Modeling trend in temperature volatility using generalized LASSO

Anonymous Author(s)
Affiliation
Address
email

Abstract

In this paper, we present methodology for estimating trends in spatio-temporal volatility. We give two algorithms for computing our estimator which are tailored for dense, gridded observations over both space and time, though these can be easily extended to other structures (time-varying network flows, neuroimaging). We motivate our methodology using by applying it to a massive climate dataset and discuss the implications for climate analysis.

1 Introduction

Estimating smooth trends over time for large collections of time series is a common problem in economics, finance, meteorology, neuroscience and more. Most of this work focuses on analyzing trends (or removing trends) in the temporal average, but trends in variance can be more relevant, especially for financial data but also for climate science.

Trends in terrestrial temperature variability are perhaps more relavant for species viability than trends in mean temperature [11], because, an increase in the temperature variability will increase the probability of extreme hot outliers [20]. Recent climate literature suggests that it is more difficult for society to adapt to these extremes than to the gradual increase in the mean temperature [8, 11]. Furthermore, the willingness of popular media to emphasize the prevalence extreme cold events coupled with a fundamental misunderstanding of the relationship between climate (the global distribution of weather over the long run) and weather (observed short-term, localized behavior) leads to public misunderstanding of climate change. In fact, increased frequency of extreme cold events in the northern hemisphere is can be partially attributed to increases in mean climate but is also due to increases in temperature variance [6, 15, 19].

Nevertheless, research examining trends in the volatility of spatio-temporal climate data is scarce. [8] studied the change in the standard deviation (SD) of the surface temperature in the NASA Goddard Institute for Space Studies gridded temperature data set by examining the empirical SD at each spatial location relative to that location's SD over a base period, and showed that these estimates are increasing. [11] took a similar approach in analyzing the ERA-40 data set. Their results showed that there still is an increase in the SDs from 1958-1970 to 1991-2001, but this is much less than what is obtained from the method used in [8]. The authors also computed the time-evolving global SD from the de-trended time-series at each position, which suggests that the global SD has been stable.

These and other related research, e.g., [14]) have several shortcomings. First, no statistical analysis has been performed to examine if the changes in the SD are statistically significant. Second, the methodologies for computing the SDs are highly sensitive to the choice of base period. Third, and most importantly, temporal and spatial correlations between the observations are ignored.

Submitted to 32nd Conference on Neural Information Processing Systems (NIPS 2018). Do not distribute.

1.1 Related work

Variance estimation for financial time series has a lengthy history, focused especially on parametric models like the generalized autoregressive conditional heteroskedasticity (GARCH) process [5] and stochastic volatility models [9]. These models (and related AR processes) are specifically for parametric modelling of short "bursts" of high volatility, behavior typical of financial instruments. Parametric models for spatial data go back at least to [2] who proposed a conditional probability model on the lattice for examining plant ecology.

More recently, nonparametric models for both spatial and temporal data have focused on using ℓ_1-regularization for trend estimation. [12] proposed ℓ_1-trend filtering for univariate time series, which forms the basis of our methods. These methods have been generalized to higher order temporal smoothness [17], graph dependencies [21], and, most recently, small, time-varying graphs [7]. Our methodology is similar in flavor to [7], though it uses a different likelihood function to emphasize variance estimation rather than trends in mean signal. Furthermore, the focus is in high-dimensional, regular data rather than a small number of changing graphs.

1.2 Main contributions

The main contribution of this work is to develop a new methodology for detecting the trend in the volatility of spatio-temporal data. In this methodology, the variance at each position and time is considered as a hidden variable. The values of these hidden variables are then estimated by maximizing the likelihood of the observed data. We show that this formulation is not appropriate for detecting the trend, so, following [17], we penalize the differences between the estimated variances which are temporally and spatially "close", resulting in a generalized LASSO problem.

Our main contributions are as follows:

1. We propose a model for nonparametric variance estimation for a spatio-temporal process (Section 2).
2. We derive two alternating direction method of multiplier algorithms (ADMM) to fit our estimator when applied to very large data (Section 3). We give situations under which each algorithm is most likely to be useful.
3. We illustrate our methods on a large global temperature dataset with the goal of tracking world-wide trends in variance as well as a simulation constructed to mimic these data's features (Section 4).

While we motivate and illustrate our methods on large, gridded climate data, we note that our algorithms are also applicable to neuroimaging data or large collections of financial instruments.

2 Estimating the variance of spatio-temporal data

ℓ_1-trend filtering was proposed by [12] as a method for estimating a smooth, time-varying trend. It is formulated as the optimization problem $\min_\beta \frac{1}{2}\sum_{t=1}^{T}(y_t-\beta_t)^2+\lambda\sum_{t=1}^{T-2}|\beta_t-2\beta_{t+1}+\beta_{t+2}|$ or equivalently:

$$\min_{\beta} \frac{1}{2}\left\|y-\beta\right\|_2^2+\lambda\left\|D\beta\right\|_1 \tag{1}$$

where y_t is an observed time-series, β is the smooth trend, D is a $(T-2)\times T$ matrix, and λ is a tuning parameter which balances fidelity to the data (small errors in the first term) with a desire for smoothness. With the penalty matrix D, the estimated β will be piecewise linear. [12] proposed a specialized primal-dual interior point (PDIP) algorithm for solving (1). From a statistical perspective, (1) is a constrained maximum likelihood problem with independent observations from a normal distribution with common variance, $y_t \sim \mathrm{N}(\beta_t, \sigma^2)$, subject to a piecewise linear constraint on β.

2.1 Estimating the variance

Inspired by the ℓ_1-trend filtering algorithm, we propose a non-parametric model for estimating the variance of a time-series. To this end, we assume that at each time step t, there is a hidden variable h_t such that conditioned on h_t the observations y_t are independent normal variables with zero mean and variance $\exp(h_t)$. The negative log-likelihood of the observed data in this model is $l(y \mid h) \propto -\sum_{t=1}^{T} h_t - y_t^2 e^{-h_t}$. Crucially, we assume that the hidden variables h_t vary smoothly. To impose this assumption, we estimate h_t by solving the penalized, negative log-likelihood:

$$\min_h -l(y \mid h) + \lambda \|Dh\|_1 \tag{2}$$

where D has the same structure as above.

As with (1), one can solve (2) using the PDIP algorithm (as in, e.g., `cvxopt` [1]). In each iteration of PDIP we need to compute a search direction by taking a Newton step on a system of nonlinear equations. Due to space limitations, we defer details to Appendix B in the Supplement.

2.2 Adding spatial constraints

The method in the previous section can be used to estimate the variance of a single time-series. In this section, we extend this method to the estimation of the variance of spatio-temporal data.

At a specific time t, the data is measured on a grid of points with n_r rows and n_c columns for a total of $S = n_r \times n_c$ spatial locations. Let y_{ijt} denote the value of the observation at time t on the i^{th} row and j^{th} column of the grid, and h_{ijt} denote the corresponding hidden variable. We seek to impose both temporal and spatial smoothness constraints on the hidden variables. Specifically, we seek a solution for h which is piecewise linear in time and piecewise constant in space (although higher-order smoothness can be imposed with minimal alterations to the methodology). We achieve this goal by solving the following optimization problem:

$$\begin{aligned}\min_h \sum_{i,j,t} h_{ijt} + y_{ijt}^2 e^{-h_{ijt}} + \lambda_1 \sum_{i,j} \sum_{t=1}^{T-2} \left| h_{ijt} - 2h_{ij(t+1)} + h_{ij(t+2)} \right| \\ + \lambda_2 \sum_{t,j} \sum_{i=1}^{n_r-1} \left| h_{ijt} - h_{(i+1)jt} \right| + \lambda_2 \sum_{t,i} \sum_{j=1}^{n_c-1} \left| h_{ijt} - h_{i(j+1)t} \right|\end{aligned} \tag{3}$$

The first term in the objective is proportional to the negative log-likelihood, the second is the temporal penalty for the time-series at each location (i, j), while the third and fourth, penalize the difference between the estimated variance of two vertically and horizontally adjacent points, respectively. The spatial component of this penalty is a special case of trend filtering on graphs [21] which penalizes the difference between the estimated values of the signal on the connected nodes. As before, we can write (3) in matrix form where h is an $T \times S$ vector and D is replaced by $D_{ST} \in \mathbb{R}^{(N_t+N_s)\times(T\cdot S)}$, where $N_t = S \cdot (T-2)$ and $N_s = T \cdot (2n_r n_c - n_r)$ are the number of temporal and spatial constraints, respectively[1]. Then, as we have two different tuning parameters for the temporal and spatial components, we write $\Lambda = \left[\lambda_1 \mathbf{1}_{N_t}^\top,\ \lambda_2 \mathbf{1}_{N_s}^\top\right]^\top$ leading to:[2]

$$\min_h -l(y \mid h) + \Lambda^\top |D_{ST} h|. \tag{4}$$

TODO: equations should have proper punctuation as if they were a sentence.

3 Proposed optimization methods

For a spatial grid of size S and T time steps, D_{ST} will have $3Tn_r n_c - 2n_r n_c - Tn_r$ rows and ST columns. For a $1^\circ \times 1^\circ$ grid over the entire northern hemisphere and daily data over 10 years, we have $n_r = 90$, $n_c = 180$, $T = 3650$ and so D_{ST} has approximately 10^8 columns and 10^8 rows. In each step of the PDIP algorithm, we need to solve a linear system of equations in A which depends on $D_{ST}^\top D_{ST}$ (see appendix A and B). Therefore, applying the PDIP directly is infeasible for our data.[3]

[1] N_s is obtained by counting the number of unique constraints at each location and at all times.

[2] Throughout the paper, we use $|x|$ for both scalars and vectors. For vectors we use this to denote a vector obtained by taking the absolute value of each entry of x.

[3] We note that this is a highly structured and sparse matrix, but, unlike trend filtering alone, it is not banded. We are unaware of general linear algebra techniques for inverting such matrix, despite our best efforts.

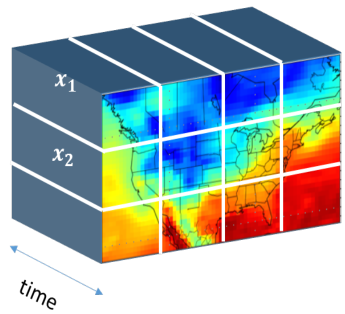

Figure 1: The cube represents the global variable h in space and time. The sub-cubes specified by the white lines are x_i.

In the next section, we develop two ADMM algorithms for solving this problem efficiently. The first casts the problem as a so-called consensus optimization problem [3] which solves smaller sub-problems using PDIP and then recombines the results. The second uses proximal methods to avoid matrix inversions.

3.1 Consensus optimization

Given an optimization problem of the form $\min_h f(h)$, where $h \in \mathbb{R}^n$ is the global variable and $f(h) : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ is convex. Consensus optimization breaks this problem into several smaller sub-problems that can be solved independently in each iteration of optimization.

Assume it is possible to define a set of local variables $x_i \in \mathbb{R}^{n_i}$ such that $f(h) = \sum_i f_i(x_i)$, where each x_i is a subset of the global variable h. More specifically, each entry of the local variables corresponds to an entry of the global variable. Therefore we can define a mapping $\mathscr{G}(i, j)$ from the local variable indices into the global variable indices: $k = \mathscr{G}(i, j)$ means that the j^{th} entry of x_i is h_k (or $(x_i)_j = h_k$). For ease of notation, define $\tilde{h}_i \in \mathbb{R}^{n_i}$ as $(\tilde{h}_i)_j = h_{\mathscr{G}(i,j)}$. Then, the original optimization problem is equivalent to the following problem:

$$\begin{aligned} \min_{\{x_1,\ldots,x_N\}} & \sum_i f_i(x_i) \\ s.t.\ & \tilde{h}_i = x_i. \end{aligned} \tag{5}$$

It is important to note that each entry of the global variable may correspond to several entries of the local variables and so the constraints $\tilde{h}_i = x_i$ enforce the consensus between the local variables corresponding to the same global variable. The augmented Lagrangian corresponding to (5) is $L_\rho(x, h, y) = \sum_i \left(f_i(x_i) + u_i^\top (x_i - \tilde{h}_i) + (\rho/2)\|x_i - \tilde{h}_i\|_2^2\right)$. Now, we can apply ADMM to L_ρ. To solve the optimization problem (4) using this method, we need to address two questions: first, how to choose the local variables x_i, and second, how to the update them.

In Figure 1, the global variable h is represented as a cube (using the subset of the US as an example). We decompose h into sub-cubes as shown by white lines. With this definition of x_i, the objective (4) decomposes as $\sum_i f_i(x_i)$ where $f_i(x_i) = -l(y_i \mid x_i) + \Lambda_{(i)}^\top |D_{(i)} x_i|$, and $\Lambda_{(i)}$ and $D_{(i)}$ contain the temporal and spatial penalties corresponding to x_i only. Thus, with this choice of the local variables x_i, we solve the x-update using the PDIP method. Algorithm 1 gives the general version of this algorithm. A more detailed discussion of the is in the Supplement.

Because consensus ADMM breaks the large optimization into sub-problems that can be solved independently, it is amenable to a split-gather parallelization strategy via, e.g., the map reduce framework. In each iteration, the computation time will be equal to the time to solve each sub-problem plus the time to communicate the solutions on the master processor and perform the consensus step. Since each sub-problem is small, with parallelization, the computation time in each iteration will be small. In addition, our experiments with several values of λ_t and λ_s showed that the algorithm converges in few hundreds iterations. However, this algorithm is only useful if we can

Algorithm 1 Consensus ADMM

Input: data y, penalty matrix D, $\epsilon, \rho, \lambda_t, \lambda_s > 0$.
Set: $h \leftarrow 0$, $z \leftarrow 0$, $u \leftarrow 0$. ▷ Initialization
repeat
$x_i \leftarrow \text{argmin}_{x_i} -l(y_i \mid x_i) + \Lambda_{(i)}^\top |D_{(i)} x_i|$
$\quad +(u_i)^\top x_i + (\rho/2)\|x_i - \tilde{h}_i\|_2^2.$ ▷ Update local vars using PDIP
$h_k \leftarrow (1/S_k) \sum_{\mathscr{G}(i,j)=k} (x_i)_j.$ ▷ Global update.
$u_i \leftarrow u_i + \rho(x_i - \tilde{h}_i).$ ▷ Dual update
until $\left\|h^{m+1} - h^m\right\|_2^2 < \epsilon$
Return: h.

parallelize the computation over several machines. In the next section, we describe another algorithm which makes the computation feasible on a single machine.

3.2 Linearized ADMM

Consider the generic optimization problem $\min_x f(x) + g(Dx)$ where $x \in \mathbb{R}^n$ and $D \in \mathbb{R}^{m \times n}$. Each iteration of the linearized ADMM algorithm [13] for solving this problem has the form

$$\begin{aligned}
x &\leftarrow \underset{\mu f}{\mathbf{prox}} \left(x - (\mu/\rho) D^\top (Dx - z + u)\right) \\
z &\leftarrow \underset{\rho g}{\mathbf{prox}} \left(z + u\right) \\
u &\leftarrow u + Dx - z
\end{aligned}$$

where the algorithm parameters μ and ρ satisfy $0 < \mu < \rho / \|D\|_2^2$, $z, u \in \mathbb{R}^m$ and the proximal operator is defined as $\mathbf{prox}_{\alpha f}(u) = \min_x \ \alpha \cdot f(x) + \frac{1}{2}\|x - u\|_2^2$.

Proximal algorithms are feasible when these proximal operators can be evaluated efficiently which, as we show next, is the case for our problem.

Lemma 1. *Let* $f(h) = \sum_k h_k + y_k^2 e^{-h_k}$ *and* $g(x) = \|x\|_1$. *Then,*

$$\begin{aligned}
[\underset{\mu f}{\mathbf{prox}}(u)]_k &= \mathscr{W}\left(\frac{y_k^2}{\mu} \exp\left(\frac{1 - \mu u_k}{\mu}\right)\right) + \frac{1 - \mu u_k}{\mu}, \\
\underset{\rho g}{\mathbf{prox}}(u) &= S_{\rho\lambda}(u)
\end{aligned}$$

where $\mathscr{W}(\cdot)$ *is the Lambert function [4],* $[S_\alpha(u)]_k = \text{sign}(u_k)(|u_k| - \alpha_k)_+$ *and* $(v)_+ = v \vee 0$.

Proof. If $f(x) = \sum_k f_k(x_k)$ then $[\mathbf{prox}_{\mu f}(x)]_k = \mathbf{prox}_{\mu f_k}(u_k)$. So $[\mathbf{prox}_{\mu f}(u)]_k = \min_{x_k} \ \mu\left(x_k + y_k^2 e^{-x_k}\right) + \frac{1}{2}(x_k - u_k)^2$. Setting the derivative to 0 and solving for u_k gives the result. Similarly, $[\mathbf{prox}_{\rho g}(u)]_\ell = \rho\lambda_\ell |z_\ell| + 1/2(z_\ell - u_\ell)^2$. This is not differentiable, but the solution must satisfy $\rho \cdot \lambda_\ell \cdot \partial\left(|z_\ell|\right) = u_\ell - z_\ell$ where $\partial\left(|z_\ell|\right)$ is the sub-differential of $|z_\ell|$. The solution is the soft-thresholding operator $S_{\rho\lambda_\ell}(u_\ell)$. □

Therefore, Algorithm 2 gives a different method for solving the same problem.

4 Empirical evaluation

In this section, we examine both simulated and real spatio-temporal climate data. All the computations were performed on a Linux machine with four 3.20GHz Intel i5-3470 cores.

4.1 Simulations

We generate observations at all time steps and all locations from independent Gaussian random variables with zero mean. However, the variance of these random variables follows a smoothly varying function in time and space

Algorithm 2 Linearized ADMM

Input: data y, penalty matrix D, $\epsilon, \rho, \lambda_t, \lambda_s > 0$.
Set: $h \leftarrow 0$, $z \leftarrow 0$, $u \leftarrow 0$. $\triangleright$ Initialization
repeat
  $h_k \leftarrow \mathscr{W}\left(\frac{y_k^2}{\mu}\exp\left(\frac{1-\mu u_k}{\mu}\right)\right) + \frac{1-\mu u_k}{\mu} \quad k = 1, \ldots TS.$ $\triangleright$ Primal update
  $z \leftarrow S_{\rho\lambda}(u)$. $\triangleright$ Elementwise soft thresholding
  $u \leftarrow u + Dh - z$. $\triangleright$ Dual update
until $\max\{\|h - z\|_2^2, \left\|z^{m+1} - z^m\right\|_2^2\} < \epsilon$
Return: z.

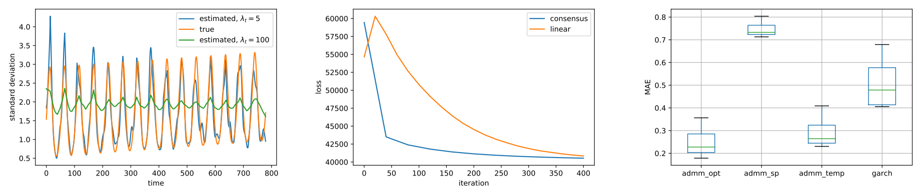

Figure 2: Left: The true (orange) and estimated standard deviation function at the location (0,0). The estimated values are obtained using linearized ADMM with $\lambda_s = 0.1$ and two values of λ_t: $\lambda_t = 5$ (blue) and $\lambda_t = 100$ (green). Middle: Convergence speed of linearized and consensus ADMM. Right: MAE for four models: admm_opt: the proposed model with optimal values of λ_t and λ_s, admm_temp: no spatial penalty, admm_sp: no temporal penalty.

$$\sigma^2(t, r, c) = \sum_{s=1}^{S} W_s(t) \cdot \exp\left(\frac{(r - r_s)^2 + (c - c_s)^2}{2\sigma_s^2}\right); \quad W_s(t) = \alpha_s \cdot t + \exp(\sin(2\pi\omega_s t + \phi_s)).$$

In words, the variance at each time and location is computed as the weighted sum of S bell-shaped functions where the weights are time-varying, consist of a linear trend $\alpha_s \cdot t$ and a periodic term $\beta_s \cdot \sin(2\pi\omega_s t + \phi_s)$. The bell-shaped functions impose the spatial smoothness, and the linear trend and the periodic terms enforce the temporal smoothness similar to the seasonal component in the real climate data. We simulated the data on a 5 by 7 grid and for 780 time steps with $S = 4$. Specific parameter choices of the variance function are shown in Table 1 in the Supplement. For illustration, we also plot the variance function for all locations at $t = 25$ and $t = 45$ in as well as the variance across time at $(0, 0)$ in Figure 1 in the Supplement.

We estimated the linearized ADMM for all combinations of values of λ_t and λ_s from the sets $\lambda_t \in \{0, 1, 5, 10, 50, 100\}$ and $\lambda_s \in \{0, 0.05, 0.1, 0.2, 0.3\}$. For each pair, we then compute the mean absolute error (MAE) between the estimated variance and the true variance at all locations and all time steps. For $\lambda_t = 5$ and $\lambda_s = 0.1$, the MAE was minimized. The left panel of Figure 2 shows the true and the estimated standard deviation at location (0,0) using $\lambda_s = 0.1$ and $\lambda_t = 5$ (blue) and $\lambda_t = 100$ (green). As we can see, larger than optimal value of λ_t leads to estimated values which are "too smooth".

The middle panel of Figure 2 shows the convergence of both methods. Each iteration of the linearized algorithm takes 0.01 seconds on average while each iteration of the consensus ADMM takes about 20 seconds.

To further examine the performance of the proposed model, we next compare it to three alternatives: a model which does not consider the spatial smoothness (equivalent to fitting the model in Section 2.1 to each time-series separately), a model which does not consider imposes only spatial smoothness, and a GARCH(1,1) model. We simulated 100 datasets using the method explained above with $\sigma_s \sim \text{uniform}(4, 7)$. The right panel of Figure 2 shows the boxplot of the MAE for these models. Interestingly, the proposed model with optimal parameters outperforms GARCH(1,1) in estimating the true value of the variance.

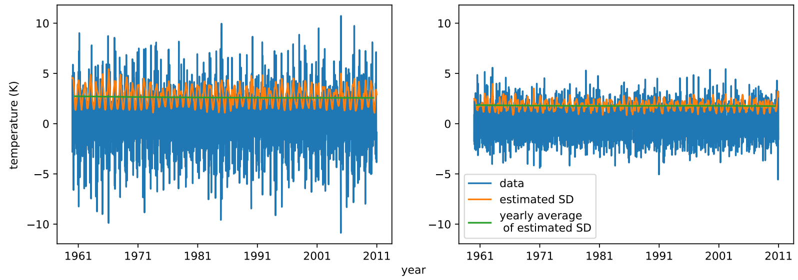

Figure 3: Detrended data and the estimated SD for a small midwestern city (left) and San Diego (right).

4.2 Data analysis

Consensus ADMM in Section 3.1 is appropriate only if we parallelize it over multiple machines, and it is significantly slower on our simulated data, so we do not pursue it further here. All the results reported in this section are obtained using Algorithm 2. We applied this algorithm to the northern hemisphere of the ERA-20C dataset available from the European Center for Medium-Range Weather Forecasts. The data are the 2 meter temperature measured daily at 12 p.m from January 1, 1960 to December 24, 2010.

The Supplement explains some preprocessing and investigates some properties of the time-series of different locations on earth. Figure 3 a processed time-series for a single location. The variance of this time-series has an irregular cyclic behavior. Additionally, the time-series of other locations show different patterns. These observations motivated the need to develop a non-parametric framework for this problem. Figure 3 also shows the estimated SD obtained using the method of Section 2.1.

Convergence As shown in Algorithm 2, we evaluated convergence using $\epsilon = 0.001\%$ of the MSE of the data. Our simulation experiments showed that the convergence speed depends on the value of λ_t and λ_s. Furthermore, using the solution obtained for smaller values of these parameters as a warm start for the larger values, the converges speed improves.

Model selection One common method for choosing the penalty parameters in the Lasso problems is to find the solution for a range of the values of these parameters and then choose the values which minimize a model selection criterion. However, such analyses needs the computation of the degrees of freedom. Several previous work have investigated the df in generalized lasso problems [10, 18, 22]. However, all these studies have considered the linear regression problem and, to the best of our knowledge, the problem of computing the df for generalized lasso with general objective function has not been considered yet.

In this paper, we use a heuristic method for choosing λ_t and λ_s: we compute the optimal solution for a range of values of these parameters and choose the values which minimize $\mathscr{L}(\lambda_t, \lambda_s) = -l(y|h) + \sum \|D_{total} h\|$. This objective is a compromise between the negative log likelihood ($-l(y|h)$) and the complexity of the solution ($\sum \|D_{total} h\|$). For smoother solutions the value of $\sum \|D_{total} h\|$ will be smaller but with the cost of larger $-l(y|h)$.

We computed the optimal solution for all the combinations of the following sets of values: $\lambda_t \in \{0, 2, 4, 8, 10, 15, 200, 1000\}$, $\lambda_s \in \{0, .1, .5, 2, 5, 10\}$. The best combination was $\lambda_t = 4$ and $\lambda_s = 2$. All the analyses in the next section are performed on the solution for these values.

Analysis of trend in temperature volatility The top row of Figure 3 shows the detrended data, the estimated standard deviation and the yearly average of these estimates for two cities in the US: a small midwestern city (left) and San Diego (right). The estimated SD captures the periodic behavior in the variance of the time-series. In addition, the number of linear segments changes adaptively in each time window depending on how fast the variance is changing.

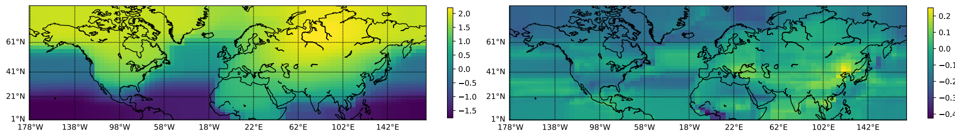

Figure 4: The average of the estimated variance over the northern hemisphere (left) and the change in the variance from 1961 to 2011 (right).

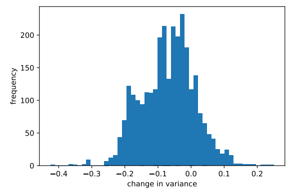

Figure 5: The histogram of changes in estimated SD.

The yearly average of the estimated SD captures the trend in the temperature volatility. For example, we can see that the variance in the midwestern city displays a small positive trend. To determine how the volatility has changed in each location, we subtract the average of the estimated variance in TODO: 1992 from the average in the following years and compute their sum. The value of this change in the variance in each location is depicted in the right panel of Figure 4. The left panel of this shows the average estimated variance in each location. Since the optimal value of the spatial penalty is rather large ($\lambda_s = 2$) the estimated variance is spatially very smooth.

It is interesting to note that the trend in volatility is almost zero over the oceans. The most positive trend can be observed in Asia and particularly in south-east Asia.

Figure 5 shows the histogram of change in the estimated SD across the northern hemisphere. The SD in most locations on the northern hemisphere had a negative trend in this time period, though spatially, this decreasing pattern is localized mainly toward the extreme northern latitudes and over oceans. In many ways, this is consistent with climate change predictions: oceans tend to operate as a local thermostat, regulating deviations in local temperature, while warming polar regions display fewer days of extreme cold.

5 Discussion

In this paper, we proposed a new method for estimating the variance of spatio-temporal data. The main idea is to cast this problem as a constrained optimization problem where the constraints enforce smooth changes in the variance for neighboring points in time and space. In particular, the solution is piecewise linear in time and piecewise constant in space. The resulting optimization is in the form of a generalized LASSO problem with high-dimension, and so applying the PDIP method directly is infeasible. We therefore developed two ADMM-based algorithms to solve this problem: the consensus ADMM and linearized ADMM.

The consensus ADMM algorithm converges in a few hundreds of iterations but each iteration takes much longer than the linearized ADMM algorithm. The appealing feature of the consensus ADMM algorithm is that if it is parallelized on enough machines the computation time per iteration remains constant as the problem size increases. The linearized ADMM algorithm on the other hand converges in a few thousand iterations but each iteration is performed in a split second. However, since the algorithm converges in many iterations it is not very appropriate for parallelization. The reason is that after each iteration the solution computed in each machine should be broadcast to the master machine and this operation takes some time which depends on the speed of the network connecting the slave machines to the master. A direction for future research would be to combine these two algorithms in the following way: the problem should be split into the sub-problems (as in the consensus ADMM) but each sub-problem can be solved using linearized ADMM.

References

[1] M. S. Andersen, J. Dahl, and L. Vandenberghe. CVXOPT: A Python package for convex optimization, version 1.1. 6. *Available at cvxopt. org 54*, 2013.

[2] J. Besag. Spatial interaction and the statistical analysis of lattice systems. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 192–236, 1974.

[3] 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*, 3(1): 1–122, 2011. ISSN 1935-8237.

[4] R. M. Corless, G. H. Gonnet, D. E. G. Hare, D. J. Jeffrey, and D. E. Knuth. On the LambertW function. *Advances in Computational Mathematics*, 5(1):329–359, Dec. 1996.

[5] R. Engle. Dynamic conditional correlation: A simple class of multivariate generalized autoregressive conditional heteroskedasticity models. *Journal of Business & Economic Statistics*, 20(3):339–350, 2002.

[6] E. M. Fischer, U. Beyerle, and R. Knutti. Robust spatially aggregated projections of climate extremes. *Nature Climate Change*, 3:1033—1038, 2013.

[7] D. Hallac, Y. Park, S. Boyd, and J. Leskovec. Network inference via the time-varying graphical lasso. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '17, pages 205–213, New York, NY, USA, 2017. ACM.

[8] J. Hansen, M. Sato, and R. Ruedy. Perception of climate change. *Proceedings of the National Academy of Sciences*, 109(37), Sept. 2012.

[9] A. Harvey, E. Ruiz, and N. Shephard. Multivariate stochastic variance models. *The Review of Economic Studies*, 61(2):247–264, 1994.

[10] Q. Hu, P. Zeng, and L. Lin. The dual and degrees of freedom of linearly constrained generalized lasso. *Computational Statistics & Data Analysis*, 86:13–26, June 2015.

[11] C. Huntingford, P. D. Jones, V. N. Livina, T. M. Lenton, and P. M. Cox. No increase in global temperature variability despite changing regional patterns. *Nature*, 500(7462):327–330, Aug. 2013.

[12] S.-J. Kim, K. Koh, S. Boyd, and D. Gorinevsky. ℓ_1 trend filtering. *SIAM Review*, 51(2):339–360, 2009.

[13] N. Parikh and S. Boyd. Proximal Algorithms. *Foundations and Trends® in Optimization*, 1(3):127–239, Jan. 2014.

[14] A. Rhines and P. Huybers. Frequent summer temperature extremes reflect changes in the mean, not the variance. *Proceedings of the National Academy of Sciences*, 110(7):E546–E546, Feb. 2013.

[15] J. A. Screen. Arctic amplification decreases temperature variance in northern mid- to high-latitudes. *Nature Climate Change*, 4:577—582, 2014.

[16] R. J. Tibshirani. *The Solution Path of the Generalized Lasso*. PhD Thesis, Stanford University, 2011.

[17] R. J. Tibshirani. Adaptive piecewise polynomial estimation via trend filtering. *Annals of Statistics*, 42: 285–323, 2014.

[18] R. J. Tibshirani and J. Taylor. Degrees of freedom in lasso problems. *The Annals of Statistics*, 40(2): 1198–1232, 2012.

[19] K. E. Trenberth, Y. Zhang, J. T. Fasullo, and S. Taguchi. Climate variability and relationships between top-of-atmosphere radiation and temperatures on earth. *Journal of Geophysical Research: Atmospheres*, 120(9):3642–3659, 2014.

[20] D. A. Vasseur, J. P. DeLong, B. Gilbert, H. S. Greig, C. D. G. Harley, K. S. McCann, V. Savage, T. D. Tunney, and M. I. O'Connor. Increased temperature variation poses a greater risk to species than climate warming. *Proceedings of the Royal Society of London B: Biological Sciences*, 281(1779), 2014.

[21] Y.-X. Wang, J. Sharpnack, A. J. Smola, and R. J. Tibshirani. Trend filtering on graphs. *Journal of Machine Learning Research*, 17(105):1–41, 2016.

[22] P. Zeng, Q. Hu, and X. Li. Geometry and Degrees of Freedom of Linearly Constrained Generalized Lasso. *Scandinavian Journal of Statistics*, 44(4):989–1008, Nov. 2017.