Artificial Intelligence 230 (2016) 1–13

Contents lists available at ScienceDirect

Artificial Intelligence

www.elsevier.com/locate/artint

Semi-supervised combination of experts for aerosol optical depth estimation [☆]

Nemanja Djuric [a,*], Lakesh Kansakar [b], Slobodan Vucetic [b]

[a] Yahoo Labs, 701 First Avenue, Sunnyvale, CA, USA
[b] Temple University, 1804 North Broad Street, Wachman Hall, Philadelphia, PA, USA

ARTICLE INFO

Article history:
Received 16 July 2014
Received in revised form 27 July 2015
Accepted 20 September 2015
Available online 1 October 2015

Keywords:
Combination of experts
Aerosol estimation
Remote sensing

ABSTRACT

Aerosols are small airborne particles produced by natural and man-made sources. Aerosol Optical Depth (AOD), recognized as one of the most important quantities in understanding and predicting the Earth's climate, is estimated daily on a global scale by several Earth-observing satellite instruments. Each instrument has different coverage and sensitivity to atmospheric and surface conditions, and, as a result, the quality of AOD estimated by different instruments varies across the globe. We present a semi-supervised method for learning how to aggregate estimations from multiple satellite instruments into a more accurate estimate, where labels come from a small number of accurate and expensive ground-based instruments. The method also accounts for the problem of missing experts, an issue inherent to the AOD estimation task. By assuming a context-dependent prior, the model is capable of incorporating additional information and providing estimates even when there are no available experts. Moreover, the proposed method uses a latent variable to partition the data, so that in each partition the expert AOD estimations are aggregated in a different, optimal way. We applied the method to combine global AOD estimations from 5 instruments aboard 4 satellites, and the results indicate it can successfully exploit labeled and unlabeled data to produce accurate aggregated AOD estimations.

1. Introduction

Aerosols are small airborne particles produced by natural and man-made sources that both reflect and absorb incoming Solar radiation. Depending on their distribution and composition, aerosols can result either in cooling or warming of the atmosphere, thus having a major role in regulating the climate system. Distribution of aerosols is measured by Aerosol Optical Depth (AOD or τ), a quantitative measure of the extinction of Solar radiation by scattering and absorption between the top of the atmosphere and the surface. Aerosols have been recognized among the most important quantities in understanding and predicting the Earth's climate by the Intergovernmental Panel on Climate Change (IPCC) [2], and have been a focal point of a number of scientific studies due to their importance and large impact on our atmosphere [3–5]. AOD is an important input to climate models, and it can significantly impact predictions of future climate changes [6]. Considering that climate predictions influence decisions of policy makers, accurate AOD estimation is a task of global significance. Moreover, in addition to its impact on large-scale terrestrial climate studies, AOD is an important quantity in estimation of air pollution that affects the well-being and quality of life of us all. For example, it was shown by [7] that AOD is an accurate

[☆] This paper is an invited revision of [1], an article published at The International Joint Conference of Artificial Intelligence (IJCAI 2013).

[*] Corresponding author.
E-mail addresses: nemanja@yahoo-inc.com (N. Djuric), lakesh@temple.edu (L. Kansakar), vucetic@temple.edu (S. Vucetic).

http://dx.doi.org/10.1016/j.artint.2015.09.010
0004-3702/© 2015 Elsevier B.V. All rights reserved.

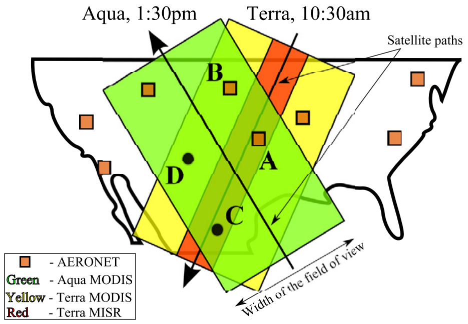


Fig. 1. Coverage of instruments over the USA.

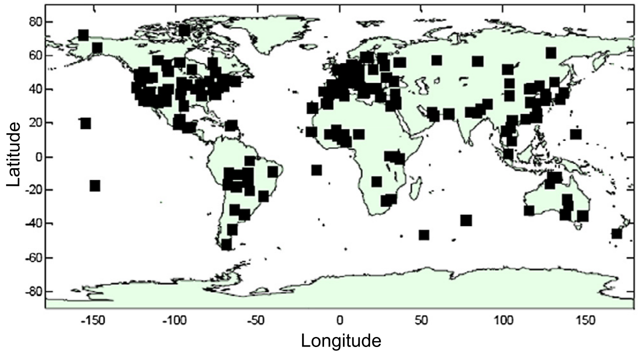


Fig. 2. Global coverage of AERONET instruments.

predictor of $PM_{2.5}$, the concentration of particulate matter with aerodynamic diameters ≤ 2.5 μm, which poses a serious health hazard to the population [8].

Currently, a number of instruments aboard several Earth-observing satellites report their AOD estimates, such as MODIS instrument aboard Terra and Aqua satellites [9], MISR aboard Terra [10], OMI aboard Aura [11], SeaWiFS aboard SeaStar [12], Cloud-Aerosol Lidar with Orthogonal Polarization (CALIOP) aboard CALIPSO [13], and others. All these instruments have a capability of providing global estimates of AOD distribution with a fine spatial (few kilometers) and temporal (few days) resolution. Each instrument has different properties and estimates AOD using a different algorithm developed by domain scientists. Coverage and quality of satellite measurements can differ from instrument to instrument for a number of reasons. As illustrated in Fig. 1, width of the field of view of MODIS instrument is 2,330 km, allowing MODIS to observe the entire Earth every day, as opposed to 360 km width of MISR instrument, which requires 9 days for global coverage. The quality of AOD estimates from different instruments varies with atmospheric and surface conditions [14]. For example, 9 cameras observing Earth at 9 different angles used by MISR allow it to be more accurate than MODIS when clouds are present, over bright surfaces, or for some types of aerosol compositions. In addition to satellite-borne sensors, AOD is also measured by a network of ground-based sensors from AERONET [15], placed at several hundred unevenly distributed locations across the globe, see Fig. 2. AERONET AOD measurements are considered a ground-truth, as they are several times more accurate than the best available satellite AOD estimations. The drawback of AERONET is that it has a very limited spatial coverage, and that it cannot be used to provide global estimation of AOD distribution required for climate models.

Different spatial and temporal coverage, design, and specific mission objectives of various satellite-borne instruments mean that they observe and measure different, possibly complementary aspects of the same phenomenon. Thus, instead of considering AOD estimates of individual instruments in isolation, combining measurements from different sources into an aggregated estimate may prove to be the best path towards obtaining a higher-quality, global estimation of aerosol distribution. A recent study by [14] confirmed this hypothesis by illustrating that even a simple average of collocated Terra MODIS and MISR AOD estimations resulted in an improved accuracy.

Apart from AOD studies, researchers have explored aggregation of measurements in other areas of Earth science as well, including work on climate change [16,17], carbon dioxide distribution [18], and sea surface height [19]. Moreover, the combination of experts that ultimately yields an estimate that is more accurate than any of the individual forecasts is a well-researched topic in many other scientific areas, such as risk analysis [20], information retrieval [21], or artificial intelligence [22]. More recently, researchers have proposed a number of state-of-the-art methods that address the problem of aggregation of discrete predictions [23–25]. However, less progress has been made to address the problem when experts provide real-valued predictions. Assuming the Gaussian distribution of prediction errors and no missing experts, in the seminal works [26] and [27] the authors described a method for learning the optimal combination of experts from labeled data.

If data set is unlabeled, in [28] the authors proposed how to learn a combination of experts by extending a classification method from [29]. The approach assumed that the experts are independent and that all experts are always available for aggregation, which may be an unrealistic assumption for the considered task of aerosol estimation.

In this paper we propose a novel method suitable for finding a linear combination of AOD estimations from multiple instruments. There are several interesting challenges specific to the aerosol domain that had to be addressed. (1) As quality of different instruments varies with atmospheric and surface conditions, it is not likely that the same linear combination would work equally well at different locations, for example, in North America and Africa [30]. Therefore, it might be needed to develop specialized combinations for different regions around the globe. (2) Number of labeled data points is relatively small. For example, in North America, thanks to a relative abundance of AERONET sites, the number of labeled data points can exceed a thousand every year, while in Africa and parts of Asia there are very few AERONET sites, and the number of labeled data points could be measured in tens every year. In addition to their small number, labeled data points might cover only a limited set of conditions observable at AERONET locations. On the other hand, the number of unlabeled data points is orders of magnitudes larger. An open question in AOD estimation is how to exploit labeled and unlabeled data. (3) As shown in Fig. 1, which illustrates daily coverage of different sensors over the USA, for most of the labeled (e.g., points A and B) and unlabeled (e.g., points C and D) data points, AOD estimations from some of the instruments are missing. For example, points A and C have AOD estimate from all 3 satellite instruments, while points B and D are just outside of MISR's field of view and do not have its AOD estimate. Moreover, even when a particular location is covered by a satellite instrument (i.e., location is within satellite's field of view), it does not mean that the instrument necessarily provides a prediction, further exacerbating the problem of missing data. Typical reasons for missing predictions are cloud contamination, sunglint, or sensor maintenance and repair. Similar holds for ground-based AERONET instrument, which does not provide a complete temporal coverage as it is strongly affected by local weather conditions and other technical issues. This opens a question of learning from data with significant amounts of missing AOD estimations.

To address the many issues plaguing remote sensing of aerosols, we assume that AOD estimation errors of satellite instruments follow multivariate Gaussian distribution, and propose a semi-supervised method that can handle missing data while being able to partition the data into homogeneous subsets on which specialized aggregators are learned. Our method can be seen as a significant generalization of the traditional supervised method for combination of experts introduced by [26] and [27], as well as of recently proposed unsupervised method for averaging of experts in regression by [28].

2. Methodology

In this section we describe the details of the proposed semi-supervised aggregation algorithm. We will see that the model is very suitable for aggregation of aerosol predictions, as it accounts for missing satellite predictions, correlated experts, as well as for unlabeled data for which AERONET failed to provide a ground-truth. In addition, by assuming a context-dependent prior, it is capable of incorporating prior knowledge and additional data sources, as well as providing predictions even when there are no available satellite predictors.

2.1. Problem setup and assumptions

Let us assume we are given a training set $\mathcal{D} = \{\mathbf{x}_i, \{\hat{y}_{ik}\}_{k=1,\dots,K}, y_i\}_{i=1,\dots,N}$, where target value y_i for the ith data point is predicted by K experts, with the kth expert providing an opinion in a form of prediction $\hat{y}_{ik}$, and $\mathbf{x}_i \in \mathbb{R}^D$ is a column-vector of explanatory features for the ith data point. For example, in the aerosol domain that we study, the experts are satellite instruments and the predictions are their individual AOD estimates, while explanatory variables can be longitude, latitude, temperature, or any other information we have readily available for the ith data point. We arrange the data set such that the first N_u data points are unlabeled, while the last N_l data points are labeled, i.e., we have a ground truth only for data points indexed by $i = (N_u + 1), \dots, N$, with $N = (N_u + N_l)$. In the remainder of the paper we use **1** to denote a column-vector of all ones of an appropriate length, **0** to denote a matrix of all zeros, and $\hat{\mathbf{y}}_i^K = [\hat{y}_{i1}, \dots, \hat{y}_{iK}]^{\mathrm{T}}$ to denote a column-vector of expert predictions for the ith data point.

We assume a linear model for the generation of true AOD values, corrupted by a stochastic process as

$$y_i \equiv f(\mathbf{x}_i, \mathbf{w}) = \mathbf{w}^{\mathrm{T}}\mathbf{x}_i + \varepsilon_i, \tag{1}$$

where $\mathbf{w} \in \mathbb{R}^D$ is a weight vector, and ε_i is a regression error due to zero-mean Gaussian noise with variance σ^2. Then, we can say that the target values y_i are sampled from the following Gaussian distribution,

$$y_i \sim \mathcal{N}(\mu_i, \sigma^2), \text{with } \mu_i = f(\mathbf{x}_i, \mathbf{w}). \tag{2}$$

Further, we assume that data points are independent and identically distributed (IID), and that expert predictions for the ith data point are sampled from a multivariate Gaussian distribution as

$$\hat{\mathbf{y}}_i^K | y_i \sim \mathcal{N}(y_i\mathbf{1}, \boldsymbol{\Sigma}_K). \tag{3}$$

This assumption allows the experts to be correlated (i.e., $\boldsymbol{\Sigma}_K$ is non-diagonal), as is the case in practice in the aerosol domain. We will first consider a case where all experts are available, and then extend the methodology to account for missing experts. Given $\mathcal{D}$, the objective is to learn $\mathbf{\Sigma}_K$, $\mathbf{w}$, and σ^2. For conciseness, in the following by $\Theta = \{\mathbf{\Sigma}_K, \mathbf{w}, \sigma^2\}$ we denote a set of parameters to be learned in a training phase.

Once Θ is learned, and given expert predictions $\hat{\mathbf{y}}_i$, aggregated prediction y_i for the ith data point can be found as a mean of the posterior distribution $y_i | \mathbf{x}_i, \hat{\mathbf{y}}_i^K \sim \mathcal{N}(\overline{y}_i, (\mathbf{1}^{\mathrm{T}}\mathbf{\Sigma}^{-1}\mathbf{1})^{-1})$, where mean $\overline{y}_i$ is computed as

$$\overline{y}_i = \frac{\hat{\mathbf{y}}_i^{\mathrm{T}}\mathbf{\Sigma}^{-1}\mathbf{1}}{\mathbf{1}^{\mathrm{T}}\mathbf{\Sigma}^{-1}\mathbf{1}}, \tag{4}$$

with $\hat{\mathbf{y}}_i = [(\hat{\mathbf{y}}_i^K)^{\mathrm{T}}, \mu_i]^{\mathrm{T}}$, and $\mathbf{\Sigma}$ is a $(K+1) \times (K+1)$ block matrix equal to

$$\mathbf{\Sigma} = \begin{bmatrix} \mathbf{\Sigma}_K & \mathbf{0} \\ \mathbf{0} & \sigma^2 \end{bmatrix}. \tag{5}$$

Interestingly, we can see that the prior mean μ_i can be viewed as an additional expert with variance σ^2, which is independent from the original K experts. Thus, to simplify the presentation, in the following we consider the prior mean as the $(K+1)$th expert that is always available to the aggregation algorithm (i.e., it can never be missing).

2.2. Semi-supervised combination of experts

Given the model parameters Θ, probability of observing the data set $\mathcal{D}$ can be written as

$$\mathbb{P}(\mathcal{D}|\Theta) = \mathbb{P}(\mathcal{D}_u|\Theta)\, \mathbb{P}(\mathcal{D}_l|\Theta), \tag{6}$$

where subscripts u and l denote unlabeled and labeled parts of the data set, respectively. Let us first consider $\mathbb{P}(\mathcal{D}_u|\Theta)$. As the data points are sampled IID, the probability factorizes over individual data points, and we can write

$$\mathbb{P}(\mathcal{D}_u|\Theta) = \prod_{i=1}^{N_u} \mathbb{P}(\hat{\mathbf{y}}_i|\Theta) = \prod_{i=1}^{N_u} \int_y \mathbb{P}(\hat{\mathbf{y}}_i^K | y, \Theta)\, \mathbb{P}(y|\Theta)\, dy. \tag{7}$$

As both probabilities under the integral are assumed Gaussian, due to (2) and (3), their product is also Gaussian. Then, by solving the integral, we obtain

$$\mathbb{P}(\mathcal{D}_u|\Theta) = \prod_{i=1}^{N_u} \Big(\sqrt{\frac{|\mathbf{\Sigma}|^{-1}}{(2\pi)^K \mathbf{1}^{\mathrm{T}}\mathbf{\Sigma}^{-1}\mathbf{1}}} \exp\big(-\frac{1}{2} (\hat{\mathbf{y}}_i - \overline{y}_i\mathbf{1})^{\mathrm{T}} \mathbf{\Sigma}^{-1} (\hat{\mathbf{y}}_i - \overline{y}_i\mathbf{1}) \big) \Big). \tag{8}$$

Moreover, likelihood of the labeled part can be written as follows,

$$\mathbb{P}(\mathcal{D}_l|\Theta) = \prod_{i=N_u+1}^{N} \mathbb{P}(\hat{\mathbf{y}}_i^K | y_i, \Theta)\, \mathbb{P}(y_i|\Theta), \tag{9}$$

which, following equations (2) and (3), is a product of N_l multivariate Gaussians with covariance matrix $\mathbf{\Sigma}$. Then, combining equations (6), (8), and (9), we can compute likelihood of the data set $\mathcal{D}$ for any given set of parameters Θ.

We employ the maximum likelihood principle to find the model parameters. After finding derivative of the log-likelihood with respect to $\mathbf{\Sigma}^{-1}$ and equating the resulting expression with zero, we obtain the following expression for $\mathbf{\Sigma}$ matrix,

$$\mathbf{\Sigma} = \frac{1}{N}\Big((\hat{\mathbf{Y}}_l - \mathbf{y}_l\mathbf{1}^{\mathrm{T}})^{\mathrm{T}} (\hat{\mathbf{Y}}_l - \mathbf{y}_l\mathbf{1}^{\mathrm{T}}) + \frac{N_u \mathbf{1}\mathbf{1}^{\mathrm{T}}}{\mathbf{1}^{\mathrm{T}}\mathbf{\Sigma}^{-1}\mathbf{1}} + \hat{\mathbf{Y}}_u^{\mathrm{T}}\hat{\mathbf{Y}}_u + \sum_{i=1}^{N_u} \big(\overline{y}_i^2 \mathbf{1}\mathbf{1}^{\mathrm{T}} - \overline{y}_i (\mathbf{1}\hat{\mathbf{y}}_i^{\mathrm{T}} + \hat{\mathbf{y}}_i\mathbf{1}^{\mathrm{T}}) \big) \Big) \odot \begin{bmatrix} \mathbb{1} & \mathbf{0} \\ \mathbf{0} & 1 \end{bmatrix}, \tag{10}$$

where $\odot$ denotes element-wise matrix multiplication operator, $\mathbb{1}$ is a $K \times K$ matrix of all ones, $\hat{\mathbf{Y}}_u$ and $\hat{\mathbf{Y}}_l$ are $N_u \times (K+1)$ and $N_l \times (K+1)$ matrices of expert predictions for unlabeled and labeled data, respectively, with each row corresponding to a single data point, and $\mathbf{y}_l$ is an $N_l \times 1$ column-vector of ground-truth values. Equation (10) yields an iterative procedure for learning $\mathbf{\Sigma}$, where $\mathbf{\Sigma}$ on the l.h.s. is a new value, and $\mathbf{\Sigma}$ on the r.h.s. is an old value.

Next, we maximize the log-likelihood of data set $\mathcal{D}$ with respect to $\mathbf{w}$. The log-likelihood, after removing all elements not dependent on $\mathbf{w}$ from equations (8) and (9), is equal to

$$\mathcal{L} = -\frac{1}{2\sigma^2} \sum_{i=1}^{N_u} \big(\overline{y}_i - f(\mathbf{x}_i, \mathbf{w}) \big)^2 - \frac{1}{2\sigma^2} \sum_{i=N_u+1}^{N_u+N_l} \big(y_i - f(\mathbf{x}_i, \mathbf{w}) \big)^2. \tag{11}$$

We can see from (11) that, in order to maximize $\mathcal{L}$ with respect to $\mathbf{w}$, we need to solve linear regression where for unlabeled points we use an estimate of a ground truth equal to $\overline{y}_i$. A closed-form solution for $\mathbf{w}$ can be found using the familiar equations for solving linear regression, computed as

$$\mathbf{w} = (\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{y}, \tag{12}$$

where $\mathbf{X}$ is an $N \times D$ matrix of explanatory features with each row corresponding to a single data point, and $\mathbf{y}$ is an N-dimensional vector with the first N_u elements equal to $\overline{y}_i, i = 1, \dots, N_u$, and the remaining N_l elements equal to $y_i, i = N_u + 1, \dots, N$.

2.3. Missing experts

Let us now consider the case where some experts are missing. For example, let us assume that the ith data point has q missing predictions. Then, we reorganize vector $\hat{\mathbf{y}}_i$ in such a way so that the first $a = (K + 1 - q)$ elements are available predictions, while the last q elements are missing predictions, i.e., $\hat{\mathbf{y}}_i = [\hat{\mathbf{y}}_{ai}^{\mathrm{T}}, \hat{\mathbf{y}}_{qi}^{\mathrm{T}}]^{\mathrm{T}}$. Similarly, we reorganize $\mathbf{\Sigma}^{-1}$ matrix so that the first a rows/columns correspond to available predictions, while the remaining q rows/columns correspond to missing predictions, or

$$\Pi_i(\mathbf{\Sigma}^{-1}) = \begin{bmatrix} \mathbf{U} & \mathbf{V} \\ \mathbf{V}^{\mathrm{T}} & \mathbf{Q} \end{bmatrix}, \tag{13}$$

where Π_i is a permutation function used to reorder both rows and columns of $\mathbf{\Sigma}^{-1}$ according to the ith data point, and $\mathbf{U}$ is an $a \times a$ matrix. Given the covariance matrix $\mathbf{\Sigma}$ and a vector of expert predictions $\hat{\mathbf{y}}_{ai}$ for the ith data point, the aggregated prediction y_i can be found as a mean of the posterior distribution $y_i | \hat{\mathbf{y}}_{ai} \sim \mathcal{N}(\overline{y}_i, (\mathbf{1}^{\mathrm{T}}\mathbf{U}_i'\mathbf{1})^{-1})$, where we introduced $\mathbf{U}' = \mathbf{U} - \mathbf{V}\mathbf{Q}^{-1}\mathbf{V}^{\mathrm{T}}$ to simplify the notation, and

$$\overline{y}_i = \frac{\hat{\mathbf{y}}_{ai}^{\mathrm{T}}\mathbf{U}_i'\mathbf{1}}{\mathbf{1}^{\mathrm{T}}\mathbf{U}_i'\mathbf{1}}. \tag{14}$$

Note that we appended subscript i to indicate that the size of a matrix $\mathbf{U}_i'$ depends on the number of available experts for the ith data point.

In the following, we derive the update equation for $\mathbf{\Sigma}$. The probability of observing the ith unlabeled point is equal to

$$\mathbb{P}(\hat{\mathbf{y}}_{ai}|\Theta) = \int\limits_{\hat{\mathbf{y}}_{qi}} \mathbb{P}([\hat{\mathbf{y}}_{ai}^{\mathrm{T}}, \hat{\mathbf{y}}_{qi}^{\mathrm{T}}]^{\mathrm{T}}|\Theta)\, d\hat{\mathbf{y}}_{qi} = \int\limits_{y} \int\limits_{\hat{\mathbf{y}}_{qi}} \mathbb{P}([\hat{\mathbf{y}}_{ai}^{\mathrm{T}}, \hat{\mathbf{y}}_{qi}^{\mathrm{T}}]^{\mathrm{T}}|y, \Theta)\, \mathbb{P}(y|\Theta)\, d\hat{\mathbf{y}}_{qi}\, dy. \tag{15}$$

Solving the equation (15) we obtain

$$\mathbb{P}(\hat{\mathbf{y}}_{ai}|\Theta) = \sqrt{\frac{|\mathbf{\Sigma}|^{-1}|\mathbf{Q}_i|^{-1}}{(2\pi)^{K+q}\mathbf{1}^{\mathrm{T}}\mathbf{U}_i'\mathbf{1}}} \exp\big(-\frac{1}{2}(\hat{\mathbf{y}}_{ai} - \overline{y}_i\mathbf{1})^{\mathrm{T}}\mathbf{U}_i'(\hat{\mathbf{y}}_{ai} - \overline{y}_i\mathbf{1})\big). \tag{16}$$

In a very similar manner we can find the probability of observing the ith labeled data point. It follows

$$\mathbb{P}(\hat{\mathbf{y}}_{ai}|y_i, \Theta) = \int\limits_{\hat{\mathbf{y}}_{qi}} \mathbb{P}([\hat{\mathbf{y}}_{ai}^{\mathrm{T}}, \hat{\mathbf{y}}_{qi}^{\mathrm{T}}]^{\mathrm{T}}|y_i, \Theta)\, d\hat{\mathbf{y}}_{qi}, \tag{17}$$

which, after solving the integral, results in

$$\hat{\mathbf{y}}_{ai}|y_i \sim \mathcal{N}(y_i\mathbf{1}, \mathbf{U}_i'^{-1}). \tag{18}$$

By combining equations (6), (16), and (18), we can find the likelihood of the data set $\mathcal{D}$. After finding derivative of the log-likelihood with respect to $\mathbf{\Sigma}^{-1}$ [31] and equating the resulting expression with zero, we obtain the following expression for computing the $\mathbf{\Sigma}$ matrix,

$$\begin{aligned} \mathbf{\Sigma} = \frac{1}{N}\Bigg(& \sum_{i=1}^{N_u}\Bigg([\![\hat{\mathbf{y}}_{ai}\hat{\mathbf{y}}_{ai}^{\mathrm{T}}]\!] + \frac{[\![\mathbf{1}\mathbf{1}^{\mathrm{T}}]\!]}{\mathbf{1}^{\mathrm{T}}\mathbf{U}_i'\mathbf{1}} + \overline{y}_i^2[\![\mathbf{1}\mathbf{1}^{\mathrm{T}}]\!] - \overline{y}_i[\![\mathbf{1}\hat{\mathbf{y}}_{ai}^{\mathrm{T}} + \hat{\mathbf{y}}_{ai}\mathbf{1}^{\mathrm{T}}]\!]\Bigg) \\ & + \sum_{i=N_u+1}^{N} [\![(\hat{\mathbf{y}}_{ai} - y_i\mathbf{1})\,(\hat{\mathbf{y}}_{ai} - y_i\mathbf{1})^{\mathrm{T}}]\!] + \sum_{i=1}^{N}\Pi_i^{-1}(\mathbf{\Psi}_i)\Bigg) \odot \begin{bmatrix} \mathbb{1} & \mathbf{0} \\ \mathbf{0} & 1, \end{bmatrix} \end{aligned} \tag{19}$$

where Π_i^{-1} is an inverse permutation function that reorders rows and columns of the matrix back to the original order of experts, symmetric $(K + 1) \times (K + 1)$ matrix $\mathbf{\Psi}_i$ is equal to

$$\mathbf{\Psi}_i = \begin{bmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{Q}_i^{-1} \end{bmatrix}, \tag{20}$$

and $[\![\mathbf{A}_i]\!]$ for some symmetric $a \times a$ matrix $\mathbf{A}_i$ denotes the following symmetric $(K+1) \times (K+1)$ matrix,

$$[\![\mathbf{A}_i]\!] = \Pi_i^{-1}\left(\begin{bmatrix} \mathbf{A}_i & -\mathbf{A}_i\,\mathbf{V}_i\,\mathbf{Q}_i^{-1} \\ -\mathbf{Q}_i^{-1}\,\mathbf{V}_i^{\mathrm{T}}\,\mathbf{A}_i & \mathbf{Q}_i^{-1}\mathbf{V}_i^{\mathrm{T}}\,\mathbf{A}_i\,\mathbf{V}_i\mathbf{Q}_i^{-1} \end{bmatrix}\right). \tag{21}$$

Lastly, after finding a derivative of the log-likelihood function with respect to $\mathbf{w}$, parameter vector of the prior model can be found as in (12).

2.4. Incorporating prior probability $\mathbb{P}(\Theta)$

Let us consider the case where we have some prior, expert knowledge about the underlining data-generation model, and would like to include this knowledge into the aggregation model. First, we write the joint probability of the data and the model as follows,

$$\mathbb{P}(\mathcal{D}, \Theta) = \mathbb{P}(\mathcal{D}|\boldsymbol{\Sigma}^{-1}, \mathbf{w})\,\mathbb{P}(\boldsymbol{\Sigma}^{-1})\,\mathbb{P}(\mathbf{w}). \tag{22}$$

Note that we defined prior $\mathbb{P}(\boldsymbol{\Sigma}^{-1})$ in terms of an inverse of the covariance matrix (i.e., in terms of a precision matrix). For the prior over the Θ parameters we choose a normal-Wishart distribution $\mathrm{NW}(\mathbf{0}, \lambda^{-1}, \mathbf{S}, n)$, a conjugate prior for multivariate Gaussian distribution, with given λ prior variance parameter, $(K+1) \times (K+1)$ scale matrix $\mathbf{S}$, and $n > K$ degrees of freedom, resulting in

$$\mathbb{P}(\boldsymbol{\Sigma}^{-1})\,\mathbb{P}(\mathbf{w}) = \frac{|\boldsymbol{\Sigma}^{-1}|^{0.5(n-K-2)}\exp\left(-0.5\,\mathrm{Tr}(\mathbf{S}^{-1}\boldsymbol{\Sigma}^{-1})\right)}{2^{0.5n(K+1)}|\mathbf{S}|^{0.5n}\Gamma_{K+1}(0.5n)}\,\mathcal{N}(\mathbf{w}|\mathbf{0}, \lambda\mathbf{I}), \tag{23}$$

where Γ_{K+1} is the multivariate gamma function, and $\mathbf{I}$ is a $(K+1) \times (K+1)$ identity matrix. After setting $n = (K+3)$ and finding the derivative of the log-likelihood with respect to $\boldsymbol{\Sigma}^{-1}$, we obtain the following update equation for the covariance matrix $\boldsymbol{\Sigma}$,

$$\begin{aligned}\boldsymbol{\Sigma} = \frac{1}{1+N}\Bigg(\mathbf{S}^{-1} &+ \sum_{i=1}^{N}\Pi_i^{-1}(\boldsymbol{\Psi}_i) + \sum_{i=N_u+1}^{N}[\![(\hat{\mathbf{y}}_{ai} - y_i\mathbf{1})\,(\hat{\mathbf{y}}_{ai} - y_i\mathbf{1})^{\mathrm{T}}]\!] \\ &+ \sum_{i=1}^{N_u}\Big([\![\hat{\mathbf{y}}_{ai}\hat{\mathbf{y}}_{ai}^{\mathrm{T}}]\!] + \frac{[\![\mathbf{1}\mathbf{1}^{\mathrm{T}}]\!]}{\mathbf{1}^{\mathrm{T}}\mathbf{U}_i'\mathbf{1}} + \overline{y}_i^2[\![\mathbf{1}\mathbf{1}^{\mathrm{T}}]\!] - \overline{y}_i[\![\mathbf{1}\hat{\mathbf{y}}_{ai}^{\mathrm{T}} + \hat{\mathbf{y}}_{ai}\mathbf{1}^{\mathrm{T}}]\!]\Big)\Bigg) \odot \begin{bmatrix} \mathbb{1} & \mathbf{0} \\ \mathbf{0} & 1 \end{bmatrix}.\end{aligned} \tag{24}$$

Similarly, after finding the derivative of the log-likelihood with respect to weight vector, $\mathbf{w}$ can be found using the closed-form solution for regularized linear regression, with regularization parameter equal to $\lambda\sigma^2$,

$$\mathbf{w} = (\mathbf{X}^{\mathrm{T}}\mathbf{X} + \lambda\sigma^2\mathbf{I})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{y}. \tag{25}$$

2.5. Data partitioning using a latent variable

It is an inherent property of the experts in the aerosol domain that they do not maintain the same quality of predictions across all observed conditions. To address this characteristic of the aggregation problem, we consider partitioning the data points into several groups, called the *regimes*, where each regime is governed by a different prior from (2) and multivariate Gaussian from (3). In the following we assume there are R regimes, and that we have available a feature vector $\widetilde{\mathbf{x}}_i \in \mathbb{R}^{\widetilde{D}}$ for the ith data point that could be used to assign it to an appropriate regime [32]. Note that we denoted feature vector $\widetilde{\mathbf{x}}_i$ used for partitioning and feature vector $\mathbf{x}_i$ used in the prior from equation (1) differently, to emphasize that they do not necessarily need to be identical.

Assuming a mixture of R regimes, probability of observing expert predictions $\hat{\mathbf{y}}_{ai}$ for the ith labeled data point can be written as follows,

$$\mathbb{P}(\hat{\mathbf{y}}_{ai}|\widetilde{\mathbf{x}}_i, y_i, \Theta) = \sum_{r=1}^{R}\mathbb{P}_r(\hat{\mathbf{y}}_{ai}|y_i)\,\pi_{ir}(\widetilde{\mathbf{x}}_i), \tag{26}$$

where $\mathbb{P}_r(\hat{\mathbf{y}}_{ai}|y_i) = \mathbb{P}(\hat{\mathbf{y}}_{ai}|\mathrm{regime}_r, \widetilde{\mathbf{x}}_i, y_i, \Theta)$, $\pi_{ir}(\widetilde{\mathbf{x}}_i) = \mathbb{P}(\mathrm{regime}_r|\widetilde{\mathbf{x}}_i, \Theta)$, and where appended subscript r denotes the rth regime. Similarly, we can write probability of observing expert predictions $\hat{\mathbf{y}}_{ai}$ for the ith unlabeled data point as

$$\mathbb{P}(\hat{\mathbf{y}}_{ai}|\widetilde{\mathbf{x}}_i, \Theta) = \sum_{r=1}^{R}\mathbb{P}_r(\hat{\mathbf{y}}_{ai})\,\pi_{ir}(\widetilde{\mathbf{x}}_i). \tag{27}$$

Probability of observing the ith unlabeled or labeled data point given that it was generated by the rth regime, $\mathbb{P}_r(\hat{\mathbf{y}}_{ai})$ or $\mathbb{P}_r(\hat{\mathbf{y}}_{ai}|y_i)$, respectively, can be computed by considering equations (16) and (18), respectively. Before computing the aggregated prediction, we first find the prior mean as

$$\mu_i \equiv \mathbb{E}[\mu_i|\widetilde{\mathbf{x}}_i, \Theta] = \sum_{r=1}^{R} \pi_{ir}(\widetilde{\mathbf{x}}_i)\, f(\mathbf{x}_i, \mathbf{w}_r), \tag{28}$$

which acts as the $(K+1)$st expert, and the aggregated prediction $\overline{y}_i$ can then be found using the following expression,

$$\overline{y}_i = \mathbb{E}[y_i|\hat{\mathbf{y}}_{ai}, \widetilde{\mathbf{x}}_i, \Theta] = \sum_{r=1}^{R} \pi_{ir}(\widetilde{\mathbf{x}}_i)\, \frac{\hat{\mathbf{y}}_{ai}^{\mathrm{T}} \mathbf{U}'_{ir} \mathbf{1}}{\mathbf{1}^{\mathrm{T}} \mathbf{U}'_{ir} \mathbf{1}}. \tag{29}$$

To facilitate model optimization, we consider regime assignments as unobserved data, and introduce a latent indicator variable z_{ir} such that the following holds,

$$z_{ir} = \begin{cases} 1 & \text{if } \hat{\mathbf{y}}_{ai} \text{ was generated by the } r\text{th regime,} \\ 0 & \text{otherwise.} \end{cases} \tag{30}$$

Further, by introducing $\mathbf{z}_i = [z_{i1}, \dots, z_{iR}]^{\mathrm{T}}$, we can write the complete-data likelihood for the ith labeled data point as

$$\mathbb{P}(\hat{\mathbf{y}}_{ai}, \mathbf{z}_i|\widetilde{\mathbf{x}}_i, y_i, \Theta) = \prod_{r=1}^{R} \big(\pi_{ir}(\widetilde{\mathbf{x}}_i)\, \mathbb{P}_r(\hat{\mathbf{y}}_{ai}|y_i)\big)^{z_{ir}}. \tag{31}$$

Note that, for conciseness, in equations (31), (32), and (34), we only give expressions for labeled data. However, when dealing with the ith unlabeled data point we simply need to replace $\mathbb{P}_r(\hat{\mathbf{y}}_{ai}|y_i)$ by $\mathbb{P}_r(\hat{\mathbf{y}}_{ai})$. Then, the complete-data log-likelihood $\mathcal{L}$ is equal to

$$\mathcal{L} = \sum_{i=1}^{N} \sum_{r=1}^{R} z_{ir} \big(\log \pi_{ir}(\widetilde{\mathbf{x}}_i) + \log \mathbb{P}_r(\hat{\mathbf{y}}_{ai}|y_i)\big). \tag{32}$$

Expectation–Maximization (EM) algorithm [33] can be used to find the parameters Θ that maximize $\mathcal{L}$ from (32).

2.5.1. EM algorithm for semi-supervised aggregation

Before moving on, we need to decide on the parameterization of the prior probability π_{ir}. We define this probability using a softmax function,

$$\pi_{ir} = \frac{\exp\big(-(\widetilde{\mathbf{x}}_i - \mathbf{q}_r)^{\mathrm{T}} \mathbf{\Lambda}_r (\widetilde{\mathbf{x}}_i - \mathbf{q}_r)\big)}{\sum_{m=1}^{R} \exp\big(-(\widetilde{\mathbf{x}}_i - \mathbf{q}_m)^{\mathrm{T}} \mathbf{\Lambda}_m (\widetilde{\mathbf{x}}_i - \mathbf{q}_m)\big)}, \tag{33}$$

where we defined a prototype vector $\mathbf{q}_r \in \mathbb{R}^{\widetilde{D}}$ and a $\widetilde{D} \times \widetilde{D}$ feature scaling matrix $\mathbf{\Lambda}_r$ for each regime, to be found during optimization, resulting in $\Theta = \{\mathbf{\Sigma}_r, \mathbf{q}_r, \mathbf{\Lambda}_r\}_{r=1,\dots,R}$.

In the E-step, we compute the current expectation of posterior probability h_{ir} that the rth regime is “responsible” for generating expert predictions for the ith labeled point as

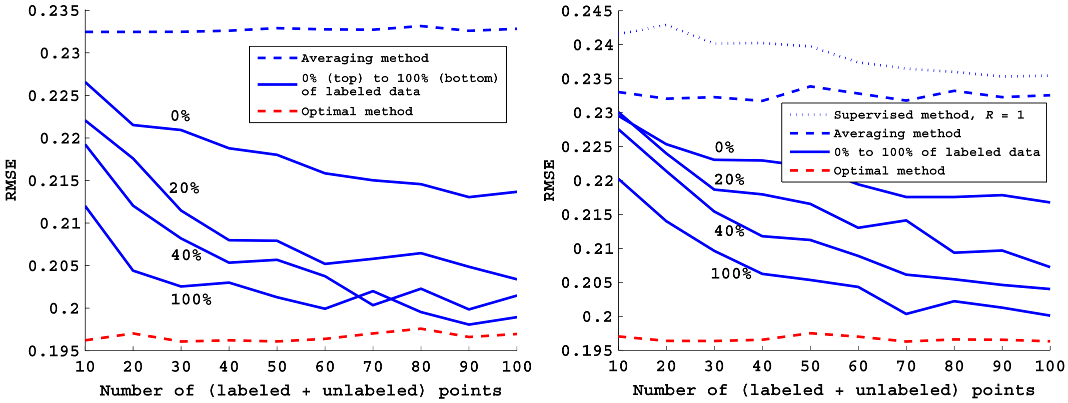
$$h_{ir} = \mathbb{E}[z_{ir}|\hat{\mathbf{y}}_{ai}, y_i, \widetilde{\mathbf{x}}_i, \Theta] = \frac{\pi_{ir}(\widetilde{\mathbf{x}}_i)\, \mathbb{P}_r(\hat{\mathbf{y}}_{ai}|y_i)}{\sum_{m=1}^{R} \pi_{im}(\widetilde{\mathbf{x}}_i)\, \mathbb{P}_m(\hat{\mathbf{y}}_{ai}|y_i)}. \tag{34}$$

In the M-step, we fix values of h_{ir} for all data points and regimes, and optimize $\mathcal{L}$ with respect to covariance matrices $\mathbf{\Sigma}_r$, weight vectors $\mathbf{w}_r$, as well as prototype vectors $\mathbf{q}_r$ and scaling matrices $\mathbf{\Lambda}_r, r = 1, \dots, R$. Note that the derivatives of $\mathcal{L}$ with respect to these two sets of variables are independent from each other, and the optimization of $\mathbf{\Sigma}_r$ and $\mathbf{w}_r$ on one side, and $\mathbf{q}_r$ and $\mathbf{\Lambda}_r$ on the other, can be easily parallelized. After derivation, the update equation for $\mathbf{\Sigma}_r$ can be written as follows,

$$\begin{aligned} \mathbf{\Sigma}_r = \frac{1}{1 + \sum_{i=1}^{N} h_{ir}} \Bigg(& \sigma_{0r}^2 \mathbf{I} + \sum_{i=1}^{N} h_{ir}\, \Pi_i^{-1}(\mathbf{\Psi}_{ir}) + \sum_{i=N_u+1}^{N} h_{ir} [\![(\hat{\mathbf{y}}_{ai} - y_i \mathbf{1})\, (\hat{\mathbf{y}}_{ai} - y_i \mathbf{1})^{\mathrm{T}}]\!]_r \\ & + \sum_{i=1}^{N_u} h_{ir} \bigg([\![\hat{\mathbf{y}}_{ai} \hat{\mathbf{y}}_{ai}^{\mathrm{T}}]\!]_r + \frac{[\![\mathbf{1}\mathbf{1}^{\mathrm{T}}]\!]_r}{\mathbf{1}^{\mathrm{T}} \mathbf{U}'_{ir} \mathbf{1}} + \overline{y}_{ir}^2 [\![\mathbf{1}\mathbf{1}^{\mathrm{T}}]\!]_r - \overline{y}_{ir} [\![\mathbf{1} \hat{\mathbf{y}}_{ai}^{\mathrm{T}} + \hat{\mathbf{y}}_{ai} \mathbf{1}^{\mathrm{T}}]\!]_r \bigg) \Bigg) \odot \begin{bmatrix} \mathbb{1} & \mathbf{0} \\ \mathbf{0} & 1 \end{bmatrix}. \end{aligned} \tag{35}$$

In order to find the weight vector $\mathbf{w}_r$, let us first write out log-likelihood $\mathcal{L}_r$, pertaining to the rth regime. After removing all elements not dependent on $\mathbf{w}_r$ from equations (16) and (18), we obtain the following expression,

$$\mathcal{L}_r = -\frac{1}{2\sigma_r^2} \sum_{i=1}^{N_u} z_{ir} \big(\overline{y}_i - f(\mathbf{x}_i, \mathbf{w}_r)\big)^2 - \frac{1}{2\sigma_r^2} \sum_{i=N_u+1}^{N_u+N_l} z_{ir} \big(y_i - f(\mathbf{x}_i, \mathbf{w}_r)\big)^2 - \frac{\lambda}{2} \mathbf{w}_r^{\mathrm{T}} \mathbf{w}_r, \tag{36}$$



(a) Data generated by $R = 1$ regime

(b) Data generated by $R = 2$ regimes

Fig. 3. Performance of the aggregation methods on the synthetic data set.

where the regularization term weighted by 0.5λ corresponds to an isotropic Gaussian prior over the weight vector $\mathbf{w}_r$, introduced in Section 2.4. A closed-form solution for $\mathbf{w}_r$ can be found using the equation for weighted, regularized linear regression, equal to

$$\mathbf{w}_r = (\mathbf{X}^{\mathrm{T}}\mathbf{H}_r\mathbf{X} + \lambda\sigma_r^2\mathbf{I})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{H}_r\mathbf{y}, \tag{37}$$

where $\mathbf{H}_r$ is a diagonal $N \times N$ weight-matrix with the ith diagonal element equal to h_{ir}.

Lastly, prototype vector $\mathbf{q}_r$ and scaling matrix $\mathbf{\Lambda}_r$ can be found through the gradient ascent optimization using the following update equations,

$$\begin{aligned}
\mathbf{q}_r^{\mathrm{new}} &= \mathbf{q}_r^{\mathrm{old}} + \eta\ \mathbf{\Lambda}_r^{\mathrm{old}} \sum_{i=1}^{N} (h_{ir} - \pi_{ir}^{\mathrm{old}})(\widetilde{\mathbf{x}}_i - \mathbf{q}_r^{\mathrm{old}}), \\
\mathbf{\Lambda}_r^{\mathrm{new}} &= \mathbf{\Lambda}_r^{\mathrm{old}} - \eta \sum_{i=1}^{N} (h_{ir} - \pi_{ir}^{\mathrm{old}})(\widetilde{\mathbf{x}}_i - \mathbf{q}_r^{\mathrm{old}})(\widetilde{\mathbf{x}}_i - \mathbf{q}_r^{\mathrm{old}})^{\mathrm{T}},
\end{aligned} \tag{38}$$

where η is an appropriately set learning rate.

3. Experiments

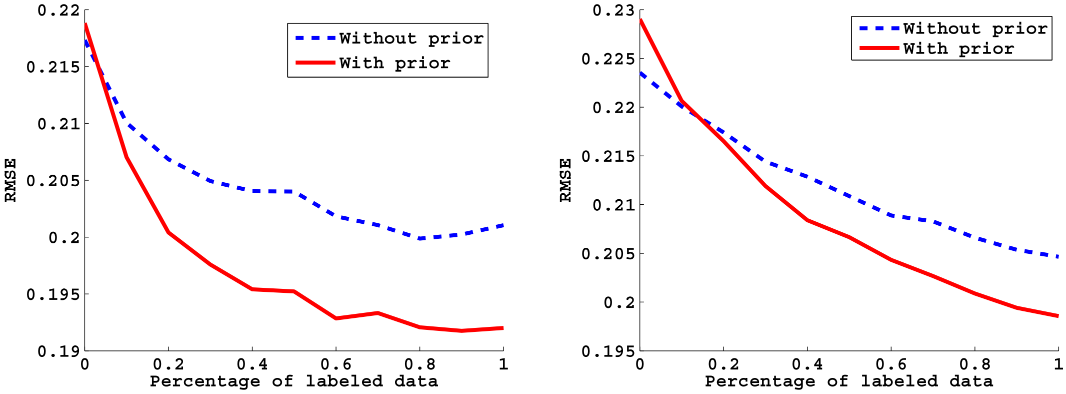
In this section, we first experimentally validate the semi-supervised aggregation on synthetic data, and then apply the method to AOD estimation using real-world aerosol data set. We used Root Mean Squared Error (RMSE) measure to report performance of various aggregation methods, commonly used in AOD research [34]. RMSE is defined as follows,

$$\mathrm{RMSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i - \overline{y}_i)^2}, \tag{39}$$

where the sum is over N labeled data points from the test set.

3.1. Validation on synthetic data

We started by evaluating our method on synthetic data generated as follows: for a given number of regimes R and experts K, we selected weight vector $\mathbf{w}$, prototype $\mathbf{q}$, and a covariance matrix for each regime $\mathbf{\Sigma}$. Then, we assigned the ith data point uniformly at random with probability $1/R$ to a regime, say the lth regime, and obtained features $\mathbf{x}_i$ by sampling from multivariate Gaussian with mean $\mathbf{q}_l$ and covariance matrix $0.5\mathbf{I}$, where for simplicity we also set $\widetilde{\mathbf{x}}_i = \mathbf{x}_i$. We sampled $\mathbf{w}_r, r = 1, \ldots, R$, from multivariate Gaussian with zero mean and unit variance. We sampled ground-truth value y_i from Gaussian with unit variance and mean computed using (1), then sampled K expert predictions from a Gaussian $\mathcal{N}(y_i\mathbf{1}, \mathbf{\Sigma}_l)$. Finally, we removed each expert's prediction with probability 0.5 to simulate missing experts. In all experiments we set $\mathbf{S} = \mathbf{I}$, and used 15 EM iterations. Learning rate η was set through cross-validation, and reported results averaged over 100 experiments. To better characterize the proposed model, in the first set of experiments we assumed uninformative prior (i.e., we set $\sigma \to \infty$ and did not learn $\mathbf{w}_r$ vectors), and evaluated benefits of the informative, context-dependent prior at the end of the section.



(a) Data generated by $R = 1$ regime

(b) Data generated by $R = 2$ regimes

Fig. 4. Performance of the proposed semi-supervised model on the synthetic data set with and without context-dependent prior.

In order to evaluate the semi-supervised method without clustering, we set $K = 5$, $R = 1$, $D = \widetilde{D} = 2$, and $\boldsymbol{\Sigma}_1 = \text{diag}([0.1, 0.2, 0.3, 0.4, 0.5])$. We compared our learning method to a baseline method that averages all available experts, as well as to the optimal predictor that computes the prediction (29) using the true $\boldsymbol{\Sigma}_1$. We increased the number of training points N from 10 to 100 in increments of 10, and for each N we experimented with percentage of labeled data points equal to 0%, 20%, 40%, and 100% (shown as four solid lines in Fig. 3).

The results in terms of RMSE, evaluated on 1,000 testing points generated in the same way as the training set, are shown in Fig. 3a. We can see that the performance of the fully unsupervised approach, given by the top-most full line, is already better than simple averaging, which further improves as the number of unlabeled data grows. Moreover, as we increase the number of labeled points, the semi-supervised method further improves the accuracy, approaching the lower bound on RMSE achieved by the optimal combination of experts.

Next, we generated the data using two regimes by setting $\mathbf{q}_1 = [1, 1]$, $\mathbf{q}_2 = [-1, -1]$, $\boldsymbol{\Sigma}_1 = \text{diag}([0.1, 0.2, 0.3, 0.4, 0.5])$, $\boldsymbol{\Sigma}_2 = \text{diag}([0.5, 0.4, 0.3, 0.2, 0.1])$, and we set $R = 2$. The results in terms of RMSE are given in Fig. 3b, where we also show accuracy of the proposed method which used only labeled data, but assumed only one cluster. The RMSE of supervised method that assumed only a single cluster is worse than simple averaging, and approached it as the data size increased. Unsupervised method using two clusters achieved better accuracy than simple averaging, and RMSE further decreased with larger data sizes. Introduction of labeled data points further decreased the RMSE.

In the following set of experiments we evaluated the benefits of the prior model introduced in (1), and learned both the feature vectors $\mathbf{w}_r$ and variances σ_r, $r = 1, \ldots, R$. We set the number of training data points to 100, and experimented with percentage of labeled data points from 0% to 100%, with 10% increments. We set $\lambda = 10$, and run experiments for $R = 1$ and $R = 2$. The results in terms of RMSE are illustrated in Fig. 4. We can see that the introduction of prior model resulted in significant decrease in the RMSE measure. It is important to note that baseline linear regression alone, trained on labeled data points only, would have RMSE of around 1, much worse than any of the experts. Interestingly, for purely unlabeled data set, the method with informative prior obtained slightly worse result than the method without, which is due to overfitting to the estimated ground-truth values for unlabeled data points (i.e., the prior model learns to predict other experts' predictions), and can be mitigated with stronger regularization when number of labeled data points is small or by increasing size of the training set. However, for both $R = 1$ and $R = 2$, the method with context-dependent prior outperformed the aggregation method with uninformative prior even for small number of available ground-truth labels. Moreover, the performance gap quickly grew as we increased the fraction of labeled data, eventually reaching better RMSE than the optimal method from Fig. 3.

3.2. Validation on aerosol data

In this section we present the results of the experiments on real-world, global aerosol data set. We first describe how the data set was generated, before moving on to the discussion of the performance results.

3.2.1. The data set

We used ground-based AERONET data [15] and global data from 5 satellite instruments spanning 5 years, from 2006 to 2010. For the ground-based AERONET sensors we downloaded data for all sites from AERONET website.[1] To obtain the target AOD values we used only Level 2.0 AOD data, the highest-quality data, as it is pre- and post-field calibrated, automatically cloud-cleared and manually inspected, and take a measurement at 10:30 am as a ground-truth AOD value. Furthermore,

[1] http://aeronet.gsfc.nasa.gov/cgi-bin/combined_data_access_new, accessed October 2015.

Table 1
Performance of the aggregation methods in terms of RMSE.

Continent	Number of sites	Baseline	Supervised	Semi-supervised
Africa	6	0.0737	0.0734	**0.0697**
Asia	11	0.0929	0.0887	**0.0847**
Europe	33	0.0873	0.0690	**0.0648**
North America	32	0.0814	0.0795	**0.0752**
South America	4	0.0916	0.0864	**0.0844**

to obtain feature vectors $\mathbf{x}_i$ we used the 2nd-generation NOAA Global Ensemble Forecast System Reforecast (GEFS/R) data,[2] and considered longitude and latitude of the location, as well as the following measurements: *wind mixing energy*, *skin temperature*, resulting in dimensionality $D = 4$. To obtain partition feature vectors $\widetilde{\mathbf{x}}_i$ we considered longitude and latitude of the location, resulting in dimensionality $\widetilde{D} = 2$.

For the satellite sensors we used Multi-sensor Aerosol Products Sampling System [35],[3] which provides satellite data only when they are collocated with AERONET locations. We used only research-quality AOD estimates, recommended by the teams responsible for the prediction algorithm of each sensor. The following satellite instruments were considered:

- **MODIS** We used Daily Level 2 aerosol product, collection 5.1 (MOD04_L2 and MYD04_L2, for Terra and Aqua, respectively). For this study we only used predictions with Quality Assurance (QA) flag equal to 3, and considered the following Scientific Data Set (SDS): *Corrected_Optical_Depth_Land*. The product was available at around 10:30 am and 1:30 pm local time, at the overpass times of Terra and Aqua satellites, respectively.
- **MISR** We used MIL2ASAE data product, a MISR Level 2 aerosol product. For this study we only used predictions with the Quality Assurance (QAb) flags equal to 0 and 1, and considered the following SDS: *RegBestEstimateSpectralOptDepth*. The product was available at around 10:30 am local time, at the overpass time of Terra satellite.
- **OMI** We used OMAERUV, a Level-2 near-UV aerosol absorption and extinction optical depth and single scattering albedo OMI data product. For this study we only used predictions with the Quality assurance for final algorithm flag (Qafaf) equal to 0, and considered the following SDS: *FinalAerosolOpticalDepth*. The product was available at around 1:30 pm local time, at the overpass time of Aura satellite.
- **SeaWiFS** We used SWDB_L2, Deep Blue Aerosol Optical Depth Daily Level 2 data product. For this study we only used predictions with Quality Assurance (QA) flag equal to 3, and considered the following SDS: *aerosol_optical_thickness_ 550_ land*. The product was available at around 12:20 pm local time, at the overpass time of SeaStar.

We note that we also considered the CALIOP data product. However, due to a very small number of collocated data points that were available, we did not include it in this study.

We considered AOD at 550 nm wavelength. If an instrument did not provide AOD at this wavelength, we performed linear interpolation or extrapolation in the log-scale of predictions at two closest wavelengths to 550 nm [36]. In particular, if τ_a and τ_b are available, where τ_x denotes AOD reported at wavelength x, τ_{550} is calculated as follows,

$$\tau_{550} = \tau_a \cdot \exp\left(\frac{\big(\log(550) - \log(a)\big)\big(\log(\tau_b) - \log(\tau_a)\big)}{\log(b) - \log(a)}\right). \tag{40}$$

We used 5 years of data from 2006 to 2010, where there were, on average, 199 working AERONET sites each year. After removing AERONET sites with too few observations, there remained 86 sites in the data set, with locations shown in Fig. 5. This resulted in a labeled data set with $N = 23{,}119$ data points, where 67% of expert predictions were missing. Both $\mathbf{x}_i$ and $\widetilde{\mathbf{x}}_i$ vectors were always available for all data points. We used this data set for three sets of experiments: (1) evaluating usefulness of unlabeled data; (2) evaluating usefulness of partitioning; and (3) evaluating usefulness of context-dependent prior over AOD values. In all sets of experiments we performed leave-one-site-out cross-validation.

3.2.2. Results

In the first set of experiments, we manually split the data into 5 per-continent subsets (2 Australian sites were assigned to Asian cluster), and then trained a separate aggregation model for each partition. We first randomly selected 2 training AERONET sites and took 50 labeled data points from each of them. Then, we selected 50 unlabeled data points from the remaining training AERONET sites, and sampled 50 labeled data points from the left-out site for validation. We trained one model which used only labeled data, and one that used both labeled and unlabeled data. We used $R = 1$ clusters in both cases, and assumed $\sigma \to \infty$, resulting in an uninformative prior over the AOD values. We compared the performance to a baseline method that takes a simple average of available expert predictions. The results in terms of RMSE are given in Table 1, clearly indicating that the proposed method successfully exploited large amounts of readily available unlabeled

[2] http://esrl.noaa.gov/psd/forecasts/reforecast2/download.html, accessed October 2015.
[3] http://disc.sci.gsfc.nasa.gov/aerosols/services/mapss/mapssdoc.html, accessed October 2015.

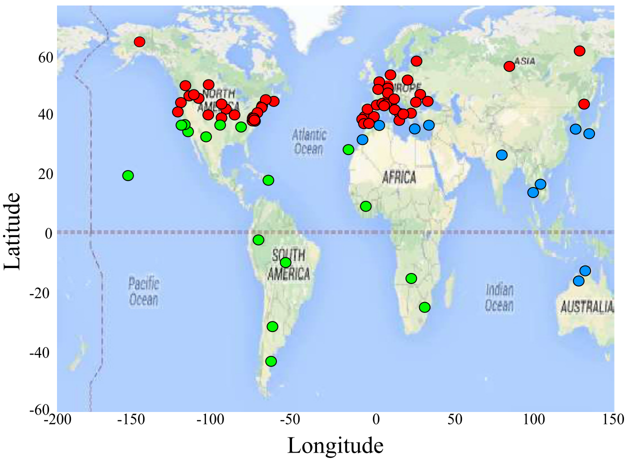


Fig. 5. Found clustering of AERONET sites.

Table 2
Results of the semi-supervised aggregation methods.

Method	# clusters	RMSE
Per-continent partition	5	0.0758
No partitioning	1	0.0728
EM-partitioning	2	0.0720
EM-partitioning	3	**0.0711**
EM-partitioning	4	0.0723
EM-partitioning	5	0.0726
EM-partitioning with prior	3	**0.0689**

data. This is particularly important for the largest continents of Asia and Africa, where the density of AERONET sites is very low and labeled data are extremely scarce. As seen in Table 1, the average improvement of semi-supervised approach over the baseline was over 10%, and around 5% over the purely supervised method.

In the second set of experiments we evaluated the benefits of data partitioning using the EM algorithm. We randomly sampled from each training site 50 labeled and 50 unlabeled data points, and from the left-out site sampled 50 labeled data points used for validation, assuming $\sigma \to \infty$. We used our proposed method with $R \in \{1, 2, 3, 4, 5\}$, repeating the experiments 5 times. RMSE is reported in the middle of Table 2, where at the top we also report the RMSE of the semi-supervised method using the per-continent partition considered earlier.

We can see that semi-supervised aggregation with partitioning using latent variables had significantly lower RMSE than the baseline where we manually split the data into per-continent subsets using domain knowledge. By increasing the number of clusters from 1 (i.e., without partitioning) to 5 we observed a drop in RMSE of around 6% over the baseline, achieving the best performance for $R = 3$ clusters. In Fig. 5 we color-code the AERONET sites according to their cluster assignments for this case. Interestingly, the shown partition roughly corresponds to clustering found by earlier studies that considered aerosol properties [30], where south-western cluster contains sites with mostly absorbing aerosols, south-eastern cluster sites with mostly non-absorbing aerosols, and northern cluster contains sites with mixed moderately absorbing and non-absorbing aerosols. Given predictions of all the experts, for the northern cluster the weights of linear combination assigned to MISR, Terra MODIS, Aqua MODIS, OMI, and SeaWiFS instruments were $[0.37, 0.19, 0.19, 0.13, 0.12]$, for the south-eastern cluster they were $[0.46, 0.18, 0.18, 0.05, 0.13]$, for the south-western cluster they were $[0.44, 0.15, 0.15, 0.10, 0.16]$, respectively. In concordance with the domain knowledge, MISR obtained the largest weights, while other instruments were given similar weights, with the exception of OMI which consistently had the lowest weight in all clusters.

Lastly, we evaluated the benefits of the linear regression prior over the AOD values, introduced in equation (1). Here we dropped the assumption that $\sigma \to \infty$, set $R = 3$ due to the results shown in the middle part of Table 2, and during training learned both weight vectors $\mathbf{w}_r$ and prior variances σ_r, $r = \{1, 2, 3\}$. In order to partition the data set we used longitude and latitude of the AERONET sites as in the previous experiment, while for explanatory features $\mathbf{x}_i$ from (1) we used 4 ground measurements discussed in Section 3.2.1. The RMSE performance of the final model is given at the bottom of Table 2, where we can see further improvements over the methods that assumed uninformative prior. In particular, compared to the best previous model, RMSE dropped by around 3.2%, and it is interesting to note that the learned weight of the linear combination for the prior model was roughly similar to the weights of Terra MODIS and Aqua MODIS. An additional benefit of the assumed prior is the ability to provide AOD estimates even when there are no available satellite experts, by using the available non-AOD ground-based measurements through equation (1). We can conclude that the results presented in this section confirm the validity of the proposed semi-supervised method for aggregation of experts, which is able to account for missing experts, find a partition of data into clusters, exploit additional available information through the prior model, and construct specialized aggregators on each cluster.

4. Conclusion and future work

Aerosol Optical Depth is an important input parameter to complex climate models developed to predict and better understand Earth's climate. Consequently, accurate AOD estimation is the problem of global significance, and in this paper we address this task by proposing a semi-supervised method for aggregation of AOD predictions from noisy satellite-borne sensors into a single, more accurate estimate. By assuming that expert predictions follow multivariate Gaussian distribution, the method accounts for both missing experts and unlabeled data in a principled manner, addressing an issue inherent to the remote sensing domain. Furthermore, by introducing a context-dependent AOD prior, the model is capable of incorporating additional data sources and providing estimates even when there are no available experts. Lastly, we also cluster the data during training by introducing a latent indicator variable for each cluster, resulting in a more interpretable model. Results on a synthetic data set and real-world aerosol data set comprising 5 satellite-borne sensors strongly indicate the benefits of the proposed aggregation method.

The described approach makes several assumptions that resulted in a simplified model of aerosol distribution, yet do not necessarily hold in practice. In particular, we assumed that the data points are sampled IID, and did not account for spatio-temporal correlations that inherently exist in the aerosol data [37,38]. In addition, the covariance matrix in this work is assumed static, and is not affected by the context in which the data points appear. Lastly, we assumed that errors in measurements are Gaussian, and that all experts have equal, zero biases. We are investigating approaches where we weaken or altogether drop these simplifying assumptions in order to extend and further improve modeling of aerosol distribution, and leave these ideas for future work.

Acknowledgements

This work was supported by the National Science Foundation grant IIS-1117433. The authors would like to thank the AERONET Principal Investigators and their staff for establishing and maintaining the sites used in this investigation.

References

[1] N. Djuric, L. Kansakar, S. Vucetic, Semi-supervised learning for integration of aerosol predictions from multiple satellite instruments, in: Proceedings of International Joint Conference on Artificial Intelligence, 2013, pp. 2797–2803.

[2] Intergovernmental Panel on Climate Change, Climate change 2007: the physical science basis, Agenda 6 (07).

[3] R.J. Charlson, J.E. Lovelock, M.O. Andreae, S.G. Warren, et al., Oceanic phytoplankton, atmospheric sulphur, cloud albedo and climate, Nature 326 (6114) (1987) 655–661.

[4] R.J. Charlson, S. Schwartz, J. Hales, R.D. Cess, J. Coakley, J. Hansen, D. Hofmann, Climate forcing by anthropogenic aerosols, Science 255 (5043) (1992) 423–430.

[5] M.O. Andreae, P.J. Crutzen, Atmospheric aerosols: biogeochemical sources and role in atmospheric chemistry, Science 276 (5315) (1997) 1052–1058.

[6] D. Randall, R. Wood, S. Bony, R. Colman, T. Fichefet, J. Fyfe, V. Kattsov, A. Pitman, J. Shukla, J. Srinivasan, et al., Climate models and their evaluation, Climate Change 323.

[7] Y. Liu, C. Paciorek, P. Koutrakis, Estimating regional spatial and temporal variability of $PM_{2.5}$ concentrations using satellite data, meteorology, and land use information, Environ. Health Perspect. 117 (6) (2009) 886.

[8] Z. Hu, K. Rao, Particulate air pollution and chronic ischemic heart disease in the eastern United States: a county level ecological study using satellite aerosol data, Environ. Health 8 (1) (2009) 26.

[9] M. King, W. Menzel, Y. Kaufman, D. Tanré, B.-C. Gao, S. Platnick, S. Ackerman, L. Remer, R. Pincus, P. Hubanks, Cloud and aerosol properties, precipitable water, and profiles of temperature and water vapor from MODIS, IEEE Trans. Geosci. Remote Sens. 41 (2) (2003) 442–458.

[10] D.J. Diner, J.C. Beckert, T.H. Reilly, C.J. Bruegge, J.E. Conel, R.A. Kahn, J.V. Martonchik, T.P. Ackerman, R. Davies, S.A.W. Gerstl, H.R. Gordon, J.P. Muller, R.B. Myneni, P.J. Sellers, B. Pinty, M.M.V. Verstraete, Multi-angle Imaging SpectroRadiometer (MISR) – instrument description and experiment overview, IEEE Trans. Geosci. Remote Sens. 36 (1998) 1072–1087.

[11] O. Torres, R. Decae, J.P. Veefkind, G. de Leeuw, OMI aerosol retrieval algorithm, in: P. Stammes (Ed.), OMI Algorithm Theoretical Basis Document, Volume III, Clouds, Aerosols, and Surface UV Irradiance, 2002, pp. 41–71.

[12] M. Wang, S. Bailey, C.R. McClain, SeaWiFS provides unique global aerosol optical property data, Eos 81 (18) (2000) 197–202.

[13] D.M. Winker, M.A. Vaughan, A. Omar, Y. Hu, K.A. Powell, Z. Liu, W.H. Hunt, S.A. Young, Overview of the CALIPSO mission and CALIOP data processing algorithms, J. Atmos. Ocean. Technol. 26 (11) (2009) 2310–2323.

[14] M.I. Mishchenko, L. Liu, I.V. Geogdzhayev, L.D. Travis, B. Cairns, A.A. Lacis, Toward unified satellite climatology of aerosol properties, 3: MODIS versus MISR versus AERONET, J. Quant. Spectrosc. Radiat. Transf. 111 (2010) 540–552.

[15] B.N. Holben, T.F. Eck, I. Slutsker, D. Tanré, J.P. Buis, A. Setzer, E. Vermote, J.A. Reagan, Y.J. Kaufman, T. Nakajima, F. Lavenu, I. Jankowiak, A. Smirnov, AERONET – a federated instrument network and data archive for aerosol characterization, Remote Sens. Environ. 66 (1) (1998) 1–16.

[16] F. Giorgi, L.O. Mearns, Probability of regional climate change based on the Reliability Ensemble Averaging (REA) method, Geophys. Res. Lett. 30 (12) (2003).

[17] C. Tebaldi, R. Knutti, The use of the multi-model ensemble in probabilistic climate projections, Philos. Trans. R. Soc., Math. Phys. Eng. Sci. 365 (1857) (2007) 2053–2075.

[18] A. Braverman, H. Nguyen, E. Olsen, C. Miller, N. Cressie, M. Katzfuss, R. Wang, A. Michalak, Geostatistical data fusion for remote sensing applications, NASA Annual Report, 2011 Report from the ESTO Advanced Information Systems Technology (AIST) Program.

[19] Aviso+: satellite altimetry data, http://www.aviso.altimetry.fr/en/data/products.html, accessed 2015-07-23.

[20] R.T. Clemen, R.L. Winkler, Combining probability distributions from experts in risk analysis, Risk Anal. 19 (2) (1999) 187–203.

[21] B.T. Bartell, G.W. Cottrell, R.K. Belew, Automatic combination of multiple ranked retrieval systems, in: Proceedings of International ACM SIGIR Conference on Research and Development in Information Retrieval, Springer-Verlag, New York, Inc., 1994, pp. 173–181.

[22] A.J.C. Sharkey, On combining artificial neural nets, Connect. Sci. 8 (3–4) (1996) 299–314.

[23] P. Welinder, S. Branson, S. Belongie, P. Perona, The multidimensional wisdom of crowds, in: Proceedings of Neural Information Processing Systems, vol. 10, 2010, pp. 2424–2432.

[24] D. Zhou, J.C. Platt, S. Basu, Y. Mao, Learning from the wisdom of crowds by minimax entropy, in: Proceedings of Neural Information Processing Systems, 2012, pp. 2204–2212.

[25] H. Kajino, Y. Tsuboi, H. Kashima, Clustering crowds, in: Proceedings of AAAI Conference on Artificial Intelligence, 2013, pp. 1120–1127.

[26] J.M. Bates, C.W.J. Granger, The combination of forecasts, Oper. Res. Q. 20 (4) (1969) 451–468.

[27] C.W.J. Granger, R. Ramanathan, Improved methods of combining forecasts, J. Forecast. 3 (2) (1984) 197–204.

[28] K. Ristovski, D. Das, V. Ouzienko, Y. Guo, Z. Obradovic, Regression learning with multiple noisy oracles, in: Proceedings of European Conference on Artificial Intelligence, 2010, pp. 445–450.

[29] V.C. Raykar, S. Yu, L.H. Zhao, G.H. Valadez, C. Florin, L. Bogoni, L. Moy, Learning from crowds, J. Mach. Learn. Res. 11 (2010) 1297–1322.

[30] R.C. Levy, L.A. Remer, O. Dubovik, Global aerosol optical properties and application to Moderate Resolution Imaging SpectroRadiometer aerosol retrieval over land, J. Geophys. Res. 112 (2007) 13,210–13,224.

[31] J. Brewer, Kronecker products and matrix calculus in system theory, IEEE Trans. Circuits Syst. 25 (9) (1978) 772–781.

[32] A.S. Weigend, M. Mangeas, N. Ashok, Nonlinear gated experts for time series: discovering regimes and avoiding overfitting, Int. J. Neural Syst. 6 (04) (1995) 373–399.

[33] A.P. Dempster, N.M. Laird, D.B. Rubin, Maximum likelihood from incomplete data via the EM algorithm, J. R. Stat. Soc. B 39 (1) (1977) 1–38.

[34] D. Chu, Y. Kaufman, C. Ichoku, L. Remer, D. Tanré, B. Holben, Validation of MODIS aerosol optical depth retrieval over land, Geophys. Res. Lett. 29 (12) (2002), MOD2-1–MOD2-4.

[35] M. Petrenko, C. Ichoku, G. Leptoukh, Multi-sensor aerosol products sampling system (MAPSS), Atmos. Meas. Tech. 5 (5) (2012) 913–926.

[36] L.A. Remer, Y.J. Kaufman, D. Tanré, S. Mattoo, D.A. Chu, J.V. Martins, R.-R. Li, C. Ichoku, R.C. Levy, R.G. Kleidman, T.F. Eck, E. Vermote, B.N. Holben, The MODIS aerosol algorithm, products, and validation, J. Atmos. Sci. 62 (2005) 947–973.

[37] N. Djuric, V. Radosavljevic, Z. Obradovic, S. Vucetic, Gaussian Conditional Random Fields for aggregation of operational aerosol retrievals, IEEE Geosci. Remote Sens. Lett. 12 (4) (2015) 761–765.

[38] V. Radosavljevic, S. Vucetic, Z. Obradovic, Continuous Conditional Random Fields for regression in remote sensing, in: Proceedings of European Conference on Artificial Intelligence, 2010, pp. 809–814.