Regression Algorithms for Large Scale Earth Science Data

Kamalika Das

SGT | NASA Ames Research Center

Kamalika.Das@nasa.gov

www.cs.umbc.edu/~kdas1

Collaborator: Dr. Ashok N. Srivastava, NASA ARC

IDU @ NASA Ames

- Group description
  - 12 members (7 Ph.D. researchers), summer interns, partners through NASA Research Announcements and SBIRs
- Develop methods that perform anomaly detection, diagnosis, and prediction within datasets that are
  - Large
  - Distributed
  - Heterogeneous---numeric (continuous, discrete) and text data

Roadmap

- Introduction
- Gaussian Process regression (GPR)
- Block GP
- Block GP experimental results
- Sparsity pattern identification in GPR
- SPI-GP for large data sets
- SPI-GP experimental results
- Conclusion

Introduction

- Desired characteristics in a regression-based model
  - Accuracy
  - Interpretability
  - Scalability
  - Confidence
- Gaussian Process Regression (GPR)
  - Predicts a distribution (mean and variance)
  - Captures non-linear relationship in data

Gaussian Process regression

Training data
- $\mathbf{X}$ data matrix of observations – n x d
- y vector of target data – n x 1

Test data
- X* matrix of new observations – n* x d

Covariance function

$$K_{ij} = k(x_i, x_j), K_{ij}^* = k(x_i^*, x_j)$$

Goal
- Predict y* corresponding to X*

Model building
- Train hyperparameters on a sample of $\mathbf{X}$
- Compute covariance matrix K (n x n)

Prediction
- Compute cross covariance matrix K* (n* x n)
- Compute mean prediction on y* using

$$\widehat{y}^* = K^*(\lambda^2 I + K)^{-1} y$$

- Compute variance of prediction using

$$C = K^{**} - K^*(\lambda^2 I + K)^{-1} K^{*T}$$

Algorithm Analysis
- Storage Complexity: Storing covariance matrix $O(n^2)$
- Time Complexity: Computing matrix inversion $O(n^3)$

Scalable GPR literature

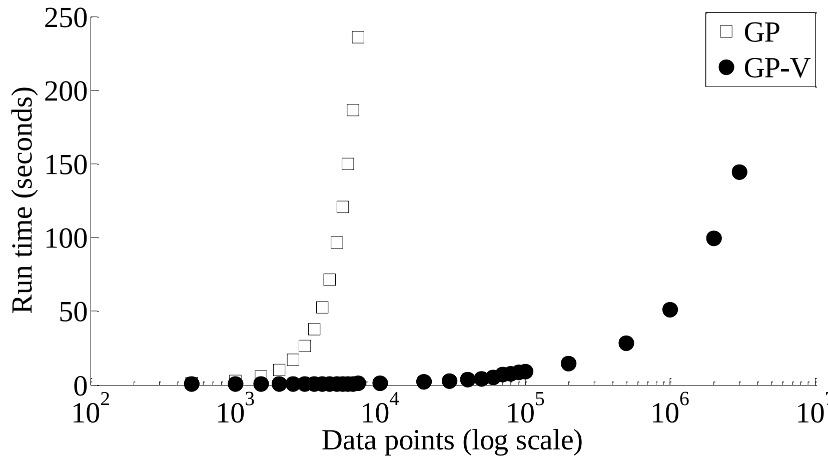
- Numerical Approximation: Subset of regressors

$$\widehat{y}_N^* = K_1^*(\lambda^2 K_{11} + K_1^T K_1)^{-1} K_1^T y$$

- where $K = \begin{pmatrix} K_{11} & K_{12} \\ K_{21} & K_{22} \end{pmatrix} = (K_1 \quad K_2), K^* = (K_1^* \quad K_2^*)$

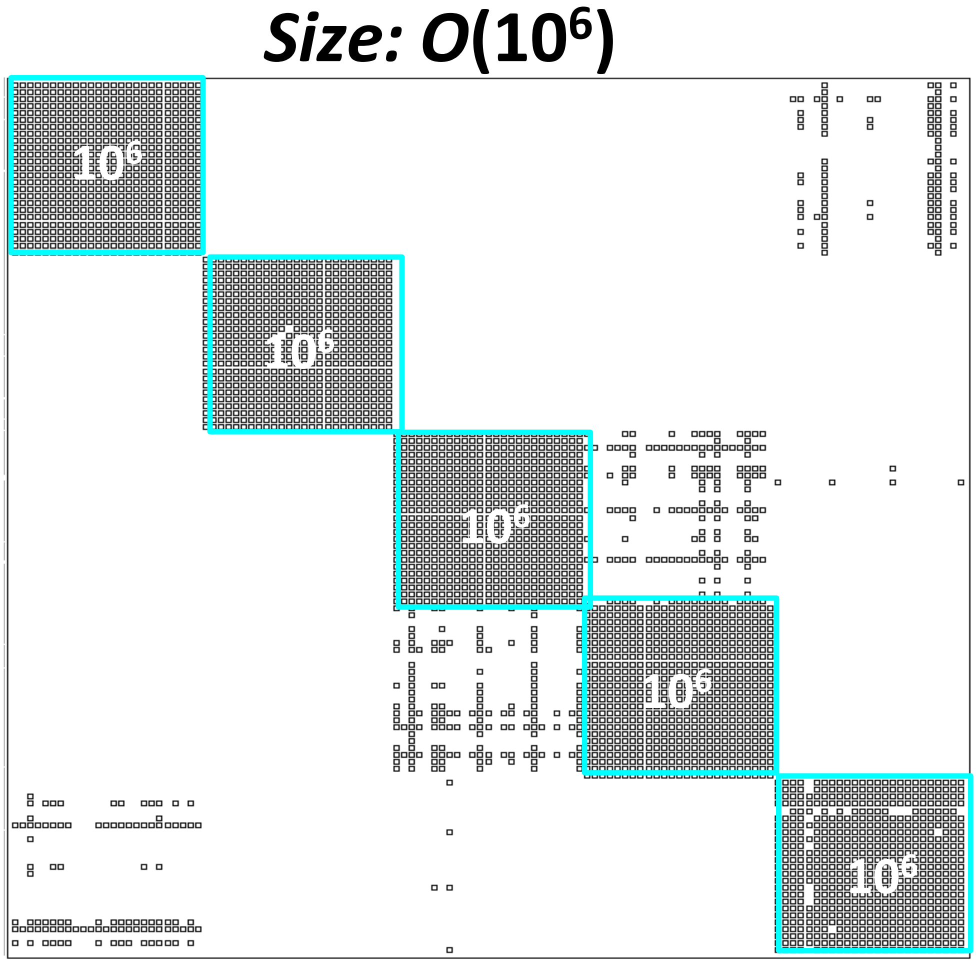
- Stable GP: Approximate $K_1 \approx V V_{11}^T$ by Cholesky factorization *with pivoting* where V is $n \times m$ and V_{11} is $m \times m$

Scalability analysis on simulated data



Graph Courtesy: Santanu Das and Ashok N. Srivastava

Illustration of GPR scalability



Mixture of experts literature

- Gaussian Process Mixture of Experts
  - Gating network decides which point is best predicted by each expert
  - Uses EM/MCMC methods for learning experts
  - All training points are used for training each experts
  - Very high convergence time and reduced scalability
- Scales up to the order of 10^3 data observations

Block GP

- Approximates Gaussian Process Mixture of Experts
  - Divides the data apriori into clusters
  - Builds separate models for each cluster/expert
  - Uses cluster membership probabilities to compute a weighted average of predictions by each cluster
  - Accounts for inter-cluster relationships

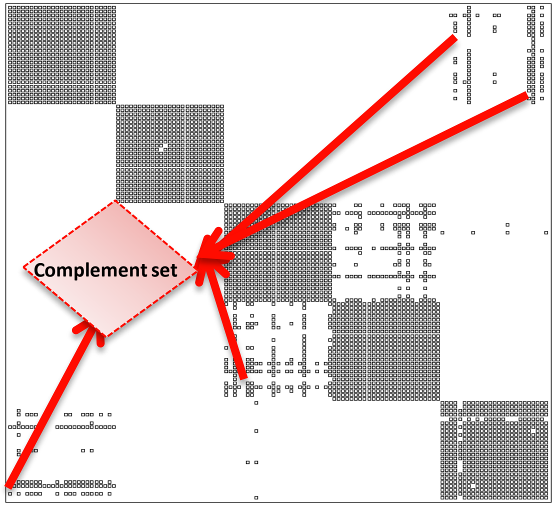
Block GP algorithm

1. Partition the data set using spectral clustering.
2. Train a GP for each partition.
3. Determine the cluster membership probability of each point for each cluster.
4. Those points that fall outside of the clusters are partitioned into a new cluster (complement set).
5. Retrain GP models for each clusters and the complement set.
6. Predicting new values using a weighted sum based on the cluster memberships and the predictions of each expert.

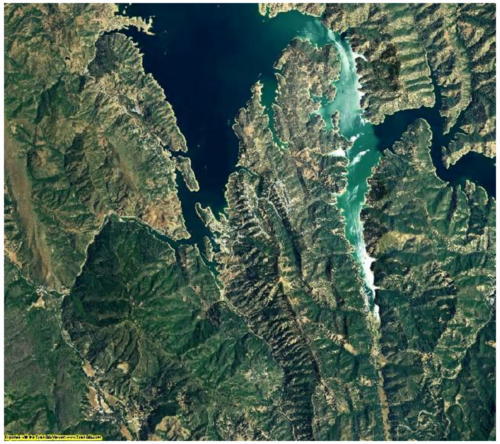
Final prediction equation is:

$$\widehat{y}^* = \sum_{i=1}^{k} h_i K_i^* (K_i + \sigma_i^2 I)^{-1} \mathbf{y}_i$$

where h_i represents the weight of the prediction by the i^{th} expert.

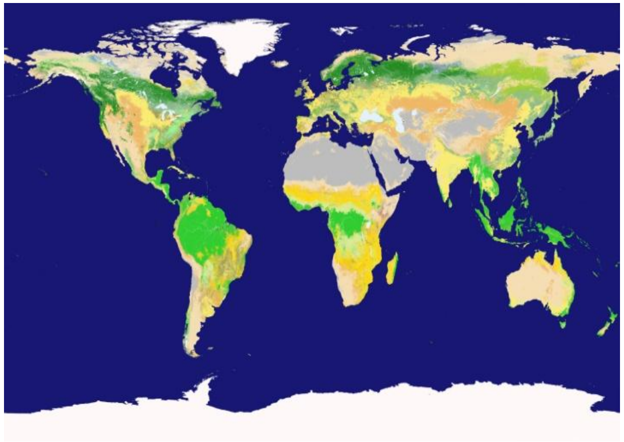


Real-life data sets: multimodality

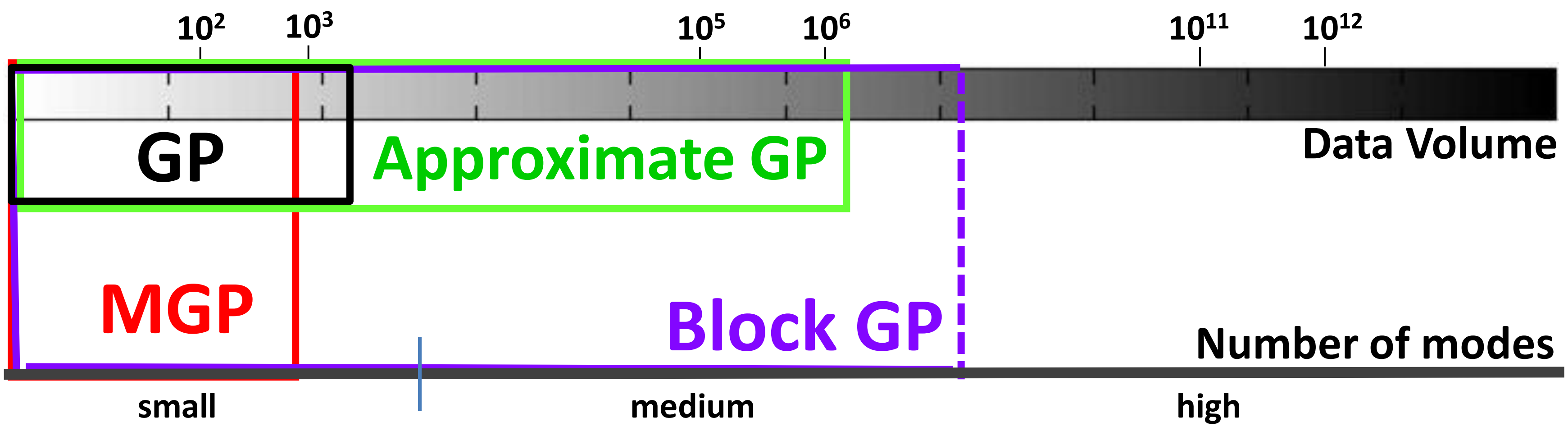


Napa

California

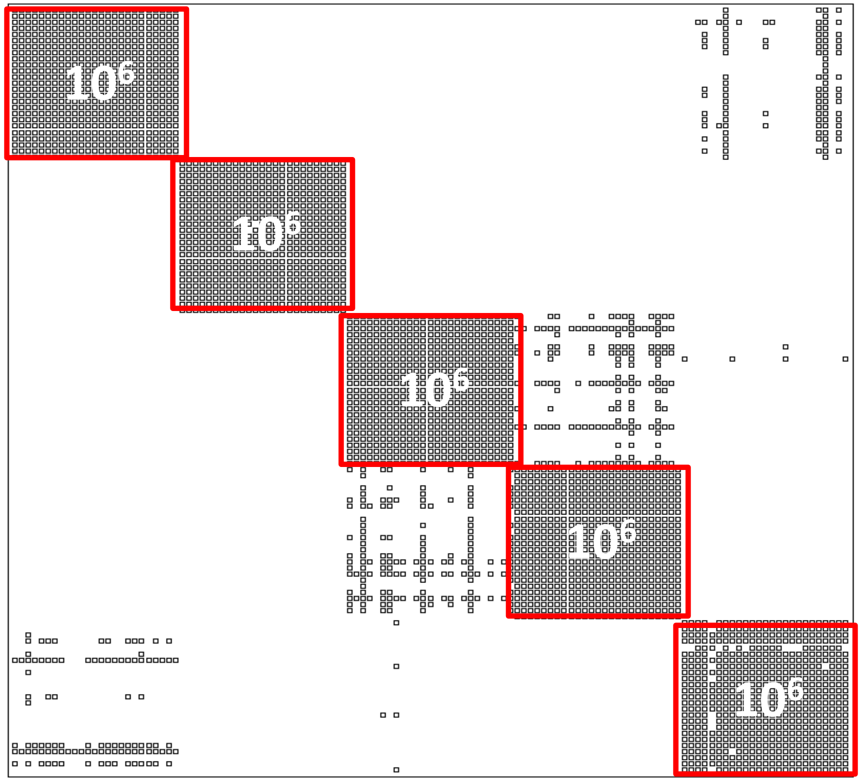


World



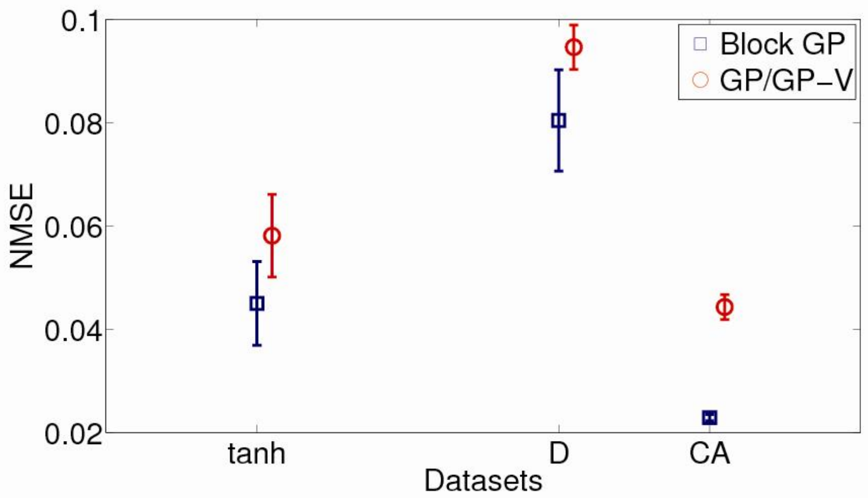
Block-GP performance analysis

- For number of modes k, number of dimensions d and maximum number of data points n_{max} prediction is $O((k+1)n_{\max}d^2))$
  - Higher scalability
  - Decomposability for distributed computation
  - Higher interpretability as different models predict different geographical regions accurately

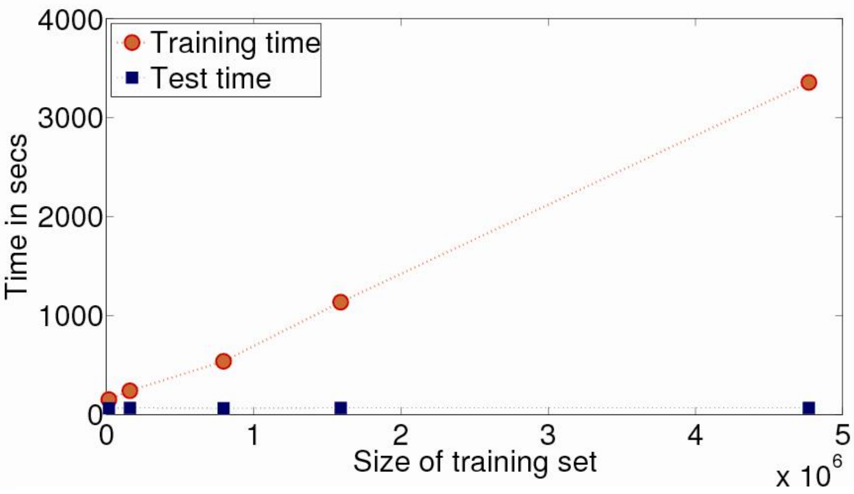


Use numerical approximation technique for each of the experts individually

Accuracy and running time



Mean and standard deviation of NMSE of Block-GP for different data sets

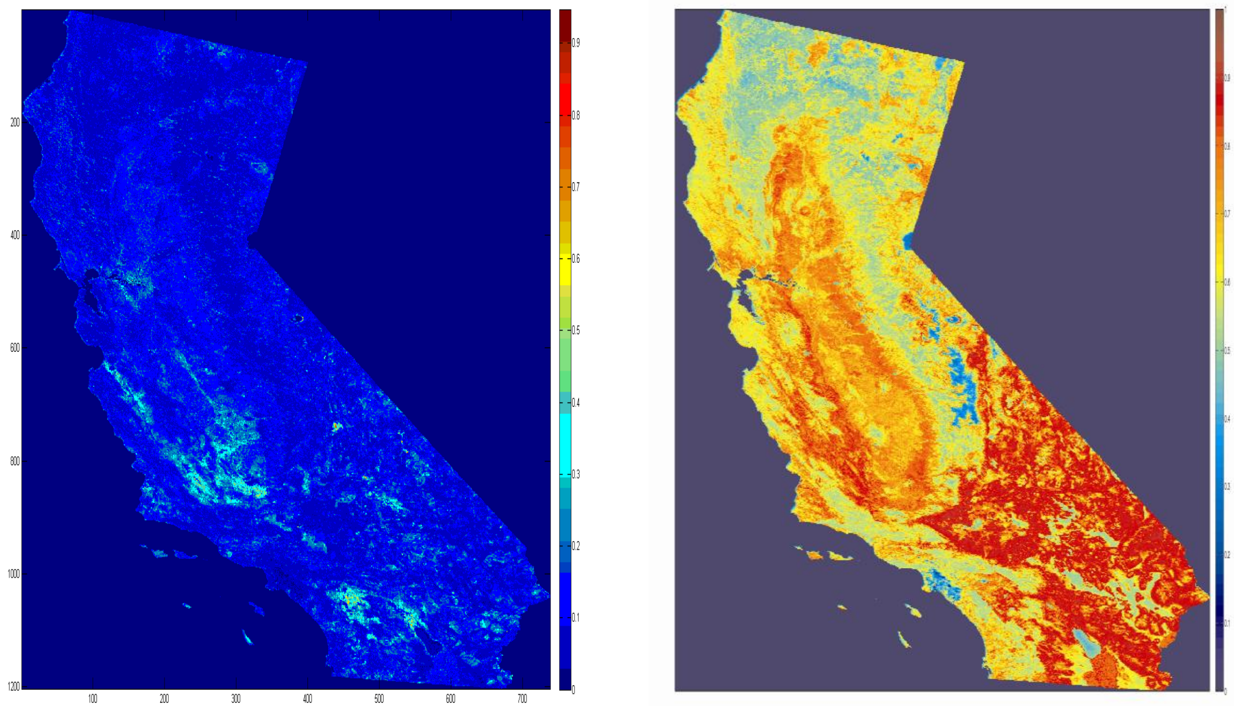


Running time of Block GP demonstrated on the California data set

Block-GP results

Data set	Modes	Size	Details
California	10	15,000,000 x 4	MODIS 8 day surface reflectance BRDF-adjusted from Terra and Aqua measured in 7 different wavelengths.

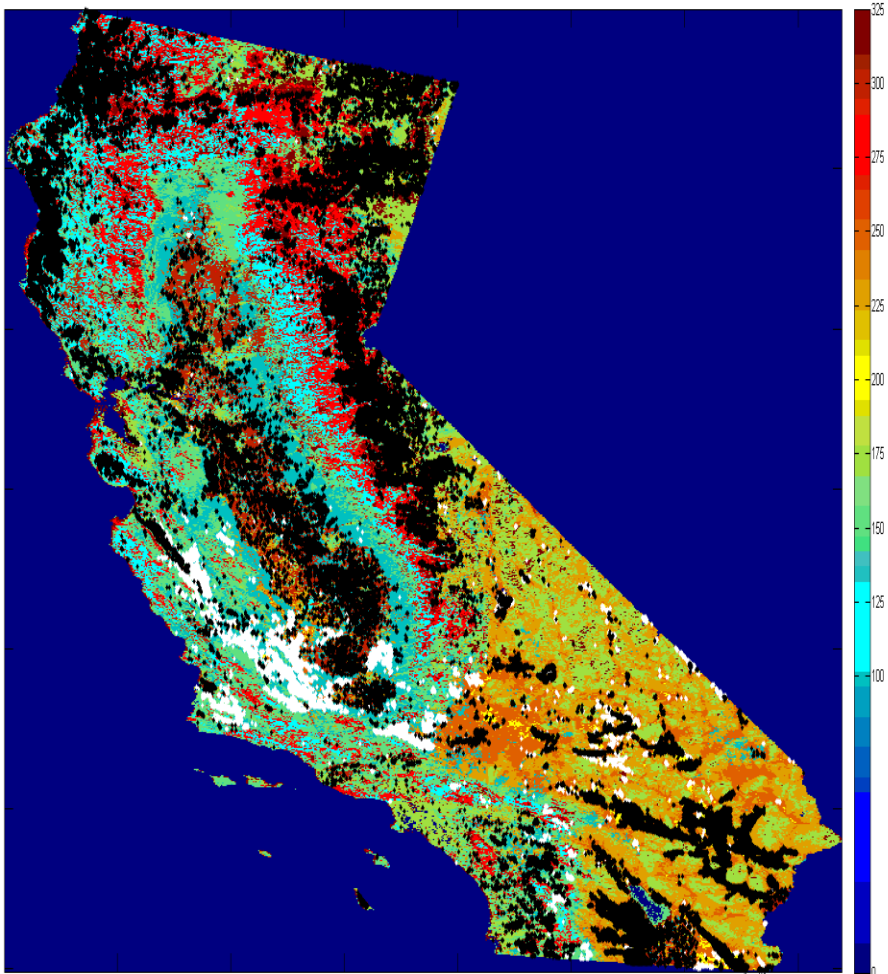
Prediction of band 6 using 1, 4 and 5



$$\frac{\mathrm{NMSE}_{\text{Block-GP}}}{\mathrm{NMSE}_{\text{low rank}}} = \frac{0.0229}{0.0443} \approx 52\%$$

Color map of normalized residual (left) and variance (right) for the prediction task

Block GP results



Top 5 percentile cases where Block-GP performed better

Top 5 percentile cases where low rank approx. performed better

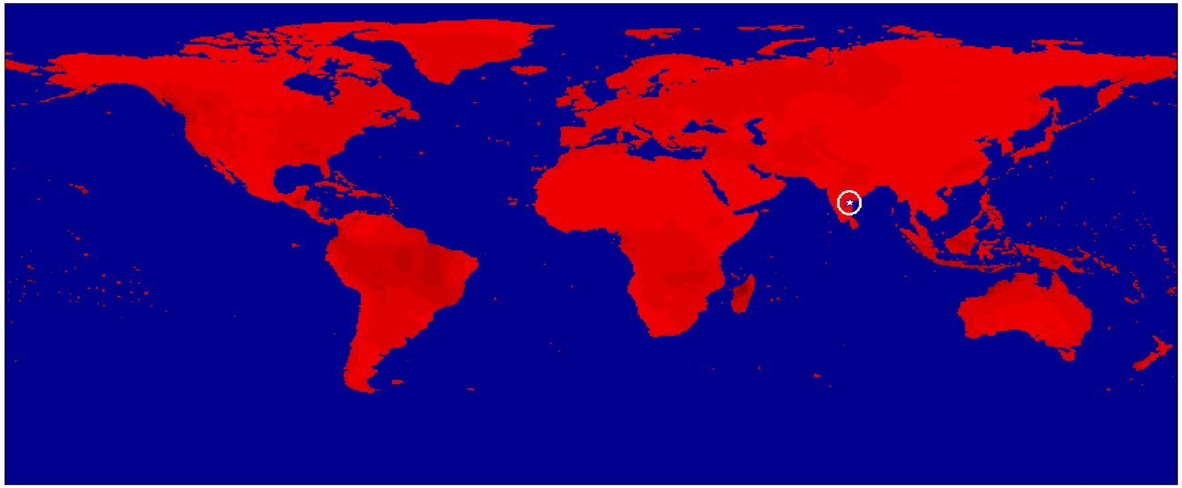
- Land cover changed with time
- Number of clusters
- Noisy target artifact

California color coded into 10 clusters based on surface reflectance using spectral clustering.

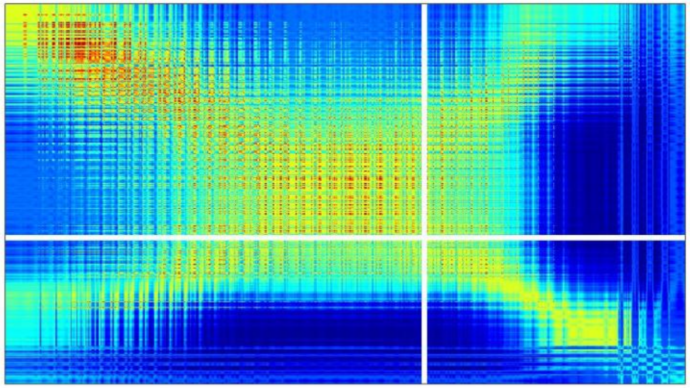
Covariance matrix structure

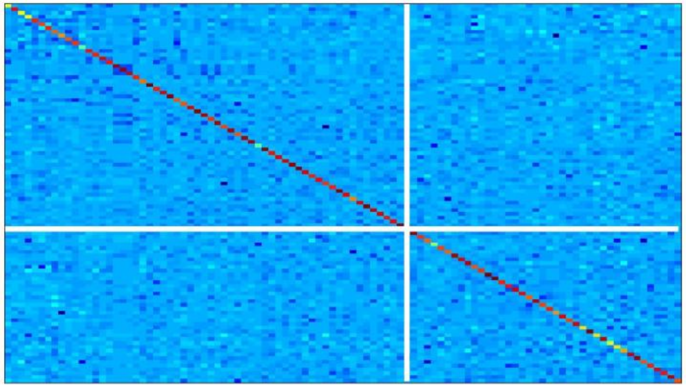
- Block GP constraints
  - Works only for block diagonal structure of covariance matrix
- Unknown sparsity structure
  - Prior assumptions can lead to erroneous results
  - Numerical approximations destroy model interpretability
  - Calculating complete covariance matrix will give much denser matrix
- Inverse covariance estimation gives relevant conditional independence information

Illustration on climate data



Precipitation data over land for the entire world





Covariance and inverse covariance matrices constructed from the above data for every pair of locations

Regularization

- Additional penalty to reduce model complexity or prevent overfitting
  - Penalty for L1: $||\beta||_1$
  - L1 regularization results in parsimonious models
- LASSO: least square regression using L1 regularization

$$||Y - X\beta||_2^2 + \lambda \, ||\beta||_1$$

  - where λ is regularization parameter

Sparse covariance selection

- Estimate sparse inverse covariance of a Gaussian distribution, given the sample mean and sample covariance matrix

Covariance selection for graphical models

⇕

Inverse covariance matrix estimation in Gaussian Process

Estimating inverse covariance

- Equivalent to inferring a graphical model
  - LASSO regression on every variable as possible target followed by AND/OR operation on pairwise relations
  - Minimize the pseudo negative log-likelihood of data; stable solution requires a L1 penalty

$$Tr(KS) - \log det(S) + \lambda \|S\|$$

  - can be solved using block-wise coordinate descent very efficiently

SPI-GP

1. Build kernel matrix
2. Use optimization to estimate sparse inverse kernel for GPR based prediction
   – Study important dependency patterns in the data
3. Compute predictions using the following equation:

$$\widehat{y}^* = K^*(\lambda^2 I + K)^{-1} y$$

ADMM for optimization

- Earth Science data - too huge to fit in memory
  - Standard optimization techniques do not work
- Alternating Direction Method of Multipliers (ADMM): decomposition algorithm for solving separable convex optimization problems
  - Based on iterative scatter and gather operations on the augmented Lagrangian

ADDM for Inverse Estimation

$$S^{t+1} = \min_{x} \left(\mathbf{Tr}(KS) - \log det(S) + \rho/2 \left\|S - Y^{t} + P^{t}\right\|_{F}\right)$$

Optimization variable

Analytical closed form requires doing eigen decomposition of matrix K

$$Y_{ij}^{t+1} = \Gamma_{\lambda/\rho}\left(S_{ij}^{t+1} + P_{ij}^{t}\right)$$

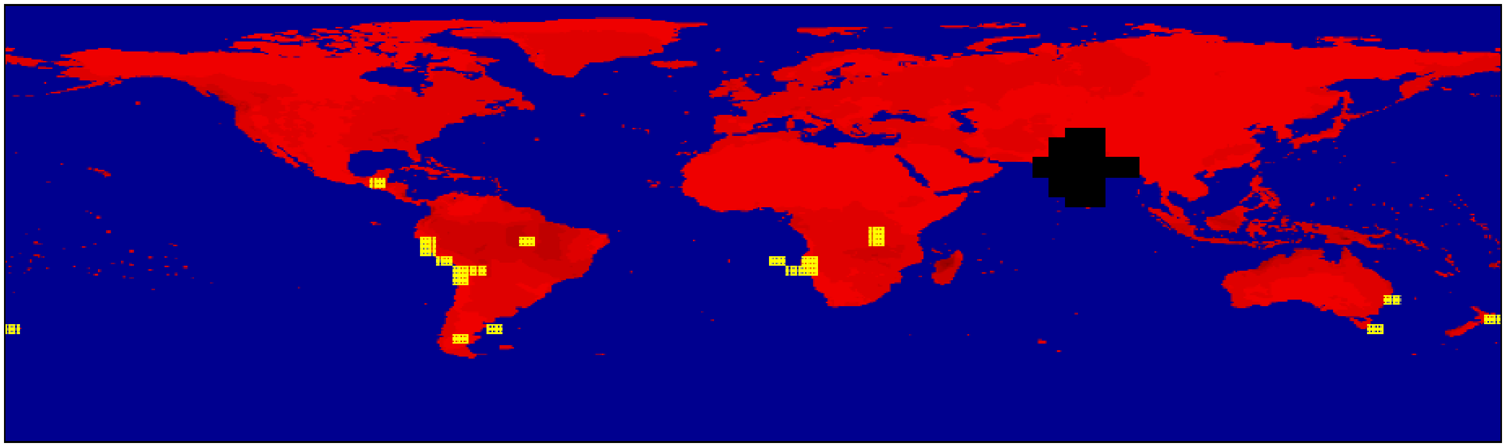
Linking /update variable

Analytical closed form is doing a soft thresholding at every step

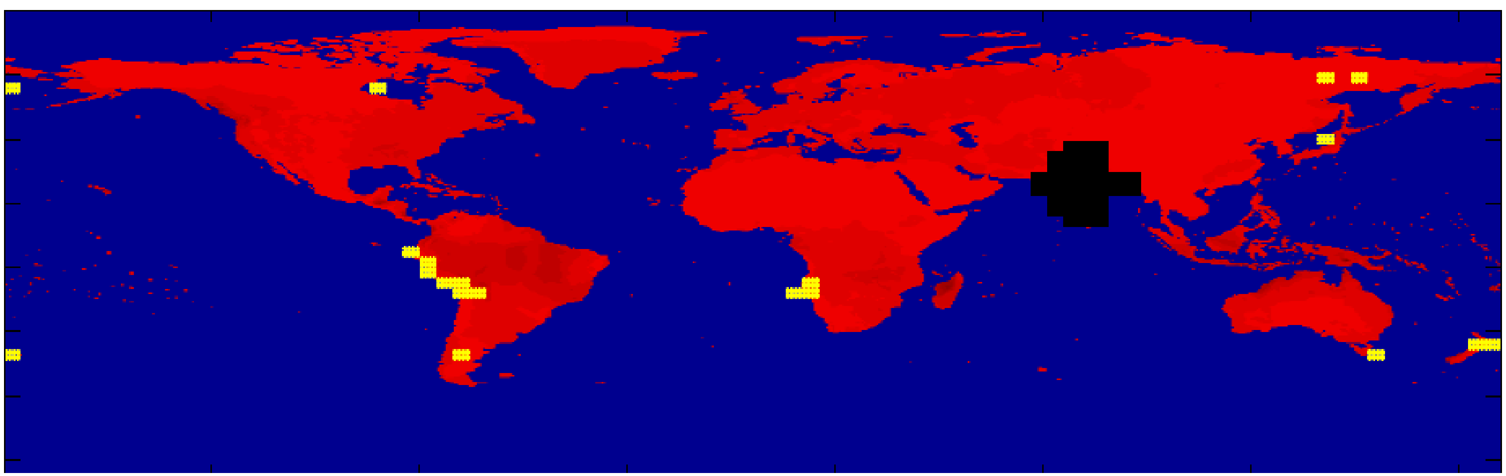
$$P^{t+1} = P^{t} + \left(S^{t+1} - Y^{t+1}\right)$$

Dual variable

SPI-GP experimental results



Climate network for years 1982 (above) and 1991 (below) based on precipitation in south Asia



Summary

- Scalable (parallelizable) Gaussian Process regression algorithm for multimodal data with scalability parameters:
  - Number of dimensions of input data
  - Number of observations
  - Number of modes in input data
- Block GP only handles approximately block diagonal covariance matrices
- SPI-GP allows identification of any sparsity pattern through inverse covariance estimation through parallelizable optimization technique
  - Able to compute (estimate) inverse kernel even when the data cannot be loaded into memory

On going research

- Method-oriented
  - Error bound on approximation for Block GP
  - Decomposable approximation for pseudo inverse
- Data oriented
  - Choice of kernel
  - Choice of number of clusters
  - Interpretation of network evolution study in terms of teleconnections

Acknowledgement

- Dr. Ramakrishna Nemani, NASA Ames
- Petr Votava, NASA Ames
- Dr. Santanu Das, NASA Ames

References

My papers:

1. K. Das, A. Srivastava. Block-GP: Scalable Gaussian Process Regression for Multimodal Data. 10th IEEE International Conference on Data Mining, Sydney, Australia. pp. 791-796. 2010.
2. K. Das, A. Srivastava. Gaussian Process Regression for climate data using network connection discovery. (in submission).

Gaussian Process Regression:

1. L. Foster, A. Waagen, N. Aijaz, M. Hurley, A. Luis, J. Rinsky, C. Satyavolu, M. Way, P. Gazis, and A. Srivastava, “Stable and Efficient Gaussian Process Calculations,” *JMLR, vol. 10,* pp. 857–882, 2009.
2. V. Tresp, “Mixtures of gaussian processes,” in *Proc. of NIPS 13, 2000, pp. 654–660.*
3. C. E. Rasmussen and Z. Ghahramani, “Infinite mixtures of gaussian process experts,” in *NIPS 14, 2001, pp. 881–888.*

Spectral Clustering:

1. W. Chen, Y. Song, H. Bai, C. Lin, and E. Y. Chang, “Parallel spectral clustering in distributed systems,” *IEEE PAMI, 2010.*
2. D. Yan, L. Huang, and M. I. Jordan, “Fast approximate spectral clustering,” in *Proc. of KDD’09, 2009, pp. 907–916.*

Inverse Covaiance Estimation:

1. O. Banerjee, L. Ghaoui, A. d’Aspremont, and G. Natsoulis. Convex optimization techniques for fitting sparse Gaussian graphical models. In *Proceedings of the 23rd International Conference on Machine Learning (ICML-06), pages 89–96, 2006.*
2. J. Friedman, T. Hastie, and R. Tibshirani. Sparse inverse covariance estimation with the graphical lasso. *Biostatistics Journal, 9(3):432–441, 2008.*
3. N. Meinshausen, P. Bhlmann, and E. Zrich. High Dimensional Graphs and Variable Selection with the Lasso.*Annals of Statistics, 34:1436–1462, 2006.*

Alternating Directions Method of Multipliers:

1. S. Boyd, N. Parikh, E. Chu, B. Peleato, and J. Eckstein. Distributed Optimization and Statistical Learning via the Alternating Direction Method of Multipliers. *Foundations and Trends in Machine Learning, 2011.*

Thank You