Graph-guided regularization for SWUS seasonal forecasting

Abby Stevens[1], Rebecca Willett[1,2], Antonios Mamalakis[3], Efi Foufoula-Georgiou[3,4], James Randerson[4], Padhraic Smyth[5], and Stephen Wright[6]

Introduction

Climate science has recently and rapidly become data-rich [1], due to a combination of

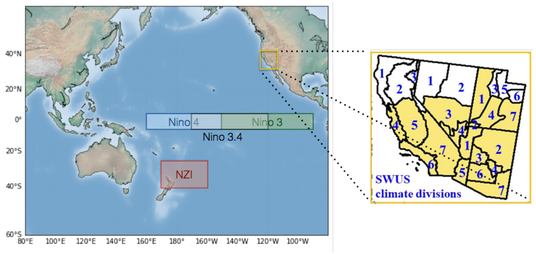
- New satellites and ground-based sensors collecting and storing massive amounts of high-resolution climate and ecosystem data
- Increased computational power to generate and archive physics-based climate model simulations

Advances in statistics and machine learning have only recently been applied to the Earth sciences, due to

- Lack of labeled training data
- Small sample sizes and high-dimensional settings
- Strong spatiotemporal dependencies among features

Seasonal Forecasting

Goal: Predict winter precipitation across the southwestern US using summer sea-surface temperatures over the Pacific



Linear model: For year i and climate division r

- $\hat{y}_r^{(i)} = X_1^{(i)}\hat{\beta}_1 + \cdots + X_p^{(i)}\hat{\beta}_p$
- $y_r^{(i)}$: winter precipitation
- $X_j^{(i)}$: sea-surface temperatures at (location, time) j
- $y \in \mathbb{R}^n$, $X \in \mathbb{R}^{n\times p}$, $p \gg n$

Graph Total Variation

GTV: Estimate sparse coefficients so that the higher the correlation of X_j and X_k (σ_{jk}) the more similar $\hat{\beta}_j$ and $\hat{\beta}_k$ [2].

$$\begin{aligned}\hat{\beta} &= \arg\min_{\beta} \|y - X\beta\|_2^2 && \textbf{Data fit}\\ &+ \lambda_{TV}\sum_{j,k\in E}\sigma_{j,k}|\beta_j - \beta_k| && \textbf{Graph total variation}\\ &+ \lambda_1\|\beta\|_1 && \textbf{Lasso}\end{aligned}$$

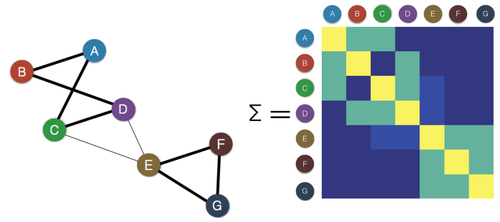
Affiliations

[1]Department of Statistics, University of Chicago, [2] Department of CS, University of Chicago, [3]Department of Civil and Environmental Engineering, University of California, Irvine, [4]Department of Earth System Science, University of California, Irvine, [5]Department of CS, University of California, Irvine, [6]CS Department, University of Wisconsin, Madison

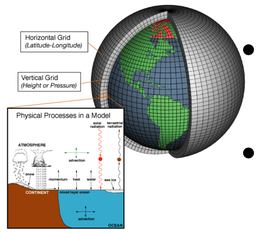
Covariance Graph

Assume $X^{(i)} \sim N(0, \Sigma)$, $X \in \mathbb{R}^{n\times p}$. Define graph $G = (V, E, W)$

- $V = \{X_1, \ldots X_p\}, \quad W_{jk} = \sigma_{jk}$
- $(j,k) \in E \iff |\Sigma_{j,k}| > 0$
- $p \gg n \implies$ sample covariance not consistent estimator for Σ



Large Ensemble Climate Simulations



www.climate.gov

CESM Large Ensemble Project (LENS)

- 40-member ensemble of physics-based climate simulations, each subject to perturbed initial conditions
- Combining observations and simulations shows potential for improving performance of ML methods in climate science
- LENS simulations are leveraged to estimate Σ, the covariance matrix of X. The simulated SSTs are interpolated onto the same spatial grid as SST observations and stacked to form a matrix $X_L \in \mathbb{R}^{40n\times p}$. Let $\hat{\Sigma}_L$ be the sample covariance of X_L. $\hat{\Sigma}_L$ is used to form the covariance graph for the GTV estimator.

MultiGTV

Responses from m different regions across the southwestern US. We assume that

- The subset of relevant features is preserved across regions
- Coefficients in each region are aligned with the covariance graph

MultiGTV: Extend GTV to multi-task setting

$$\begin{aligned}\hat{B} &= \arg\min_{B} \|Y - XB\|_F^2 && \textbf{Data fit}\\ &+ \lambda_1\sum_{r=1}^{m}\sum_{j,k\in E}\sigma_{j,k}|\beta_j^{(r)} - \beta_k^{(r)}| && \textbf{Within-region GTV}\\ &+ \lambda_2\sum_{i=1}^{p}\left\|\beta_i^{(:)}\right\|_2 && \textbf{Group sparsity}\end{aligned}$$

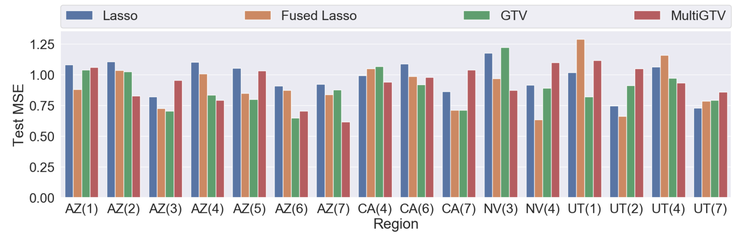
$$Y = [y_1, y_2, \ldots, y_m] \qquad B = \left[\beta^{(1)}, \beta^{(2)}, \ldots, \beta^{(m)}\right]$$

Experiments

Data: Precipitation and sea-surface temperatures from 1940-2014 ($y \in \mathbb{R}^{75}$, $X \in \mathbb{R}^{75\times 904}$)

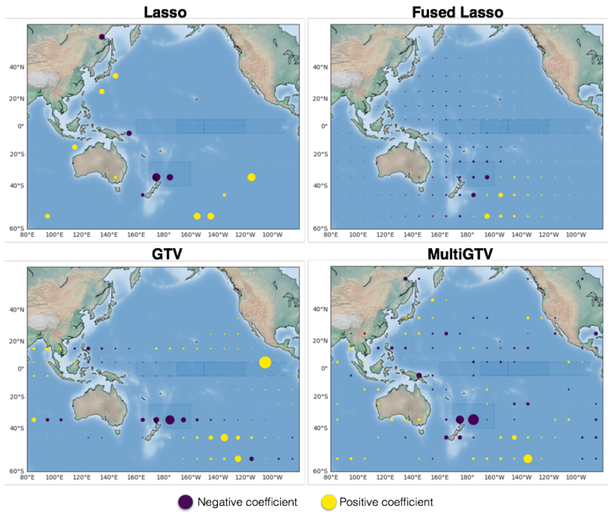
- Precipitation over 16 climate divisions across California, Nevada, Utah, and Arizona, aggregated over November-March
- Sea-surface temperatures aggregated over $10^\circ \times 10^\circ$ regions across the Pacific in July, August, September, and October
- Models trained on the first 50 years of data and errors reported on the held out last 25 years of data

Predictive performance



Model selection

- Markers indicate location and relative magnitude of selected coefficients in October for different methods
- The area-weighted average precipitation across all SWUS regions is used as the response for Lasso, Fused Lasso, and GTV, and for MultiGTV we aggregate estimates across regions



References

[1] A. R. Goncalves, A. Banerjee, V. Sivakumar, , and S. Chatterjee, "Structured estimation in high dimensions: applications in climate," in *Large-scale machine learning in the earth sciences*. CRC Press, 2017, ch. 2, pp. 13–32.

[2] Y. Li, B. Mark, G. Rasutti, and R. Willett, "Graph-based regularization for regression problems with highly-correlated designs," *arXiv preprint arXiv:1410.5093*, 2018.