Accurate Integration of Aerosol Predictions by Smoothing on a Manifold

Shuai Zheng **James T. Kwok**
Department of Computer Science and Engineering
Hong Kong University of Science and Technology
Hong Kong
{szhengac, jamesk}@cse.ust.hk

Abstract

Accurately measuring the aerosol optical depth (AOD) is essential for our understanding of the climate. Currently, AOD can be measured by (i) satellite instruments, which operate on a global scale but have limited accuracies; and (ii) ground-based instruments, which are more accurate but not widely available. Recent approaches focus on integrating measurements from these two sources to complement each other. In this paper, we further improve the prediction accuracy by using the observation that the AOD varies slowly in the spatial domain. Using a probabilistic approach, we impose this smoothness constraint by a Gaussian random field on the Earth's surface, which can be considered as a two-dimensional manifold. The proposed integration approach is computationally simple, and experimental results on both synthetic and real-world data sets show that it significantly outperforms the state-of-the-art.

Introduction

Aerosols are fine solid airborne particles or liquid droplets present throughout our environment. They possess different forms, such as dust, haze, mist, smog and smokes. A good understanding of the aerosol characteristics can enable us a better understanding in the formation of clouds, rain drops, snow flakes and ice crystals in the atmosphere. Besides, aerosols are useful in predicting climatic effects (Watson et al. 1990) and the estimation of air pollution such as $\mathrm{PM}_{2.5}$ (Wang and Christopher 2003). Aerosols also play a critical role in the radiative forcing in the Earth's atmosphere system (Abdou et al. 2005). Moreover, aerosol properties have a significant impact on industries such as manufacturing and transport. Hence, not surprisingly, there have been a lot of studies on aerosols in the past decades.

The distribution of aerosols is measured by the aerosol optical depth (AOD), and can be assessed by different instruments on-board satellites. For example, the Terra satellite is equipped with the MODerate resolution Imaging Spectroradiometer (MODIS) and Multiangle Imaging SpectroRadiometer (MISR), the Aqua satellite is equipped with MODIS, the Aura satellite with the Ozone Monitoring Instrument (OMI), and SeaStar with the SEA-viewing Wide Field-of-view Sensor (SeaWiFS). These instruments vary in their retrieval algorithms, coverage, sensor characteristics, and thus also accuracies. As such, additional surface sensors from the Aerosol Robotic Network (AERONET) are often needed for validation. While these ground-based instruments have higher temporal and spectral resolutions, they are unevenly distributed, with most being installed in North America and Europe.

To have a better AOD estimate, one useful approach is to integrate measurements from multiple satellite measurements. For example, Mishchenko et al. (2010) average measurements from the MODIS and MISR. Recently, Djuric, Kansakar, and Vucetic (2013) aggregate both ground-based and satellite instruments in a semi-supervised learning manner. In particular, they use locations with both ground-based and satellite AOD measurements (considered as "labeled" data) to help in the AOD estimation of locations that have only satellite measurements ("unlabeled" data).

However, in (Djuric, Kansakar, and Vucetic 2013), the borrowing of strength from labeled data to unlabeled data is achieved via a shared correlation matrix of the satellite measurements. The AOD values at different locations, however, are assumed to be independent of each other. This is at odds with the observation that the AOD indeed varies slowly on scales of tens of kilometers (Chu et al. 2002; Koelemeijer, Homan, and Matthijsen 2006). An example on the MODIS data is shown in Figure 1.

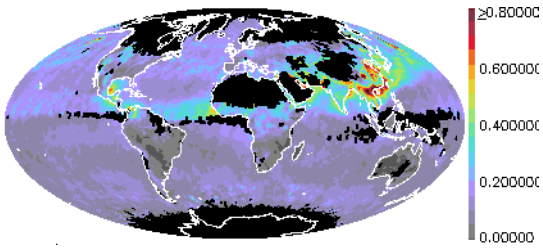

Figure 1: The smooth distribution of the average monthly AOD (reproduced with permission from NASA). As can be seen, the AOD's at nearby locations are close to each other.

In this paper, we attempt to further improve AOD estimation by exploring such spatial correlations. In particular,

we require the AOD predictions to be smooth on a manifold defined over all the locations. The use of manifolds has been popularly used in semi-supervised learning (Belkin, Niyogi, and Sindhwani 2006), and has been successfully used in tasks including dimensionality reduction (Belkin and Niyogi 2003), classification (Belkin and Niyogi 2002), clustering (Chapelle, Weston, and Schölkopf 2002) and ranking (Zhou et al. 2004b). The proposed approach is able to utilize the slow spatial variation of AOD, while preserving the computational simplicity in (Djuric, Kansakar, and Vucetic 2013). Experiments on both synthetic and real-world data sets demonstrate that the proposed method is much more accurate than the state-of-the-art.

Notations: In the sequel, the transpose of vector/matrix is denoted by the superscript ^{T}, tr($\mathbf{A}$) denotes the trace of matrix $\mathbf{A}$, $\mathbf{I}$ is the identity matrix, $\mathbf{0}$ is the zero matrix, $\mathbf{1}$ is the vector of all ones, and $\mathcal{N}(\cdot,\cdot)$ is the scalar/vector normal distribution.

Related Work

At a particular location i, let its vector of K satellite measurements be $\hat{\mathbf{y}}_i = [\hat{y}_{i1}, \ldots, \hat{y}_{iK}]^T$. In (Djuric, Kansakar, and Vucetic 2013), the $\{\hat{\mathbf{y}}_i\}_{i=1}^N$ from N locations are assumed to be generated i.i.d. as

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

where y_i is the underlying ground truth AOD at location i, and $\mathbf{\Sigma}$ captures the correlation among satellite measurements. The y_i's are also assumed to be generated i.i.d., as

$$y_i \sim \mathcal{N}(u, \sigma^2), \tag{2}$$

where u can be regarded as the default AOD, and σ^2 the corresponding variance across locations. From (1) and (2), it can be easily shown that the joint measurements $\hat{\mathbf{Y}} = [\hat{\mathbf{y}}_1, \hat{\mathbf{y}}_2, \ldots, \hat{\mathbf{y}}_N]^T \in \mathbb{R}^{N \times K}$ is distributed as

$$\hat{\mathbf{Y}} | \mathbf{y} \sim \mathcal{MN}_{N,K}(\mathbf{y}\mathbf{1}^T, \mathbf{\Sigma} \otimes \mathbf{I}), \tag{3}$$

$$\mathbf{y} \sim \mathcal{N}(u\mathbf{1}, \sigma^2 \mathbf{I}). \tag{4}$$

where $\mathbf{y} = [y_1, y_2, \ldots, y_N]^T$, $\otimes$ is the Kronecker product, and $\mathcal{MN}_{N,K}(\cdot,\cdot)$ is the $N \times K$ matrix-variate normal distribution[1] (Gupta and Nagar 2000).

At any location, its surface AOD measurement (e.g., from AERONET) can be considered as ground truth. However, typically these are only available at, say, N_l, locations. Without loss of generality, assume that $\mathbf{y}$ is reordered as $[\mathbf{y}_u^T, \mathbf{y}_l^T]^T$, where $\mathbf{y}_l \in \mathbb{R}^{N_l}$ is the subvector for locations with known ground truths, and $\mathbf{y}_u \in \mathbb{R}^{N_u}$ (where $N_u = N - N_l$) is for locations with unknown ground truths. In the sequel, these will be referred to as the labeled and unlabeled locations, respectively. Similarly, $\hat{\mathbf{Y}}$ is reordered as $\begin{bmatrix} \hat{\mathbf{Y}}_u \\ \hat{\mathbf{Y}}_l \end{bmatrix}$, with $\hat{\mathbf{Y}}_u \in \mathbb{R}^{N_u \times K}$ for the unlabeled locations, and $\hat{\mathbf{Y}}_l \in \mathbb{R}^{N_l \times K}$ for the labeled locations. The task is to infer the unknown $\mathbf{y}_u$ from $\hat{\mathbf{Y}}$ and $\mathbf{y}_l$.

In (Djuric, Kansakar, and Vucetic 2013), they first estimate $\mathbf{\Sigma}$, and then, on using (1), (2) (or, equivalently, (3), (4)), the posterior of $\mathbf{y}_u$ can be shown to follow

$$\mathcal{N}\left(\left(\bar{\Sigma} + \frac{1}{\sigma^2}\right)^{-1}\left(\hat{\mathbf{Y}}_u \mathbf{\Sigma}^{-1}\mathbf{1} + \frac{u}{\sigma^2}\mathbf{1}\right), \left(\bar{\Sigma} + \frac{1}{\sigma^2}\right)^{-1}\mathbf{I}\right), \tag{5}$$

where

$$\bar{\Sigma} = \mathbf{1}^T \mathbf{\Sigma}^{-1} \mathbf{1}. \tag{6}$$

The mean of (5), i.e., $\mathbf{y}_u$, is used as the AOD prediction. Note that if $\mathbf{\Sigma} = \mathbf{I}$ and $\sigma^2 \to \infty$, this reduces to $\frac{1}{K}\hat{\mathbf{Y}}_u \mathbf{1}$. In other words, the prediction at location i is simply the average of its K satellite measurements.

As discussed in (Djuric, Kansakar, and Vucetic 2013), as satellites have limited daily coverage, a location may not always have all its K satellite measurements. When some measurements are missing, the handling of the resultant incomplete $\hat{\mathbf{Y}}$ matrix is also considered in (Djuric, Kansakar, and Vucetic 2013). Besides, as can be seen from Figure 1, the AOD in North America is considerably lower than most parts of Asia. Thus, it is sometimes useful to partition the globe into, say, R, regimes. In (Djuric, Kansakar, and Vucetic 2013), location i is assigned to regime r with probability π_{ir} according to the following softmax function

$$\pi_{ir} = \frac{\exp(-(\mathbf{x}_i - \mathbf{q}_r)^T \mathbf{S}_r (\mathbf{x}_i - \mathbf{q}_r))}{\sum_{m=1}^R \exp(-(\mathbf{x}_i - \mathbf{q}_m)^T \mathbf{S}_m (\mathbf{x}_i - \mathbf{q}_m)}, \tag{7}$$

where $\mathbf{x}_i \in \mathbb{R}^d$ is the feature vector for location i (typically, its latitude and longitude), $\mathbf{q}_r$ is the prototype vector for regime r, and $\mathbf{S}_r$ a scaling matrix. Different regimes are assumed to have different $\mathbf{\Sigma}$'s, and all the parameters are then learned as in a standard Gaussian mixture model.

Integration of AOD Predictions with Manifold

Recall that in (Djuric, Kansakar, and Vucetic 2013), the ground truth AODs from the various locations are assumed to be independent. However, as discussed in the introduction, the AOD should vary slowly in the spatial domain. Such a smoothness property can be easily imposed with the use of a manifold, a notion that has been popularly used in the semi-supervised learning literature (Belkin, Niyogi, and Sindhwani 2006; Zhou et al. 2004a; Zhu, Ghahramani, and Lafferty 2003; Zhu 2007).

Definition of the Manifold

A manifold $\mathcal{M}$ is often represented by a weighted graph $\mathcal{G}$. Here, every geographical location is a node. For simplicity, we assume that the nodes are fully connected. For two locations i, j, with latitude-longitude values (ϕ_i, λ_i) and (ϕ_j, λ_j) respectively, their great-circle distance (i.e., shortest distance over the earth's surface) d_{ij} is given by the Haversine formula (Sinnott 1984):

$$d_{ij} = 2r \arcsin\Big(\big(\sin^2((\phi_i - \phi_j)/2) + \cos(\phi_i)\cos(\phi_j)\sin^2((\lambda_i - \lambda_j)/2)\big)^{\frac{1}{2}}\Big), \tag{8}$$

[1] A random variable $\mathbf{X} \in \mathbb{R}^{m \times n}$ follows the *matrix-variate normal distribution* $\mathcal{MN}_{m,n}(\mathbf{M}, \mathbf{\Sigma} \otimes \mathbf{\Psi})$ with mean $\mathbf{M} \in \mathbb{R}^{m \times n}$ and covariance matrix $\mathbf{\Sigma} \otimes \mathbf{\Psi}$ (where $\mathbf{\Psi} \in \mathbb{R}^{m \times m}$ and $\mathbf{\Sigma} \in \mathbb{R}^{n \times n}$) if its pdf is given by $(2\pi)^{-\frac{mn}{2}} |\mathbf{\Psi}|^{-\frac{n}{2}} |\mathbf{\Sigma}|^{-\frac{m}{2}} \exp(-\frac{1}{2}\text{tr}[\mathbf{\Psi}^{-1}(\mathbf{X} - \mathbf{M})\mathbf{\Sigma}^{-1}(\mathbf{X} - \mathbf{M})^T])$.

where r is the Earth's radius. Obviously, the similarity (or weight w_{ij}) between i and j should decrease with d_{ij}. Using the local scaling approach in (Zelnik-Manor and Perona 2004), we define it as

$$w_{ij} = \begin{cases} \exp\left(-\frac{d_{ij}}{\sqrt{s_i}\sqrt{s_j}}\right) & i \neq j, \\ 0 & i = j, \end{cases} \quad (9)$$

where s_i, s_j automatically rescale d_{ij} based on the local statistics of the neighborhoods of i and j. Typically, s_i is set to the distance between i and its Hth neighbor (with $H = 5$ in the experiments). Given $\mathbf{W} = [w_{ij}]$, the Laplacian matrix of $\mathcal{M}$ is then defined as $\mathbf{L} = \mathbf{D} - \mathbf{W}$, where $\mathbf{D}$ is the diagonal matrix with diagonal elements $d_i = \sum_j w_{ij}$. To ensure that the Laplacian is non-singular, it is customary to add some regularization (Verbeek and Vlassis 2006; Zhu, Lafferty, and Ghahramani 2003), leading to $\tilde{\mathbf{L}} = \mathbf{L} + \alpha\mathbf{I}$, where $\alpha > 0$ is a small number.

Instead of assuming that the ground truth AODs (y_i's) are generated independently as in (Djuric, Kansakar, and Vucetic 2013), we assume that they are formed by a Gaussian random field defined on the manifold $\mathcal{M}$ (Zhu, Ghahramani, and Lafferty 2003):

$$\mathbf{y} \sim \mathcal{N}(u\mathbf{1}, \sigma^2\tilde{\mathbf{L}}^{-1}). \quad (10)$$

Here, σ^2 controls the scale of the covariance. Essentially, (10) implies that when w_{ij} is large, the corresponding y_i and y_j should be close to each other.

Prediction of $\mathbf{y}_u$

Given the satellite measurement's covariance matrix $\mathbf{\Sigma}$ (which can be either fixed or learned), the posterior of $\mathbf{y}_u$ can be obtained as

$$\begin{aligned} P(\mathbf{y}_u|\mathbf{y}_l, \hat{\mathbf{Y}}) &= \frac{P(\mathbf{y}_u, \mathbf{y}_l|\hat{\mathbf{Y}})}{P(\mathbf{y}_l|\hat{\mathbf{Y}})} = \frac{P(\mathbf{y}|\hat{\mathbf{Y}})}{P(\mathbf{y}_l|\hat{\mathbf{Y}})} \\ &= \frac{P(\hat{\mathbf{Y}}|\mathbf{y})P(\mathbf{y})}{P(\hat{\mathbf{Y}}, \mathbf{y}_l)}. \end{aligned} \quad (11)$$

Now, $P(\hat{\mathbf{Y}}|\mathbf{y})$ and $P(\mathbf{y})$ are defined in (3) and (10); while

$$P(\hat{\mathbf{Y}}, \mathbf{y}_l) = \int P(\hat{\mathbf{Y}}|\mathbf{y})P(\mathbf{y})d\mathbf{y}_u \quad (12)$$

can be evaluated in closed-form as both $\hat{\mathbf{Y}}|\mathbf{y}$ and $\mathbf{y}$ follow the normal distribution. Assume that the Laplacian matrix $\mathbf{L}$ (and similarly its regularized version $\tilde{\mathbf{L}}$) has been reordered and partitioned as $\mathbf{L} = \begin{bmatrix} \mathbf{L}_l & \mathbf{L}_{lu} \\ \mathbf{L}_{lu}^T & \mathbf{L}_u \end{bmatrix}$, where the subscripts l and u denote the parts corresponding to the labeled and unlabeled locations, respectively. It can be shown that the conditional distribution of $\mathbf{y}_u$ is also a normal distribution:

$$\mathbf{y}_u|\mathbf{y}_l, \hat{\mathbf{Y}} \sim \mathcal{N}\left(\bar{\mathbf{y}}, \left(\bar{\Sigma}\mathbf{I} + \frac{\tilde{\mathbf{L}}_u}{\sigma^2}\right)^{-1}\right), \quad (13)$$

where $\bar{\Sigma}$ is as defined in (6),

$$\begin{aligned} \bar{\mathbf{y}} &= \left(\bar{\Sigma}\mathbf{I} + \frac{\tilde{\mathbf{L}}_u}{\sigma^2}\right)^{-1}\left(\breve{\mathbf{y}} - \frac{\mathbf{L}_{lu}^T\mathbf{y}_l}{\sigma^2} + \frac{u\mathbf{L}_{lu}^T\mathbf{1}}{\sigma^2} + \frac{u\tilde{\mathbf{L}}_u\mathbf{1}}{\sigma^2}\right), \\ \breve{\mathbf{y}} &= \hat{\mathbf{Y}}_u\mathbf{\Sigma}^{-1}\mathbf{1}, \end{aligned} \quad (14)$$

Hence, we can use $\bar{\mathbf{y}}$ as the prediction for $\mathbf{y}_u$.

Remark. *In the absence of the manifold,* $\mathbf{W} = \mathbf{D} = \mathbf{L} = \mathbf{0}$*, and (13) reduces to* $\mathcal{N}\left((\bar{\Sigma} + \frac{\alpha}{\sigma^2})^{-1}(\breve{\mathbf{y}} + \frac{u\alpha}{\sigma^2}\mathbf{1}), (\bar{\Sigma} + \frac{\alpha}{\sigma^2})^{-1}\mathbf{I}\right)$*. This is the same as (5) on setting* $\alpha = 1$*.*

In situations where the positive semidefinite (psd) matrix $\mathbf{\Sigma}$ is not known, it can be learned by maximizing the likelihood $P(\hat{\mathbf{Y}}, \mathbf{y}_l)$ in (12) w.r.t. $\mathbf{\Sigma}$ using projected gradient (Bertsekas 2004). Since $\mathbf{\Sigma} \in \mathbb{R}^{K\times K}$ and K is typically small (equal to 5 in the experiments), projection onto the psd cone in each iteration is computationally inexpensive. Similarly, u and σ^2 in (10) can also be learned by maximizing $P(\hat{\mathbf{Y}}, \mathbf{y}_l)$. In particular, u can be obtained in closed-form as

$$u = \frac{\left(\mathbf{1}^T\tilde{\mathbf{L}}_u + \mathbf{1}^T\mathbf{L}_{lu}\right)\bar{\mathbf{y}}^* + \left(\mathbf{1}^T\tilde{\mathbf{L}}_l + \mathbf{1}^T\mathbf{L}_{lu}^T\right)\mathbf{y}_l}{\mathbf{1}^T\tilde{\mathbf{L}}\mathbf{1} - \frac{1}{\sigma^2}\left(\mathbf{1}^T\mathbf{L}_{lu} + \mathbf{1}\tilde{\mathbf{L}}_u\right)\left(\bar{\Sigma}\mathbf{I} + \frac{\tilde{\mathbf{L}}_u}{\sigma^2}\right)^{-1}\left(\mathbf{L}_{lu}^T\mathbf{1} + \tilde{\mathbf{L}}_u\mathbf{1}\right)},$$

where $\bar{\mathbf{y}}^* = \left(\bar{\Sigma}\mathbf{I} + \frac{\tilde{\mathbf{L}}_u}{\sigma^2}\right)^{-1}\left(\breve{\mathbf{y}} - \frac{\mathbf{L}_{lu}^T\mathbf{y}_l}{\sigma^2}\right)$.

Missing Satellite Measurements

In this section, we consider the case where each location may have some missing satellite measurements. Without loss of generality, we rearrange each satellite measurement vector $\hat{\mathbf{y}}_i$ as $\hat{\mathbf{y}}_i = [(\hat{\mathbf{y}}_i^{(a)})^T, (\hat{\mathbf{y}}_i^{(q)})^T]^T$ such that $\hat{\mathbf{y}}_i^{(a)}$ contains the a_i measurements available at location i, while $\hat{\mathbf{y}}_i^{(q)}$ is for the $q_i = K - a_i$ missing measurements. Similarly, for each i, $\mathbf{\Sigma}^{-1}$ is reordered as

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

where $\mathbf{U}_i \in \mathbb{R}^{a_i\times a_i}, \mathbf{V}_i \in \mathbb{R}^{a_i\times q_i}$, and $\mathbf{Q}_i \in \mathbb{R}^{q_i\times q_i}$, so that the first a_i rows/columns of $\Pi_i(\mathbf{\Sigma}^{-1})$ correspond to the available measurements, while the remaining q_i rows/columns correspond to the missing measurements.

Analogous to (11), the posterior probability of $\mathbf{y}_u$ (given $\mathbf{y}_l$ and the available satellite measurements $\hat{\mathbf{y}}_i^{(a)}$'s) is

$$P(\mathbf{y}_u|\mathbf{y}_l, \hat{\mathbf{y}}_1^{(a)}, \ldots, \hat{\mathbf{y}}_N^{(a)}) = \frac{P(\hat{\mathbf{y}}_1^{(a)}, \ldots, \hat{\mathbf{y}}_N^{(a)}|\mathbf{y})P(\mathbf{y})}{P(\hat{\mathbf{y}}_1^{(a)}, \ldots, \hat{\mathbf{y}}_N^{(a)}, \mathbf{y}_l)}. \quad (15)$$

The $P(\hat{\mathbf{y}}_1^{(a)}, \ldots, \hat{\mathbf{y}}_N^{(a)}|\mathbf{y})$ term in the numerator can be obtained by marginalizing the missing $\hat{\mathbf{y}}_i^{(q)}$'s from $P(\hat{\mathbf{Y}}|\mathbf{y})$, as

$$\begin{aligned} &P(\hat{\mathbf{y}}_1^{(a)}, \ldots, \hat{\mathbf{y}}_N^{(a)}|\mathbf{y}) \\ &= \int\ldots\int P(\hat{\mathbf{Y}}|\mathbf{y})d\hat{\mathbf{y}}_1^{(q)}\ldots d\hat{\mathbf{y}}_N^{(q)} \\ &= \prod_{i=1}^{N}\left(\frac{\exp(-\frac{1}{2}((\hat{\mathbf{y}}_i^{(a)} - y_i\mathbf{1})^T\mathbf{C}_i(\hat{\mathbf{y}}_i^{(a)} - y_i\mathbf{1}))}{(2\pi)^{\frac{a_i}{2}}|\mathbf{C}_i^{-1}|^{\frac{1}{2}}}\right), \end{aligned}$$

where

$$\mathbf{C}_i = \mathbf{U}_i - \mathbf{V}_i\mathbf{Q}_i^{-1}\mathbf{V}_i^T$$

is the Schur complement of the block $\mathbf{Q}_i$ of $\Pi_i(\mathbf{\Sigma}^{-1})$. Similarly, one can compute the denominator $P(\hat{\mathbf{y}}_1^{(a)}, \ldots, \hat{\mathbf{y}}_N^{(a)}, \mathbf{y}_l)$ in (15) as

$\int P(\hat{\mathbf{y}}_1^{(a)}, \ldots, \hat{\mathbf{y}}_N^{(a)}|\mathbf{y})P(\mathbf{y})d\mathbf{y}_u$. After tedious derivation, it can be shown that the conditional distribution of $\mathbf{y}_u$ is again a normal distribution:

$$\mathbf{y}_u|\mathbf{y}_l, \hat{\mathbf{y}}_1^{(a)}, \ldots, \hat{\mathbf{y}}_N^{(a)} \sim \mathcal{N}\left(\mathbf{m}, \left(\mathbf{A} + \frac{\tilde{\mathbf{L}}_u}{\sigma^2}\right)^{-1}\right), \quad (16)$$

where

$$\mathbf{m} = \left(\mathbf{A} + \frac{\tilde{\mathbf{L}}_u}{\sigma^2}\right)^{-1}\left(\tilde{\mathbf{y}} - \frac{\mathbf{L}_{lu}^T\mathbf{y}_l}{\sigma^2} + \frac{u\mathbf{L}_{lu}^T\mathbf{1}}{\sigma^2} + \frac{u\tilde{\mathbf{L}}_u\mathbf{1}}{\sigma^2}\right), \quad (17)$$

$$\mathbf{A} = \text{diag}\left(\mathbf{1}^T\mathbf{C}_1\mathbf{1}, \ldots, \mathbf{1}^T\mathbf{C}_{N_u}\mathbf{1}\right), \quad (18)$$

$$\tilde{\mathbf{y}} = \left[\hat{\mathbf{y}}_1^{(a)T}\mathbf{C}_1\mathbf{1}, \ldots, \hat{\mathbf{y}}_{N_u}^{(a)T}\mathbf{C}_{N_u}\mathbf{1}\right]^T. \quad (19)$$

Hence, the prediction for $\mathbf{y}_u$ is $\mathbf{m}$. Moreover, as in the previous section, when $\mathbf{\Sigma}$ is not known, we can compute its maximum likelihood estimate (MLE) by projected gradient.

Remark. *When there is no missing measurement, all the $\mathbf{U}_i$'s and $\mathbf{C}_i$'s become $\mathbf{\Sigma}^{-1}$. Thus, $\mathbf{A}$ reduces to $\bar{\Sigma}\mathbf{I}$, $\tilde{\mathbf{y}}$ reduces to $\breve{\mathbf{y}}$ in (14), and (16) reduces to the normal distribution in (13).*

Remark. *In the presence of missing measurements, the covariance matrix of the prediction changes from $\left(\bar{\Sigma}\mathbf{I} + \frac{\tilde{\mathbf{L}}_u}{\sigma^2}\right)^{-1}$ in (13) to $\left(\mathbf{A} + \frac{\tilde{\mathbf{L}}_u}{\sigma^2}\right)^{-1}$ in (16). The following Proposition shows that the variance on the predictions is increased, which agrees with our intuition.*

Proposition 1. *The diagonal elements of $\left(\mathbf{A} + \frac{\tilde{\mathbf{L}}_u}{\sigma^2}\right)^{-1}$ are larger than the corresponding elements of $\left(\bar{\Sigma}\mathbf{I} + \frac{\tilde{\mathbf{L}}_u}{\sigma^2}\right)^{-1}$.*

Proof. It can be easily shown that

$$[\bar{\Sigma}\mathbf{I}]_{ii} - [\mathbf{A}]_{ii} = \bar{\Sigma} - \mathbf{1}^T\mathbf{C}_i\mathbf{1} = \mathbf{1}^T\mathbf{B}_i\mathbf{1}, \quad (20)$$

where $\mathbf{B}_i = \begin{bmatrix} \mathbf{V}_i\mathbf{Q}_i^{-1}\mathbf{V}_i^T & \mathbf{V}_i \\ \mathbf{V}_i^T & \mathbf{Q}_i \end{bmatrix}$. Obviously, the Schur complement of $\mathbf{Q}_i$ in $\mathbf{B}_i$ is zero, and thus psd. Hence, $\mathbf{B}_i$ is also psd, which implies $\mathbf{1}^T\mathbf{B}_i\mathbf{1} \geq 0$. From (20), $\mathbf{A}$ thus has smaller diagonal entries than $\bar{\Sigma}\mathbf{I}$, and so $\bar{\Sigma}\mathbf{I} - \mathbf{A} \succeq 0$. Moreover,

$$\left(\mathbf{A} + \frac{\tilde{\mathbf{L}}_u}{\sigma^2}\right)^{-1} - \left(\bar{\Sigma}\mathbf{I} + \frac{\tilde{\mathbf{L}}_u}{\sigma^2}\right)^{-1} = \left(\mathbf{A} + \frac{\tilde{\mathbf{L}}_u}{\sigma^2}\right)^{-1}\mathbf{J}\left(\mathbf{A} + \frac{\tilde{\mathbf{L}}_u}{\sigma^2}\right)^{-1},$$

where $\mathbf{J} = (\bar{\Sigma}\mathbf{I} - \mathbf{A}) - (\bar{\Sigma}\mathbf{I} - \mathbf{A})\left(\bar{\Sigma}\mathbf{I} + \frac{\tilde{\mathbf{L}}_u}{\sigma^2}\right)^{-1}(\bar{\Sigma}\mathbf{I} - \mathbf{A})$ is the Schur complement of $\bar{\Sigma}\mathbf{I} + \frac{\tilde{\mathbf{L}}_u}{\sigma^2}$ in $\mathbf{F} = \begin{bmatrix} \bar{\Sigma}\mathbf{I} + \frac{\tilde{\mathbf{L}}_u}{\sigma^2} & \bar{\Sigma}\mathbf{I} - \mathbf{A} \\ \bar{\Sigma}\mathbf{I} - \mathbf{A} & \bar{\Sigma}\mathbf{I} - \mathbf{A} \end{bmatrix}$. It can be seen that $\mathbf{F}$ is psd, as the generalized Schur complement of $\bar{\Sigma}\mathbf{I} - \mathbf{A}$ in $\mathbf{F}$ is $\mathbf{A} + \frac{\tilde{\mathbf{L}}_u}{\sigma^2}$ (pd) and $\left(\mathbf{I} - (\bar{\Sigma}\mathbf{I} - \mathbf{A})(\bar{\Sigma}\mathbf{I} - \mathbf{A})^\dagger\right)(\bar{\Sigma}\mathbf{I} - \mathbf{A}) = \mathbf{0}$ (Zhang 2005). Hence, $\mathbf{J}$ is psd, and $\left(\mathbf{A} + \frac{\tilde{\mathbf{L}}_u}{\sigma^2}\right)^{-1}\mathbf{J}\left(\mathbf{A} + \frac{\tilde{\mathbf{L}}_u}{\sigma^2}\right)^{-1}$ is psd, which shows that the diagonal entries of $\left(\mathbf{A} + \frac{\tilde{\mathbf{L}}_u}{\sigma^2}\right)^{-1} - \left(\bar{\Sigma}\mathbf{I} + \frac{\tilde{\mathbf{L}}_u}{\sigma^2}\right)^{-1}$ are non-negative. □

Multiple Data Regimes

Recall that in (Djuric, Kansakar, and Vucetic 2013), different regimes are assumed to have different $\mathbf{\Sigma}$'s. In other words, correlations among the satellite measurements are assumed to be different in different geographical regions.

In this paper, we instead take the more plausible assumption that the satellite correlations are independent of the geographical location. Instead, while (10) assumes that all the y_i's are sampled from the same value u, we now assume that the ground truths at different locations are generated from different values. Specifically,

$$\mathbf{y} \sim \mathcal{N}(\mathbf{u}, \sigma^2\tilde{\mathbf{L}}^{-1}), \quad (21)$$

where $\mathbf{u} = [u_1, u_2, \ldots, u_N]^T$, with u_i's generated from a mixture with R components

$$u_i = \sum_{r=1}^{R} \pi_{ir}\mu_r. \quad (22)$$

Here, μ_r is the default AOD value at the rth regime, and π_{ir}, as defined in (7), is the probability that location i belongs to regime r. This model is also in line with the observation in (Levy, Remer, and Dubovik 2007) that the aerosol sources and atmospheric composition differ in different regions.

As in the previous sections, it can be shown that the posterior of $\mathbf{y}_u$ is again a normal distribution

$$\mathbf{y}_u|\mathbf{y}_l, \hat{\mathbf{y}}_1^{(a)}, \ldots, \hat{\mathbf{y}}_N^{(a)} \sim \mathcal{N}\left(\mathbf{z}, \left(\mathbf{A} + \frac{\tilde{\mathbf{L}}_u}{\sigma^2}\right)^{-1}\right),$$

where $\mathbf{z} = \left(\mathbf{A} + \frac{\tilde{\mathbf{L}}_u}{\sigma^2}\right)^{-1}\left(\tilde{\mathbf{y}} - \frac{\mathbf{L}_{lu}^T\mathbf{y}_l}{\sigma^2} + \frac{\mathbf{L}_{lu}^T\mathbf{u}_l}{\sigma^2} + \frac{\tilde{\mathbf{L}}_u\mathbf{u}_u}{\sigma^2}\right)$, $\mathbf{A}, \tilde{\mathbf{y}}$ are as defined in (18) and (19), and $\mathbf{u}_l, \mathbf{u}_u$ are the subvectors of $\mathbf{u}$ corresponding to the labeled and unlabeled data, respectively. In particular, when $R = 1$ (i.e., there is only one regime), $\mathbf{u} = u_1\mathbf{1}$ and $\mathbf{z}$ reduces to $\mathbf{m}$ in (17).

Note that $\mathbf{u}$, in turn, depends on μ_r's and π_{ir}'s (which are defined by $\mathbf{q}_r$'s and $\mathbf{S}_r$'s). All these parameters can be obtained by maximum likelihood. In particular, the MLE of μ_r can be given in closed-form; the MLEs of $\{\mathbf{q}_r, \mathbf{S}_r\}_{r=1,\ldots,R}$ can be obtained by gradient ascent on $P(\hat{\mathbf{y}}_1^{(a)}, \ldots, \hat{\mathbf{y}}_N^{(a)}, \mathbf{y}_l)$, while those of $\mathbf{S}_r$'s (which are psd) by projected gradient ascent. Again, note that as $\mathbf{S}_r \in \mathbb{R}^{2\times 2}$, the projection onto the psd cone is computationally inexpensive.

Multiple Time Points

Typically, measurements are collected over a long period of time. Hence, in the real-world data sets, it is common for a location to have multiple measurements from the same satellite collected at different time points. In (Djuric, Kansakar, and Vucetic 2013), these measurements are simply taken as

independent. This can be problematic when the manifold is introduced. Specifically, though satellites have large spatial coverage, they still take a relatively long period of time to scan the whole globe (e.g., the MISR satellite takes 9 days (Meloni et al. 2004)). For a particular satellite, it is thus unlikely that measurements for the same location collected at different time points are similar to each other. In other words, measurements from different time stamps may not be smooth on the manifold.

To alleviate this problem, we extend the definition of manifold as follows. First, each node in the graph $\mathcal{G}$ is no longer a location, but a (location, time) pair. The weight between two such nodes (ℓ_i, t_i) and (ℓ_j, t_j) is defined as

$$w_{ij} = \begin{cases} \exp\left(-\frac{d_{\ell_i,\ell_j}}{\sqrt{s_{\ell_i}}\sqrt{s_{\ell_j}}}\right) & \ell_i \neq \ell_j \text{ and } t_i = t_j, \\ 0 & \text{otherwise.} \end{cases}$$

Experiments

In this section, experiments are performed on both synthetic data and real-world data from the ground-based AERONET measurements and five satellite measurements.

Synthetic Data

The experimental setup is similar to that in (Djuric, Kansakar, and Vucetic 2013). First, we consider the case with only one regime (the black cluster over North America in Figure 2). The locations are generated from the normal distribution $\mathcal{N}(\mathbf{q}_1, \mathbf{S}_1)$ with $\mathbf{q}_1 = [38, -100]^T$, and $\mathbf{S}_1 = \text{diag}([20, 60])$. The weight function on the manifold is defined by (9), where the radius r in (8) is set to a thousandth of that of the Earth (i.e., $r = 6.371$). For simplicity, all the scaling parameter s_i's in (9) are fixed to 1. The ground truth $\mathbf{y}$ vector is generated from (10), with $u = 0.1$, $\sigma^2 = 0.01$ and $\alpha = 1$. The number of satellites K is 5, and their measurements are sampled using (3), with $\mathbf{\Sigma} = \text{diag}([0.01, 0.02, 0.03, 0.04, 0.05])$. To simulate missing data, we remove each satellite measurement randomly with probability $p = 0.5$.

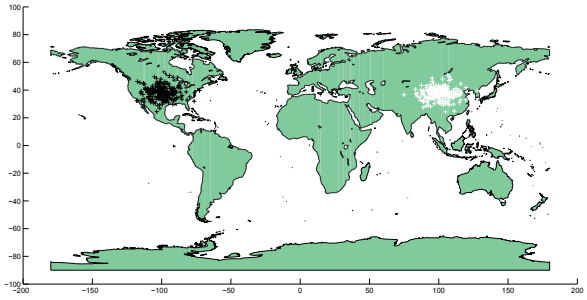

Figure 2: Synthetic data set showing the geographical locations for the two regimes. The first regime is colored in black and the second one in white.

In the sequel, the proposed algorithm will be called APM (Aerosol Prediction using Manifold). The following algorithms will be compared in the experiments:

1. APM with $\mathbf{\Sigma} = \mathbf{I}$;
2. APM with a learnable diagonal $\mathbf{\Sigma}$;
3. APM with a learnable full $\mathbf{\Sigma}$;
4. DKV, which is the algorithm[2] proposed in (Djuric, Kansakar, and Vucetic 2013); and
5. simple averaging of the available satellite measurements.

For performance evaluation, we use the root mean squared error (RMSE), $\sqrt{\frac{1}{N_u}\sum_{i=1}^{N_u}(y_i - f_i)^2}$, evaluated on a set of $N_u = 200$ unlabeled samples. Here, y_i is the ground truth and f_i the corresponding prediction. The number of labeled locations is varied from 0 to 200. To reduce statistical variability, results are averaged over 100 repetitions.

Prediction Accuracy Figure 3 shows the RMSEs. As can be seen, with the use of manifold, APM significantly outperforms both DKV and averaging[3]. This holds even when there is no labeled data, and the performance improves as the amount of labeled data increases. Moreover, note that APM with a learnable diagonal $\mathbf{\Sigma}$ achieves the lowest RMSE, which is consistent with the data generation process.

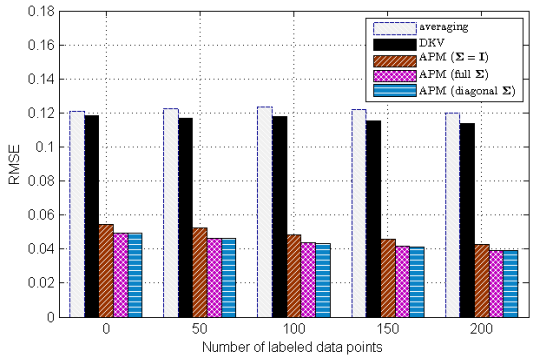

Figure 3: RMSE's obtained on the synthetic data set.

Manifold Noise In this experiment, we inject each label y_i with noise ξ_i, which is generated from the normal distribution with mean 0 and variance in $\{0.01, 0.1, 1\}$. The corresponding signal-to-noise ratios (averaged over the 100 repetitions) are 12.63, 1.27, and 0.13, respectively. To avoid clutterness, we only show the performance of APM with a learnable diagonal $\mathbf{\Sigma}$. Results are shown in Figure 4. As expected, the RMSE of APM increases as the manifold gets noisier, and becomes comparable with DKV and averaging only when the signal-to-noise ratio is as low as 0.13.[4]

[2]The code is provided by Djuric, Kansakar, and Vucetic.

[3]The improvements are statistically significant according to the pairwise t-test with 99% confidence.

[4]When the variance is 0.01 and 0.1, the improvements of APM over DKV and averaging are always statistically significant according to the pairwise t-test with 99% confidence. When the variance equals 1, the improvements of APM are statistically significant only with 0 and 50 labeled data points.

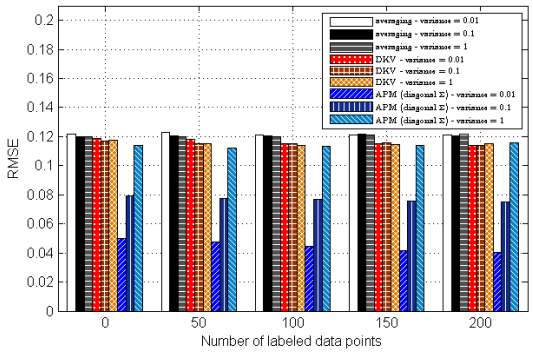

Figure 4: RMSE's with different amounts of manifold noise.

Varying the Number of Mixture Components In this experiment, we demonstrate the effect of using the mixture. First, we add one more regime (the white cluster over Asia in Figure 2), which generates locations from $\mathcal{N}(\mathbf{q}_2, \mathbf{S}_2)$ with $\mathbf{q}_2 = [38, 100]^T$, $\mathbf{S}_2 = \text{diag}([20, 60])$. Both regimes have equal probabilities of generating locations. The true μ_1, μ_2 values in (22) are 0.1 and 0.2, respectively. We evaluate the performance with 1, 2 and 3 mixture components.

Results are shown in Figure 5. As can be seen, APM again outperforms DKV and averaging.[5]. Moreover, the results of APM having 2 and 3 components are very similar. Indeed, when $R = 3$, two of the components obtained by APM have very similar μ_r values.

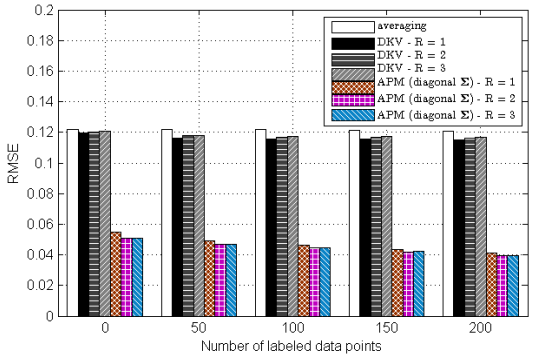

Figure 5: RMSE's with different mixture components.

Real-World Aerosol Data

In this section, we perform experiments using the ground-based AERONET data[6] (10:00-11:00am local time) and five satellite measurements[7] (including two from Terra MODIS and MISR at 10:00-11:00am local time; and three from Aqua MODIS, OMI and SeaWiFS at 1:00-2:00pm local time) from the years 2004-2010.

[5]The improvements are statistically significant according to the pairwise t-test with 99% confidence.

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

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

USA Data

As in (Djuric, Kansakar, and Vucetic 2013), we first experiment with data on the United States. After removing days with fewer than 10 observations, we obtain 2,382 data points spanning 206 time points and 86 locations. Overall, around 70% of the satellite measurements are missing. To reduce statistical variability, results are averaged over 100 repetitions. In each repetition, we randomly sample 150 time points, each with 10 locations. 5 of these are used as labeled locations, while the remaining 5 are unlabeled locations.

Table 1 shows the RMSE's obtained. As can be seen, APM has significantly lower RMSE compared with the other methods. In particular, the best performance is obtained with a diagonal $\mathbf{\Sigma}$, suggesting that the satellite measurements are indeed not strongly correlated, contrary to the assumption in (Djuric, Kansakar, and Vucetic 2013).

Table 1: RMSE on the real-world aerosol data sets. The improvements are statistically significant according to the pairwise t-test with 99% confidence.

		USA	Europe
averaging		0.0968 ± 0.0029	0.0784 ± 0.0032
DKV	$R = 1$	0.0968 ± 0.0031	0.0789 ± 0.0032
	$R = 2$	0.0967 ± 0.0030	0.0784 ± 0.0032
	$R = 3$	0.0970 ± 0.0030	0.0786 ± 0.0032
APM ($\mathbf{\Sigma} = \mathbf{I}$)	$R = 1$	0.0907 ± 0.0057	0.0646 ± 0.0015
	$R = 2$	0.0894 ± 0.0056	0.0642 ± 0.0015
	$R = 3$	0.0893 ± 0.0056	0.0642 ± 0.0015
APM (full $\mathbf{\Sigma}$)	$R = 1$	0.0690 ± 0.0028	0.0539 ± 0.0014
	$R = 2$	0.0676 ± 0.0028	0.0535 ± 0.0014
	$R = 3$	0.0676 ± 0.0028	0.0536 ± 0.0014
APM (diagonal $\mathbf{\Sigma}$)	$R = 1$	0.0688 ± 0.0028	0.0538 ± 0.0009
	$R = 2$	$\mathbf{0.0674 \pm 0.0027}$	$\mathbf{0.0535 \pm 0.0010}$
	$R = 3$	$\mathbf{0.0674 \pm 0.0027}$	0.0535 ± 0.0014

Europe Data

Next, we perform experiments with data on Europe. After removing days with fewer than 10 observations, we obtain 4,879 data points spanning 397 time points and 90 locations. Around 71% of the satellite measurements are missing. The rest of the experimental setup is the same as that for the USA data. The RMSE results are shown in Table 1. Again, APM outperforms the other methods.

Conclusion

In this paper, we proposed an enhanced probabilistic approach to integrate AOD measurements from satellite instruments and ground-based sensors. By considering the Earth's surface as a two-dimensional manifold, a Gaussian random field is used to enforce spatial smoothness of the AOD predictions. The resultant model allows simple probabilistic inference, and can handle missing satellite measurements and the division of locations into regimes. Experimental results on both synthetic and real-world data sets show that it significantly outperforms the state-of-the-art.

Acknowledgments

This research was supported in part by the Research Grants Council of the Hong Kong Special Administrative Region

(Grant 614012).

References

Abdou, W.; Diner, D.; Martonchik, J.; Bruegge, C.; Kahn, R.; Gaitley, B.; Crean, K.; Remer, L.; and Holben, B. 2005. Comparison of coincident Multiangle Imaging Spectroradiometer and Moderate Resolution Imaging Spectroradiometer aerosol optical depths over land and ocean scenes containing Aerosol Robotic Network sites. *Journal of Geophysical Research* 110(D10).

Belkin, M., and Niyogi, P. 2002. Using manifold structure for partially labeled classification. In *Advances in Neural Information Processing Systems 14*.

Belkin, M., and Niyogi, P. 2003. Laplacian eigenmaps for dimensionality reduction and data representation. *Neural Computation* 15(6):1373–1396.

Belkin, M.; Niyogi, P.; and Sindhwani, V. 2006. Manifold regularization: A geometric framework for learning from labeled and unlabeled examples. *Journal of Machine Learning Research* 7:2399–2434.

Bertsekas, D. 2004. *Nonlinear Programming*. Athena Scientific, 2nd edition.

Chapelle, O.; Weston, J.; and Schölkopf, B. 2002. Cluster kernels for semi-supervised learning. In *Advances in Neural Information Processing Systems 14*.

Chu, D.; Kaufman, Y.; Ichoku, C.; Remer, L.; Tanré, D.; and Holben, B. 2002. Validation of MODIS aerosol optical depth retrieval over land. *Geophysical Research Letters* 29(12):8007.

Djuric, N.; Kansakar, L.; and Vucetic, S. 2013. Semi-supervised learning for integration of aerosol predictions from multiple satellite instruments. In *Proceedings of the International Joint Conference on Artificial Intelligence*, 2797–2803.

Gupta, A., and Nagar, D. 2000. *Matrix Variate Distributions*. Chapman & Hall/CRC.

Koelemeijer, R.; Homan, C.; and Matthijsen, J. 2006. Comparison of spatial and temporal variations of aerosol optical thickness and particulate matter over Europe. *Atmospheric Environment* 40(27):5304–5315.

Levy, R.; Remer, L.; and Dubovik, O. 2007. Global aerosol optical properties and application to Moderate Resolution Imaging Spectroradiometer aerosol retrieval over land. *Journal of Geophysical Research* 112(D13210).

Meloni, D.; Di Sarra, A.; Di Iorio, T.; and Fiocco, G. 2004. Direct radiative forcing of saharan dust in the Mediterranean from measurements at Lampedusa Island and MISR space-borne observations. *Journal of Geophysical Research* 109(D8).

Mishchenko, M.; Liu, L.; Geogdzhayev, I.; Travis, L.; Cairns, B.; and Lacis, A. 2010. Toward unified satellite climatology of aerosol properties.: 3. MODIS versus MISR versus AERONET. *Journal of Quantitative Spectroscopy and Radiative Transfer* 111(4):540–552.

Sinnott, R. 1984. Virtues of the Haversine. *Sky and Telescope* 68(2):158–159.

Verbeek, J., and Vlassis, N. 2006. Gaussian fields for semi-supervised regression and correspondence learning. *Pattern Recognition* 39(10):1864–1875.

Wang, J., and Christopher, S. 2003. Intercomparison between satellite-derived aerosol optical thickness and PM2.5 mass: Implications for air quality studies. *Geophysical Research Letters* 30(21):2095.

Watson, R.; Rodhe, H.; Oeschger, H.; and Siegenthaler, U. 1990. Greenhouse gases and aerosols. *Climate Change: The IPCC Scientific Assessment* 1:17.

Zelnik-Manor, L., and Perona, P. 2004. Self-tuning spectral clustering. In *Advances in Neural Information Processing Systems 17*, 1601–1608.

Zhang, F. 2005. *The Schur Complement and Its Applications*. Springer.

Zhou, D.; Bousquet, O.; Lal, T.; Weston, J.; ; and Schölkopf, B. 2004a. Learning with local and global consistency. In *Advances in Neural Information Processing Systems 16*.

Zhou, D.; Weston, J.; Gretton, A.; Bousquet, O.; and Schölkopf, B. 2004b. Ranking on data manifolds. In *Advances in Neural Information Processing Systems 16*, 169–176.

Zhu, X.; Ghahramani, Z.; and Lafferty, J. 2003. Semi-supervised learning using Gaussian fields and harmonic functions. In *Proceedings of the 20th International Conference on Machine Learning*, 912–919.

Zhu, X.; Lafferty, J.; and Ghahramani, Z. 2003. Semi-supervised learning: From Gaussian fields to Gaussian processes. Technical Report CMU-CS-03-175, Carnegie Mellon University.

Zhu, X. 2007. Semi-supervised learning literature survey. Technical Report 1530, Department of Computer Sciences, University of Wisconsin - Madison.