Semi-Supervised Learning for Integration of Aerosol Predictions from Multiple Satellite Instruments

Nemanja Djuric, Lakesh Kansakar, Slobodan Vucetic
Department of Computer and Information Sciences, Temple University, USA
{nemanja.djuric, lakesh, slobodan.vucetic}@temple.edu

Abstract

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 method for learning how to aggregate AOD estimations from multiple satellite instruments into a more accurate estimation. The proposed method is semi-supervised, as it is able to learn from a small number of labeled data, where labels come from a few accurate and expensive ground-based instruments, and a large number of unlabeled data. The 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 AOD estimations from 5 instruments aboard 4 satellites, and the results indicate that 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. AOD is an important input to climate models, and it can significantly impact predictions of future climate changes [Randall *et al.*, 2007]. Considering that climate predictions influence decisions of policy makers, accurate AOD estimation is a task of global significance. In addition to its impact on climate studies, AOD is an important quantity in estimation of air pollution. For example, it was shown in [Liu *et al.*, 2009] that AOD is an accurate predictor of $PM_{2.5}$, the concentration of particulate matter with aerodynamic diameters $\leq 2.5\mu m$, which poses a serious health hazard to the population [Hu and Rao, 2009].

Currently, a number of instruments aboard several Earth-observing satellites report their AOD estimates, such as MODIS instrument aboard Terra and Aqua satellites [King *et al.*, 2003], MISR aboard Terra [Diner *et al.*, 1998], OMI aboard Aura [Torres *et al.*, 2002], SeaWiFS aboard SeaStar [Wang *et al.*, 2000], 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 Figure 1, width of the field of view of MODIS instrument is $2{,}330km$, allowing MODIS to observe the entire Earth every day, as opposed to $360km$ 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 [Mishchenko *et al.*, 2010]. 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 [Holben *et al.*, 1998], placed at several hundred unevenly distributed locations across the globe, see Figure 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 the satellite-borne instruments mean that they observe and measure different, possibly complementary aspects of the same phenomenon. 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 AOD estimation. A recent study by [Mishchenko *et al.*, 2010] confirmed this hypothesis by illustrating that simple average of collocated Terra MODIS and MISR AOD estimations resulted in improved accuracy.

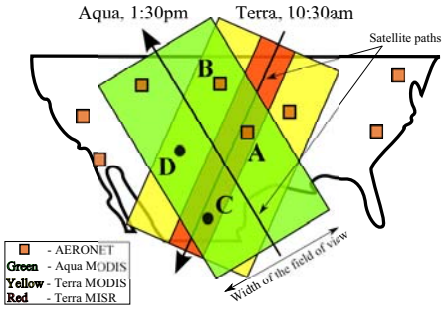


Figure 1: Coverage of instruments over the USA

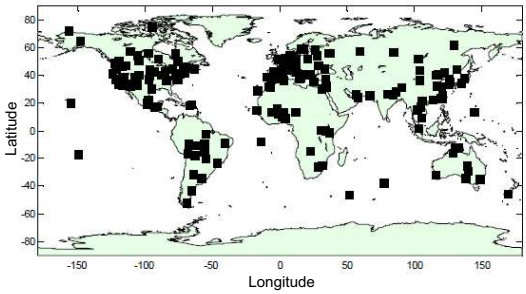


Figure 2: Global coverage of AERONET

The combination of experts that ultimately yields an estimate that is more accurate than any of the individual forecasts is a well-researched topic. Assuming the Gaussian distribution of prediction errors of AOD and no missing experts, [Bates and Granger, 1969; Granger and Ramanathan, 1984] proposed how to learn the optimal combination of experts from labeled data. If data set is unlabeled, [Ristovski *et al.*, 2010] proposed how to learn a combination of experts by extending the classification-based method from [Raykar *et al.*, 2009]. However, the approach assumed that experts are independent, and that all experts are available for aggregation.

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 [Levy *et al.*, 2007]. 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 Figure 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. This opens a question of learning from data with significant amount of missing AOD estimations.

In this paper, we assume that estimation errors of individual satellite instruments have 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. We note that our method can be seen as a significant generalization of the traditional supervised method for combination of experts by [Bates and Granger, 1969; Granger and Ramanathan, 1984], as well as of recently proposed unsupervised method for averaging of experts in regression by [Ristovski *et al.*, 2010].

2 Methodology

2.1 Problem setup and assumptions

Let us assume that we have a training data set $\mathcal{D} = \{\{\hat{y}_{ik}\}_{k=1,\ldots,K}, y_i\}_{i=1,\ldots,N}$, where target value y_i for the i^{th} data point is predicted by K experts, with the k^{th} expert providing an opinion in a form of prediction $\hat{y}_{ik}$. For example, in the aerosol domain that we study, the experts are satellite instruments and the predictions are their individual AOD estimates. We assume that data points are independent and identically distributed (IID), and that ground truth y_i is normally distributed with mean μ_y and variance σ_y^2,

$$y_i \sim \mathcal{N}(\mu_y, \sigma_y^2). \tag{1}$$

We also assume 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 $i = (N_u + 1), \ldots, N$, with $N = (N_u + N_l)$. We use $\mathbf{1}$ to denote a column-vector of all ones, $\mathbf{0}$ to denote a matrix of all zeros, and $\hat{\mathbf{y}}_i = [\hat{y}_{i1}, \ldots, \hat{y}_{iK}]^{\mathrm{T}}$ to denote a column-vector of expert predictions for the i^{th} data point. We assume expert predictions for the i^{th} data point are sampled from a multivariate normal distribution as

$$\hat{\mathbf{y}}_i | y_i \sim \mathcal{N}(y_i \mathbf{1}, \mathbf{\Sigma}). \tag{2}$$

This assumption allows the experts to be correlated (i.e., $\mathbf{\Sigma}$ is non-diagonal), as is the case in aerosol domain. We 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 $\mathbf{\Sigma}$, μ_y, and σ_y^2. By $\Theta = \{\mathbf{\Sigma}, \mu_y, \sigma_y^2\}$ we denote a set of parameters to be learned.

Once Θ is learned, and given expert predictions $\hat{\mathbf{y}}_i$, aggregated prediction y_i for the i^{th} data point can be found as a mean of the posterior distribution $y_i | \hat{\mathbf{y}}_i \sim \mathcal{N}(\overline{y}_i, (\mathbf{1}^{\mathrm{T}} \mathbf{\Sigma'}^{-1} \mathbf{1})^{-1})$, where mean $\overline{y}_i$ is computed as

$$\overline{y}_i = \frac{\hat{\mathbf{y}}_i'^{\mathrm{T}} \mathbf{\Sigma'}^{-1} \mathbf{1}}{\mathbf{1}^{\mathrm{T}} \mathbf{\Sigma'}^{-1} \mathbf{1}}, \tag{3}$$

$\hat{\mathbf{y}}'_i = [\hat{\mathbf{y}}_i^{\mathrm{T}}, \mu_y]^{\mathrm{T}}$, and $\boldsymbol{\Sigma}'$ is a $(K+1) \times (K+1)$ block matrix

$$\boldsymbol{\Sigma}' = \begin{bmatrix} \boldsymbol{\Sigma} & \mathbf{0} \\ \mathbf{0} & \sigma_y^2 \end{bmatrix}. \tag{4}$$

2.2 Semi-supervised combination of experts

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

$$\mathbb{P}(\mathcal{D}|\Theta) = \mathbb{P}(\mathcal{D}_u|\Theta) \cdot \mathbb{P}(\mathcal{D}_l|\Theta), \tag{5}$$

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

$$\mathbb{P}(\mathcal{D}_u|\Theta) = \prod_{i=1}^{N_u} \mathbb{P}(\hat{\mathbf{y}}_i|\Theta) = \prod_{i=1}^{N_u} \int_y \mathbb{P}(\hat{\mathbf{y}}_i|y,\Theta)\, \mathbb{P}(y|\Theta)\, dy. \tag{6}$$

As both probabilities under the integral are assumed Gaussian, by solving the integral we obtain

$$\begin{aligned} \mathbb{P}(\mathcal{D}_u|\Theta) = \prod_{i=1}^{N_u} \Big(& \sqrt{\frac{|\boldsymbol{\Sigma}'|^{-1}}{(2\pi)^{K-1}\mathbf{1}^{\mathrm{T}}\boldsymbol{\Sigma}'^{-1}\mathbf{1}}} \\ & \exp\big(-\frac{1}{2}(\hat{\mathbf{y}}'_i - \overline{y}_i\mathbf{1})^{\mathrm{T}}\boldsymbol{\Sigma}'^{-1}(\hat{\mathbf{y}}'_i - \overline{y}_i\mathbf{1})\big)\Big). \end{aligned} \tag{7}$$

Prior parameters μ_y and σ_y^2, appearing in $\hat{\mathbf{y}}'_i$ and $\boldsymbol{\Sigma}'$, respectively, can be fixed to some values, such as mean and variance of the available target values, or could be learned. In order to keep the notation simple, in the remainder of the section we assume $\sigma_y^2 \to \infty$, which amounts to an uninformative prior $\mathbb{P}(y|\Theta)$. We note that it is straightforward to modify the following expressions for finite σ_y^2, or to derive a learning rule.

Likelihood of the labeled part can be written as

$$\mathbb{P}(\mathcal{D}_l|\Theta) = \prod_{i=N_u+1}^{N} \mathbb{P}(\hat{\mathbf{y}}_i|y_i,\Theta), \tag{8}$$

which, due to (2), is a product of N_l multivariate Gaussians. Then, combining equations (5), (7), and (8), we can compute the likelihood of the data set $\mathcal{D}$. After finding the derivative of the log-likelihood with respect to $\boldsymbol{\Sigma}^{-1}$ and equating the resulting expression with zero, we obtain the following expression for computing $\boldsymbol{\Sigma}$ matrix,

$$\begin{aligned} \boldsymbol{\Sigma} = \frac{1}{N}\Big(& (\hat{\mathbf{Y}}_l - \mathbf{y}_l\,\mathbf{1}^{\mathrm{T}})^{\mathrm{T}}(\hat{\mathbf{Y}}_l - \mathbf{y}_l\,\mathbf{1}^{\mathrm{T}}) + \hat{\mathbf{Y}}_u^{\mathrm{T}}\hat{\mathbf{Y}}_u + \\ & \frac{N_u\mathbf{1}\mathbf{1}^{\mathrm{T}}}{\mathbf{1}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\mathbf{1}} + \sum_{i=1}^{N_u}\big(\overline{y}_i^2\mathbf{1}\mathbf{1}^{\mathrm{T}} - \overline{y}_i(\mathbf{1}\hat{\mathbf{y}}_i^{\mathrm{T}} + \hat{\mathbf{y}}_i\mathbf{1}^{\mathrm{T}})\big)\Big), \end{aligned} \tag{9}$$

where $\hat{\mathbf{Y}}_u$ and $\hat{\mathbf{Y}}_l$ are $N_u \times K$ and $N_l \times K$ 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 (9) yields an iterative procedure for learning $\boldsymbol{\Sigma}$, where $\boldsymbol{\Sigma}$ on the l.h.s. is a new value, and $\boldsymbol{\Sigma}$ on the r.h.s. is an old value.

2.3 Missing experts

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

$$\Pi_i(\boldsymbol{\Sigma}^{-1}) = \begin{bmatrix} \mathbf{U} & \mathbf{V} \\ \mathbf{V}^{\mathrm{T}} & \mathbf{Q} \end{bmatrix}, \tag{10}$$

where Π_i is a permutation function used to reorder both rows and columns of $\boldsymbol{\Sigma}^{-1}$ according to the i^{th} data point, and $\mathbf{U}$ is an $a \times a$ matrix. Given the covariance matrix $\boldsymbol{\Sigma}$ and a vector of expert predictions $\hat{\mathbf{y}}_{ai}$ for the i^{th} 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

$$\overline{y}_i = \frac{\hat{\mathbf{y}}_{ai}^{\mathrm{T}}\mathbf{U}'_i\mathbf{1}}{\mathbf{1}^{\mathrm{T}}\mathbf{U}'_i\mathbf{1}}. \tag{11}$$

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 i^{th} data point.

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

$$\begin{aligned} \mathbb{P}(\hat{\mathbf{y}}_{ai}|\Theta) &= \int_{\hat{\mathbf{y}}_{qi}} \mathbb{P}([\hat{\mathbf{y}}_{ai}^{\mathrm{T}}, \hat{\mathbf{y}}_{qi}^{\mathrm{T}}]^{\mathrm{T}}|\Theta)\, d\hat{\mathbf{y}}_{qi} \\ &= \int_y \int_{\hat{\mathbf{y}}_{qi}} \mathbb{P}([\hat{\mathbf{y}}_{ai}^{\mathrm{T}}, \hat{\mathbf{y}}_{qi}^{\mathrm{T}}]^{\mathrm{T}}|y,\Theta)\, \mathbb{P}(y|\Theta)\, d\hat{\mathbf{y}}_{qi}\, dy. \end{aligned} \tag{12}$$

Solving the equation (12) we obtain

$$\begin{aligned} \mathbb{P}(\hat{\mathbf{y}}_{ai}|\Theta) = & \sqrt{\frac{|\boldsymbol{\Sigma}|^{-1}|\mathbf{Q}_i|^{-1}}{(2\pi)^{K+q-1}\mathbf{1}^{\mathrm{T}}\mathbf{U}'_i\mathbf{1}}} \\ & \exp\big(-\frac{1}{2}(\hat{\mathbf{y}}_{ai} - \overline{y}_i\mathbf{1})^{\mathrm{T}}\mathbf{U}'_i(\hat{\mathbf{y}}_{ai} - \overline{y}_i\mathbf{1})\big). \end{aligned} \tag{13}$$

In a very similar manner we can find the probability of observing the i^{th} labeled data point. It follows

$$\mathbb{P}(\hat{\mathbf{y}}_{ai}|y_i,\Theta) = \int_{\hat{\mathbf{y}}_{qi}} \mathbb{P}([\hat{\mathbf{y}}_{ai}^{\mathrm{T}}, \hat{\mathbf{y}}_{qi}^{\mathrm{T}}]^{\mathrm{T}}|y_i,\Theta)\, d\hat{\mathbf{y}}_{qi}, \tag{14}$$

which, after solving the integral, results in

$$\hat{\mathbf{y}}_{ai}|y_i \sim \mathcal{N}(y_i\,\mathbf{1}, \mathbf{U}'^{-1}_i). \tag{15}$$

By combining equations (5), (13), and (15), we can find the likelihood of the data set $\mathcal{D}$. After finding derivative of the log-likelihood with respect to $\boldsymbol{\Sigma}^{-1}$ [Brewer, 1978] and equating the resulting expression with zero, we obtain the following expression for computing $\boldsymbol{\Sigma}$ matrix

$$\begin{aligned} \boldsymbol{\Sigma} = \tfrac{1}{N}\Bigg(& \textstyle\sum_{i=1}^{N} \Pi_i^{-1}(\boldsymbol{\Psi}_i) + \sum_{i=N_u+1}^{N} [\![(\hat{\mathbf{y}}_{ai} - y_i\,\mathbf{1})\,(\hat{\mathbf{y}}_{ai} - y_i\,\mathbf{1})^{\mathrm{T}}]\!] + \\ & \textstyle\sum_{i=1}^{N_u}\Big([\![\hat{\mathbf{y}}_{ai}\hat{\mathbf{y}}_{ai}^{\mathrm{T}}]\!] + \frac{[\![\mathbf{1}\mathbf{1}^{\mathrm{T}}]\!]}{\mathbf{1}^{\mathrm{T}}\mathbf{U}'_i\mathbf{1}} + \overline{y}_i^2[\![\mathbf{1}\mathbf{1}^{\mathrm{T}}]\!] - \overline{y}_i[\![\mathbf{1}\hat{\mathbf{y}}_{ai}^{\mathrm{T}} + \hat{\mathbf{y}}_{ai}\mathbf{1}^{\mathrm{T}}]\!]\Big)\Bigg), \end{aligned} \tag{16}$$

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 $K \times K$ matrix $\mathbf{\Psi}_i$ is equal to

$$\mathbf{\Psi}_i = \begin{bmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{Q}_i^{-1} \end{bmatrix}, \tag{17}$$

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

$$[\![\mathbf{A}_i]\!] = \Pi_i^{-1}\Big(\begin{bmatrix} \mathbf{A}_i & -\mathbf{A}_i\,\mathbf{V}_i\,\mathbf{Q}_i^{-1} \\ -\mathbf{Q}_i^{-1}\,\mathbf{V}_i^{\mathrm{T}}\,\mathbf{A}_i & \mathbf{Q}_i^{-1}\mathbf{V}_i^{\mathrm{T}}\,\mathbf{A}_i\,\mathbf{V}_i\mathbf{Q}_i^{-1} \end{bmatrix} \Big). \tag{18}$$

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

Let us consider the case where we have some prior knowledge about $\mathbf{\Sigma}$, and would like to include this knowledge into the model. As we assumed $\sigma_y^2 \to \infty$, it follows $\Theta = \{\mathbf{\Sigma}\}$, and we can write

$$\mathbb{P}(\mathcal{D}, \Theta) = \mathbb{P}(\mathcal{D}|\mathbf{\Sigma}^{-1})\,\mathbb{P}(\mathbf{\Sigma}^{-1}). \tag{19}$$

Note that we defined prior $\mathbb{P}(\mathbf{\Sigma}^{-1})$ in terms of an inverse of the covariance matrix (i.e., in terms of a precision matrix). For the precision matrix $\mathbf{\Sigma}^{-1}$ we choose the prior as a Wishart distribution $\mathcal{W}(\mathbf{S}, n)$ with given $K \times K$ scale matrix $\mathbf{S}$ and $n > (K-1)$ degrees of freedom, resulting in

$$\mathbb{P}(\mathbf{\Sigma}^{-1}) = \frac{|\mathbf{\Sigma}^{-1}|^{0.5(n-K-1)} \exp\big(-0.5\,\mathrm{Tr}(\mathbf{S}^{-1}\mathbf{\Sigma}^{-1})\big)}{2^{0.5nK}|\mathbf{S}|^{0.5n}\Gamma_K(0.5n)}, \tag{20}$$

which is a conjugate prior for multivariate Gaussian distribution, and where Γ_K is the multivariate gamma function. After choosing $n = (K+2)$ and finding the derivative of the log-likelihood with respect to $\mathbf{\Sigma}^{-1}$, we obtain the following update equation for the covariance matrix $\mathbf{\Sigma}$

$$\mathbf{\Sigma} = \frac{\mathbf{S}^{-1} + \sum_{i=1}^{N} \Pi_i^{-1}(\mathbf{\Psi}_i) + \sum_{i=N_u+1}^{N} [\![(\hat{\mathbf{y}}_{ai} - y_i\,\mathbf{1})\,(\hat{\mathbf{y}}_{ai} - y_i\,\mathbf{1})^{\mathrm{T}}]\!]}{1+N} + \frac{\sum_{i=1}^{N_u}([\![\hat{\mathbf{y}}_{ai}\hat{\mathbf{y}}_{ai}^{\mathrm{T}}]\!] + \frac{[\![\mathbf{1}\mathbf{1}^{\mathrm{T}}]\!]}{\mathbf{1}^{\mathrm{T}}\mathbf{U}_i'\mathbf{1}} + \overline{y}_i^2[\![\mathbf{1}\mathbf{1}^{\mathrm{T}}]\!] - \overline{y}_i[\![\mathbf{1}\hat{\mathbf{y}}_{ai}^{\mathrm{T}} + \hat{\mathbf{y}}_{ai}\mathbf{1}^{\mathrm{T}}]\!])}{1+N}. \tag{21}$$

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 multivariate Gaussian from (2) [Weigend *et al.*, 1995]. In the following we assume there are R regimes, and that we have available a feature vector $\mathbf{x}_i$ for the i^{th} data point that could be used to assign it to an appropriate regime.

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

$$\mathbb{P}(\hat{\mathbf{y}}_{ai}|\mathbf{x}_i, y_i, \Theta) = \sum_{r=1}^{R} \mathbb{P}_r(\hat{\mathbf{y}}_{ai}|y_i)\;\pi_{ir}(\mathbf{x}_i), \tag{22}$$

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

$$\mathbb{P}(\hat{\mathbf{y}}_{ai}|\mathbf{x}_i, \Theta) = \sum_{r=1}^{R} \mathbb{P}_r(\hat{\mathbf{y}}_{ai})\;\pi_{ir}(\mathbf{x}_i). \tag{23}$$

Probability of observing the i^{th} unlabeled or labeled data point given that it was generated by the r^{th} 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 (13) and (15), respectively. The aggregated prediction $\overline{y}_i$ can be found as

$$\overline{y}_i = \mathbb{E}[y_i|\hat{\mathbf{y}}_{ai}, \mathbf{x}_i, \Theta] = \sum_{r=1}^{R} \pi_{ir}(\mathbf{x}_i)\;\frac{\hat{\mathbf{y}}_{ai}^{\mathrm{T}}\mathbf{U}_{ir}'\mathbf{1}}{\mathbf{1}^{\mathrm{T}}\mathbf{U}_{ir}'\mathbf{1}}. \tag{24}$$

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

$$z_{ir} = \begin{cases} 1 & \text{if } \hat{\mathbf{y}}_{ai} \text{ was generated by the } r^{\text{th}} \text{ regime,} \\ 0 & \text{otherwise.} \end{cases} \tag{25}$$

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

$$\mathbb{P}(\hat{\mathbf{y}}_{ai}, \mathbf{z}_i|\mathbf{x}_i, y_i, \Theta) = \prod_{r=1}^{R} \big(\pi_{ir}(\mathbf{x}_i)\;\mathbb{P}_r(\hat{\mathbf{y}}_{ai}|y_i)\big)^{z_{ir}}. \tag{26}$$

Note that, due to the lack of space, in equations (26), (27), and (29), we only give expressions for labeled data. However, when dealing with the i^{th} 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

$$\mathcal{L} = \sum_{i=1}^{N}\sum_{r=1}^{R} z_{ir}\big(\log \pi_{ir}(\mathbf{x}_i) + \log \mathbb{P}_r(\hat{\mathbf{y}}_{ai}|y_i)\big). \tag{27}$$

Expectation-Maximization (EM) algorithm [Dempster *et al.*, 1977] can be used to find the parameters Θ that maximize $\mathcal{L}$ from (27).

EM algorithm for semi-supervised aggregation

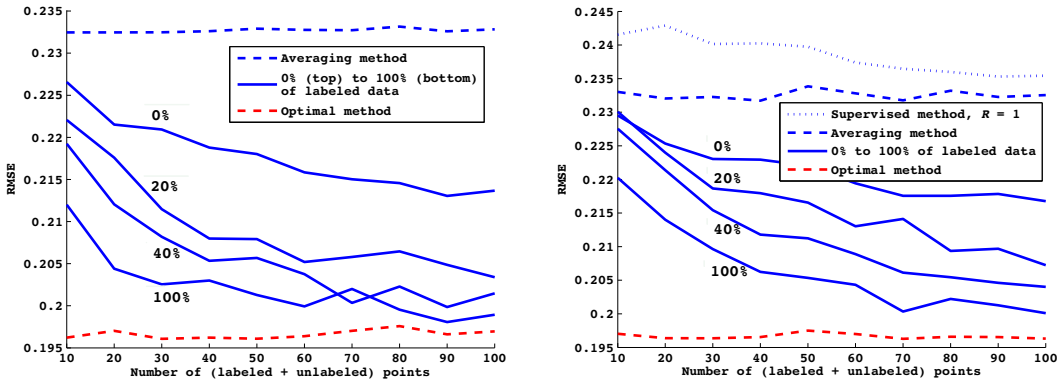
Before moving on, we need to decide on the parameterization of the prior probability π_{ir}. We define this probability using a softmax function,

$$\pi_{ir} = \frac{\exp\big(-(\mathbf{x}_i - \mathbf{q}_r)^{\mathrm{T}}\mathbf{\Lambda}_r(\mathbf{x}_i - \mathbf{q}_r)\big)}{\sum_{m=1}^{R}\exp\big(-(\mathbf{x}_i - \mathbf{q}_m)^{\mathrm{T}}\mathbf{\Lambda}_m(\mathbf{x}_i - \mathbf{q}_m)\big)}, \tag{28}$$

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

In the E-step, we compute the current expectation of posterior probability h_{ir} that the r^{th} regime is "responsible" for generating expert predictions for the i^{th} labeled data point as

$$h_{ir} = \mathbb{E}[z_{ir}|\hat{\mathbf{y}}_{ai}, y_i, \mathbf{x}_i, \Theta] = \frac{\pi_{ir}(\mathbf{x}_i)\;\mathbb{P}_r(\hat{\mathbf{y}}_{ai}|y_i)}{\sum_{m=1}^{R}\pi_{im}(\mathbf{x}_i)\;\mathbb{P}_m(\hat{\mathbf{y}}_{ai}|y_i)}. \tag{29}$$



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

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

Figure 3: Results on the synthetic data set

Then, 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 $\mathbf{\Sigma}_r$ and prototype vectors and scaling matrices $\mathbf{q}_r, \mathbf{\Lambda}_r, r = 1, \ldots, R$. Note that the derivatives of $\mathcal{L}$ with respect to these two sets of variables are independent from each other, and the optimization of $\mathbf{\Sigma}_r$ 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

$$\begin{aligned}\mathbf{\Sigma}_r = &\frac{1}{1+\sum_{i=1}^{N} h_{ir}}\Bigg(\sigma_{0r}^2\mathbf{I} + \sum_{i=1}^{N} h_{ir}\,\Pi_i^{-1}(\mathbf{\Psi}_{ir}) + \\ &\sum_{i=N_u+1}^{N} h_{ir}[\![(\hat{\mathbf{y}}_{ai} - y_i\,\mathbf{1})\,(\hat{\mathbf{y}}_{ai} - y_i\,\mathbf{1})^{\mathrm{T}}]\!]_r + \\ &\sum_{i=1}^{N_u} h_{ir}\Bigg([\![\hat{\mathbf{y}}_{ai}\hat{\mathbf{y}}_{ai}^{\mathrm{T}}]\!]_r + \frac{[\![\mathbf{1}\mathbf{1}^{\mathrm{T}}]\!]_r}{\mathbf{1}^{\mathrm{T}}\mathbf{U}'_{ir}\mathbf{1}} + \\ &\overline{y}_{ir}^2[\![\mathbf{1}\mathbf{1}^{\mathrm{T}}]\!]_r - \overline{y}_{ir}[\![\mathbf{1}\hat{\mathbf{y}}_{ai}^{\mathrm{T}} + \hat{\mathbf{y}}_{ai}\mathbf{1}^{\mathrm{T}}]\!]_r\Bigg)\Bigg),\end{aligned} \tag{30}$$

while prototype vector $\mathbf{q}_r$ and scaling matrix $\mathbf{\Lambda}_r$ are found through the gradient ascent optimization as

$$\begin{aligned}\mathbf{q}_r^{\text{new}} &= \mathbf{q}_r^{\text{old}} + \eta\,\mathbf{\Lambda}_r^{\text{old}} \sum_{i=1}^{N}(h_{ir} - \pi_{ir}^{\text{old}})(\mathbf{x}_i - \mathbf{q}_r^{\text{old}}), \\ \mathbf{\Lambda}_r^{\text{new}} &= \mathbf{\Lambda}_r^{\text{old}} + \eta \sum_{i=1}^{N}(h_{ir} - \pi_{ir}^{\text{old}})(\mathbf{x}_i - \mathbf{q}_r^{\text{old}})(\mathbf{x}_i - \mathbf{q}_r^{\text{old}})^{\mathrm{T}},\end{aligned} \tag{31}$$

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.

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 a prototype and a covariance matrix for each regime. Then, we assigned the i^{th} data point uniformly at random with probability $1/R$ to a regime, say the l^{th} regime, and obtained features $\mathbf{x}_i$ by sampling from multivariate Gaussian with mean $\mathbf{q}_l$ and diagonal covariance matrix $0.5\mathbf{I}$. We sampled ground-truth value y_i from zero-mean Gaussian with unit-variance, 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.

First, in order to evaluate the semi-supervised method without clustering, we set $K = 5$, $R = 1$, and $\mathbf{\Sigma}_1 = \mathrm{diag}([0.1, 0.2, 0.3, 0.4, 0.5])$. 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 (24) using the true $\mathbf{\Sigma}_1$. 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 Figure 3). The results in terms of Root Mean Squared Error (RMSE), evaluated on $1{,}000$ testing points generated in the same way as the training set and averaged over 100 experiments, are shown in Figure 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 $\mathbf{q}_1 = [1, 1]$, $\mathbf{q}_2 = [-1, -1]$, $\mathbf{\Sigma}_1 = \mathrm{diag}([0.1, 0.2, 0.3, 0.4, 0.5])$, $\mathbf{\Sigma}_2 = \mathrm{diag}([0.5, 0.4, 0.3, 0.2, 0.1])$, and we set $R = 2$. The results in terms of RMSE are given in Figure 3b, where we also show accuracy of the proposed method

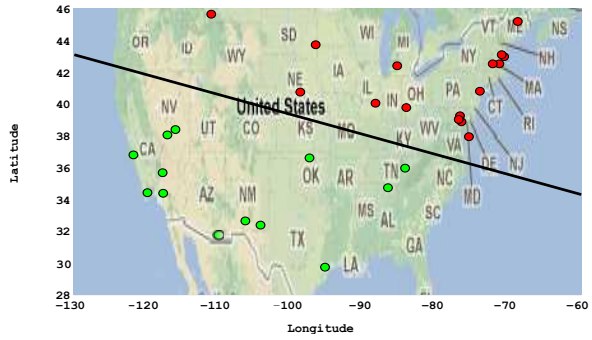


Figure 4: Found clustering of AERONET sites in the USA

Table 1: Performance of the aggregation methods

Method	# clusters	RMSE
Averaging	–	0.0818
All sites, semi-super.	1	0.0677
All sites, semi-super.	2	0.0648
2 sites, supervised	2	0.0795
2 sites, semi-super.	2	0.0752
4 sites, supervised	2	0.0728
4 sites, semi-super.	2	0.0704
6 sites, supervised	2	0.0694
6 sites, semi-super.	2	0.0688

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.

3.2 Validation on aerosol data

To construct an aerosol data set we considered ground-based AERONET data[1] from the United States and collocated data from 5 satellite instruments[2] spanning years 2006 to 2010. We considered AERONET measurements at 10:30am local time as the ground truth. Among the $K = 5$ instruments, two measured AOD at around 10:30am local time (Terra MODIS, MISR), and three at around 1:30pm local time (Aqua MODIS, OMI, SeaWiFS). After removing AERONET sites with too few observations, there remained 33 sites in the data set, with locations shown in Figure 4. This resulted in a labeled data set with $N = 6{,}913$ data points, where 58% of expert predictions were missing. We used this data set for two sets of experiments: (1) evaluating usefulness of partitioning; and (2) evaluating usefulness of unlabeled data.

In both sets of experiments we performed leave-one-site-out cross-validation. In the first set of experiments, from each training site we randomly sampled 100 data points, and assumed that 50 of them are labeled and 50 unlabeled. From the left-out site we sampled 100 data points. We used the geographic location (i.e., longitude and latitude) of the corresponding AERONET site as a feature vector $\mathbf{x}_i$ for the i^{th} data point. We used our proposed method with $R = 1$ and $R = 2$, repeating the experiments 5 times. We compared the performance to a baseline method that takes a simple average of available expert predictions. RMSE is reported at the top of Table 1. We can see that semi-supervised aggregation for both $R = 1$ and $R = 2$ had significantly lower RMSE than the baseline. Moreover, by increasing the number of clusters from 1 (i.e., without clustering) to 2 we observed a drop in RMSE of nearly 5%. In Figure 4 we color-code the AERONET sites according to their cluster assignments. Interestingly, the clustering roughly corresponds to partitioning proposed by domain scientists [Levy *et al.*, 2007]. For the south-western cluster the weights of linear combination assigned to MISR, Terra MODIS, Aqua MODIS, OMI, and SeaWiFS instruments, given predictions of all the experts, were $[0.51, 0.31, 0.09, 0.01, 0.08]$, while for the north-eastern cluster they were $[0.24, 0.27, 0.21, 0.15, 0.12]$, respectively. Consistent with the domain knowledge, MISR obtained the largest weight in the first cluster, while all instruments were given similar weights in the second cluster.

In the second set of experiments, we simulated conditions consistent with aerosol data availability in Africa and large parts of Asia, where very few AERONET sites are available and labeled data are very scarce. We randomly selected 2, 4, or 6 training AERONET sites and took 100 data points from each of them as labeled data. Then, we selected 100 data points from the remaining training AERONET sites and treated them as unlabeled data. We trained one model which used only labeled data, and one that used both labeled and unlabeled data. We used $R = 2$ clusters in both cases. The results given at the bottom of Table 1 show that the RMSE of purely supervised approach decreased with the number of AERONET sites. More importantly, unlabeled data were helpful and led to significant reductions in RMSE. This benefit of unlabeled data increased with the decrease in the number of labeled data points. The results in Table 1 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, and construct specialized aggregators on each cluster.

4 Conclusion

We proposed 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. Moreover, we also cluster the data during training by introducing a latent indicator variable for each cluster, resulting in a more interpretable model. Results on synthetic and real-world aerosol data comprising 5 satellite-borne sensors indicate the benefits of the proposed approach.

Acknowledgments

This work was supported by NSF grant IIS-1117433.

[1] aeronet.gsfc.nasa.gov/cgi-bin/combined_data_access_new

[2] disc.sci.gsfc.nasa.gov/aerosols/services/mapss/mapssdoc

References

[Bates and Granger, 1969] J. M. Bates and Clive W. J. Granger. The Combination of Forecasts. *Operational Research Quarterly*, 20(4):451–468, 1969.

[Brewer, 1978] J. Brewer. Kronecker products and matrix calculus in system theory. *IEEE Transactions on Circuits and Systems*, 25(9):772–781, 1978.

[Dempster *et al.*, 1977] A.P. Dempster, N.M. Laird, and D.B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 1–38, 1977.

[Diner *et al.*, 1998] 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, and M. M. V. Verstraete. Multi-angle Imaging SpectroRadiometer (MISR) - instrument description and experiment overview. *IEEE Transactions on Geoscience and Remote Sensing*, 36:1072–1087, 1998.

[Granger and Ramanathan, 1984] Clive W. J. Granger and Ramu Ramanathan. Improved methods of combining forecasts. *Journal of Forecasting*, 3(2):197–204, 1984.

[Holben *et al.*, 1998] 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, and A. Smirnov. AERONET - A Federated Instrument Network and Data Archive for Aerosol Characterization. *Remote Sensing of Environment*, 66(1):1–16, 1998.

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

[King *et al.*, 2003] M.D. King, W.P. Menzel, Y.J. Kaufman, D. Tanré, Bo-Cai Gao, S. Platnick, S.A. Ackerman, L.A. Remer, R. Pincus, and P.A. Hubanks. Cloud and aerosol properties, precipitable water, and profiles of temperature and water vapor from MODIS. *IEEE Transactions on Geoscience and Remote Sensing*, 41(2):442–458, feb. 2003.

[Levy *et al.*, 2007] Robert C. Levy, Lorraine A. Remer, and Oleg Dubovik. Global aerosol optical properties and application to Moderate Resolution Imaging SpectroRadiometer aerosol retrieval over land. *Journal of Geophysical Research*, 112:13,210–13,224, 2007.

[Liu *et al.*, 2009] Y. Liu, C.J. Paciorek, and P. Koutrakis. Estimating regional spatial and temporal variability of $PM_{2.5}$ concentrations using satellite data, meteorology, and land use information. *Environmental health perspectives*, 117(6):886, 2009.

[Mishchenko *et al.*, 2010] Michael I. Mishchenko, Li Liu, Igor V. Geogdzhayev, Larry D.Travis, Brian Cairns, and Andrew A. Lacis. Toward unified satellite climatology of aerosol properties. 3. MODIS versus MISR versus AERONET. *Journal of Quantitative Spectroscopy & Radiative Transfer*, 111:540–552, 2010.

[Randall *et al.*, 2007] D.A. Randall, R.A. 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, 2007.

[Raykar *et al.*, 2009] V.C. Raykar, S. Yu, L.H. Zhao, A. Jerebko, C. Florin, G.H. Valadez, L. Bogoni, and L. Moy. Supervised learning from multiple experts: Whom to trust when everyone lies a bit. In *Proceedings of the 26th Annual International Conference on Machine Learning*, pages 889–896. ACM, 2009.

[Ristovski *et al.*, 2010] K. Ristovski, D. Das, V. Ouzienko, Y. Guo, and Z. Obradovic. Regression Learning with Multiple Noisy Oracles. In *19th European Conference on Artificial Intelligence*, pages 445–450, 2010.

[Torres *et al.*, 2002] O. Torres, R. Decae, J. P. Veefkind, and G. de Leeuw. OMI aerosol retrieval algorithm. In P. Stammes, editor, *OMI Algorithm Theoretical Basis Document, Volume III, Clouds, Aerosols, and Surface UV Irradiance*. 2002.

[Wang *et al.*, 2000] M. Wang, S. Bailey, and C. R. Mcclain. SeaWiFS Provides Unique Global Aerosol Optical Property Data. *EOS*, 81(18), 2000.

[Weigend *et al.*, 1995] A. S Weigend, M. Mangeas, and N.S. Ashok. Nonlinear gated experts for time series: Discovering regimes and avoiding overfitting. *International Journal of Neural Systems*, 6(04):373–399, 1995.