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

Nemanja Djuric, Lakesh Kansakar, Slobodan Vucetic

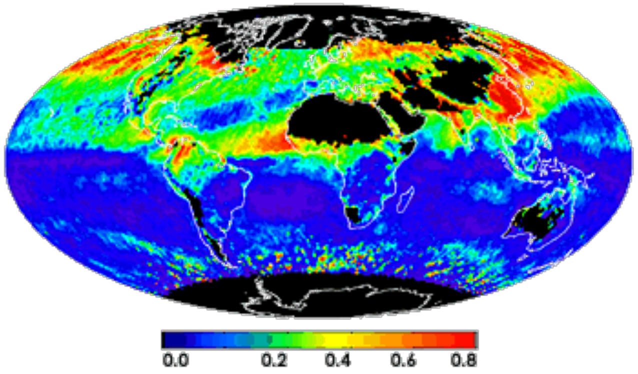
Temple University, Philadelphia

Aerosols

- Aerosols are small particles suspended in the atmosphere, originating from natural and man-made sources
  - Smoke, sea salt, dust, volcano ash, fossil fuel burning
- Negative effect on public health
  - Lung cancer, asthma, birth defects
- Profound effect on Earth's radiation budget
  - Absorb and reflect sunlight
  - Can have either cooling or heating effect on the Earth

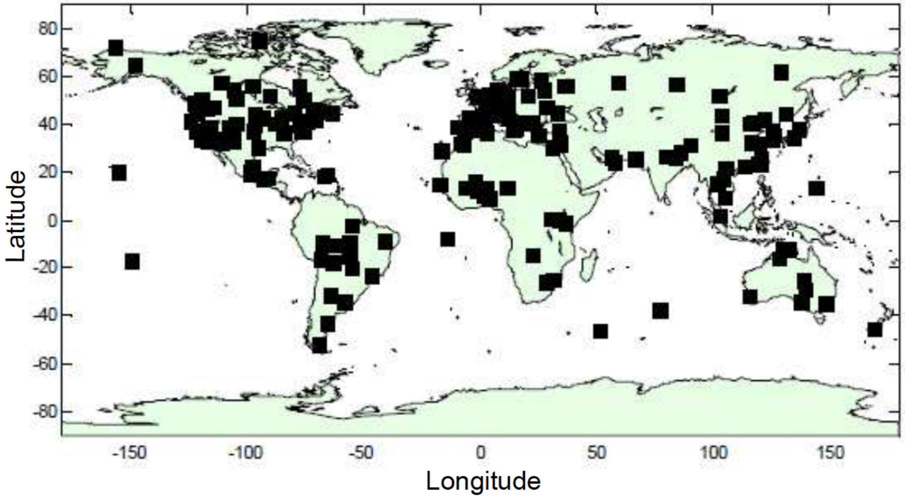
Aerosols

- Estimation of global aerosol distribution is one of the biggest challenges in climate research
  - United Nations Intergovernmental Panel on Climate Change: Aerosols are one of the major sources of uncertainty in climate models
- Standard measure of aerosol distribution is Aerosol Optical Depth (AOD)
  - AOD measures extinction of Solar radiation within the atmosphere
  - Higher AOD ➔ higher aerosol concentration



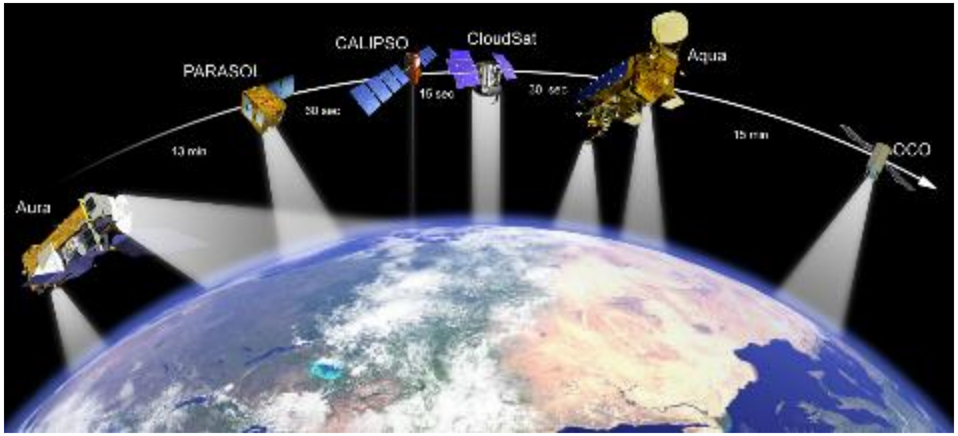
Measurement of AOD

- Ground-based sensors (Sun photometers)
  - High cost of installment and maintenance
  - High accuracy of AOD estimates
  - AERONET network of instruments
    - sparse and uneven distribution



Measurement of AOD

- Satellite-based sensors
  - Instruments aboard Terra, Aqua, Aura, Calipso, SeaStar, and other satellites
  - Lower accuracy of AOD estimation
  - Global daily coverage



Satellite-based AOD measurement

- Different satellite sensors have different:
  - Spatial coverage
  - Accuracy
  - Sensitivity to atmospheric and ground conditions
- Climate scientists typically choose one of the satellites for their climate models
- Combining different satellite measurements into a single, more accurate aggregated estimate possibly the best path towards high-quality, global AOD estimation

Problem setting

- We are given training data set consisting of targets y_i (*AERONET*) and of estimates of y_i by K different experts (*satellites*), with N_u unlabeled and N_l labeled data points

$$D = D_u \cup D_l = \{\{\hat{y}_{ik}\}_{k=1,\ldots,K}\}_{i=1,\ldots,N_u} \cup \{y_i, \{\hat{y}_{ik}\}_{k=1,\ldots,K}\}_{i=N_u+1,\ldots,N_u+N_l}$$

- ***OBJECTIVE:*** *find an optimal linear combination of available satellite measurements, using scarce AERONET measurements as a ground-truth AOD during training*

Issues inherent to remote sensing

1. Satellite prediction errors are **correlated**
2. Satellite predictions may be **missing** (due to lack of coverage or due to presence of clouds)
3. Number of **labeled** data points is small and orders of magnitude less than number of **unlabeled** data points
4. Satellites should be **combined differently** for different parts of the world (e.g., MISR does not maintain the same quality of AOD estimates across the globe)

Related work: Combination of experts

- Bates and Granger, 1969;
  Granger and Ramanathan, 1984
  - Supervised method, no missing data allowed
- Raykar et al., 2009; Ristovski et al., 2010
  - Unsupervised methods, no missing data allowed
  - Experts assumed independent
- The proposed semi-supervised method presents a significant generalization of the two approaches
  - Allows missing data, correlated experts, and finds different data-generating regimes

Assumptions

- Data points sampled IID, and target follows normal distribution,

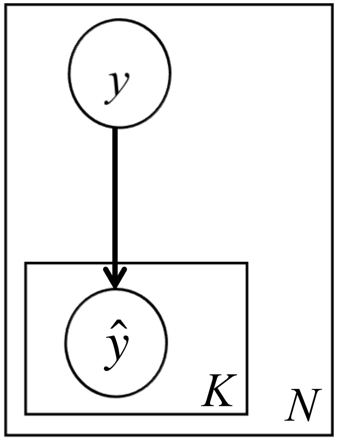
$$y_i \sim Norm(\mu_y, \sigma_y^2)$$

- Denote by $\hat{\mathbf{y}}_i$ a K-dimensional vector of expert predictions, sampled from multivariate Gaussian,

$$\hat{\mathbf{y}}_i \mid y_i \sim Norm(y_i \mathbf{1}, \Sigma)$$

- Training task is to find the parameters

$$\Theta = \{\Sigma, \mu_y, \sigma_y^2\}$$



Inference

- Once the training is completed, aggregated prediction can be found as a mean of the posterior distribution

$$y_i \mid \hat{\mathbf{y}}_i \sim Norm(\bar{y}_i, (\mathbf{1}^{\mathrm{T}} \Sigma'^{-1} \mathbf{1})^{-1})$$

where the mean can be computed as follows,

$$\bar{y}_i = \frac{\hat{\mathbf{y}}_i'^{\mathrm{T}} \Sigma'^{-1} \mathbf{1}}{\mathbf{1}^{\mathrm{T}} \Sigma'^{-1} \mathbf{1}}, \quad \text{with} \quad \hat{\mathbf{y}}_i' = [\hat{\mathbf{y}}_i^{\mathrm{T}}, \mu_y]^{\mathrm{T}} \quad \text{and} \quad \Sigma' = \begin{bmatrix} \Sigma & \mathbf{0} \\ \mathbf{0} & \sigma_y^2 \end{bmatrix}$$

Training – No missing experts

- We write probability of the training data as follows

$$P(D \mid \Theta) = P(D_u \mid \Theta) \cdot P(D_l \mid \Theta)$$

- Learning by maximizing likelihood of the training data
  - Before considering more general setting, we first derive equations for the case where all experts are available

- The probability of unlabeled data set is equal to

$$P(D_u \mid \Theta) = \prod_{i=1}^{N_u} P(\hat{\mathbf{y}}_i \mid \Theta) = \prod_{i=1}^{N_u} \int_y P(\hat{\mathbf{y}}_i \mid y, \Theta) P(y \mid \Theta)\, dy$$

$$= \prod_{i=1}^{N_u} \left(\sqrt{\frac{|\Sigma'|^{-1}}{(2\pi)^{K-1} \mathbf{1}^{\mathrm{T}} \Sigma'^{-1} \mathbf{1}}} \right) \exp\left(-\frac{1}{2} (\hat{\mathbf{y}}'_i - \bar{y}_i \mathbf{1})^{\mathrm{T}} \Sigma'^{-1} (\hat{\mathbf{y}}'_i - \bar{y}_i \mathbf{1})\right)$$

Training – No missing experts

- Further, the probability of labeled data can be written as

$$P(D_l \mid \Theta) = \prod_{i=N_u+1}^{N} P(\hat{\mathbf{y}}_i \mid y_i, \Theta) = \prod_{i=N_u+1}^{N} \frac{1}{(2\pi)^{K/2} \mid \Sigma \mid^{0.5}} \exp(-0.5(\hat{\mathbf{y}}_i - y_i \mathbf{1})^{\mathrm{T}} \Sigma^{-1} (\hat{\mathbf{y}}_i - y_i \mathbf{1}))$$

- To simplify the equations, in the following we assume that $\sigma_y^2 \to \infty$, which amounts to an uninformative prior over the target variable

- After finding derivative of the data log-likelihood with respect to $\boldsymbol{\Sigma}^{-1}$, we obtain the iterative update equation,

$$\Sigma = \frac{1}{N}(\underbrace{(\hat{\mathbf{Y}}_l - \mathbf{y}_l \mathbf{1}^{\mathrm{T}})^{\mathrm{T}} (\hat{\mathbf{Y}}_l - \mathbf{y}_l \mathbf{1}^{\mathrm{T}})}_{\text{Bates and Granger, 1969}} + \underbrace{\hat{\mathbf{Y}}_u^{\mathrm{T}} \hat{\mathbf{Y}}_u + \frac{N_u \mathbf{1}\mathbf{1}^{\mathrm{T}}}{\mathbf{1}^{\mathrm{T}} \Sigma^{-1} \mathbf{1}} + \sum_{i=1}^{N_u} (\bar{y}_i^2 \mathbf{1}\mathbf{1}^{\mathrm{T}} - \bar{y}_i (\mathbf{1} \hat{\mathbf{y}}_i^{\mathrm{T}} + \hat{\mathbf{y}}_i \mathbf{1}^{\mathrm{T}}))}_{\text{Ristovski et al., 2010}})$$

Inference – Missing experts

- Assume that the i^{th} data point has q out of K experts missing
- We reorganize vector $\hat{\mathbf{y}}_i$ so the first a elements are from available experts, and the last q elements are missing

$$\hat{\mathbf{y}}_i = [\hat{\mathbf{y}}_{ai}^{\mathrm{T}}, \hat{\mathbf{y}}_{qi}^{\mathrm{T}}]^{\mathrm{T}}$$

- We similarly reorganize precision matrix, so that the first a rows/columns correspond to available experts

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

- Given the learned covariance matrix and $\hat{\mathbf{y}}_{ai}$, it follows

$$y_i \mid \hat{\mathbf{y}}_{ai} \sim Norm(\bar{y}_i, (\mathbf{1}^{\mathrm{T}}\mathbf{U}'_i\mathbf{1})^{-1}), \text{ where } \bar{y}_i = \frac{\hat{\mathbf{y}}_{ai}^{\mathrm{T}}\mathbf{U}'_i\mathbf{1}}{\mathbf{1}^{\mathrm{T}}\mathbf{U}'_i\mathbf{1}} \text{ and } \mathbf{U}' = \mathbf{U} - \mathbf{V}\mathbf{Q}^{-1}\mathbf{V}^{\mathrm{T}}$$

Training – Missing experts

- We again derive the equations for probabilities of unlabeled and labeled parts of the training set
- Probability of the i^{th} unlabeled data point can be found as

$$P(\hat{\mathbf{y}}_{ai} \mid \Theta) = \iint_{y,\hat{\mathbf{y}}_{qi}} P([\hat{\mathbf{y}}_{ai}^{\mathrm{T}}, \hat{\mathbf{y}}_{qi}^{\mathrm{T}}]^{\mathrm{T}} \mid y, \Theta) P(y \mid \Theta)\, dy\; d\hat{\mathbf{y}}_{qi}$$

$$= \left(\sqrt{\frac{|\Sigma|^{-1} |\mathbf{Q}_i|^{-1}}{(2\pi)^{K+q-1} \mathbf{1}^{\mathrm{T}} \mathbf{U}'_i \mathbf{1}}}\right) \exp\left(-\frac{1}{2} (\hat{\mathbf{y}}_{ai} - \overline{y}_i \mathbf{1})^{\mathrm{T}} \mathbf{U}'_i (\hat{\mathbf{y}}_{ai} - \overline{y}_i \mathbf{1})\right)$$

- Probability of the i^{th} labeled data point can be found as

$P(\hat{\mathbf{y}}_{ai} \mid y_i, \Theta) = \int_{\hat{\mathbf{y}}_{qi}} P([\hat{\mathbf{y}}_{ai}^{\mathrm{T}}, \hat{\mathbf{y}}_{qi}^{\mathrm{T}}]^{\mathrm{T}} \mid y_i, \Theta) d\hat{\mathbf{y}}_{qi}$, resulting in $\hat{\mathbf{y}}_{ai} \mid y_i \sim Norm(y_i \mathbf{1}, \mathbf{U}'^{-1}_i)$

Training – Missing experts

- We find the derivative of data log-likelihood with respect to precision matrix Σ^{-1} to obtain the update equation,

$$\Sigma = \frac{1}{N}\Big(\sum_{i=1}^{N}\Pi_i^{-1}(\Psi_i) + \sum_{i=N_u+1}^{N}\left\langle(\hat{\mathbf{y}}_{ai} - y_i\mathbf{1})(\hat{\mathbf{y}}_{ai} - y_i\mathbf{1})^{\mathrm{T}}\right\rangle +$$

$$\sum_{i=1}^{N_u}\Big(\left\langle\hat{\mathbf{y}}_{ai}\hat{\mathbf{y}}_{ai}^{\mathrm{T}}\right\rangle + \frac{\left\langle\mathbf{1}\mathbf{1}^{\mathrm{T}}\right\rangle}{\mathbf{1}^{\mathrm{T}}\mathbf{U}'_i\mathbf{1}} + \overline{y}_i^2\left\langle\mathbf{1}\mathbf{1}^{\mathrm{T}}\right\rangle - \overline{y}_i\left\langle\mathbf{1}\hat{\mathbf{y}}_{ai}^{\mathrm{T}} + \hat{\mathbf{y}}_{ai}\mathbf{1}^{\mathrm{T}}\right\rangle\Big)\Big),$$

where $\left\langle\mathbf{A}_i\right\rangle = \Pi_i^{-1}\left(\begin{bmatrix}\mathbf{A}_i & -\mathbf{A}_i\mathbf{V}_i\mathbf{Q}_i^{-1}\\ -\mathbf{Q}_i^{-1}\mathbf{V}_i^{\mathrm{T}}\mathbf{A}_i & -\mathbf{Q}_i^{-1}\mathbf{V}_i^{\mathrm{T}}\mathbf{A}_i\mathbf{V}_i\mathbf{Q}_i^{-1}\end{bmatrix}\right)$

$$\Psi_i = \begin{bmatrix}\mathbf{0} & \mathbf{0}\\ \mathbf{0} & \mathbf{Q}_i^{-1}\end{bmatrix}$$

Including prior knowledge

- Assume we have prior knowledge about experts' correlation, we can write the joint probability of data and parameters as

$$P(D,\Theta) = P(D \mid \Sigma^{-1}) P(\Sigma^{-1})$$

- For the prior on precision matrix, we assume Wishart distribution

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

- This results in the following update rule (after setting $n = K + 2$)

$$\Sigma = \frac{1}{N+1}(\mathbf{S}^{-1} + \sum_{i=1}^{N} \Pi_i^{-1}(\Psi_i) + \sum_{i=N_u+1}^{N} \left\langle (\hat{\mathbf{y}}_{ai} - y_i\mathbf{1})(\hat{\mathbf{y}}_{ai} - y_i\mathbf{1})^{\mathrm{T}} \right\rangle +$$

$$\sum_{i=1}^{N_u} (\left\langle \hat{\mathbf{y}}_{ai}\hat{\mathbf{y}}_{ai}^{\mathrm{T}} \right\rangle + \frac{\left\langle \mathbf{1}\mathbf{1}^{\mathrm{T}} \right\rangle}{\mathbf{1}^{\mathrm{T}}\mathbf{U}'_i\mathbf{1}} + \bar{y}_i^2 \left\langle \mathbf{1}\mathbf{1}^{\mathrm{T}} \right\rangle - \bar{y}_i \left\langle \mathbf{1}\hat{\mathbf{y}}_{ai}^{\mathrm{T}} + \hat{\mathbf{y}}_{ai}\mathbf{1}^{\mathrm{T}} \right\rangle))$$

Mixture of regimes

- Let us assume that the experts do not maintain the same level of accuracy across all data points
- We derive an approach for partitioning data into several **regimes**, where expert predictions within each regime are sampled from a different multivariate Gaussian
- We assume existence of feature vectors $\mathbf{x}_i$, which can be used to assign examples to different regimes (e.g., time and/or location information in AOD estimation task)

Inference – Mixture of regimes

- Assuming a mixture of R regimes, probability of expert predictions for the i^{th} labeled data point can be written as

$$P(\hat{\mathbf{y}}_{ai} \mid y_i, \mathbf{x}_i, \Theta) = \sum_{r=1}^{R} \pi_{ir}(\mathbf{x}_i) P_r(\hat{\mathbf{y}}_{ai} \mid y_i)$$

- Similarly, probability of the unlabeled data point is

$$P(\hat{\mathbf{y}}_{ai} \mid \mathbf{x}_i, \Theta) = \sum_{r=1}^{R} \pi_{ir}(\mathbf{x}_i) P_r(\hat{\mathbf{y}}_{ai})$$

- Then, given a trained model, the aggregated prediction can be found as

$$\bar{y}_i = E[y_i \mid \hat{\mathbf{y}}_{ai}, \mathbf{x}_i, \Theta] = \sum_{r=1}^{R} \pi_{ir}(\mathbf{x}_i) \frac{\hat{\mathbf{y}}_{ai}^{\mathrm{T}} \mathbf{U}'_{ir} \mathbf{1}}{\mathbf{1}^{\mathrm{T}} \mathbf{U}'_{ir} \mathbf{1}}$$

Training – Mixture of regimes

- However, not easy to maximize log-likelihood due to the sum
- To address this issue, we introduce R latent binary variables z_{ir}, indicating whether or not the i^{th} data point was generated by the r^{th} regime, resulting in

$$P(\hat{\mathbf{y}}_{ai}, \mathbf{z}_i \mid y_i, \mathbf{x}_i, \Theta) = \prod_{r=1}^{R} (\pi_{ir}(\mathbf{x}_i) P_r(\hat{\mathbf{y}}_{ai} \mid y_i))^{z_{ir}}$$

- We define prior probability over regimes using softmax

$$\pi_{ir} = \frac{\exp(-(\mathbf{x}_i - \mathbf{q}_r)^{\mathrm{T}} \Lambda_r (\mathbf{x}_i - \mathbf{q}_r))}{\sum_{m=1}^{R} \exp(-(\mathbf{x}_i - \mathbf{q}_m)^{\mathrm{T}} \Lambda_m (\mathbf{x}_i - \mathbf{q}_m))}$$

- The log-likelihood is now much easier to maximize, equaling

$$L = \sum_{i=1}^{N} \sum_{r=1}^{R} z_{ir} (\log \pi_{ir}(\mathbf{x}_i) + \log P_r(\hat{\mathbf{y}}_{ai} \mid y_i))$$

Mixture of regimes – EM algorithm

- E-step:

$$h_{ir} = E[z_{ir} \mid \hat{\mathbf{y}}_{ai}, y_i, \mathbf{x}_i, \Theta] = \frac{\pi_{ir}(\mathbf{x}_i) P_r(\hat{\mathbf{y}}_{ai} \mid y_i)}{\sum_{m=1}^{R} \pi_{im}(\mathbf{x}_i) P_m(\hat{\mathbf{y}}_{ai} \mid y_i)}$$

- M-step:

$$\Sigma_r = \frac{1}{1 + \sum_{i=1}^{N} h_{ir}} (\mathbf{S}_r^{-1} + \sum_{i=1}^{N} h_{ir} \Pi_i^{-1}(\Psi_{ir}) + \sum_{i=N_u+1}^{N} h_{ir} \left\langle (\hat{\mathbf{y}}_{ai} - y_i \mathbf{1})(\hat{\mathbf{y}}_{ai} - y_i \mathbf{1})^{\mathrm{T}} \right\rangle_r +$$

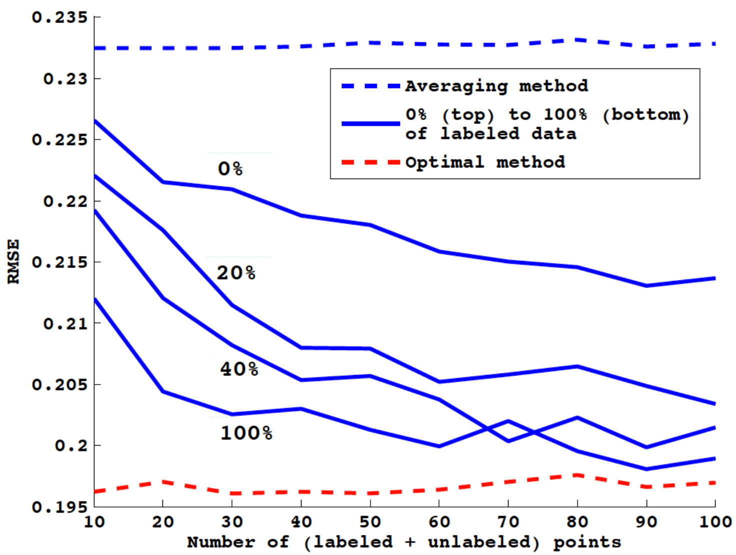
$$\sum_{i=1}^{N_u} h_{ir} (\left\langle \hat{\mathbf{y}}_{ai} \hat{\mathbf{y}}_{ai}^{\mathrm{T}} \right\rangle_r + \frac{\left\langle \mathbf{1}\mathbf{1}^{\mathrm{T}} \right\rangle_r}{\mathbf{1}^{\mathrm{T}} \mathbf{U}'_{ir} \mathbf{1}} + \bar{y}_{ir}^2 \left\langle \mathbf{1}\mathbf{1}^{\mathrm{T}} \right\rangle_r - \bar{y}_{ir} \left\langle \mathbf{1} \hat{\mathbf{y}}_{ai}^{\mathrm{T}} + \hat{\mathbf{y}}_{ai} \mathbf{1}^{\mathrm{T}} \right\rangle_r))$$

$$\mathbf{q}_r^{new} = \mathbf{q}_r^{old} + \eta \Lambda_r^{old} \sum_{i=1}^{N} (h_{ir} - \pi_{ir}^{old})(\mathbf{x}_i - \mathbf{q}_r^{old})$$

$$\Lambda_r^{new} = \Lambda_r^{old} + \eta \sum_{i=1}^{N} (h_{ir} - \pi_{ir}^{old})(\mathbf{x}_i - \mathbf{q}_r^{old})(\mathbf{x}_i - \mathbf{q}_r^{old})^{\mathrm{T}}$$

Experiments – Synthetic data

- Ground truth was sampled from zero mean, unit variance Gaussian, and we assumed $K = 5$ experts, each missing with 50% probability
  - For $R = 1$, we set $\Sigma = \text{diag}([0.1, 0.2, 0.3, 0.4, 0.5])$

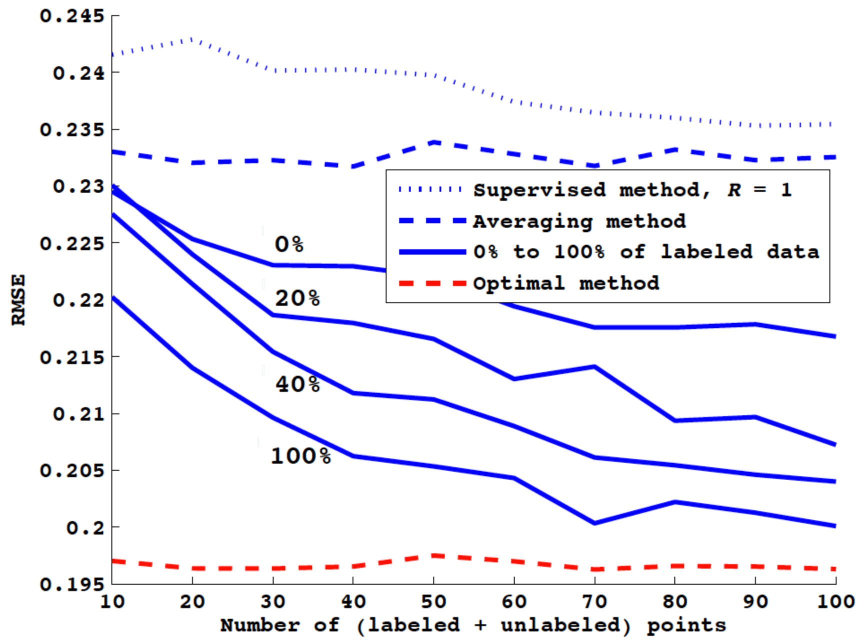


(a) Data generated by $R = 1$ regime

- We compared to averaging and optimal aggregation methods
- More unlabeled data leads to improved performance
- Small number of labeled data suffices

Experiments – Synthetic data

- For $R = 2$, we set $\mathbf{q}_1$=[1, 1], $\mathbf{q}_2$=[-1, -1], and Σ_1 = diag([0.1, 0.2, 0.3, 0.4, 0.5]), Σ_2 = diag([0.5, 0.4, 0.3, 0.2, 0.1])



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

- Wrong number of regimes leads to even worse performance
- EM-algorithm successfully found per-regime parameters

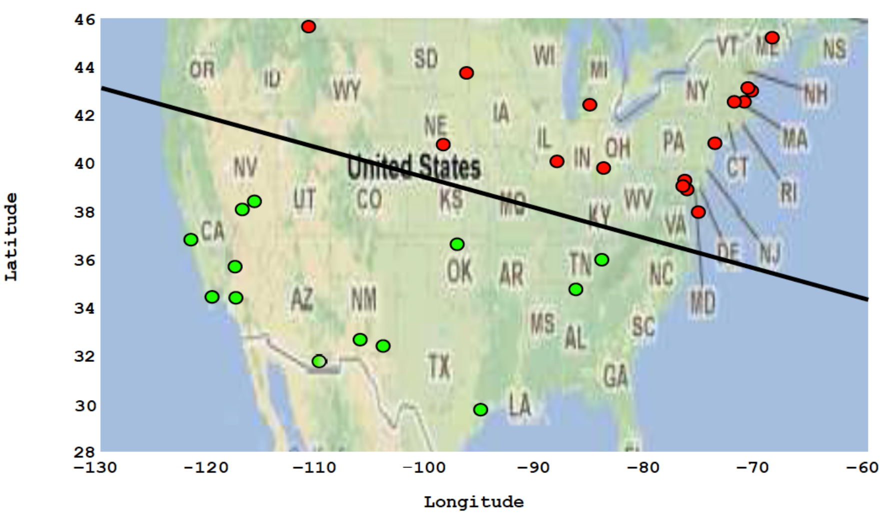
Experiments – Aerosol data

- We used 5 years of aerosol data from 33 AERONET US locations, and predictions from 5 experts (MISR, Terra MODIS, Aqua MODIS, OMI, SeaWiFS)
- Training data set with 6,913 examples (roughly 200 examples per site)
  - 58% of satellite predictions missing
  - Longitude and latitude used as $\mathbf{x}_i$ feature vectors

Experiments – Aerosol data

- Evaluating usefulness of partitioning
  - From each site we randomly sampled 100 points, and assumed that 50 are labeled and 50 unlabeled

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



Experiments – Aerosol data

- Evaluating usefulness of unlabeled data
  - Randomly selected 2, 4, and 6 sites and took 100 points from each as labeled data; then, we selected 100 points from each remaining site and treated them as unlabeled

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

- Simulates large areas where just few AERONET sites are available
- Unlabeled data helpful, although benefit decreased when larger amounts of labeled data points available

Conclusion

- The proposed semi-supervised method combines noisy expert predictions
  - Accounts for correlations between expert predictions
  - Accounts for unlabeled data, as well as for missing expert predictions
  - Separates training data into clusters, and finds different linear combinations for each cluster
- Future work
  - Model AERONET measurements as noisy observations
  - Allow prior parameters on target variable to be functions of $\mathbf{x}_i$
  - Extend the model to account for spatio-temporal correlations

Thank you!

- Questions?