obtained with a mini-batch size of 500 and $\kappa = 0.6$. The *delay* was kept at 1. The results are shown in Fig. 4 for different numbers of topics, where we can see that the proposed model outperforms all the baselines, being the *svi* version the one that performs best.

In order to assess the computational advantages of the stochastic variational inference (*svi*) over the *batch* algorithm, the log marginal likelihood (or log evidence) was plotted against the number of iterations. Fig. 5 shows this comparison. Not surprisingly, the *svi* version converges much faster to higher values of the log marginal likelihood when compared to the *batch* version, which reflects the efficiency of the *svi* algorithm.

5.1.2 Amazon Mechanical Turk

In order to validate the proposed classification model in real crowdsourcing settings, Amazon Mechanical Turk was used to obtain labels from multiple annotators for two popular datasets: Reuters-21578 [31] and LabelMe [32].

The Reuters-21578 is a collection of manually categorized newswire stories with labels such as Acquisitions, Crude-oil, Earnings or Grain. For this experiment, only the documents belonging to the ModApte split were considered with the additional constraint that the documents should have no more than one label. This resulted in a total of 7,016
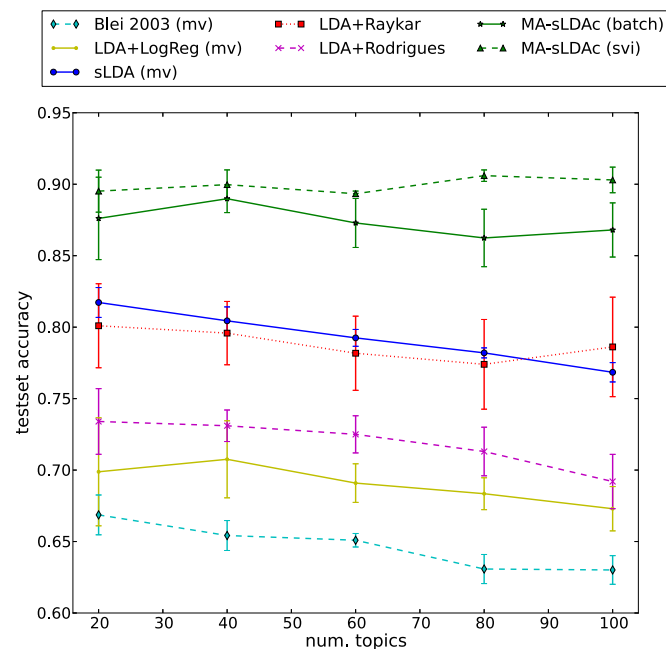
documents distributed among 8 classes. Of these, 1,800 documents were submitted to AMT for multiple annotators to label, giving an average of approximately three answers per document (see Table 2 for further details). The remaining 5,216 documents were used for testing. The collected answers yield an average worker accuracy of 56.8 percent. Applying majority voting to these answers reveals a ground truth accuracy of 71.0 percent. Fig. 6 shows the boxplots of the number of answers per worker and their accuracies. Observe how applying majority voting yields a higher accuracy than the median accuracy of the workers.

The results obtained by the different approaches are given in Fig. 7, where it can be seen that the proposed model (MA-sLDAc) outperforms all the other approaches. For this dataset, the *svi* algorithm is using mini-batches of 300 documents.
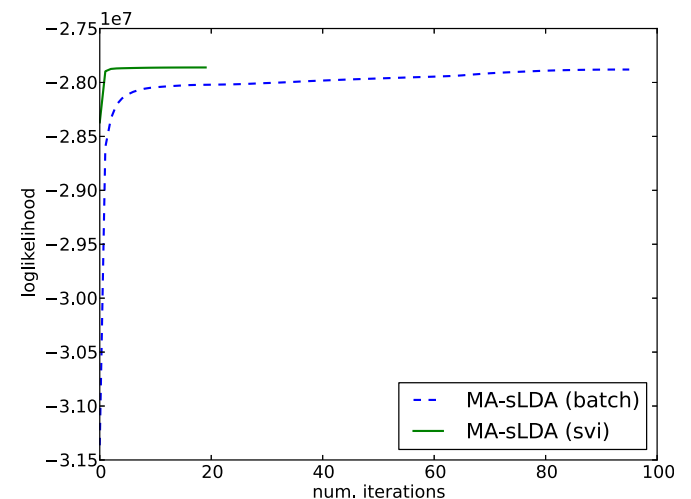

Fig. 5. Comparison of the log marginal likelihood between the *batch* and the stochastic variational inference (*svi*) algorithms on the 20-newsgroups corpus.

Fig. 4. Average testset accuracy (over five runs; $\pm$ stddev.) of the different approaches on the 20-newsgroups data.
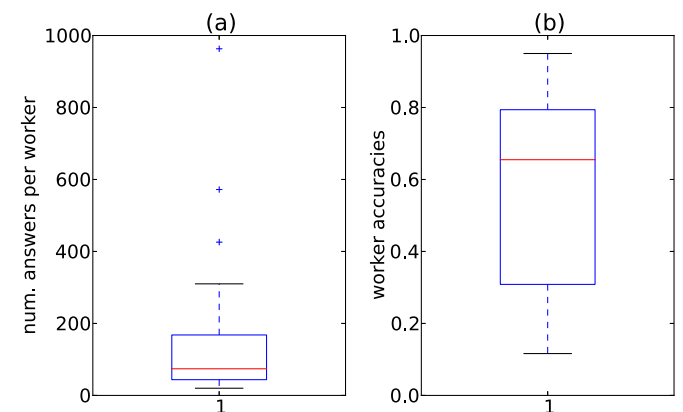

Fig. 6. Boxplot of the number of answers per worker (a) and their respective accuracies (b) for the reuters dataset.

Fig. 7. Average testset accuracy (over 30 runs; $\pm$ stddev.) of the different approaches on the reuters data.
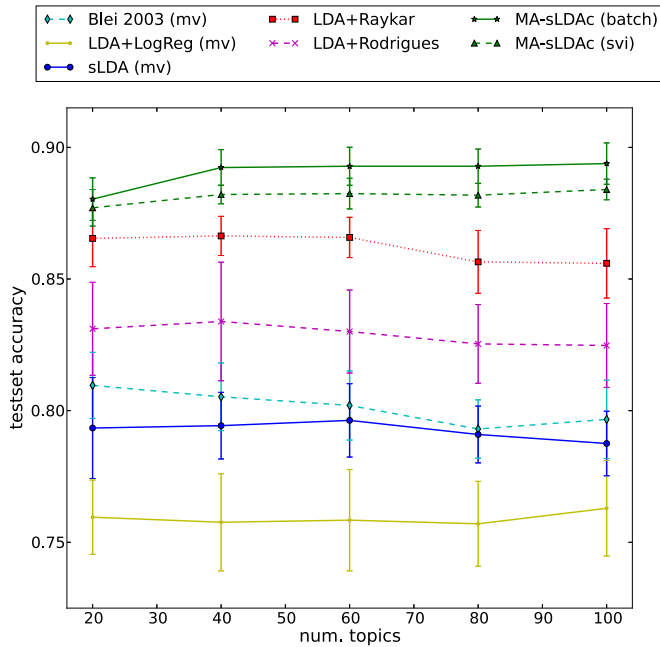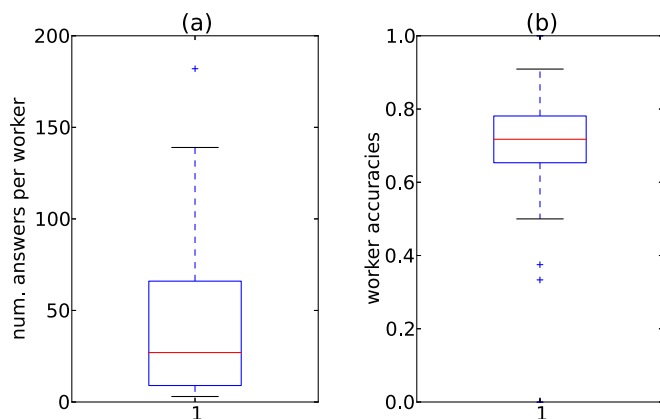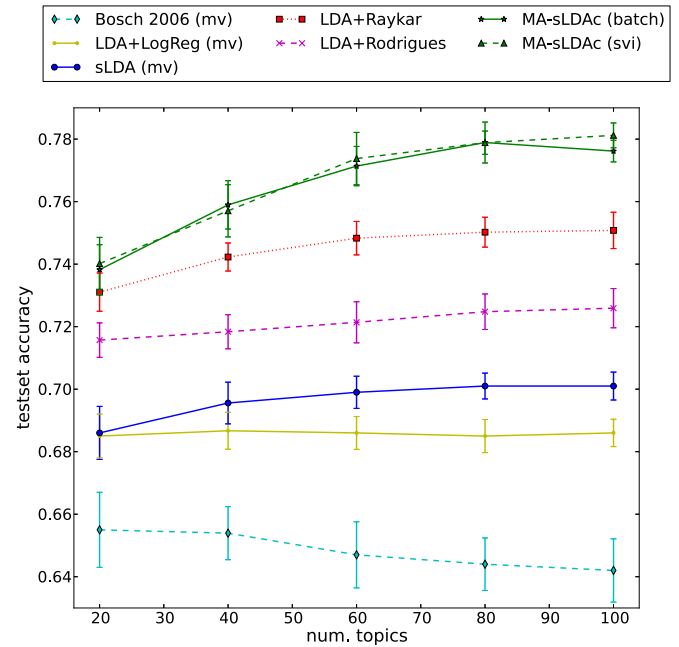

Fig. 9. Average testset accuracy (over 30 runs; $\pm$ stddev.) of the different approaches on the LabelMe data.

The proposed model was also validated using a dataset from the computer vision domain: LabelMe [32]. In contrast to the Reuters and Newsgroups corpora, LabelMe is an open online tool to annotate images. Hence, this experiment allows us to see how the proposed model generalizes beyond non-textual data. Using the Matlab interface provided in the projects' website, we extracted a subset of the LabelMe data, consisting of all the 256×256 images with the categories: "highway", "inside city", "tall building", "street", "forest", "coast", "mountain" or "open country". This allowed us to collect a total of 2,688 labeled images. Of these, 1,000 images were given to AMT workers to classify with one of the classes above. Each image was labeled by an average of 2.547 workers, with a mean accuracy of 69.2 percent. When majority voting is applied to the collected answers, a ground truth accuracy of 76.9 percent is obtained. Fig. 8 shows the boxplots of the number of answers per worker and their accuracies. Interestingly, the worker accuracies are much higher and their distribution is much more concentrated than on the Reuters-21578 data

(see Fig. 6), which suggests that this is an easier task for the AMT workers.

The preprocessing of the images used is similar to the approach in [2]. It uses 128-dimensional SIFT [33] region descriptors selected by a sliding grid spaced at one pixel. This sliding grid extracts local regions of the image with sizes uniformly sampled between 16×16 and 32×32 pixels. The 128-dimensional SIFT descriptors produced by the sliding window are then fed to a k-means algorithm (with k=200) in order construct a vocabulary of 200 "visual words". This allows us to represent the images with a bag of visual words model.

With the purpose of comparing the proposed model with a popular state-of-the-art approach for image classification, for the LabelMe dataset, the following baseline was introduced:

- *Bosch 2006 (mv)*: This baseline is similar to one in [34]. The authors propose the use of pLSA to extract the latent topics, and the use of k-nearest neighbor (kNN) classifier using the documents' topics distributions. For this baseline, unsupervised LDA is used instead of pLSA, and the labels from the different annotators for kNN (with $k = 10$) are aggregated using majority voting (mv).

The results obtained by the different approaches for the LabelMe data are shown in Fig. 9, where the *svi* version is using mini-batches of 200 documents.

Analyzing the results for the Reuters-21578 and LabelMe data, we can observe that MA-sLDAc outperforms all the baselines, with slightly better accuracies for the *batch* version, especially in the Reuters data. Interestingly, the second best results are consistently obtained by the multi-annotator approaches, which highlights the need for accounting for the noise and biases of the answers of the different annotators.

In order to verify that the proposed model was estimating the (normalized) confusion matrices π^r of the different workers correctly, a random sample of them was plotted against the true confusion matrices (i.e., the normalized

Fig. 8. Boxplot of the number of answers per worker (a) and trespective accuracies (b) for the LabelMe dataset.

(a) annotator 1 (b) annotator 2 (c) annotator 3
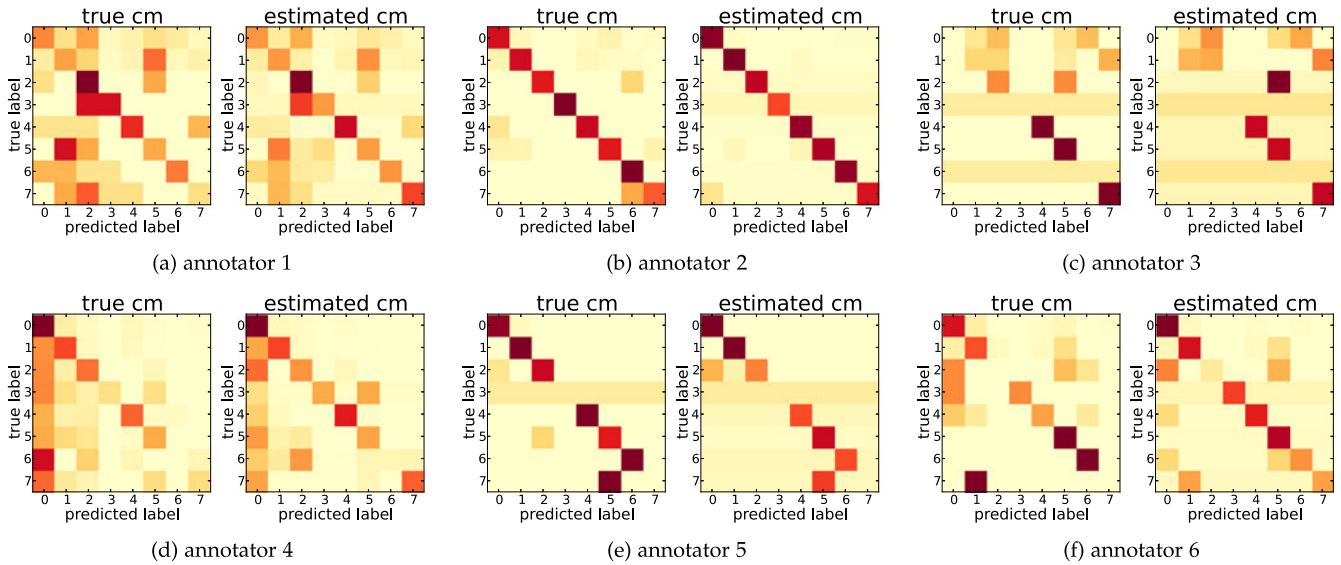
(d) annotator 4 (e) annotator 5 (f) annotator 6

Fig. 10. True versus estimated confusion matrix (cm) of six different workers of the reuters-21,578 dataset.

confusion matrices evaluated against the true labels). Fig. 10 shows the results obtained with 60 topics on the Reuters-21578 dataset, where the color intensity of the cells increases with the magnitude of the value of $p(y^{d,r} = l | c^d) = \pi_{c,l}^r$ (the supplementary material, available online provides a similar figure for the LabelMe dataset). Using this visualization we can verify that the AMT workers are quite heterogeneous in their labeling styles and in the kind of mistakes they make, with several workers showing clear biases (e.g., workers 3 and 4), while others made mistakes more randomly (e.g., worker 1). Nevertheless, the proposed is able to capture these patterns correctly and account for effect.

To gain further insights, Table 3 shows four example images from the LabelMe dataset, along with their true labels, the answers provided by the different workers, the true label inferred by the proposed model and the likelihood of the different possible answers given the true label for each annotator ($\pi_{c^d,l}^r$ for $l \in \{1, \ldots, C\}$) using a color-coding scheme similar to Fig. 10. In the first example, although majority voting suggests "inside city" to be the correct label, we can see that the model has learned that annotators 32 and 43 are very likely to provide the label "inside city" when the true label is actually "street", and it is able to leverage that fact to infer that the correct label is "street". Similarly, in the second image the model is able to infer the correct true label from three conflicting labels. However, in the third image the model is not able to recover the correct true class, which can be explained by it not having enough evidence about the annotators and their reliabilities and biases (likelihood distribution for these cases is uniform). In fact, this raises interesting questions regarding requirements for the minimum number of labels per annotator, their reliabilities and their coherence. Finally, for the fourth image, somehow surprisingly, the model is able to infer the correct true class, even though all three annotators labeled it as "inside city".

5.2 Regression

5.2.1 Simulated Annotators

As for proposed classification model, we start by validating MA-sLDAr using simulated annotators on a popular corpus
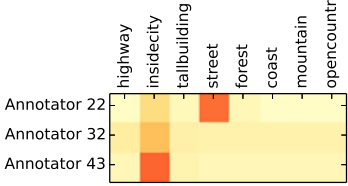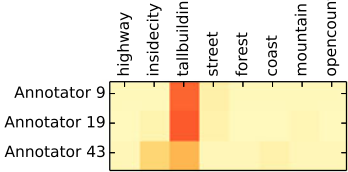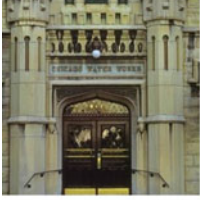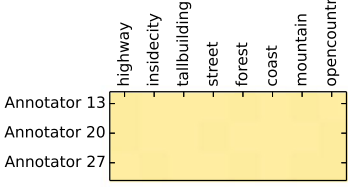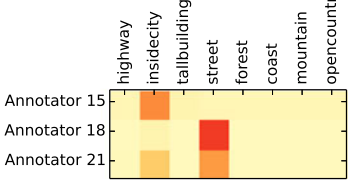
where the documents have associated targets that we wish to predict. For this purpose, we shall consider a dataset of user-submitted restaurant reviews from the website we8there.com. This dataset was originally introduced in [35] and it consists of 6,260 reviews. For each review, there is a five-star rating on four specific aspects of quality (food, service, value, and atmosphere) as well as the overall experience. Our goal is then to predict the overall experience of the user based on his comments in the review. We apply the same preprocessing as in [19], which consists in tokenizing the text into bigrams and discarding those that appear in less than ten reviews. The preprocessing of the documents consisted of stemming and stop-words removal. After that, 75 percent of the documents were randomly selected for training and the remaining 25 percent for testing.

As with the classification model, we seek to simulate an heterogeneous set of annotators in terms of reliability and bias. Hence, in order to simulate an annotator r, we proceed as follows: let x^d be the true review of the restaurant; we start by assigning a given bias b^r and precision p^r to the reviewers, depending on what type of annotator we wish to simulate (see Fig. 2); we then sample a simulated answer as $y^{d,r} \sim \mathcal{N}(x^d + b^r, 1/p^r)$. Using this procedure, we simulated 5 annotators with the following (bias, precision) pairs: (0.1, 10), (−0.3, 3), (−2.5, 10), (0.1, 0.5) and (1, 0.25). The goal is to have two good annotators (low bias, high precision), 1 highly biased annotator and 2 low precision annotators where one is unbiased and the other is reasonably biased. The coefficients of determination (R^2) of the simulated annotators are: [0.940, 0.785, −2.469, −0.131, −1.749]. Computing the mean of the answers of the different annotators yields a R^2 of 0.798. Table 4 gives an overview on the statistics of datasets used in the regression experiments.

We compare the proposed model (MA-sLDAr) with the two following baselines:

- *LDA + LinReg (mean)*: This baseline corresponds to applying unsupervised LDA to the data, and learning a linear regression model on the inferred topics distributions of the documents. The answers from

TABLE 3
Results for Four Example LabelMe Images

True label: street
Annotator 22: street
Annotator 32: inside city
Annotator 43: inside city

Lik. of different answers given true class
Inferred ground truth: street

True label: tall building
Annotator 9: tall building
Annotator 19: street
Annotator 43: inside city

Lik. of different answers given true class
Inferred ground truth: tall building

True label: inside city
Annotator 13: inside city
Annotator 20: tall building
Annotator 27: tall building

Lik. of different answers given true class
Inferred ground truth: tall building

True label: street
Annotator 15: inside city
Annotator 18: inside city
Annotator 21: inside city

Lik. of different answers given true class
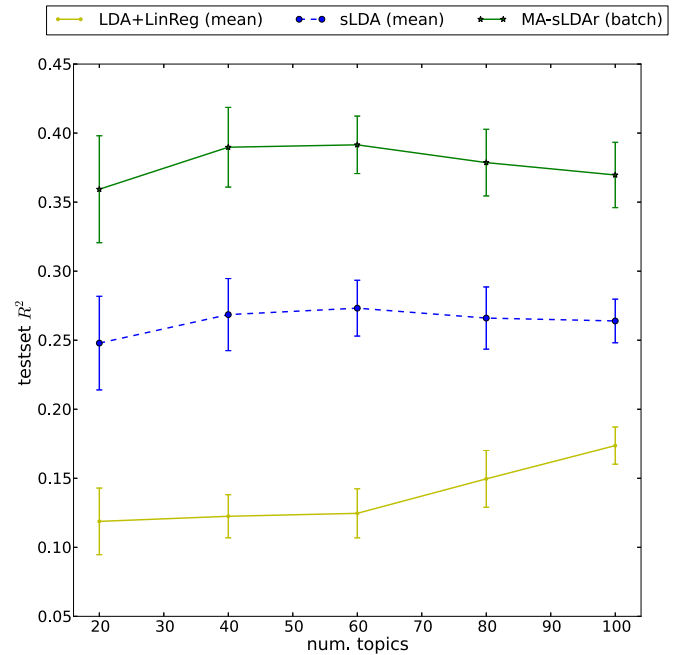Inferred ground truth: street

Fig. 11. Average testset R^2 (over 30 runs; $\pm$ stddev.) of the different approaches on the we8there data.

times and report the average R^2 obtained with the corresponding standard deviation. Since the regression datasets that are considered in this article are not large enough to justify the use of a stochastic variational inference (svi) algorithm, we only made experiments using the batch algorithm developed in Section 4.2. The results obtained clearly show the improved performance of MA-sLDAr over the other methods.

5.2.2 Amazon Mechanical Turk

The proposed multi-annotator regression model (MA-sLDAr) was also validated with real annotators by using AMT. For that purpose, the movie review dataset from [36] was used. This dataset consists of 5,006 movie reviews along with their respective star rating (from 1 to 10). The goal of this experiment is then predict how much a person liked a movie based on what she says about it. We ask workers to guess how much they think the writer of the review liked the movie based on her comments. An average of 4.96 answers per-review was collected for a total of 1,500 reviews. The remaining reviews were used for testing. In average, each worker rated approximately 55 reviews. Using the mean answer as an estimate of the true rating of the movie yields a R^2 of 0.830. Table 4 gives an overview of the statistics of this data. Fig. 12 shows boxplots of the number of answers per worker, as well as boxplots of their respective biases (b^r) and variances (inverse precisions, $1/p^r$).

The preprocessing of the text consisted of stemming and stop-words removal. Using the preprocessed data, the

the different annotators were aggregated by computing the mean.

- *sLDA (mean)*: This corresponds to using the regression version of sLDA [7] with the target variables obtained by computing the mean of the annotators' answers.

Fig. 11 shows the results obtained for different numbers of topics. Do to the stochastic nature of both the annotators simulation procedure and the initialization of the variational Bayesian EM algorithm, we repeated each experiment 30

TABLE 4
Overall Statistics of the Regression Datasets Used in the Experiments

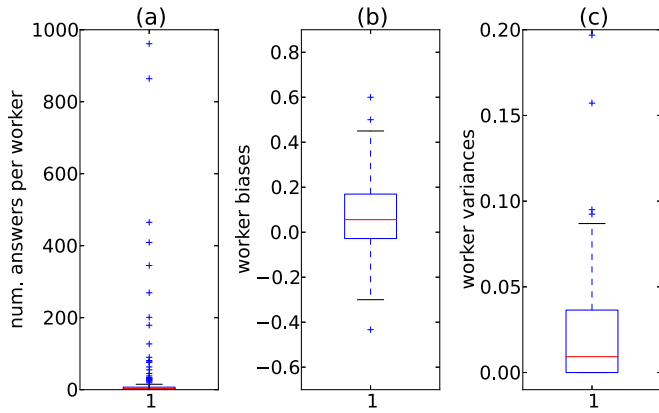| Dataset	Train/test sizes	Annotators source	Num. answers per instance ($\pm$ stddev.)	Num. answers per worker ($\pm$ stddev.)	Mean annotators R^2 ($\pm$ stddev.)	Mean answer R^2
we8there	4,624/1,542	Simulated	5.000 $\pm$ 0.000	4,624.000 $\pm$ 0.000	-0.525 ± 1.364	0.798
movie reviews	1,500/3,506	Mech. Turk	4.960 $\pm$ 0.196	55.111 $\pm$ 171.092	-0.387 ± 1.267	0.830

Fig. 12. Boxplot of the number of answers per worker (a) and their respective biases (b) and variances (c) for the movie reviews dataset.

proposed MA-sLDAr model was compared with the same baselines that were used with the we8there dataset in Section 5.2.1. Fig. 13 shows the results obtained for different numbers of topics. These results show that the proposed model outperforms all the other baselines.

With the purpose of verifying that the proposed model is indeed estimating the biases and precisions of the different workers correctly, we plotted the true values against the estimates of MA-sLDAr with 60 topics for a random subset of 10 workers. Fig. 14 shows the obtained results, where higher color intensities indicate higher values. Ideally, the colour of two horizontally-adjacent squares would then be of similar shades, and this is indeed what happens in practice for the majority of the workers, as Fig. 14 shows. Interestingly, the figure also shows that there are a couple of workers that are considerably biased (e.g., workers 6 and 8) and that those biases are being correctly estimated, thus justifying the inclusion of a bias parameter in the proposed model, which contrasts with previous works [22], [24].
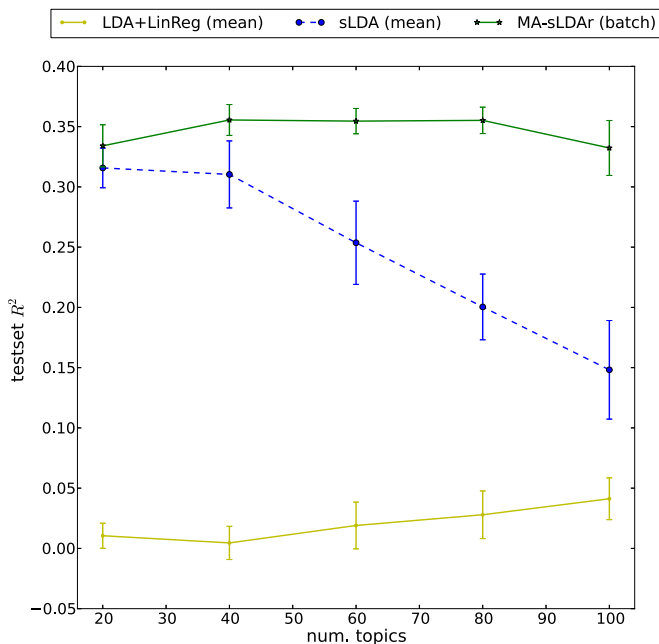

Fig. 13. Average testset R^2 (over 30 runs; $\pm$ stddev.) of the different approaches on the movie reviews data.
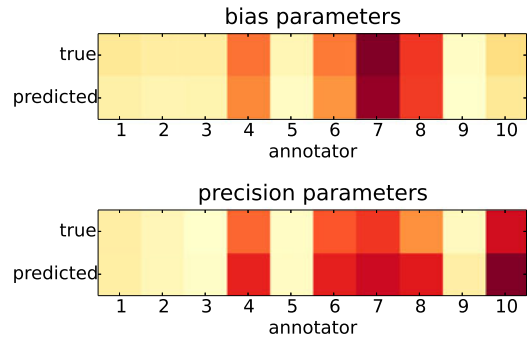

Fig. 14. True versus predicted biases and precisions of 10 random workers of the movie reviews dataset.

6 CONCLUSION

This article proposed a supervised topic model that is able to learn from multiple annotators and crowds, by accounting for their biases and different levels of expertise. Given the large sizes of modern datasets, and considering that the majority of the tasks for which crowdsourcing and multiple annotators are desirable candidates, generally involve complex high-dimensional data such as text and images, the proposed model constitutes a strong contribution for the multi-annotator paradigm. This model is then capable of jointly modeling the words in documents as arising from a mixture of topics, as well as the latent true target variables and the (noisy) answers of the multiple annotators. We developed two distinct models, one for classification and another for regression, which share similar intuitions but that inevitably differ due to the nature of the target variables. We empirically showed, using both simulated and real annotators from Amazon Mechanical Turk that the proposed model is able to outperform state-of-the-art approaches in several real-world problems, such as classifying posts, news stories and images, or predicting the number of stars of restaurant and the rating of movie based on their reviews. For this, we use various popular datasets from the state-of-the-art, that are commonly used for benchmarking machine learning algorithms. Finally, an efficient stochastic variational inference algorithm was described, which gives the proposed models the ability to scale to large datasets.

REFERENCES

[1] D. Blei, A. Ng, and M. I. Jordan, "Latent Dirichlet allocation," *J. Mach. Learn. Res.*, vol. 3, pp. 993–1022, 2003.
[2] L. Fei-Fei and P. Perona, "A Bayesian hierarchical model for learning natural scene categories," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2005, vol. 2, pp. 524–531.
[3] C. Wang, D. Blei, and L. Fei-Fei, "Simultaneous image classification and annotation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2009, pp. 1903–1910.
[4] J. Niebles, H. Wang, and L. Fei-Fei, "Unsupervised learning of human action categories using spatial-temporal words," *Int. J. Comput. Vis.*, vol. 79, no. 3, pp. 299–318, 2008.

[5] E. Erosheva, S. Fienberg, and C. Joutard, "Describing disability through individual-level mixture models for multivariate binary data," *Ann. Appl. Statist.*, vol. 1, no. 2, 2007, Art. no. 346.

[6] E. Airoldi, D. Blei, S. Fienberg, and E. Xing, "Combining stochastic block models and mixed membership for statistical network analysis," in *Statistical Network Analysis*. Berlin, Germany: Springer, 2007, pp. 57–74.

[7] J. Mcauliffe and D. Blei, "Supervised topic models," in *Proc. Advances Neural Inf. Process. Syst.*, 2008, pp. 121–128.

[8] J. Zhu, A. Ahmed, and E. Xing, "MedLDA: Maximum margin supervised topic models," *J. Mach. Learn. Res.*, vol. 13, no. 1, pp. 2237–2278, 2012.

[9] 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 *Proc. Conf. Empirical Methods Natural Language Process.*, 2008, pp. 254–263.

[10] F. Rodrigues, F. Pereira, and B. Ribeiro, "Learning from multiple annotators: distinguishing good from random labelers," *Pattern Recognit. Lett.*, vol. 34, pp. 1428–1436, 2013.

[11] P. Welinder, S. Branson, P. Perona, and S. Belongie, "The multidimensional wisdom of crowds," in *Proc. Advances Neural Inf. Process. Syst.*, 2010, pp. 2424–2432.

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

[13] F. Rodrigues, M. Lourenço, F. Pereira, and B. Ribeiro, "Learning supervised topic models from crowds," in *Proc. 3rd AAAI Conf. Human Comput. Crowdsourcing*, 2015.

[14] M. Hoffman, D. Blei, C. Wang, and J. Paisley, "Stochastic variational inference," *J. Mach. Learn. Res.*, vol. 14, pp. 1303–1347, 2013.

[15] S. Lacoste-Julien, F. Sha, and M. Jordan, "DiscLDA: Discriminative learning for dimensionality reduction and classification," in *Proc. Advances Neural Inf. Process. Syst.*, 2009, pp. 897–904.

[16] D. Ramage, D. Hall, R. Nallapati, and C. Manning, "Labeled LDA: A supervised topic model for credit attribution in multi-labeled corpora," in *Proc. Conf. Empirical Methods Natural Language Process.*, 2009, pp. 248–256.

[17] D. Mimno and A. McCallum, "Topic models conditioned on arbitrary features with Dirichlet-multinomial regression," in *Proc. 24th Conf. Uncertainty Artif. Intell.*, 2008, pp. 411–418.

[18] M. Rabinovich and D. Blei, "The inverse regression topic model," in *Proc. 31st Int. Conf. Mach. Learn.*, 2014, pp. 199–207.

[19] M. Taddy, "Multinomial inverse regression for text analysis," *J. Amer. Statist. Assoc.*, vol. 108, no. 503, pp. 755–770, 2013.

[20] A. P. Dawid and A. M. Skene, "Maximum likelihood estimation of observer error-rates using the EM algorithm," *J. Roy. Statist. Soc. Series C*, vol. 28, no. 1, pp. 20–28, 1979.

[21] P. Smyth, U. Fayyad, M. Burl, P. Perona, and P. Baldi, "Inferring ground truth from subjective labelling of venus images," in *Proc. Advances Neural Inf. Process. Syst.*, 1995, pp. 1085–1092.

[22] V. Raykar, et al., "Learning from crowds," *J. Mach. Learn. Res.*, vol. 11, pp. 1297–1322, 2010.

[23] F. Rodrigues, F. Pereira, and B. Ribeiro, "Gaussian process classification and active learning with multiple annotators," in *Proc. 31st Int. Conf. Mach. Learn.*, 2014, pp. 433–441.

[24] P. Groot, A. Birlutiu, and T. Heskes, "Learning from multiple annotators with Gaussian processes," in *Proc. Int. Conf. Artif. Neural Netw.*, 2011, vol. 6792, pp. 159–164.

[25] Y. Yan, et al., "Modeling annotator expertise: Learning when everybody knows a bit of something," *J. Mach. Learn. Res.*, vol. 9, pp. 932–939, 2010.

[26] M. I. Jordan, Z. Ghahramani, T. Jaakkola, and L. Saul, "An introduction to variational methods for graphical models," *Mach. Learn.*, vol. 37, no. 2, pp. 183–233, 1999.

[27] J. Chuang, S. Gupta, C. Manning, and J. Heer, "Topic model diagnostics: Assessing domain relevance via topical alignment," in *Proc. 30th Int. Conf. Mach. Learn.*, 2013, pp. 612–620.

[28] J. Nocedal and S. Wright, *Numerical Optimization*. Singapore: World Scientific, 2006.

[29] H. Robbins and S. Monro, "A stochastic approximation method," *Ann. Math. Statist.*, vol. 22, no. 3, pp. 400–407, Sep. 1951.

[30] K. Lang, "NewsWeeder: Learning to filter netnews," in *Proc. Int. Conf. Mach. Learn.*, 1995, pp. 331–339.

[31] D. Lewis, "Reuters-21578 text categorization test collection," 1997.

[32] B. Russell, A. Torralba, K. Murphy, and W. Freeman, "LabelMe: A database and web-based tool for image annotation," *Int. J. Comput. Vis.*, vol. 77, no. 1–3, pp. 157–173, 2008.

[33] D. Lowe, "Object recognition from local scale-invariant features," in *Proc. Int. Conf. Comput. Vis.*, 1999, vol. 2, pp. 1150–1157.

[34] A. Bosch, A. Zisserman, and X. Muñoz, "Scene classification via pLSA," in *Proc. Eur. Conf. Comput. Vis.*, 2006, pp. 517–530.

[35] D. Mauá and F. Cozman, "Representing and classifying user reviews," 2009.

[36] B. Pang and L. Lee, "Seeing stars: Exploiting class relationships for sentiment categorization with respect to rating scales," in *Proc. 43rd Annu. Meeting Assoc. Comput. Linguistics*, 2005, pp. 115–124.

Filipe Rodrigues received the PhD degree in information science and technology from the University of Coimbra, Portugal, where he developed probabilistic models for learning from crowdsourced and noisy data. He is postdoctoral fellow with Technical University of Denmark, where he is working on machine learning models for understanding urban mobility and the behavior of crowds, with emphasis on the effect of special events in mobility and transportation systems. His research interests include machine learning, probabilistic graphical models, natural language processing, intelligent transportation systems, and urban mobility.

Mariana Lourenço received the MSc degree in informatics engineering from the University of Coimbra, Portugal. Her thesis presented a supervised topic model that is able to learn from crowds and she took part in a research project whose primary objective was to exploit online information about public events to build predictive models of flows of people in the city. Her main research interests are machine learning, pattern recognition, and natural language processing.

Bernardete Ribeiro received the MSc degree in computer science, the DSc degree in informatics engineering, and the PhD degree in electrical engineering, speciality of Informatics from the Informatics Engineering Department, University of Coimbra, Portugal. She is an associate professor in the Informatics Engineering Department, University of Coimbra, Portugal. Her research interests include of machine learning, pattern recognition and signal processing and their applications to a broad range of fields. She was responsible/participated in several research projects in a wide range of application areas such as Text Classification, Financial, Biomedical, and Bioinformatics. She is a senior member of the IEEE, member of the IARP International Association of Pattern Recognition, and the ACM.

Francisco C. Pereira received the master's and PhD degrees in computer science from the University of Coimbra, in 2000 and 2005, respectively. He is full professor with Technical University of Denmark (DTU), where he leads the Smart Mobility Research Group. His main research focus is on applying machine learning and pattern recognition to the context of transportation systems with the purpose of understanding and predicting mobility behavior, and modeling and optimizing the transportation system as a whole. He has authored/co-authored more than 70 journal and conference papers in areas such as pattern recognition, transportation, knowledge based systems and cognitive science. He was previously research scientist with MIT and an assistant professor with the University of Coimbra. He received several prestigious prizes, including an IEEE Achievements award, in 2009, the Singapore GYSS Challenge in 2013, and the Pyke Johnson award from Transportation Research Board, in 2015. He is a member of the IEEE.

▷ **For more information on this or any other computing topic, please visit our Digital Library at** www.computer.org/publications/dlib.