Sparse Group Lasso: Consistency and Climate Applications

Soumyadeep Chatterjee* Karsten Steinhaeuser * Arindam Banerjee *
Snigdhansu Chatterjee† Auroop Ganguly‡

Abstract

The design of statistical predictive models for climate data gives rise to some unique challenges due to the high dimensionality and spatio-temporal nature of the datasets, which dictate that models should exhibit parsimony in variable selection. Recently, a class of methods which promote *structured sparsity* in the model have been developed, which is suitable for this task. In this paper, we prove theoretical statistical consistency of estimators with *tree-structured* norm regularizers. We consider one particular model, the *Sparse Group Lasso* (SGL), to construct predictors of land climate using ocean climate variables. Our experimental results demonstrate that the SGL model provides better predictive performance than the current state-of-the-art, remains climatologically interpretable, and is robust in its variable selection.

Keywords: Sparse Group Lasso, climate prediction, statistical consistency

1 Introduction

The success of data mining techniques in complementing and supplementing findings from several topics of scientific research is well documented [2, 11, 9]. However, climate science problems have some singular challenges, which makes the issue of scientifically meaningful prediction a complex process. Several climate variables are observed at various location on the planet on multiple occasions, thus creating a very large dataset. These variables are dependent between themselves, and across space. However, scientific interpretability and parsimony demands that any discovered relationship among climate variables be simultaneously eclectic and selective. It is not viable to work out such complex dependencies from the first principles of physics, and data mining discovery of potential climate variable relations can be of immense benefit to the climate science community.

The *sparse group lasso* (SGL hereafter) method is of considerable importance in this context. For a target climate variable in a given location, it allows the selection of other locations that may have an influence through one or more variables, and then allows for a choice of variables at that location. Inherent in this technique is the notion of sparsity, by which only important variables at important locations are selected, from the plethora of potential covariates at various spatial locations.

Recent work in statistical modeling has proved the utility of having parsimony in the inferred dependency structure. Efforts in this direction have been successful in developing *sparse* models, which promote sparsity within the dependencies characterized by the model. These models have been applied successfully in a number of fields, such as signal processing [4], bioinformatics [10], computer vision [26] etc. Incorporating sparsity within a statistical model provides a natural control over the complexity of the model achieved through training.

The classical statistical model trains from the training data at hand by defining a loss function to measure the discrepancy between its predictions and observations of the response variables. Optimization routines are used to obtain an optimal parameter set for the model so that the loss function is minimized. *Sparsity* is induced within the optimal parameter set by adding a *sparsity-inducing* regularizer function to the loss and optimizing this combination over the parameter set. The regularizer is usually a norm function of the parameter vector. This construction gives rise to a family of *sparse statistical models* with a convex loss function and a convex norm regularizer [13, 23]. Building on this literature, recent work has shown the utility of imposing *structure* among the dependencies through the use of group [27] and hierarchical norm regularizers [14, 12]. These structures can be learnt from some external sources, such as domain experts, and are useful in obtaining more robust and interpretable predictive models. Efficient optimization algorithms have been proposed to solve such estimation problems [17]. Recent results [19, 28] have proved statistical consistency guarantees for a class of *sparse estimators* under fairly mild conditions.

In this paper, using the analysis method developed in [19], we have proved statistical consistency guarantees for the class of tree-structured hierarchical norm regularized estimation problems [17]. We have applied sparse modeling to one particular climate prediction task - prediction of land climate variables from measurements of ocean

*Department of CSE, University of Minnesota, Twin Cities
†School of Statistics, University of Minnesota, Twin Cities
‡Department of Civil & Env. Engg., Northeastern University, Boston

climate variables. Assuming a linear regression model, we have used a recently proposed *group-structured* sparse method, called *Sparse Group Lasso* (SGL) for the prediction tasks. Our main contributions in this paper are as follows:

1. We provide statistical consistency bounds for a general class of hierarchical sparsity inducing norm regularized estimation problems.

2. We show that SGL provides better predictive accuracy and a more interpretable prediction model than the state-of-the-art in climate science.

3. We show that SGL is robust in covariate selection through an empirical analysis of its *regularization path*.

We formally describe the predictive problem from a climate perspective in Section 2. Consistency of hierarchical sparsity inducing norm regularized estimators is proved in Section 3. We discuss optimization methods for SGL in Section 4. The dataset and methodology is described in detail in Section 5. Sections 6 - 8 present our experimental results using SGL on climate data. Finally, we conclude with discussion in Section 9.

2 Problem Statement

We consider the task of predicting climate variables over "target" regions on land by using information from 6 climate variables over oceans. In particular, we choose temperature and precipitation as response variables on the chosen target regions. A similar task was performed by [22] using "climate clusters" in a linear regression model. The authors have also recently compared the performance of linear predictive models with a number of nonlinear predictive models [21] and the analysis shows that the they typically have similar performances.

The statistical model that we use is linear and can be defined as:

$$y \sim X\theta^* + w, \tag{2.1}$$

where $y \in \mathbb{R}^n$ is the n-dimensional vector of observations of a climate variable at a target region, $\theta^* \in \mathbb{R}^p$ is the coefficient associated with all p variables at all locations, $X \in \mathbb{R}^{n\times p}$ is the covariate matrix and $w \in \mathbb{R}^n$ is the noise vector. Our goal is two-fold:

1. understand which covariates are relevant/important for predicting the target variable, and

2. build a suitable regressor based on these relevant variables.

Assuming that the noise vector w follows a Gaussian distribution, estimating the vector θ^* amounts to solving the "ordinary least squares" (OLS) problem:

$$\hat{\theta}_{OLS} = \operatorname*{argmin}_{\theta\in\mathbb{R}^p} \left\{ \frac{1}{2n}\|y - X\theta\|_2^2 \right\}. \tag{2.2}$$

Clearly, when $n < p$, the system is unidentifiable and we will obtain multiple solutions $\hat{\theta}_{OLS}$. Moreover, in general, all coefficients of $\hat{\theta}_{OLS}$ will be non-zero, signifying statistical dependency of the "target" variable on all variables over all oceans. As is well known in statistical literature [13], the OLS estimate has large variance and hence, is not robust. Also, the estimate is not interpretable in terms of climate science due to the presence of many spurious dependencies.

In such cases, a regularizer $r(\theta)$ is added to the squared loss function in order to have a more robust estimate of θ^* [13]. In many applications, such as climate, the dependencies are, in general, *sparse*, meaning that most of the coefficients of $\hat{\theta}$ are 0 [25, 24]. To promote sparsity in the estimate, sparsity-inducing convex norm regularizers are commonly used [23, 1]. These sparse methods offer significant computational benefits over traditional feature selection methods and some have been proven to be statistically consistent [28, 1].

As mentioned earlier, the covariates in our problem are 6 climate variables measured at ocean locations over the globe. This spatial structure of the data indicates a natural "grouping" of the variables at each ocean location. Simple sparse regularizers, such as the LASSO penalty [23] do not respect this structure inherent in the data. Therein arises the need to have regularizers which impose *structured sparsity* that respects this spatial nature. The model that we use incorporates such a regularizer and is called *Sparse Group Lasso* (SGL) [12]. The next subsection describes the model.

SGL and Hierarchical Norms: Our motivation in promoting structured sparsity is drawn from the fact that for predicting a target variable, if a particular location on oceans is *irrelevant*, then coefficients of all 6 variables at that location should be zero. Furthermore, if a particular location is deemed 'relevant', then we should be able to select the "most important" variable(s) at that location to be considered for prediction.

To formalize the problem, let T be the total number of locations over oceans. Therefore, we have $p = 6T$ variables as covariates in the regression problem. Then, for a penalty parameter λ, the SGL estimator is given by

$$\hat{\theta}_{SGL} = \operatorname*{argmin}_{\theta\in\mathbb{R}^p} \left\{ \frac{1}{2N}\|y - X\theta\|_2^2 + \lambda r(\theta) \right\}, \tag{2.3}$$

where r is the SGL regularizer given by

$$r(\theta) := r_{(1,\mathcal{G}_2,\alpha)} = \alpha\|\theta\|_1 + (1-\alpha)\|\theta\|_{1,\mathcal{G}} \ , \tag{2.4}$$

where

$$\begin{aligned} \|\theta\|_1 &= \sum_{i=1}^{p} |\theta_i| \ , \\ \|\theta\|_{1,\mathcal{G}} &= \sum_{k=1}^{T} \|\theta_{G_k}\|_2 \end{aligned}$$

and $\mathcal{G} = \{G_1, \ldots, G_T\}$ are the groups of variables at the T locations considered. The mixed norm $\|\theta\|_{1,\mathcal{G}}$ penalizes groups of variables at irrelevant locations, while the L_1 norm $\|\theta\|_1$ promotes sparsity among variables chosen at selected locations.

The SGL regularizer belongs to a general class of convex norm regularizers $r(\cdot)$ which impose a tree-structured hierarchical structure in the sparsity induced [17, 14]. Such norms impose a hierarchy among groups formed from the index set $\{1, \ldots, p\}$ in the following way. Given a tree with p leaves, let the nodes of the tree denote groups of indices from the index set. The root of the tree denotes a single group containing all p indices, while each leaf denotes a single index. Now, the constraint imposed is that for any node of the tree, the elements (indices) contained in it should be a subset of the elements (indices) contained in its parent node.

In the next section, following the analysis technique developed in [19], we prove that, under fairly general conditions, hierarchical tree-structured norm regularized estimation is statistically consistent in estimating the true parameter θ^* of the distribution from which the data samples (X, y) were generated. We illustrate the SGL regularizer to be a special case of such norms and provide explicit bounds for the consistency of SGL.

3 Consistency of Sparse Group Lasso

3.1 Formulation: Let $Z_1^n := \{Z_1, \ldots, Z_n\}$ denote n observations drawn i.i.d. according to some distribution $\mathbb{P}$, and suppose that we are interested in estimating some parameter θ of the distribution $\mathbb{P}$. Let $\mathcal{L} : \mathbb{R}^p \times \mathcal{Z}^n \mapsto \mathbb{R}$ be some convex loss function that, for a given set of observations Z_1^n, assigns a cost $\mathcal{L}(\theta; Z_1^n)$ to any parameter $\theta \in \mathbb{R}^p$. We assume that that the population risk $R(\theta) = E_{Z_1^n}[\mathcal{L}(\theta; \mathcal{Z}_1^n)]$ is independent of n, and we let $\theta^* \in \operatorname{argmin}_{\theta \in \mathbb{R}^p} R(\theta)$ be a minimizer of the population risk. As is standard in statistics, in order to estimate the parameter vector θ^* from the data Z_1^n, we solve a convex program that combines the loss function with a regularizer. For the regularization function $r : \mathbb{R}^p \mapsto \mathbb{R}$, consider the regularized M-estimator given by

$$\hat{\theta}_{\lambda_n} \in \operatorname*{argmin}_{\theta \in \mathbb{R}^p} \ \{\mathcal{L}(\theta; Z_1^n) + \lambda_n r(\theta)\} \ , \tag{3.5}$$

where $\lambda_n > 0$ is a user-defined regularization penalty.

For the purpose of this paper, we consider linear models based on n observations $Z_i = (x_i, y_i) \in \mathbb{R}^p \times \mathbb{R}$ of covariate-response pairs as given in (2.1). We assume that the noise vector w is zero mean and has sub-Gaussian tails, i.e., there is a constant $\sigma > 0$ such that for any v, $\|v\|_2 = 1$, we have

$$\mathbb{P}(|\langle v, w\rangle| \geq \delta) \leq 2\exp\left(-\frac{\delta^2}{2\sigma^2}\right) \ , \quad \text{for all } \delta > 0 \ . \tag{3.6}$$

The condition holds in the special case of Gaussian noise; it also holds whenever the noise vector w consists of independent bounded random variables.

3.2 Assumptions on Regularizer and Loss Function: Following [19], the first key requirement for the analysis is a property of the regularizer r. The regularizer is defined to be *decomposable* w.r.t. a subspace pair $A \subseteq B \subseteq \mathbf{R}^p$ if, for any $\alpha \in A$ and $\beta \in B^\perp$, where $B^\perp$ is the orthogonal space of B,

$$r(\alpha + \beta) = r(\alpha) + r(\beta) \ . \tag{3.7}$$

Let us define the error vector $\hat{\Delta}_{\lambda_n} := \hat{\theta}_{\lambda_n} - \theta^*$, and the projection operator $\Pi_A : \mathbf{R}^p \mapsto A$, such that

$$\Pi_A(u) = \operatorname*{argmin}_{v \in A} \|u - v\|_* \ ,$$

for some given error norm $\|\cdot\|_*$. Then, if the loss $\mathcal{L}$ is convex, we can define the set

$$\begin{aligned} &\mathbb{C}(A, B; \theta^*) := \\ &\quad \{\Delta \in \mathbf{R}^p | r(\Pi_{B^\perp}(\Delta)) \leq 3r(\Pi_B(\Delta)) + 4r(\Pi_{A^\perp}(\theta^*))\} \ , \end{aligned} \tag{3.8}$$

which contains the error $\hat{\Delta}$ for any λ_n satisfying

$$\lambda_n \geq 2r^*(\nabla\mathcal{L}(\theta^*; Z_1^n)) \ .$$

A formal proof of the statement is provided in [19].

The second key requirement, as stated in [19], is that the loss function $\mathcal{L}$ should satisfy the Restricted Strong Convexity (RSC) property. Let us define $\delta\mathcal{L}(\Delta, \theta^*) := \mathcal{L}(\theta^* + \Delta; Z_1^n) - \mathcal{L}(\theta^*; Z_1^n) - \langle\nabla\mathcal{L}(\theta^*; Z_1^n), \Delta\rangle$. $\mathcal{L}$ satisfies RSC with curvature $\kappa_{\mathcal{L}} > 0$ and tolerance function $\tau_{\mathcal{L}}$ if, for all $\Delta \in \mathbb{C}(A, B; \theta^*)$,

$$\delta\mathcal{L}(\Delta, \theta^*) \geq \kappa_{\mathcal{L}}\|\Delta\|_*^2 - \tau_{\mathcal{L}}^2(\theta^*) \ . \tag{3.9}$$

Further, [19] defines a subspace compatibility constant with respect to the pair $(r, \|\cdot\|_*)$ for any subspace $B \subseteq \mathbb{R}^p$ as follows:

$$\Psi(B) := \sup_{u \in B \backslash \{0\}} \frac{r(u)}{\|u\|_*} \ . \tag{3.10}$$

Based on the assumption that $\mathcal{L}$ is convex and differentiable, and the norm regularizer r is decomposable w.r.t. a subspace pair $A \subseteq B$, [19] presents the following key result:

Theorem 1 *Consider the convex program in (3.5) based on a strictly positive regularization constant*

$$\lambda_n \geq 2r^*(\nabla\mathcal{L}(\theta^*; Z_1^n)) \ . \tag{3.11}$$

Then any optimal solution $\hat{\theta}_{\lambda_n}$ to (3.5) satisfies the bound
(3.12)

$$\|\hat{\theta}_{\lambda_n} - \theta^*\|_*^2 \leq 9\frac{\lambda_n^2}{\kappa_{\mathcal{L}}^2}\Psi^2(B) + \frac{\lambda_n}{\kappa_{\mathcal{L}}}\left\{2\tau_{\mathcal{L}}^2(\theta^*) + 4r(\Pi_{A^\perp}(\theta^*))\right\} .$$

3.3 Analysis for Hierarchical Tree-Structured Norm:

We now provide statistical consistency analysis of the hierarchical tree-structured norm regularizer described in Section (2). Let the height of the tree be $h+1$, with the leaves having a height 0 and the root having a height $h+1$. Let the maximum size of a group at height i be m_i. Let the nodes (groups) at height i be denoted by $\{G_j^i\}$, $j = 1, \ldots, n_i$. Note that $n_0 = p$ and $m_0 = 1$. The group norm at height i is computed as:

$$\|\theta_{\mathcal{G}^i}\|_{(1,\nu)} := \sum_{j=1}^{n_i} \|\theta_{G_j^i}\|_\nu \tag{3.13}$$

For any $(\alpha_0, \alpha_1, \ldots, \alpha_h)$ such that $1 > \alpha_i > 0$, $\forall i$ and $\alpha_0 + \alpha_1 + \ldots + \alpha_h = 1$, the tree-norm regularizer is formally defined as

$$r(\theta) := r_{tree}(\theta) := \sum_{i=0}^{h} \alpha_i \|\theta_{\mathcal{G}^i}\|_{(1,\nu)} \ . \tag{3.14}$$

Our analysis consists of three key parts: (i) Showing that the regularizer r_{tree} is decomposable, (ii) Showing that the loss function satisfies the RSC condition, and (iii) Choosing a λ_n which satisfies the prescribed lower bound.

Following [19], we assume that for each $k = 1, \ldots, p$

$$\frac{\|X_k\|_2}{\sqrt{n}} \leq 1 \ . \tag{3.15}$$

Note that the assumption can be satisfied by simply rescaling the data, and is hence without loss of generality. Further, the above assumption implies that

$$\frac{\|X_{G_j^i}\|_{\nu\to 2}}{\sqrt{n}} \leq 1 \ , \tag{3.16}$$

where the operator norm

$$\|X_{G_t}\|_{\nu\to 2} := \max_{\|\theta\|_\nu = 1} \|X_{G_t}\theta\|_2 \ .$$

3.3.1 Decomposability of Regularizer:

We may note that the group norm at a particular height in the tree, $\|\theta_{\mathcal{G}^i}\|_{(1,\nu)}$ is over groups which are disjoint. Hence it decomposes over the subspace spanned by each group. Therefore, following the definitions and arguments in [19], the tree-norm is decomposable.

3.3.2 Restricted Strong Convexity:

As shown in [19], RSC for the loss function $\mathcal{L}$ is equivalent to a *restricted eigenvalue* condition on the covariate matrix X. If X is formed by sampling each row $X_i \sim N(0, \Sigma)$, referred to as the Σ-*Gaussian ensemble*, then with high probability $\mathcal{L}$ satisfies RSC. It has been shown [29] that the guarantee extends to sub-Gaussian designs as well.

3.3.3 Bounds for λ_n:

Recall from Theorem 1 that the λ_n needs to satisfy the following lower bound:

$$\lambda_n \geq 2r_{tree}^*(\nabla\mathcal{L}(\theta^*; Z_1^n)) \ . \tag{3.17}$$

A key issue with the above lower bound is that it is a random variable depending on Z_1^n. A second issue is that the conjugate r_{tree}^* for the mixed norm $r_{tree}(v)$ may not be obtainable in closed (non-variational) form. So we first obtain an upper bound $\bar{r}_{tree}^*$ on r_{tree}^*, and choose a λ_n which will satisfy the lower bound in (3.17) with high probability over choices of Z_1^n.

By definition

$$\begin{aligned}
& r_{tree}^*(v) \\
& \quad = \sup_{u\in\mathbb{R}^p\setminus\{0\}} \frac{\langle u, v\rangle}{r_{tree}(u)} \\
& \quad = \sup_{u\in\mathbb{R}^p\setminus\{0\}} \frac{\langle u, v\rangle}{\sum_{i=0}^{h} \alpha_i \|u_{\mathcal{G}^i}\|_{(1,\nu)}} \\
& \quad \overset{(a)}{\leq} \sup_{u\in\mathbb{R}^p\setminus\{0\}} \left[\sum_{i=0}^{h} \alpha_i \frac{\langle u, v\rangle}{\|u_{\mathcal{G}^i}\|_{(1,\nu)}}\right] \\
& \quad \leq \sum_{i=0}^{h} \alpha_i \sup_{u\in\mathbb{R}^p\setminus\{0\}} \frac{\langle u, v\rangle}{\|u_{\mathcal{G}^i}\|_{(1,\nu)}} \\
& \quad = \sum_{i=0}^{h} \alpha_i r_{\mathcal{G}_\nu^i}^*(v) = \bar{r}_{tree}^*(v) \ ,
\end{aligned} \tag{3.18}$$

where (a) follows from Jensen's inequality and $r_{\mathcal{G}_\nu^i}^*$ is the conjugate norm of $r_{\mathcal{G}_\nu^i}(v) = \sum_{j=1}^{n_i} \|v_{G_j^i}\|_\nu$ given by

$$r_{\mathcal{G}_\nu^i}^*(v) = \max_{j=1,\ldots,n_i} \|v_{G_j^i}\|_{\nu^*} \ , \tag{3.19}$$

where $\nu^* > 0$ satisfies $\frac{1}{\nu} + \frac{1}{\nu^*} = 1$.

By definition, we have $\nabla L(\theta^*; Z_1^n) = \frac{X^T w}{n}$ where $w = y - X\theta^*$ is a zero mean sub-Gaussian random variable. As a result, it is sufficient to choose λ_n satisfying:

$$\lambda_n \geq 2\left[\sum_{i=0}^{h} \alpha_i \left(\max_{j=1,\ldots,n_i} \left\| \frac{X_{G_j^i}^T w}{n} \right\|_{\nu^*} \right)\right] . \tag{3.20}$$

For any $j \in \{1, \ldots, n_i\}$, consider the random variable:

$$Y_j = Y_j(w) = \sum_{i=0}^{h} \alpha_i \left\| \frac{X_{G_j^i}^T w}{n} \right\|_{\nu^*} .$$

Following exactly similar arguments as in [19], we can show that $Y_j(w)$ is a Lipschitz function of w with constant $\frac{1}{\sqrt{n}}$. It follows that

$$\mathbb{P}\left[Y_j(w) \geq E[Y_j(w)] + \delta\right] \leq 2\exp(-\frac{n\delta^2}{2\sigma^2}) \tag{3.21}$$

Suitably applying the Sudakov-Fernique comparison principle [15, 7] shows that:

$$E\left[\left\| \frac{X_{G_j^i}^T w}{n} \right\|_{\nu^*}\right] \leq 2\sigma \frac{m_i^{1-\mathbb{1}/\nu}}{\sqrt{n}} , \tag{3.22}$$

so that we have

$$\begin{aligned} E[Y_j(w)] &= \sum_{i=0}^{h} \alpha_i E\left[\left\| \frac{X_{G_j^i}^T w}{n} \right\|_{\nu^*}\right] \\ &\leq 2\sigma \frac{\sum_{i=0}^{h} \alpha_i m_i^{1-1/\nu}}{\sqrt{n}} . \end{aligned} \tag{3.23}$$

Substituting everything in (3.21), we obtain

$$\begin{aligned} \mathbb{P}\left\{\sum_{i=0}^{h} \alpha_i \left\| \frac{X_{G_j^i}^T w}{n} \right\|_{\nu^*} \geq 2\sigma \frac{\sum_{i=0}^{h} \alpha_i m_i^{1-1/\nu}}{\sqrt{n}} + \delta\right\} \\ \leq 2\exp\left(-\frac{n\delta^2}{2\sigma^2}\right) . \end{aligned} \tag{3.24}$$

Applying the union bound over $j = 1, \ldots, n_i$ and $i = 0, \ldots, h$ we obtain

$$\begin{aligned} \mathbb{P}\Bigg\{\sum_{i=0}^{h} \alpha_i \left(\max_{j=1,\ldots,n_i} \left\| \frac{X_{G_j^i}^T w}{n} \right\|_{\nu^*} \right) \\ \geq 2\sigma \frac{\sum_{i=0}^{h} \alpha_i m_i^{1-1/\nu}}{\sqrt{n}} + \delta\Bigg\} \\ \leq 2\exp\left(-\frac{n\delta^2}{2\sigma^2} + \log(\prod_{i=0}^{h} n_i)\right) . \end{aligned} \tag{3.25}$$

For any $k > 0$, choosing

$$\delta = \sigma\sqrt{\frac{2(k+1)\log(\prod_{i=0}^{h} n_i)}{n}} , \tag{3.26}$$

we get the following result:

Lemma 1 *If*

$$\lambda_n \geq 2\sigma\left\{\frac{2\sum_{i=0}^{h} \alpha_i m_i^{1-1/\nu}}{\sqrt{n}} + \frac{\sqrt{2(k+1)\log(\prod_{i=0}^{h} n_i)}}{\sqrt{n}}\right\} ,$$

then

$$\mathbb{P}\left[\lambda_n \geq 2r^*(\nabla\mathcal{L}(\theta^*; Z_1^n))\right] \geq 1 - \frac{2}{(\prod_{i=0}^{h} n_i)^k} .$$

A Simplified Bound: We may make the observation that because of the tree-structure, the max-norm of groups at a higher level in the tree dominate the max-norm of groups at lower levels. Specifically, for any v and $i \geq l$,

$$\max_{j=1,\ldots,n_i} \|v_{G_j^i}\|_{\nu^*} \geq \max_{j=1,\ldots,n_l} \|v_{G_j^l}\|_{\nu^*} . \tag{3.27}$$

Hence, the right hand side of (3.20) is upper bounded by the max-norm of groups at height h. Since $\sum_{i=0}^{h} \alpha_i = 1$, we need to choose λ_n satisfying:

$$\lambda_n \geq 2 \max_{j=1,\ldots,n_h} \left\| \frac{X_{G_j^h}^T w}{n} \right\|_{\nu^*} . \tag{3.28}$$

Following the previous construction, we can show:

$$\begin{aligned} \mathbb{P}\left\{\max_{j=1,\ldots,n_h} \left\| \frac{X_{G_j^h}^T w}{n} \right\|_{\nu^*} \geq 2\sigma \frac{m_h^{1-1/\nu}}{\sqrt{n}} + \delta\right\} \\ \leq 2\exp\left(-\frac{n\delta^2}{2\sigma^2} + \log(n_h)\right) . \end{aligned} \tag{3.29}$$

Then, for any $k > 0$, choosing

$$\delta = \sigma\sqrt{\frac{2(k+1)\log(n_h)}{n}} , \tag{3.30}$$

we get the following simplified bound:

Lemma 2 *If*

$$\lambda_n \geq 2\sigma\left\{2\frac{m_h^{1-1/\nu}}{\sqrt{n}} + \sqrt{\frac{2(k+1)\log n_h}{n}}\right\} ,$$

then

$$\mathbb{P}\left[\lambda_n \geq 2r^*(\nabla\mathcal{L}(\theta^*; Z_1^n))\right] \geq 1 - \frac{2}{(n_h)^k} .$$

3.4 Analysis of Sparse Group Lasso: The SGL regularizer can be easily seen to be a special case of the tree-structured hierarchical norm regularizer, when the height of the tree is 2. The first level of the tree contains nodes corresponding to the T disjoint groups $\mathcal{G} = \{G_1, \ldots, G_T\}$, while the second level contains the singletons. It combines a group-structured norm with a element-wise norm (2.4). For ease of exposition, we assume the groups G_t are of the same size, say of m indices, so we have T groups of size m, and $p = Tm$.

A direct analysis for SGL using the proof method just described provides the following lemma:

Lemma 3 *If*

$$\lambda_n \geq 2\sigma \left\{ \frac{2(1 + m^{1-\mathbb{1}/\nu})}{\sqrt{n}} + \frac{\sqrt{2(k+1)(2\log T + \log m)}}{\sqrt{n}} \right\}, \tag{3.31}$$

then

$$\mathbb{P}\left[\lambda_n \geq 2r^*(\nabla\mathcal{L}(\theta^*; Z_1^n))\right] \geq 1 - \frac{2}{(pT)^k} . \tag{3.32}$$

3.4.1 Main Result: A direct application of Theorem 1 now gives the following result:

Theorem 2 *Let A be any subspace of $\mathbb{R}^p$ of dimension s_A. Let θ^* be the optimal (unknown) regression parameter, and let $r^{A^\perp}_{(1,\mathcal{G}_2,\alpha)}(\theta^*)$ be the sparse-group lasso norm of θ^* restricted to $A^\perp$, the orthogonal subspace of A. Then, if λ_n satisfies the lower bound in Lemma 3, with probability at least $(1 - \frac{2}{(pT)^k})$, we have*

$$\|\hat{\theta}_{\lambda_n} - \theta^*\|_2^2 \leq \frac{9\lambda_n^2}{k_{\mathcal{L}}^2} s_A + \frac{\lambda_n}{k_{\mathcal{L}}} \{2\tau_{\mathcal{L}}^2(\theta^*) + 4r^{A^\perp}_{(1,\mathcal{G}_2,\alpha)}(\theta^*)\} , \tag{3.33}$$

where $\hat{\theta}_{\lambda_n}$ is the SGL estimator in (2.3).

Corollary 1 *If the optimal parameter θ^* is in the subspace A, then*

$$\|\hat{\theta}_{\lambda_n} - \theta^*\|_2^2 \leq \frac{9\lambda_n^2}{k_{\mathcal{L}}^2} s_A + \frac{2\lambda_n}{k_{\mathcal{L}}} \tau_{\mathcal{L}}^2(\theta^*) = O\left(\frac{\log p}{n}\right) . \tag{3.34}$$

4 Optimization Method

Our analysis in the previous section illustrates that SGL encodes a tree-structured hierarchy in grouping covariates which leads to sparsity at two levels: groups and singletons. The different sparsity structures induced by hierarchical norms have been explored in [14] and [17]. The authors have independently proposed methods for optimization.

We follow the method proposed in [17] which is a sub-gradient based approach to solve the primal SGL problem. It may be noted that (2.3) is the sum of two convex functions [20], where the squared loss $\mathcal{L}$ is smooth and the regularizer r is non-smooth. The proposed algorithm iteratively computes the gradient update w.r.t. the loss $\mathcal{L}$ and minimization w.r.t. the regularizer r.

The minimization associated with the non-smooth function r is done as a sequential update of the coefficient vector during one pass of the nodes of the hierarchical tree. At each node of the tree, a shrinkage operation is executed on the elements of the coefficient vector indexed by the node. The algorithm is initialized at the leaves of the tree and terminates at the root. The authors prove that at termination, the vector returned by the algorithm is the unique solution to the regularization update step. The gradient update is computed using accelerated gradient descent [3]. Since SGL constitutes of a depth-2 hierarchical tree, the proposed algorithm is expected to be fast. Theoretically, it achieves a global convergence rate of $O(\frac{1}{k})$ after k iterations.

The authors of [17] have done an efficient implementation of their algorithm in a MATLAB interfaced module called SLEP [16]. We utilized SLEP to conduct all experiments in this paper.

5 Experimental Dataset and Methodology

5.1 Dataset: We used the NCEP/NCAR Reanalysis 1 dataset, where we considered the monthly means for 1948-present [18]. The data is arranged as points(locations) on the globe and is available at a $2.5° \times 2.5°$ resolution level. Our main goal is to highlight the utility of using sparse methods to model complex dependencies in climate. Since we trying to model dependencies between target variables and ocean regions, we coarsened the data to $10° \times 10°$ resolution. In total we have data over $N = 756$ time steps. The 6 variables over oceans, considered as covariates, are (i)Temperature, (ii) Sea Level Pressure, (iii) Precipitation, (iv) Relative Humidity, (v) Horizontal Wind Speed and (vi)Vertical Wind Speed.

We considered 9 "target regions" on land, viz., Brazil, Peru, Western USA, Eastern USA, Western Europe, Sahel, South Africa, Central India and Southeast Asia, as shown in Fig. 1. Prediction was done for air temperature and precipitation at each of these 9 locations. So, in total, we had 18 response variables. These regions were

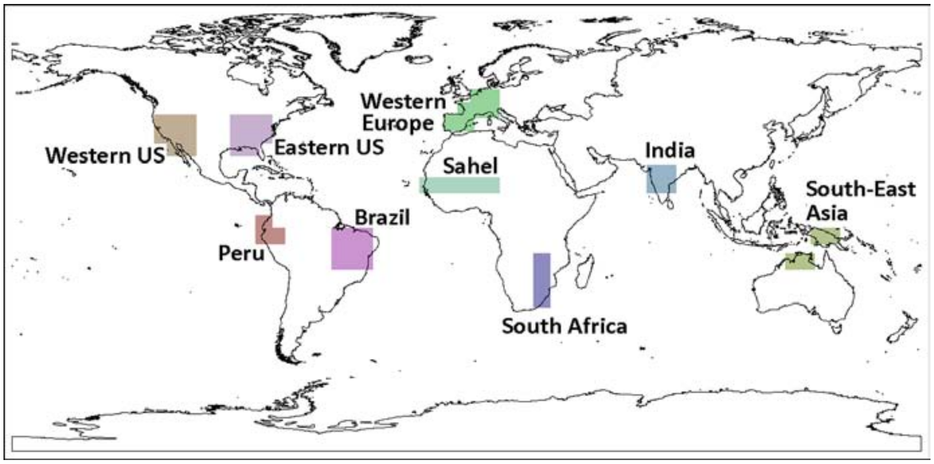


Figure 1: Land regions chosen for predictions (picture from [22]).

chosen following [22] because of their diverse geological properties and their impact on human interests.

In total, the dataset contains $L = 439$ locations on the oceans, so that we had $p = 6 \times L = 2634$ covariates in our regression model. We considered the data from January,1948 - December,1997 as the training data and from January,1998 - December,2007 as the test data in our experiments. So, our training set had $n_{train} = 600$ samples and the test set had $n_{test} = 120$ samples.

As mentioned earlier, we used the SLEP package [16] for MATLAB to run Sparse Group Lasso on our dataset. It may be noted that we do not take into account temporal relationships that exist in climate data. Moreover, since we consider monthly means, temporal lags of less than a month are typically not present in the data. However, the data does allow us to capture more long-term dependencies present in climate.

5.2 Removing Seasonality and Trend: As illustrated in [22], seasonality and autocorrelation within climate data at different time points often dominate the signal present in it. Hence, when trying to utilize such data to capture dependency, we look at series of *anomaly* values, i.e., the deviation at a location from the 'normal' value. Firstly, we remove the seasonal component present in the data by subtracting the monthly mean from each datapoint and then normalize by dividing it by the monthly standard deviation. At each location we calculate the monthly mean μ_m and standard deviation σ_m for each month $m = 1, \ldots, 12$ (i.e. separately for January, February,... etc.) for the entire time series. Finally, we obtain the anomaly series for location A as the z-score of the variable at location A for month m over the time series.

Further, we need to detrend the data to remove any trend components in the time-series, which might also dominate the signal present in it and bias our regression estimate. Therefore, we fit a linear trend to the anomaly series at each location over the entire time period 1948-2010 and take the residuals by subtracting the trend. We use this deseasonalized and detrended residuals as the dataset for all our subsequent experiments.

5.3 Choice of penalty parameter (λ)**:** The choice of the penalty parameters plays a crucial role in the performance of the sparse regression method. Following our analysis in section (3), we need to have $\lambda \geq 1.32$. To empirically compute the optimal choice of $(\lambda_1 = \alpha\lambda, \lambda_2 = (1-\alpha)\lambda)$, where λ_1 is the penalty for $\|\theta\|_1$ and λ_2 is that for $\|\theta\|_{1,\mathcal{G}}$, we ran hold-out cross-validation experiments on the training set for each response variable for choices of $10^{-4} \leq \lambda_1, \lambda_2 \leq 10^3$, in increments of 10, since the results were insensitive to similar penalty values. Table 1 shows the optimal choices obtained from cross-validation. The values for different target variables are similar, with the exception of three, which correspond to precipitation in Africa and East USA.

6 Prediction Accuracy

Evaluation of our predictions was done by computing the root mean square errors (RMSE) on the test data and comparing the results against those obtained in [22] and using OLS estimates. [22] uses a correlation based approach to separately cluster each ocean variable into *regions* using a k-means clustering algorithm. The *regions* (78 clusters in total) for all ocean variables are used as covariates for doing linear regression on response variables. Their model is referred to as the *Network Clusters* model. RMSE values were computed, as mentioned earlier, by predicting monthly mean anomaly for each response variable over

Table 1: Optimal Choices of (λ_1, λ_2) obtained through 20-fold Cross-Validation.

Region	Variable	λ_1	λ_2
Brazil	Temperature	1	1
	Precipitation	1	1
Peru	Temperature	1	1
	Precipitation	1	1
Western USA	Temperature	1	1
	Precipitation	1	1
Eastern USA	Temperature	1	1
	Precipitation	10	10
Western Europe	Temperature	1	1
	Precipitation	1	1
Sahel	Temperature	1	1
	Precipitation	10	10
South Africa	Temperature	1	1
	Precipitation	10	10
Central India	Temperature	1	1
	Precipitation	1	1
SE Asia	Temperature	1	1
	Precipitation	1	1

the test set for 10 years. The RMSE scores are summarized in Table 2. We observed that SGL consistently performs better than both the *Network Clusters* method and the OLS method. The anomalies for temperature and precipitation in the test set lie in the range of $[-2.5, +2.5]$. Therefore, SGL obtained a gain of $8\% - 14\%$ in accuracy over *Network Clusters* and a $7\% - 15\%$ gain in accuracy over OLS.

The higher prediction accuracy might be explained through the model parsimony that SGL provides. Applying SGL, only the most relevant predictor variables are given non-zero coefficients and any irrelevant variable is considered as noise and suppressed. Since such parsimony will be absent in OLS, the noise contribution is large and therefore the predictions are more erroneous. We elaborate on this aspect in the next section.

7 Variable Selection by SGL

The high prediction accuracy of SGL brings to light the inherent power of the model to select appropriate variables (or features) from the covariates during its training phase. To quantitatively elaborate on this aspect, we select two scenarios: (i) Temperature prediction in Brazil and (ii) Temperature prediction in India.

In order to evaluate the covariates which consistently get selected from the set, we repeat the hold out cross-validation experiment with the optimal choices of (λ_1, λ_2) determined earlier for each scenario. During the training phase, an ocean variable was considered *selected*, if it had a corresponding non-zero coefficient. So, in each run of cross-validation, some of the covariates were selected, while others were not. We illustrate our findings in the following subsections.

Table 2: RMSE scores for prediction of air-temperature and precipitation using SGL, network clusters [22] and OLS.

Variable	Region	**SGL**	Network Clusters	OLS
Air Temperature	Brazil	**0.198**	0.534	0.348
	Peru	**0.247**	0.468	0.387
	West USA	**0.270**	0.767	0.402
	East USA	**0.304**	0.815	0.348
	W Europe	**0.379**	0.936	0.493
	Sahel	**0.320**	0.685	0.413
	S Africa	**0.136**	0.726	0.267
	India	**0.205**	0.649	0.3
	SE Asia	**0.298**	0.541	0.383
Precipitation	Brazil	**0.261**	0.509	0.413
	Peru	**0.312**	0.864	0.523
	West USA	**0.451**	0.605	0.549
	East USA	**0.365**	0.686	0.413
	W Europe	**0.358**	0.450	0.551
	Sahel	**0.427**	0.533	0.523
	S Africa	**0.235**	0.697	0.378
	India	**0.146**	0.672	0.264
	SE Asia	**0.159**	0.665	0.312

7.1 Region: Brazil In Fig.4, we plot, in descending order of magnitude, the number of times each covariate was selected during cross-validation for temperature prediction in Brazil.

We observe that there are ~ 60 covariates among the 2634 covariates that are selected in every single run of cross-validation. In Fig. 2, we plot the covariates which are given high coefficient magnitudes by SGL by training on the training dataset from years 1948-1997, in order to illustrate that SGL consistently selects *relevant* covariates. It turns out that these covariates are exactly those which were selected in every cross-validation run. Most of these covariates lie off the coast of Brazil. The influences of horizontal wind speed and pressure is captured, which is consistent with the fact that the ocean currents affect land climate typically through horizontal wind. The tropical climate over Brazil is expected to be influenced by the Inter-tropical Convergence from the north, Polar Fronts from the south, and disturbances in ocean currents from the west, as well as the influence of Easterlies from the east and immediate south. It is interesting to see that SGL model captures these influences, as well as the spatial autocorrelation present in climate data, without having any explicit assumptions.

7.2 Region: India Similarly as before, for temperature prediction in India, we construct a histogram of the

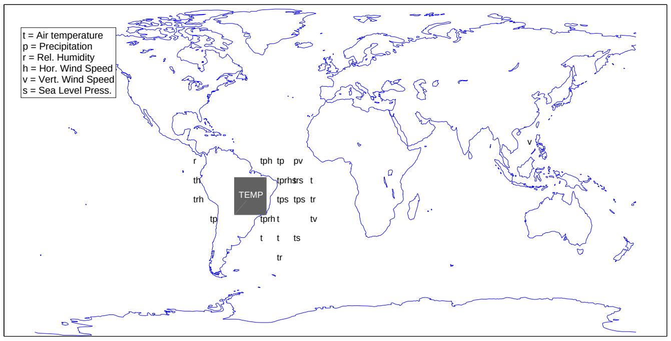


Figure 2: Temperature prediction in Brazil: Variables chosen through cross-validation.

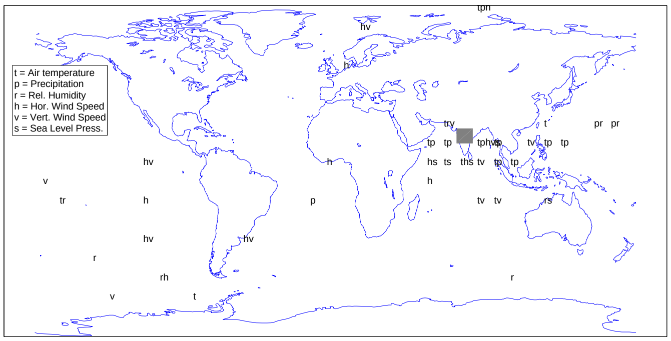


Figure 3: Temperature prediction in India: Variables chosen through cross-validation.

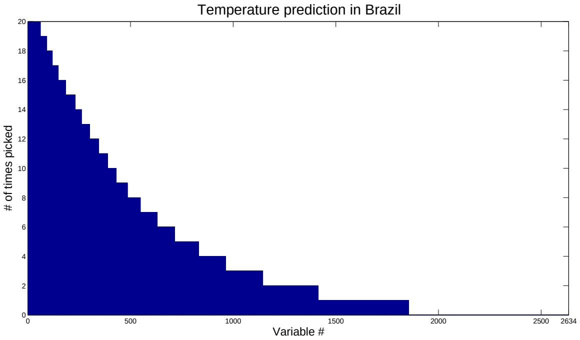


Figure 4: Temperature prediction in Brazil: Variables vs. No. of times selected.

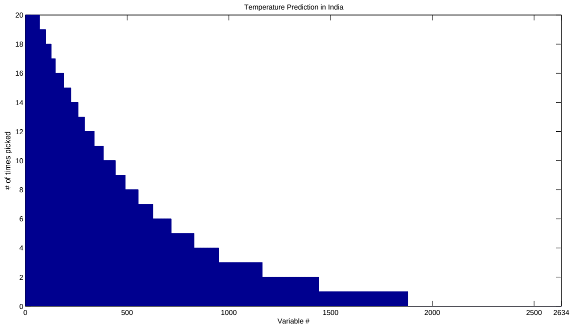
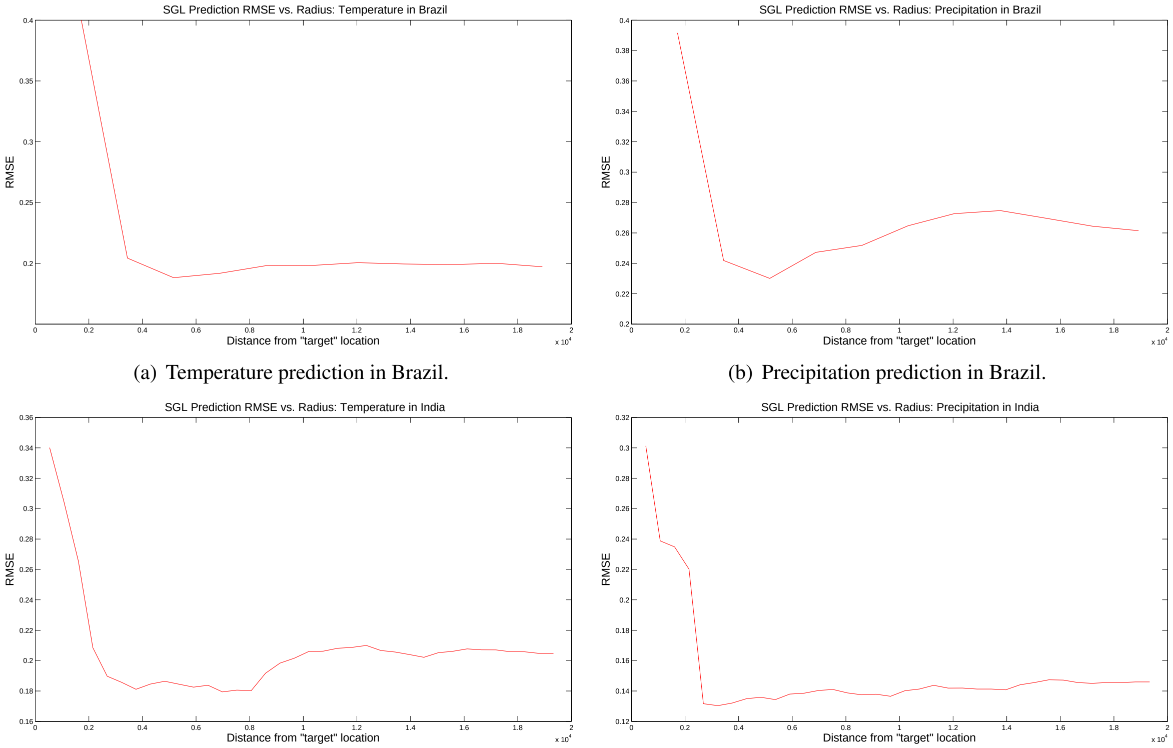


Figure 5: Temperature prediction in India: Variables vs. No. of times selected.

number of times covariates get selected during cross-validation (Fig. 5).

Among the 2634 covariates considered, in this case ~ 65 covariates were chosen in every single run of cross-validation. These are plotted in Fig. 3. Again, these were the covariates with largest coefficient magnitudes during training on the entire training-set. We observe the impact of Arabian Sea and Bay of Bengal on the Indian climate. Interestingly, there are some teleconnections which are captured by SGL over the Pacific Ocean, which may be due to the connections between Indian Monsoon and El-Nino [6] and SE Asian and Australian monsoons. This may be an interesting observation for further investigation by domain scientists.

It should be noted that the dataset is a set of discrete samples from variables which vary continuously over space and time. This gives rise to 'sampling noise', which is manifested in some variables being selected by SGL, which might not have physical interpretations. Handling



(a) Temperature prediction in Brazil.

(b) Precipitation prediction in Brazil.

(c) Temperature prediction in India.

(d) Precipitation prediction in India.

Figure 6: SGL RMSE vs. Radius R.

such data appropriately is a topic of future research.

7.3 Neighborhood Influence in Linear Prediction: The previous discussion indicates that neighborhood sea locations play one of the most crucial roles in determining climate on land. We further investigate this fact through the following experiments.

We observe the RMSE on the test set from SGL regression by considering only those ocean variables which lie within a certain (geodesic) distance R from the target land region. We increase R from the 'smallest' distance, where only immediate neighborhood ocean locations of the target land region are considered, to the 'largest', when all locations on the earth are considered and note the change in RMSE of SGL prediction. Figs.6(a)-6(b) show the plots obtained for temperature and precipitation prediction in Brazil, while Figs.6(c)-6(d) show the same for India. The x-axis denotes the geodesic radius in kilometers from the target region within which all ocean covariates are considered, while disregarding all other ocean covariates outside this radius. The y-axis denotes the corresponding RMSE.

The plots show that the least error in prediction is obtained when we include covariates in locations which are in the immediate neighborhood of the target variable. Omitting some of the locations leads to a sharp decrease in predictive power. This is consistent with our previous observation that SGL captures high proximity-dependence of the target variables. Covariates which are far away lead to a small increase in RMSE. It may be because most of these covariates are *irrelevant* to our prediction task and appear as "noise". However, the power of the SGL model lies in the fact that it can "filter" out this noise by having much smaller weight on some of these covariates and zero weight on others. The RMSE curve shows a number of 'dips', which might denote that there exist covariates with high predictive power at that distance, which, on being included, increase predictive accuracy of the model.

8 Regularization Paths

As we noted earlier, the regularization parameters (α, λ) play a crucial role in variable selection. It is, therefore, noteworthy to study how variable selection changes with the change in the parameter values. For each covariate, we can compute and plot the coefficient value for a set of chosen (α, λ). Thus, this plot, referred to as the *Regularization Path* of the SGL solutions [8], illustrates how the coefficient values change with change in penalty λ acts as a "tuning" parameter for the model. With

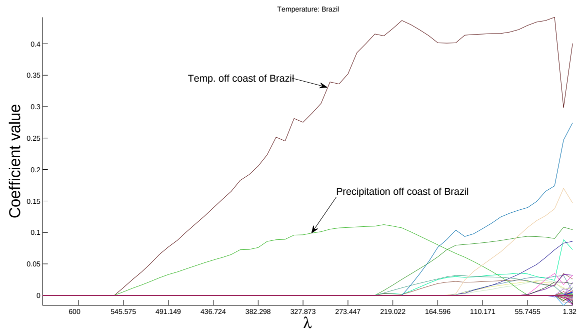


Figure 7: Temp. prediction in Brazil: Regularization path.

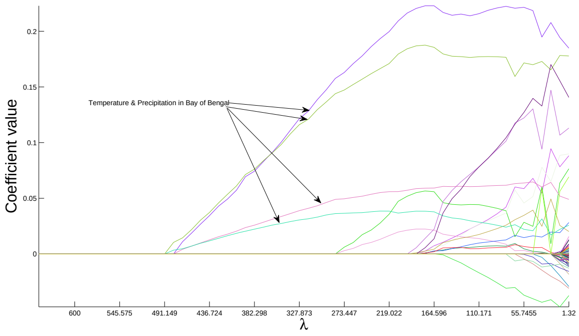


Figure 9: Temp. prediction in India: Regularization path.

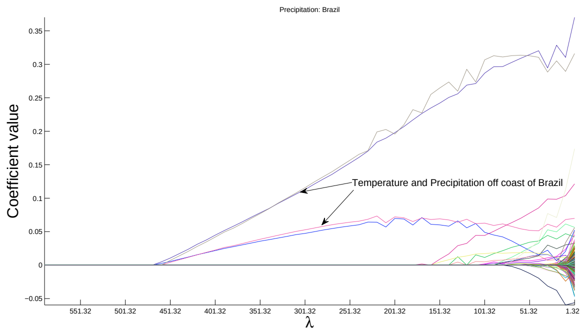


Figure 8: Precip. prediction in Brazil: Regularization Path.

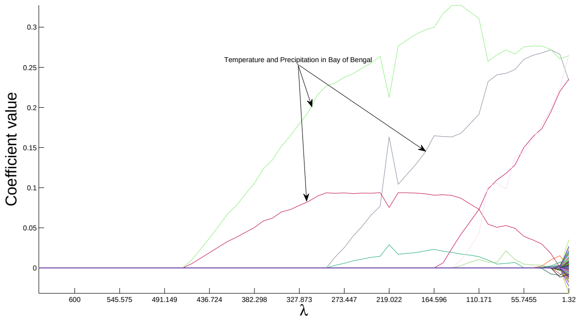


Figure 10: Precip. prediction in India: Regularization path.

higher penalties, we obtain a sparser model. However, it usually corresponds to a gain in RMSE. Most importantly, though, we obtain a quantitative view of the complexity of the model. In particular, the covariates which persist over considerably large ranges of λ and α are the most *robust* covariates in our regression task.

For the chosen training and test datasets, we compute the regularization path for temperature and precipitation predictions in Brazil and India. We fix $\alpha = 0.5$, so that $\lambda_1 = \lambda_2 = \frac{\lambda}{2}$. Figs.7 - 8 show the regularization paths for prediction in Brazil. The most 'stable' covariates, viz. temperature and precipitation in location(s) just off the coast of Brazil, have been earlier reported as among the most *relevant* covariates obtained through cross-validation on the training set.

The regularization paths for prediction in India are plotted in Figs.9 - 10. We observe that in this case too, the most stable covariates are among the *relevant* ones obtained through cross-validation. It is interesting to note that in all the plots, for low values in penalty, a mild increase in penalty dramatically changes the selected model. However, in higher ranges, since the only covariates which survive are the relevant and stable ones, the change in model selection is more gradual.

9 Conclusion

In this paper, we have proved statistical consistency guarantees for a general class hierarchical tree-structured norm regularized estimators. It follows that SGL, which belongs to this class, has statistical consistency of estimation. Application of SGL for predictive modeling of land climate variables has shown that it inherently captures important dependencies that exist between land and ocean variables. In terms of prediction accuracy, SGL is empirically found to outperform the state-of-the-art models in climate. We observe that parsimony in covariate selection improves predictive performance and SGL is robust in its selection of covariates.

Hierarchical tree-structured norm regularized estimators provide a powerful tool for various sparse regression problems in climate, such as hurricane prediction and modeling climate extremes. The results motivate us to build sparsity inducing regularizers that capture more complex dependency structures that are known to climate science, e.g., ocean currents, climate cycles etc. We also want to incorporate temporal lags, which are known to affect the climate system, into our model. Our ultimate goal is to design statistical models which, when incorporated with the existing physical models of climate [5], can provide reasonable predictions of the chaotic climate system.

Acknowledgments: This research was supported in part by NSF Grants IIS-1029711, IIS-0916750, SES-0851705, IIS-0812183, and NSF CAREER Grant IIS-0953274. The authors are grateful for technical support from University of Minnesota Supercomputing Institute (MSI).

References

[1] F. Bach, "Consistency of the group lasso and multiple kernel learning," *JMLR*, vol. 9, pp. 1179–1225, 2008.

[2] P. Baldi and S. Bruna, *Bioinformatics - The Machine Learning Approach*. MIT Press, 1998.

[3] A. Beck and M. Teboulle, "A fast iterative shrinkage-thresholding algorithm for linear inverse problems," *SIAM Journal on Imaging Sciences*, vol. 2 1, pp. 183–202, 2009.

[4] E. Candes and T. Tao, "Decoding by linear programming," *IEEE Trans. on Info. Th.*, vol. 5112, pp. 4203–4215, 2005.

[5] CESM. Community earth system model. [Online]. Available: http://www.cesm.ucar.edu/models/cesm1.0/

[6] D. P. Chambers, B. D. Tapley, and R. H. Stewart, "Anomalous warming in the indian ocean coincident with el nio," *J. of Geoph. Res.*, vol. 104, pp. 3035–3047, 1999.

[7] S. Chatterjee, "An error bound in the Sudakov-Fernique inequality," Arxiv, Tech. Rep., 2008, http://arxiv.org/abs/math/0510424.

[8] B. Efron, T. Hastie, I. Johnstone, and R. Tibshirani, "Least angle regression," *Ann. of Stats.*, vol. 32, pp. 407–499, 2002.

[9] L. et. al, Ed., *Intelligent Data Analysis in Medicine and Pharmacology*. Kluwer, 1997.

[10] L. Evers and C. Messow, "Sparse kernel methods for high-dimensional survival data," *Bioinformatics*, vol. 24, no. 14, pp. 1632–1638, 2008.

[11] U. M. Fayyad, S. G. Djorgovski, and N. Weir, "Automating the analysis and cataloging of sky surveys," in *Advances in Knowledge Discovery and Data Mining*, 1996.

[12] J. Friedman, T. Hastie, and R. Tibshirani, "A note on the group lasso and a sparse group lasso," *Preprint*, 2010.

[13] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning; Data mining, Inference and Prediction*. Springer Verlag, New York, 2001.

[14] R. Jenatton, J. Mairal, G. Obozinski, and F. Bach, "Proximal methods for sparse hierarchical dictionary learning," in *ICML*, June 2010, pp. 487–494.

[15] M. Ledoux and M. Talagrand, *Probability in Banach Spaces: Isoperimetry and Processes*. Springer, 2002.

[16] J. Liu, S. Ji, and J. Ye, *SLEP: Sparse Learning with Efficient Projections*. [Online]. Available: http://www.public.asu.edu/~jye02/Software/SLEP

[17] J. Liu and J. Ye, "Moreau-yosida regularization for grouped tree structure learning," in *NIPS*, 2010.

[18] NCEP/NCAR Reanalysis 1: Surface air temperature (0.995 sigma level). [Online]. Available: http://www.esrl.noaa.gov/psd/data/gridded/data.ncep.reanalysis.surface.html

[19] S. Negahban, P. Ravikumar, M. J. Wainwright, and B. Yu, "A unified framework for high-dimensional analysis of m-estimators with decomposable regularizers," *Arxiv*, 2010, http://arxiv.org/abs/1010.2731v1.

[20] R. T. Rockafellar, *Convex Analysis*. Princeton University Press, 1996.

[21] K. Steinhaeuser, N. Chawla, and A. Ganguly, "Comparing predictive power in climate data: Clustering matters," in *Advances in Spatial and Temporal Databases*, vol. 6849, 2011, pp. 39–55.

[22] ——, "Complex networks as a unified framework for descriptive analysis and predictive modeling in climate science," *Statistical Analysis and Data Mining*, vol. 4, no. 5, pp. 497–511, 2011.

[23] R. Tibshirani, "Regression shrinkage and selection via the lasso," *J. Royal. Statist. Soc. B*, vol. 58, pp. 267–288, 1996.

[24] A. Tsonis, G. Wang, K. Swanson, F. Rodrigues, and L. Costa, "Community structure and dynamics in climate networks," *Climate Dynamics*, 2010.

[25] D. Watts and S. Strogatz, "Collective dynamics of small-world networks," *Nature*, vol. 393, pp. 440–442, 1999.

[26] J. Wright, Y. Ma, J. Mairal, G. Sapiro, T. Huang, and S. Yan, "Sparse representation for computer vision and pattern recognition," *Proc. IEE*, vol. 98, no. 6, pp. 1031–1044, 2010.

[27] M. Yuan and Y. Lin, "Model selection and estimation in regression with grouped variables," *Journal of Royal Statistical Society Series B*, vol. 68 (1), pp. 49–67, 2006.

[28] P. Zhao and B. Yu, "On model selection consistency of lasso," *JMLR*, vol. 7, pp. 2541–2563, 2006.

[29] S. Zhou, J. Lafferty, and L. Wasserman, "Time varying undirected graphs," *Machine Learning*, vol. 80, pp. 295–319, 2010.