Proceedings of the Twenty-Fourth International Joint Conference on Artificial Intelligence (IJCAI 2015)

Bayesian Modelling of Community-Based Multidimensional Trust in Participatory Sensing under Data Sparsity

Matteo Venanzi, W. T. Luke Teacy, Alex Rogers, Nicholas R. Jennings
University of Southampton
Southampton, UK
{mv1g10, wtlt, acr, nrj}@ecs.soton.ac.uk

Abstract

We propose a new Bayesian model for reliable aggregation of crowdsourced estimates of real-valued quantities in participatory sensing applications. Existing approaches focus on probabilistic modelling of user's reliability as the key to accurate aggregation. However, these are either limited to estimating discrete quantities, or require a significant number of reports from each user to accurately model their reliability. To mitigate these issues, we adopt a community-based approach, which reduces the data required to reliably aggregate real-valued estimates, by leveraging correlations between the reporting behaviour of users belonging to different communities. As a result, our method is up to 16.6% more accurate than existing state-of-the-art methods and is up to 49% more effective under data sparsity when used to estimate Wi-Fi hotspot locations in a real-world crowdsourcing application.

1 Introduction

Crowdsourcing has become a viable way of providing fast inexpensive services by engaging collectives of untrained users to perform micro-tasks, such as image labelling or text classification. A key application of this paradigm is *participatory sensing* in which people perform sensing tasks using devices such as cameras and GPS sensors embedded in smartphones to report *estimates* of continuous-valued quantities, e.g., locations, radioactivity levels and temperatures, that include both reported measurements and precisions. This can facilitate large-scale information gathering more efficiently than can otherwise be achieved by a single individual or organisation. For example, in the DARPA Red Balloons challenge, ten balloons dispersed across the US were found within nine hours [Pickard *et al.*, 2011]; and in large–scale network coverage mapping, cell towers can be located using GPS and signal strength measurements from mobile phones [Hall and Jordan, 2010]. However, such applications raise issues of trust, due to the unknown incentives of the participants and reliability of their devices [Naroditskiy *et al.*, 2012]. For this reason, fusing crowd responses into a single reliable estimate is a non–trivial problem, for which several solutions have been proposed.

In particular, several methods have been proposed to compute crowd consensus from reported data, considering factors such as user trustworthiness [Raykar *et al.*, 2010], user's biases [Piech *et al.*, 2013], and task difficulty [Bachrach *et al.*, 2012]. These methods have proved more accurate than others that treat participants as equally trustworthy, such as majority voting [Tran-Thanh *et al.*, 2013] and covariance intersection [Julier and Uhlmann, 1997]. Alternatives such as reputation systems [Despotovic and Aberer, 2005; Teacy *et al.*, 2012] or gold–standard driven trust mechanisms [Oleson *et al.*, 2011] are limited by the assumption that the true value of estimated quantities (such as the measurable qualities of a product or service) are eventually revealed, and can therefore be used to assess the reliability of their reported values post hoc. Unfortunately, this is not usually possible in participatory sensing, because the true values of estimated quantities (e.g. cell tower locations) are rarely revealed, and so cannot be used to verify reports directly. Similar methods presented in sensor fusion, such as Kalman filtering and covariance union, have been proved less effective in crowdsourcing domains due to the different type of noise in 'human' sensors compared to traditional sensors [Hall and Jordan, 2010]. In contrast, the best methods for fusing crowdsourced data do not require eventual knowledge of true values, but rely only on statistical correlations between reports from different users to assess each user's reliability [Dawid and Skene, 1979; Salek *et al.*, 2013]. However, this typically requires a significant number of reports from each participant about each item being estimated — a requirement that is hard to support in crowdsourcing, where users often provide only a few reports about a small number of items.

Recently, *community based models* has proved effective at mitigating this issue in applications such as image labelling, galaxy classification and Web retrieval [Li *et al.*, 2014; Venanzi *et al.*, 2014; Qui *et al.*, 2013]. Here, the key idea is to model users with similar behaviour as *communities*, which naturally form within a crowd [Simpson *et al.*, 2013]. For example, when reporting measurements taken by their mobile phones, users with similar devices are likely to have similar reliability. However, community based methods have so far been limited to reports about discrete quantities (e.g. image labels) for which reliability can be represented as a confusion matrix expressing the probability of user's judgments conditioned on a discrete set of possible labels. Unfortunately, this

approach cannot be applied to continuous-valued estimates, and cannot be trivially extended to account for other factors influencing reliability, such as bias and calibration issues, or a user's own confidence in their estimates.

In this paper, we address the above limitations by defining a new set of community based models for aggregating crowdsourced continuous estimates. The key innovation of our approach compared to the existing ones is to provide simultaneous learning of the latent (unobserved) user's multidimensional reliability (bias and trust) and users' communities to improve robustness against data sparsity for fusing multivariate crowdsourced estimates. Specifically, our method is based on a novel hierarchical probabilistic model of user's reliability defined in terms of the latent precision and biases within different communities of users. By generalising from the reporting behaviour of a community as a whole, we can make robust inference about user reliability, even when reports from any given individual user are sparse. In particular, we make the following three contributions to the state-of-the-art: (1) we define the first community based trust model for aggregating datasets of crowdsourced estimates of continuous quantities reported by uncalibrated users in participatory sensing settings; (2) we present an extension of our initial model to deal with more complex datasets affected by community's biases and precision errors for more general crowdsourcing settings; and (3) using real data from Android phones,[1] we show that our method is up to 16.6% more accurate at estimating Wi-Fi hotspot locations, and is up to 49% more robust than existing methods when only a few reports per user are observed.

In the remainder of the paper, we first describe the preliminaries of modelling user's reliability in crowdsourcing. We then describe our community models along with the details of its probabilistic inference. Subsequently, we present our empirical results and outline directions for future work.

2 Background

We now summarise the state–of–the-art for aggregating crowdsourced estimates by modelling user reliabilities. These provide the basis to develop our proposed community based aggregation model. Here, we will adopt the standard notation of using bold symbols for vector random variables, sets and matrices.

Suppose there are N multivariate items (such as GPS locations) to be estimated given reports from a crowd of K users. For each item i, we define $\boldsymbol{\mu}_i \in \mathbb{R}^n$ to be its true latent value, for which we receive a set of $p_{k,i}$ observations from each user k. In each case, the j-th observation from k about i is a pair, $\langle \boldsymbol{x}_{k,i,j}, \theta_{k,i,j} \rangle$, where $\boldsymbol{x}_{k,i,j}$ is an estimate of $\boldsymbol{\mu}_i$ with reported precision $\theta_{k,i,j}$. Intuitively, $\theta_{k,i,j}$ quantifies k's confidence in the estimate, which may be set by self-appraisal or by the precision of the sensor used to make the observation (e.g. for GPS locations). When it is not feasible for precisions to be determined by the user for each observation, a default value may be used.

[1]Data supplied by OpenSignal (`opensignal.com`).

2.1 Modelling Users' Trust

To capture uncertainty about report reliability, each user is assigned a trust parameter $t_k \in \mathbb{R}^+$, which models the accuracy of k in providing observations. In particular, t_k close to 0 means that k is unreliable and overestimates its precisions; while values close to 1 mean that k is trustworthy and so accurately reports its precision. In contrast, $t_k > 1$ means that k is conservative and tends to underestimate the precision of its estimates. One standard way to capture this intuition is to define t_k as a *scaling parameter* of the precisions reported by k, such that the true precision of any given estimate, $\boldsymbol{x}_{k,i,j}$, is $t_k\theta_{k,i,j}$ [Venanzi *et al.*, 2013]. Assuming Gaussian noise, $\boldsymbol{x}_{k,i,j}$ is thus normally distributed with noise proportional to the user's scaled precision:

$$\boldsymbol{x}_{k,i,j}|\boldsymbol{\mu}_i, \theta_{k,i,j}, t_k \sim \mathcal{N}(\boldsymbol{x}_{k,i,j}|\boldsymbol{\mu}_{k,i,j}, t_k\theta_{k,i,j}I) \quad (1)$$

where I is the n-dimensional identity matrix. This means that users are assumed to observe items with uncorrelated (diagonal) noise proportional to the reported precision scaled by their trustworthiness parameter. This model is adopted by the *MaxTrust* method [Venanzi *et al.*, 2013] that estimates the trust parameters using a maximum likelihood approach. As a result, t_k is computed by MaxTrust as an exact value while Bayesian approaches like ours are able to estimate the full uncertainty around these parameters expressed as probability distributions.

2.2 Modelling Users' Biases

To model uncertainty about individual biases, [Piech *et al.*, 2013] propose to use an extra multivariate parameter $\boldsymbol{b}_k \in \mathbb{R}$ to represent the bias of k. Referring to this as the *Student's peer grading* model, k is assumed to draw its observations of the items from a Gaussian distribution with biased mean $\boldsymbol{\eta}_{k,i} = \boldsymbol{\mu}_i + \boldsymbol{b}_k$ and diagonal precision $t_k\theta_{k,i,j}$:

$$\boldsymbol{x}_{k,i,j}|\boldsymbol{\mu}_i, \theta_{k,i,j}, t_k \sim \mathcal{N}(\boldsymbol{\mu}_{k,i,j}|\boldsymbol{\mu}_i + \boldsymbol{b}_k, t_k\theta_{k,i,j}I) \quad (2)$$

Let $\boldsymbol{x}$, $\boldsymbol{\theta}$, $\boldsymbol{t}$, $\boldsymbol{b}$ $\boldsymbol{\mu}$ be vectors comprising all reported estimates, precisions, trust values, biases and item values respectively. The joint likelihood of all the reported estimates $\boldsymbol{x}$ is then:

$$p(\boldsymbol{x}|\boldsymbol{\mu}, \boldsymbol{\theta}, \boldsymbol{t}, \boldsymbol{b}) = \prod_{i=1}^{N}\prod_{k=1}^{K}\prod_{j=1}^{p_{k,i}}\mathcal{N}(\boldsymbol{x}_{k,i,j}|\boldsymbol{\mu}_i + \boldsymbol{b}_k, t_k\theta_{k,i,j}I)$$

To enable tractable inference, the unknown parameters are assigned standard Bayesian conjugate priors [Bishop, 2006]. and using a Markov chain Monte Carlo (MCMC) inference algorithm [Gelfand and Smith, 1990] it is possible to compute the approximate inference of the marginal distributions of each random variable through sampling based methods.

Crucially, both the described models assume that the users' reliability are all independent, i.e., no community structures are assumed to exist within the crowd. However, this assumption goes against a readily observable fact that users' reliability tends to follow community patterns in which groups of users share similar reliability, as shown in several empirical studies of human crowd behaviour [Simpson *et al.*, 2013; Li *et al.*, 2014]. These works show that reliabilities are in fact highly correlated and this is not taken into account for aggregating estimates, which is the major drawback of existing methods.

3 Our Community–Based Model for Aggregating Crowdsourced Estimates

To account for extra correlations between users, we now describe our Community–Based Bayesian Aggregation model for Crowdsourced Estimates (CBACE). In this description, we first consider the case of the estimates' noise w.r.t the users' misreported precisions.

In detail, assume there are M communities of users within the crowd where M is unknown. Each community m is associated with a precision τ_m that represents the average precision of its members and $\boldsymbol{\tau} = \{\tau_1, \ldots, \tau_M\}$ is the vector of all the community precisions. Moreover, assume each user, k, belongs to exactly one community, c_k, drawn from a categorical distribution with parameters $\boldsymbol{c}$:

$$c_k|\boldsymbol{c} \sim \text{Cat}(c_k|\boldsymbol{c}) \qquad \forall k$$

where $\boldsymbol{c}$ is a vector specifying the probability for each possible assignment of c_k in the set of M communities.[2] Further, we assume that each user has a reliability (precision) that is equal to a perturbation of the reliability of its community. More formally, we assume that t_k is drawn from a Log–Gaussian distribution with mean τ_{c_k} and precision α_0:

$$t_k|\boldsymbol{\tau}, c_k \sim \text{Log}\,\mathcal{N}(t_k|\tau_{c_k}, \alpha_0) \qquad \forall k \tag{3}$$

where α_0 is a fixed hyperparameter that expresses the variability of user's reliability within the community. Notice our choice of using a Log-Gaussian distribution to build a hierarchy over t_k as opposed to the Gamma prior adopted by the Student's peer grading model. This allows us to reduce the complexity of inference using conjugate normal priors without restricting the expressibility of our model, as both the Log-Gaussian and the Gamma distribution have support in $\mathbb{R}^+$. In addition, c_k used as a subscript to select the parameters that generate t_k defines a Gaussian mixture model over user precisions and biases using the community parameters as mixture components.

Then, we assume that the user's reports are generated according to the true value of the items and the precision of the user as described by Equation 1. Based on these assumptions, the likelihood is obtained as:

$$p(\boldsymbol{x}, \boldsymbol{\theta}|\boldsymbol{\mu}, \boldsymbol{t}, \boldsymbol{\tau}, \boldsymbol{c}) = \prod_{k=1}^{K} \Big\{ \text{Cat}(c_k|\boldsymbol{c}) \text{Log}\,\mathcal{N}(t_k|\tau_{c_k}, \alpha_0) \prod_{i=1}^{N} \prod_{j=1}^{p_k} \mathcal{N}(\boldsymbol{x}_{k,i,j}|\boldsymbol{\mu}_i, t_k\theta_{k,i,j}I) \Big\} \tag{4}$$

where $\boldsymbol{\tau}$ is the vector comprising all the precisions and the biases of the communities. Given a set of reports, we can perform inference over all the latent variables of CBACE using a principled Bayesian approach. Specifically, we use a conjugate multivariate Gaussian prior for $\boldsymbol{\mu}_i$ with mean $\boldsymbol{\mu}_0$ and precision matrix $\boldsymbol{H}_0$:

$$(\text{True value prior})\ \boldsymbol{\mu}_i \sim \mathcal{N}(\boldsymbol{\mu}_i|\boldsymbol{\mu}_0, \boldsymbol{H}_0) \qquad \forall i$$

with $\boldsymbol{\mu}_0$ and $\boldsymbol{H}_0$ as fixed hyperparameters. The prior governing the community membership probabilities $\boldsymbol{c}$ is Dirichlet distributed with hyperparameter $\boldsymbol{p}$:

$$(\text{Community membership prior})\ \boldsymbol{c} \sim \text{Dir}(\boldsymbol{c}|\boldsymbol{p})$$

The community precision has Gaussian priors:

$$(\text{Community precision prior})\ \tau_m \sim \mathcal{N}(\tau_m|\tau_0, \gamma_0) \qquad \forall m$$

with hyperparameters τ_0 and γ_0. We apply Bayes theorem to derive the posterior distribution as proportional to the likelihood (Equation 4) multiplied by the priors as follows:

$$p(\boldsymbol{\mu}, \boldsymbol{t}, \boldsymbol{\tau}, \boldsymbol{c}|\boldsymbol{x}, \boldsymbol{\theta}) \propto \text{Dir}(\boldsymbol{c}|\boldsymbol{p}) \prod_{m=1}^{M} \mathcal{N}(\tau_m|\tau_0, \gamma_0) \prod_{k=1}^{K} \Big\{ \text{Cat}(c_k|\boldsymbol{c}) \text{Log}\,\mathcal{N}(t_k|\tau_{c_k}, \alpha_m) \prod_{i=1}^{N} \mathcal{N}(\boldsymbol{\mu}_i|\boldsymbol{\mu}_0, \boldsymbol{H}_0) \prod_{j=1}^{p_k} \mathcal{N}(\boldsymbol{x}_{k,i,j}|\boldsymbol{\mu}_i, t_k\theta_{k,i,j}I) \Big\} \tag{5}$$

Then, we can compute the marginal distribution of each parameter by integrating out all the other variables from the posterior. Unfortunately, there is no tractable close–form solution of these integrals for CBACE. Therefore, we use a tractable approximation scheme based on variational inference [Wainwright and Jordan, 2008]. This is based on approximating the posterior distribution of CBACE with a factorised variational distribution that eases the computation of the posterior updates. Specifically, we use the variational message passing (VMP) algorithm [Winn *et al.*, 2005]. [3] provided by the Infer.NET framework [Minka *et al.*, 2014] that enables us to compute the marginal posterior distributions of all the variables of CBACE within minutes with datasets in the order of hundreds of thousands points on a standard 4 cores i5 CPU, 8 GB RAM laptop.

Inferring Latent Communities

A crucial step of CBACE is how to infer the optimal number of communities within the crowd. We do so by tuning the community count parameter, M via a line search maximisation of the marginal log-likelihood function of our model. This function expresses the probability of the data with all the parameters integrated out: $M^* = \arg\max_M \int_{\boldsymbol{\Theta}} p(\boldsymbol{R}|\boldsymbol{\Theta}, M)p(\boldsymbol{\Theta}|M)$where $\boldsymbol{R}$ is the set of the reports and $\boldsymbol{\Theta}$ is the set of all the parameters. In particular, an approximate estimate of the marginal log-likelihood function can be computed using the standard Bernoulli approximation that computes this function as the log odds of an auxiliary Bernoulli variable [Wang and Wand, 2011]. This approximation is also computed for CBACE by the Infer.NET inference engine. To illustrate this, Figure 1 shows the plot of the approximate marginal log-likelihood function with the number

[2] In practice, the possibility of k belonging to multiple communities can be expressed by spreading the probabilities in c_k amongst the communities to which k belongs.

[3] All the models presented in this paper are implemented using the Infer.NET framework and the code is available at `eprints.soton.ac.uk/376365`. The OpenSignal dataset is also available at `eprints.soton.ac.uk/376373` (DOI: 10.5258/SOTON/376373).

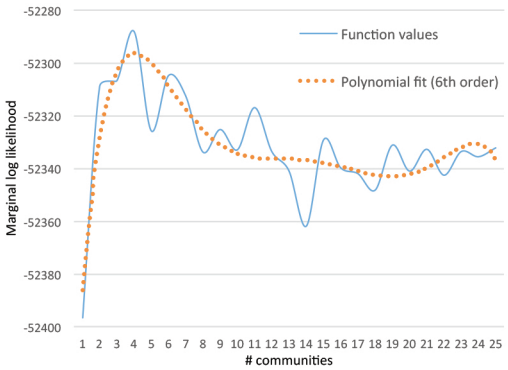

Figure 1: Approximate marginal likelihood of CBACE computed for synthetic reports.

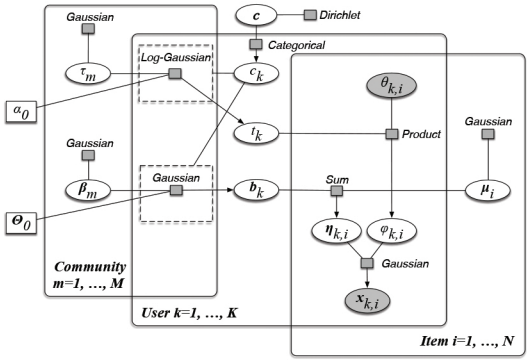

Figure 2: Factor graph of CBACE with communities' bias.

of communities varying from 1 to 25. This function is computed for a dataset of 100 synthetic reports that was generated from four communities (see Section 5.4 for the specific setting). Therefore, we have the ground truth for the number of communities in this dataset. The irregular shape of the function is due to the typical error introduced by the Bernoulli approximation computed by the VMP algorithm. However, it can be noted that the approximate function values (solid line) has the maximum peak at 4 communities that is indeed the real number of communities in the dataset. This fact is even more evident when these values are interpolated through a 6-th order polynomial function (dotted line) that correctly identifies the local maximum of the function.

It is worth noting that alternative approaches can be used to infer the number of communities that change over time, for example using non–parametric infinite mixture modelling based on Dirichlet processes. While these approaches can potentially provide more accurate results, they are also computationally more complex and therefore they can be more constraining for practical use in real–world applications. Furthermore, several regularisation methods that introduce a penalty term to balance the model evidence with its complexity can make the community selection more robust against the risk of overfitting a particular dataset. Since our focus is not to take these optimisation methods to a further level – and we did not experience overfitting problems in our settings – we opt for a simple and more practical maximum likelihood-based model evidence approach for community selection.

4 CBACE with Communities' bias

In many crowdsourcing tasks, such as the crowdsourced peer reviewing of students' assignments in MOOCs [Piech *et al.*, 2013], it is likely that estimates might be affected by individual biases of the users in addition to their precision errors. In fact, student's ratings might be biased towards higher or lower scores. Here we show how CBACE can be easily extended to consider user's trust w.r.t communities' latent biases. We will refer to this extension as the CBACE+Bias model.

In detail, assume that each community, m has an extra parameter $\boldsymbol{\beta}_m$ denoting the average bias of its members and let $\boldsymbol{\beta} = \{\beta_1, \ldots, \beta_M\}$ be the vector of the biases of all the communities. Then, k has an individual bias, $\boldsymbol{b}_k$, that is drawn from $\boldsymbol{\beta}_{c_k}$ with multivariate Gaussian noise with precision matrix $\boldsymbol{\Theta_0}$:

$$\boldsymbol{b}_k|\boldsymbol{\beta}, c_k \sim \mathcal{N}(\boldsymbol{b}_k|\boldsymbol{\beta}_{c_k}, \boldsymbol{\Theta}_0) \qquad \forall k$$

where c_k is the random variable that selects of the user's community. Using conjugate Gaussian priors for $\boldsymbol{\beta}_m$:

$$\text{(Community bias prior) } \boldsymbol{\beta}_m \sim \mathcal{N}(\boldsymbol{\beta}_m|\boldsymbol{\beta}_0, \boldsymbol{B}_0) \qquad \forall m$$

we obtain the new likelihood function:

$$p(\boldsymbol{x}|\boldsymbol{\mu}, \boldsymbol{\theta}, \boldsymbol{t}, \boldsymbol{b}, \boldsymbol{\tau}, \boldsymbol{\beta}, \boldsymbol{c}) = \prod_{k=1}^{K} \Big\{ \text{Cat}(c_k|c)\mathcal{N}(\boldsymbol{b}_k|\boldsymbol{\beta}_{c_k}, \boldsymbol{\Theta}_0)$$
$$\text{Log}\,\mathcal{N}(t_k|\tau_{c_k}, \alpha_m) \prod_{i=1}^{N} \prod_{j=1}^{p_k} \mathcal{N}(\boldsymbol{x}_{k,i,j}|\boldsymbol{\mu}_i + \boldsymbol{b}_k, t_k\theta_{k,i,j}I) \Big\}$$

Then, following the same process outlined for CBACE, we can derive the joint posterior distribution from which we can compute approximate predictions of each latent variable using variational Bayesian inference. The number of communities can also be estimated using the same model evidence optimisation technique described for CBACE. In more detail, Figure 2 shows the factor graph of CBACE+Bias using the plates notation, where the plates (boxes) show replicated parts of the graph for the variables associated with users and items. The graph uses the *gate* notation [Minka and Winn, 2009] (dashed boxes) to represent the mixture models of user's parameters conditioned on the user's community membership.

5 Evaluation

We evaluate the performance of our two proposed algorithms in an important participatory sensing application of Wi-Fi hotspots localisation that looks at using crowdsourced reports to build Wi-Fi maps for improving mobile localisation services [Lim *et al.*, 2007]. We also consider datasets reported by simulated reporters in various settings to extend the generality and the scale of our experimental analysis.

5.1 Dataset

We used a real-world dataset of crowdsourced Wi-Fi hotspots detections provided by OpenSignal. This dataset includes

	OpenSignal–Wi-Fi dataset
Wi-Fi hotspots	208
- with ground truth	13
Android devices	49
Reports	7464
- Max. reports per device	1762
- Q_3 reports per device	77
- Q_2 reports per device	14
- Q_1 reports per device	3
- Min. reports per device	1

Table 1: Statistics of the OpenSignal–Wi-Fi dataset.

7464 reports from 49 Android devices for 208 Wi-Fi hotspots. Each report provides (i) the SSID and BSSID of the detected Wi-Fi hotspot, (ii) the GPS location (latitude and longitude) of the detecting device and (iii) the precision of the GPS fix (in meters). We have ground truth for the location of 13 Wi-Fi hotspots acquired from the British Telecom (BT) Wi-Fi network database (`btWi-Fi.com`). The statistics of this dataset are reported in Table 1. In this dataset, we observe that the distribution of reports per device is mostly skewed towards the low counts: 50% of the devices have less than 14 reports (Q_2), 25% of the devices have more than 77 reports and only 6% of the devices have very high counts (more than 1000). This shows that, in practice, data sparsity is indeed a realistic case for data fusion in crowdsourcing settings.

5.2 Benchmarks

In our evaluation we consider the following three benchmarks as the state-of-the-art rival methods for multivariate fusion of crowdsourced estimates: (1) *Covariance Intersection (CI)* [Julier and Uhlmann, 1997], this method is a standard linear combination of equally trustworthy Gaussian estimates used in Kalman filtering, (2) *MaxTrust* [Venanzi *et al.*, 2013], this method merges the estimates by learning the precision of each users using the uncertainty scaling Gaussian model described in Section 2, (3) *Student's peer grading* [Piech *et al.*, 2013], this method merges estimates while simultaneously learning the precision and the biases *individually* for each user as also described in Section 2. Thus, we evaluate the performance of five methods: {CI, MaxTrust, Student's peer grading, CBACE, CBACE+Bias}

To reproduce a typical scenario where no prior information is available about the users, the communities and the items as in the case of our Wi-Fi dataset, we use uninformative priors for all the parameters, with the community bias prior set to be a multivariate standard Gaussian distribution, i.e., $\boldsymbol{\beta} = 0$ and $\boldsymbol{B}_0 = \text{diag}(1)$. The community's trust prior is set to be a log–Gaussian distribution with mean $\tau_0 = 1$ and precision $\gamma_0 = 0.1$. The noise precision of the communities are set to $\boldsymbol{\Theta}_0 = \text{diag}(0.1)$ and $\alpha_0 = 0.1$. The priors of the items are also set to a multivariate standard Gaussian distribution. Furthermore, the hyperparameters of all the other methods are set to values equivalent to the priors of CBACE.

To define a setting suitable for applying our methods to this dataset, we convert the GPS location of each report from the provided geographical coordinates (latitude and longitude) to planar coordinates (km) using the standard UTM Mercator projection. This location is taken as the mean value $\boldsymbol{x}_{k,i,j}$ of the report of the Wi-Fi hotspot. The precision of each report is estimated from the reported GPS precision that we consider as the 95% precision (i.e., two standard deviations) around the reported GPS fix. Adding the standard error of 0.1km, which is the default range of Wi-Fi access points, the report's precision is: $\theta_{k,i,j} = (\text{GPS\_precision} * 2 \times 10^{-3} + 0.1)^{-2}$ This provides a setting for this dataset suitable for application of our models. For example, Figure 4 shows the subset of 27 reports for one of our Wi-Fi hotspots with ground truth. Each report is plotted as a Gaussian circle centred on $\boldsymbol{x}_{k,i,j}$ with the radius of two standard deviations given by $2\theta_{k,i,j}$

Figure 4: The reports for a sample Wi-Fi hotspots taken from the OpenSignal–Wi-Fi dataset with the true Wi-Fi hotspot position marked on the map. Each circle shows the value and the precision (2 standard deviations) reported by the user.

5.3 Accuracy Metrics

To measure the accuracy of predictions about items, we compute the *root mean square error* (RMSE) between $E[\boldsymbol{\mu}_i]$ and the true location of i, $\hat{\boldsymbol{\mu}}_i$ for i averaged over all the hotspots with known ground truth location: $\text{RMSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(E[\boldsymbol{\mu}_i] - \hat{\boldsymbol{\mu}}_i)^2}$ Furthermore, we wish to score the methods based on the predictive uncertainty of their estimates in such a way that the highest score is assigned to the predictor with both the lowest expected error and the lowest uncertainty. Thus, we consider the *mean energy score* (MES) [Gneiting *et al.*, 2008], that is a density–based scoring rule that generalises the continuous rank probability score (CRPS) [Matheson and Winkler, 1976] – widely used in statistics and increasingly in AI – to multivariate sampled probability distributions. In detail, let $\boldsymbol{\mu}_{s,i}$ be the s-th sample from a chain of independent samples drawn from the predictive estimate of $\boldsymbol{\mu}_i$. Then, $\text{MES} = \frac{1}{N}\sum_{i=1}^{N}\frac{1}{S}\sum_{s=1}^{S}||\boldsymbol{\mu}_{s,i} - \hat{\boldsymbol{\mu}}_i|| - \frac{1}{2(S-1)}\sum_{s=2}^{S-1}||\boldsymbol{\mu}_{s,i} - \boldsymbol{\mu}_{s+1,i}||$ where the first term is the expected error from $\hat{\boldsymbol{\mu}}$ and the second term scores the variability between the samples.

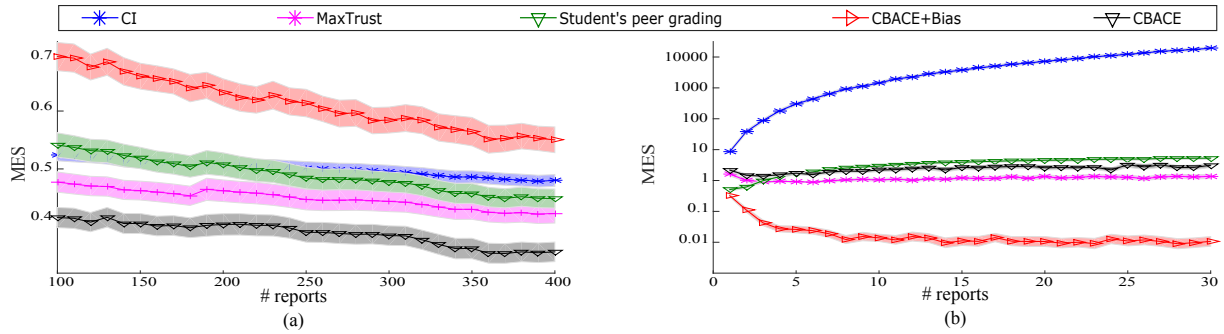

Figure 3: Accuracy of the five methods over number of reports measured on (a) OpenSignal Wi–Fi data and (b) synthetic data.

5.4 Results

Prediction Accuracy Table 2 reports the scores (RMSE, km and MES) of the four methods for the OpenSignal–Wi-Fi dataset. The results show that CBACE has the lowest RMSE of 0.1432km that improves by 6.7% the accuracy of the second best method, MaxTrust. This improvement is even more evident in terms of MES where CBACE outperforms MaxTrust by 16.6% (0.1559 vs. 0.1870). This means that the community based model of CBACE provides substantially more informative location estimates. Interestingly, both of the two community methods (CBACE and CBACE+Bias) found that $M = 3$ is the most likely number of communities of devices in this dataset. Crucially, we observe that the two models that consider user's biases, Student's peer grading and our CBACE+Bias, do not provides competitive estimates of the Wi-Fi hotspots. This is explained by the fact that Wi-Fi location reports are unlikely to have strong biases, unless due to manufacturing defects in the device's hardware, while they are more likely to be corrupted by precision errors that may be due to the limited sensitivity of their GPS sensors. For this reason, it is reasonable to expect some degradation in performance for CBACE+Bias in cases where no bias is present given the prior belief over the community biases. To clarify this aspect, we will extend our analysis by testing our models in settings with stronger user's biases in our next experiments.

Robustness to Data Sparsity To test the robustness of the methods under data sparsity, we measure their accuracy using only a portion of the report set. Using the same number of three communities for CBACE and CBACE-Bias, we start with an initial set of 10% of reports of our Wi-Fi dataset and simulate 30 rounds where 10 new reports are randomly selected and included in the training set at each round. Figure 3a shows the plots of the MES of each method averaged over 40 runs. The shaded areas of each line represents the standard deviation of their mean error. While the error of all the methods progressively decrease as more reports are added to the set, CBACE is consistently the method with the lowest error. At iteration 1, it has 49% higher accuracy than MaxTrust. This shows the capability of CBACE of bootstrapping learning of the users' reliability through the communities when the report set is very sparse[4].

	RMSE (km)	MES
CI	0.1629	0.2359
MaxTrust	0.1518	0.1870
Student's peer grading	0.1903	0.2090
CBACE+Bias	0.3058	0.3108
CBACE	**0.1432**	**0.1559**

Table 2: Accuracy of the four methods on the OpenSignal-Wi-Fi dataset. The best scores are highlighted in bold

To test the performance of the methods in the different setting that emulates the MOOC's peer reviewing application in which the reviewer (user)'s estimates contain both biases and calibration errors, we run a second experiment using a synthetic dataset of univariate estimates generated from ten users for two items. Specifically, we randomly generate estimates assuming an heterogeneous set of four communities: (1) calibrated-low bias, $\tau_1 = 1, \beta_1 = 1$; (2) uncalibrated-low bias, $\tau_2 = 0.1, \beta_2 = 1$; (3) calibrated-high bias, $\tau_3 = 1, \beta_3 = 5$; (iv) uncalibrated-high bias, $\tau_4 = 0.1\beta = 5$. The true value of each item was randomly sampled from a standard Gaussian distribution. The community membership of each users was randomly sampled with 0.55 probability for community 1 and uniform probabilities for the other communities. We run the experiment for 30 rounds by randomly selecting a new report to be added to the training set at each round. Figure 3b shows the MES of each method averaged over 40 runs. Here, we see that CBACE+Bias is now the best method that converges faster to the best accuracy due to its correctly learning of the precisions and biases of the users based on the underlying community structures of this dataset. All the other methods show a much slower progress in terms of accuracy as they struggle to learn accurate user's models with only a few reports per user. In particular, CBACE is now less accurate due to the fact that it can only learn precisions and therefore it ignores the users' biases. Also, the error of CI grows linearly in the number of reports as a result of considering all the users as equally trustworthy. Globally, this second experiment shows that CBACE+Bias is an effective method for merging estimates in sparse datasets with stronger user's biases. In practice, a natural way for choosing the best model between CBACE and CBACE+Bias for a given dataset may

[4]The statistical significance of all these results was tested with a paired t-test at the significance level of 0.05.

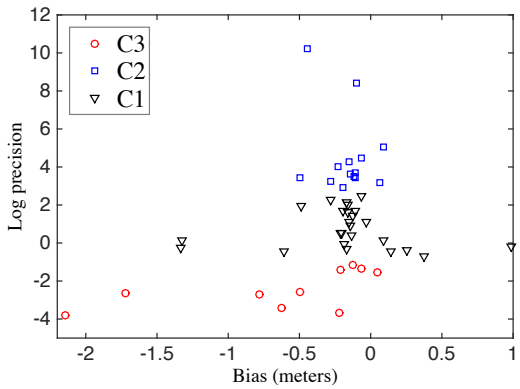

Figure 5: The plot of the biases (x-axis), the log-precisions (y-axis) and community assignments inferred by CBACE+Bias for the devices of the OpenSignal–Wi-Fi dataset.

be through cross–validation by holding out a subset of estimates for validation and selecting the model with the lowest error in predicting the estimates in the validation set.

Community Learning Through our models, we are able to analyse the communities of devices inferred from the OpenSignal–Wi-Fi dataset. In particular, Table 3 shows the three communities detected from CBACE. These communities correspond to a group of *calibrated* users with low bias. This includes 51% of the users. The second community is a group of *conservative* users that have high latent precisions, i.e., they tend to underestimate the actual precisions of their reports, and low bias. This includes 31% of the users. The third group is a minority (19%) of *uncalibrated* user that tend to overestimate the precision of their reports and they also have low bias. Figure 5 shows the community assignments as well as the individual biases and precisions of the devices inferred by CBACE+Bias. In particular, having all the communities with zero biases is meaningful to further highlight that precision errors have much stronger influence in this Wi-Fi dataset. To validate these results, we run k-means to cluster the precisions of the users estimated by the Student's peer grading model. In particular, we set $k = 3$ to match number of communities of CBACE. The three clusters computed by $k-$means are reported in Table 4. It can be seen that both these methods show similar types of communities. In particular, the k-means clusters are comparable to CBACE in terms of centroid precisions – except for C2 whose the k-means centroid precision is smaller compared to the value estimated by CBACE for the same community. However, such a difference in precision values for C2 is less significant in terms of standard deviations, i.e., the inverse square root of the precision value, where the values are std=0.12 for CBACE+bias and std=0.43 for k-means. The two methods also have similar results for the estimated proportions of users for each cluster. However, unlike k-means, CBACE efficiently exploits the inferred communities to improve the accuracy of the aggregated estimates.

CBACE+Bias communities	*Prec.*	*Bias (meters)*	*%*
C1 [calibrated, low bias]	2.03	-0.03	51
C2 [conservative, low bias]	61.31	-0.02	29
C3 [uncalibrated, low bias]	0.12	-0.06	20

Table 3: Community detection results for CBACE on the OpenSignal–Wi-Fi dataset. The rows show the *mean* value of the precision and the bias, and the estimated proportions of users (last column) in each community.

k-means *clusters*	*Prec.*	*%*
C1 [calibrated]	1.08	59
C2 [conservative]	5.4	27
C3 [uncalibrated]	0.12	14

Table 4: The k-means clustering results of the precisions estimated from the Student's peer grading model on the OpenSignal–Wi-Fi dataset.

6 Conclusions

We proposed two community based Bayesian models for reliable aggregation of crowdsourced continuous estimates. The key innovation of our models is a method for probabilistic reasoning about latent communities that makes them more effective at aggregating continuous-valued sparse data, that often occur in crowdsourcing settings. This advancement is achieved through hierarchical modelling of communities of users and their reliability, as part of the inference of aggregated estimates. By doing so, our models not only perform more accurate aggregations but also learn valuable information about different types of users' reliability, which is useful to identify groups of good users and their propensity to specific tasks. We showed that our first model, CBACE, is 16.6% more accurate in estimating the Wi-Fi hotspots locations and is up to 49% more accurate at making predictions from sparse data. We also described a second model, CBACE+Bias, that is more suitable for merging reports when their noise relates to both users' biases and precisions. Finally, we showed that our algorithms provide community learning outputs comparable to standard clustering algorithms, which makes them suitable to analyse group behaviours from crowd generated sensor measurements.

However, several aspects of our current model outline promising directions for future work. For example, the time–dependent aspects of user's reliability and the task's difficulty can be taken into account to potentially improve the quality of the inference. More broadly, our approach could be extended to data aggregation problems in the general context of human–centred information systems, such as web recommendation and peer review systems, to merge subjective inputs from human users.

7 Acknowledgments

The authors gratefully acknowledge fundings from the UK Research Council for the ORCHID project, grant EP/I011587/1. Thanks to OpenSignal for supplying the data and to John Guiver (Microsoft Research) for discussions and feedbacks about the model.

References

[Bachrach *et al.*, 2012] Y. Bachrach, T. Graepel, T. Minka, and J. Guiver. How to grade a test without knowing the answers — a bayesian graphical model for adaptive crowdsourcing and aptitude testing. In *Proceedings of the 29th Int. Conf. on Machine Learning (ICML-12)*, pages 1183–1190, July 2012.

[Bishop, 2006] C. Bishop. *Pattern recognition and machine learning*, volume 4. Springer New York, 2006.

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

[Despotovic and Aberer, 2005] Z. Despotovic and K. Aberer. Probabilistic prediction of peers' performance in p2p networks. *Engineering Applications of Artificial Intelligence*, 18(7):771–780, 2005.

[Gelfand and Smith, 1990] A. Gelfand and A. Smith. Sampling-based approaches to calculating marginal densities. *Journal of the American statistical association*, 85(410):398–409, 1990.

[Gneiting *et al.*, 2008] T. Gneiting, L. Stanberry, E. Grimit, L. Held, and N. Johnson. Assessing probabilistic forecasts of multivariate quantities, with an application to ensemble predictions of surface winds. *Test*, 17(2):211–235, 2008.

[Hall and Jordan, 2010] D. Hall and J. M Jordan. *Human-centered information fusion*. Artech House, 2010.

[Julier and Uhlmann, 1997] S. J. Julier and J. K. Uhlmann. A non-divergent estimation algorithm in the presence of unknown correlations. In *Proceedings of the American Control Conf.*, pages 2369–2373, 1997.

[Li *et al.*, 2014] Hongwei Li, Bo Zhao, and A. Fuxman. The wisdom of minority: discovering and targeting the right group of workers for crowdsourcing. In *Proceedings of the 23rd Int. Conf. on World Wide Web*, pages 165–176, 2014.

[Lim *et al.*, 2007] Y. Lim, C.and Wan, B. Ng, and C. See. A real-time indoor wifi localization system utilizing smart antennas. *IEEE Transactions on Consumer Electronics*, 53(2):618–622, 2007.

[Matheson and Winkler, 1976] J. E Matheson and R. Winkler. Scoring rules for continuous probability distributions. *Management science*, 22(10):1087–1096, 1976.

[Minka and Winn, 2009] T. Minka and J. Winn. Gates. In *Advances in Neural Information Processing Systems*, pages 1073–1080, 2009.

[Minka *et al.*, 2014] T. Minka, J. Winn, J. Guiver, and D. Knowles. Infer.NET 2.6. Microsoft Research Cambridge, 2014.

[Naroditskiy *et al.*, 2012] V. Naroditskiy, I. Rahwan, M. Cebrian, and N. R. Jennings. Verification in referral-based crowdsourcing. *PloS One*, 7(10):e45924, 2012.

[Oleson *et al.*, 2011] D. Oleson, A. Sorokin, G. Laughlin, V. Hester, J. Le, and L. Biewald. Programmatic gold: Targeted and scalable quality assurance in crowdsourcing. *Human computation*, 11(11), 2011.

[Pickard *et al.*, 2011] G. Pickard, W. Pan, I. Rahwan, M. Cebrian, R. Crane, A. Madan, and A. Pentland. Time-critical social mobilization. *Science*, 334(6055):509–512, 2011.

[Piech *et al.*, 2013] C. Piech, J. Huang, Z. Chen, C. Do, A. Ng, and D. Koller. Tuned models of peer assessment in moocs. In *Proceedings of the 6th Int. Conf. on Educational Data Mining (EDM 2013)*, pages 153–160, 2013.

[Qui *et al.*, 2013] G. Qui, C. Aggarwal, J. Han, and T. Huang. Mining collective intelligence in diverse groups. In *22nd Int. Conf. on World Wide Web (WWW 2013)*, pages 1041–1052, 2013.

[Raykar *et al.*, 2010] V. Raykar, S. Yu, L. Z., G. Valadez, C. Florin, L. B., and L. Moy. Learning from crowds. *The Journal of Machine Learning Research*, 11:1297–1322, 2010.

[Salek *et al.*, 2013] M. Salek, Y. Bachrach, and P. Key. Hotspotting–a probabilistic graphical model for image object localization. In *Proceedings of the 27th Conf. on Artificial Intelligence (AAAI)*, pages 1156–1162, 2013.

[Simpson *et al.*, 2013] E. Simpson, S. Roberts, I. Psorakis, and A. Smith. Dynamic bayesian combination of multiple imperfect classifiers. In *Decision Making and Imperfection*, pages 1–35. Springer, 2013.

[Teacy *et al.*, 2012] W. T. L. Teacy, M. Luck, A. Rogers, and N. R. Jennings. An efficient and versatile approach to trust and reputation using hierarchical bayesian modelling. *Artificial Intelligence*, 193:149–185, 2012.

[Tran-Thanh *et al.*, 2013] L. Tran-Thanh, M. Venanzi, A. Rogers, and N. R. Jennings. Efficient budget allocation with accuracy guarantees for crowdsourcing classification tasks. In *Proceedings of the 2013 Int. Conf. on Autonomous Agents and Multi-agent Systems*, AAMAS '13, pages 901–908, 2013.

[Venanzi *et al.*, 2013] M. Venanzi, A. Rogers, and N. R. Jennings. Trust-based fusion of untrustworthy information in crowdsourcing applications. In *Proceedings of the 2013 Int. Conf. on Autonomous Agents and Multi-agent Systems*, AAMAS '13, pages 829–836, Richland, SC, 2013.

[Venanzi *et al.*, 2014] M. Venanzi, J. Guiver, G. Kazai, P. Kohli, and M. Shokouhi. Community-based bayesian aggregation models for crowdsourcing. In *Proceedings of the 23rd Int. Conf. on World Wide Web*, pages 155–164, 2014.

[Wainwright and Jordan, 2008] M. Wainwright and M. Jordan. Graphical models, exponential families, and variational inference. *Foundations and Trends® in Machine Learning*, 1(1-2):1–305, 2008.

[Wang and Wand, 2011] S. Wang and M. P. Wand. Using infer. net for statistical analyses. *The American Statistician*, 65(2), 2011.

[Winn *et al.*, 2005] J. Winn, C. Bishop, and T. Jaakkola. Variational message passing. *Journal of Machine Learning Research*, 6(4), 2005.