A Sparse Bayesian Model for Dependence Analysis of Extremes: Climate Applications

Debasish Das, Auroop R. Ganguly D.DAS , A.GANGULY@NEU.EDU
Sustainability and Data Sciences Lab, Northeastern University,
360 Huntington Avenue, Boston, MA, USA

Zoran Obradovic ZORAN@IST.TEMPLE.EDU
Center for Data Analysis and Biomedical Informatics,Temple University,
1805 N. Broad St., Philadelphia, PA, USA

Abstract

In many real applications, such as climate, finance and social media among others, we are often interested in extreme events. An important part of modeling extremes is discovery of covariates on which the quantities related to the extremes are dependent, as this may lead to improved understanding and the discovery of new causal drivers of extremes. Despite developments in sparse covariate discovery algorithms, adaptations to extremes can fail because the tail attributes do not follow a Gaussian distribution. In this paper, we proposed a sparse Bayesian framework for discovery of covariates that influences the frequency of extremes based on the Poisson description of extremes frequency and hierarchical Bayesian description of a sparse regression model. We developed an efficient approximation algorithm based on the variational Bayes approach to estimate the distribution over regression coefficients that indicate dependence of extremes on the corresponding covariates. Experiments with synthetic data demonstrate the ability of the approach to accurately characterize dependence structures. Applications to rainfall extremes suggest new insights relevant for improved understanding of hydrological extremes under climate variability and change.

1. Introduction

Descriptive analysis and predictive modeling of extreme values are growing in importance and urgency across societal priorities, ranging from natural hazards and climate change to security of physical and cyber systems, as well as financial markets, telecommunication signals (Reiss & Thomas, 2007) and even relatively newer fields like social media. Extremes are particularly relevant for climate due to the connection of consistent change in extreme weather patterns with the warming climate, as identified by Intergovernmental Panel on Climate Change (IPCC) in their Special Report on EXtremes (SREX) (Field et al., 2012).

Characterizing the dependence of extreme values on multiple covariates is often a first step in developing causal insights involving physical and empirical relations. Sparse regression techniques based on L_1-regularizers have gained prominence as a tool for finding dependence structures between variables in many application areas due to the development of efficient algorithms (Tibshirani, 1994; Friedman et al., 2010) that can handle millions of potential feature variables. However, application of such methods for dependence discovery of extremes is not straightforward since the distribution of extreme values are not Gaussian. This violates the basic assumption of Gaussianity inherent to the sparse regression techniques. The statistical approach developed to model extremes is the Extreme Value Theory (EVT) (Coles, 2001) that provides a natural family of probability distributions for modeling of extremes. Despite their importance, problems related to extremes were rarely addressed in knowledge discovery community. In (Lozano et al., 2009) a Group Elastic Net method was described that discovers Granger causality between a set of time-series which was used for climate change attribution. However, it is still a

Presented at the International Conference on Machine Learning (ICML) workshop on *Inferning: Interactions between Inference and Learning*, Atlanta, Georgia, USA, 2013. Copyright 2013 by the author(s).

method for analyzing dependence among Gaussian distributed variables. In (Liu et al., 2012) a latent space model was used to obtain dependence between a set of extremes time series which is a very different problem from the one we are addressing. We propose a general Bayesian framework to adapt the L1-regularized sparse regression (LASSO) approach for dependence discovery for frequency of extremes. We adopted a latent space model of (Liu et al., 2012) and regarded the average rates of the extremes frequency as latent variable and estimated their linear dependence on the covariates using the Bayesian version of the L_1-regularized regression (Park & Casella, 2008). We derived a fully Bayesian description of the joint distribution of observed frequencies and all latent variables and developed an efficient variational approximation algorithm to obtain the posterior distribution of the regression coefficients. This enabled us to discover the relevant features along with an estimate of uncertainty. The proposed framework can be applied as a statistical tool for dependence analysis of extremes in many other fields where extremes are important.

The rest of the paper is organized as following. In section 2, we describe the proposed Bayesian framework for dependence discovery of extremes frequencies along with some basics of extreme value modeling. In section 3 we describe related work and in section 4 we presented and discussed the experimental results. Finally in section 5 we conclude and propose future work.

2. Methodology

2.1. Basics of Extreme Value Modeling

Extremes events are rare and their probabilities are represented by the tail of distributions. As mentioned earlier, the statistical properties of extremes are described by Extreme Value Theory (Coles, 2001; Reiss & Thomas, 2007). Extreme events are generally characterized by their intensity, duration or frequency. Intensity of extremes may be described by either Generalized Extreme Value (GEV) distribution (Coles, 2001) or Generalized Pareto distribution (GPD) (Coles, 2001) depending on whether samples are obtained using block maxima (e.g. annual maxima) or peaks-over-threshold approach (Coles, 2001). On the other hand, frequency and the inter-arrival time of extremes are modeled by Poisson point process approach where the number of extremes N within a fixed period are described by Poisson distribution (Coles, 2001), which is given by the following pdf.

$$Pois(N|\lambda) = \frac{\lambda^N e^{-\lambda}}{N!} \qquad (1)$$

where λ is the average rate of arrival of extremes.

2.2. Sparse Regression for Extremes Dependence Analysis

Let us denote the observed numbers of occurrences (frequencies) of extremes by $N_1, N_2, ..., N_n$ in n equal period of time (e.g. 1 year). We also have n observations of p covariates that may or may not influence the frequency of extremes. They are given by $\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_n$ where $\mathbf{x}_i$ is a p-dimensional vector. Now we assume each of these frequencies are drawn from distinct Poisson distributions with fixed but unknown average rates $\lambda_1, \lambda_2, ..., \lambda_n$. Given the above observations, our goal is to find a sparse subset of covariates (we may use "feature variables" alternately to mean covariates) that influence the frequencies of extremes. In order to obtain a robust and interpretable model, a natural choice is a sparse linear model. However, directly modeling the relation between observed frequencies and the feature variables using a linear model may not be appropriate since the the relation may well be non-linear. So, we use a latent space model similar to (Liu et al., 2012), where the average rate parameters of Poisson distributions are regarded as latent variables that are more stable than the actual observed frequencies and can be described reasonably well by a linear model. Therefore, instead of modeling the actual observations, we model the relationship between the latent variables and the feature variables using a sparse linear regression model. However, we model $\log(\lambda_i)$ instead of λ_i in order to enforce positivity on λ_i (also, $\log(\lambda_i)$ is a natural parameter for Poisson distribution). The final linear model is given by

$$\log(\lambda_i) = \boldsymbol{\beta}^\top \mathbf{x}_i + \epsilon, \quad \text{subject to} \quad ||\boldsymbol{\beta}||_1^1 \leq t \qquad (2)$$

where $\epsilon \sim \mathcal{N}(\mu, \tau^{-1})$. Here $\boldsymbol{\beta} = [\beta_1 \beta_2 ... \beta_p]$ is the vector of regression coefficients corresponding to p different features. Now we denote $\psi_i = \log(\lambda_i)$ and the joint distribution of the frequencies N_i and average rate parameters λ_i is given by

$$p(N_i, \psi_i|\boldsymbol{\beta}, \tau, \mathbf{X}) = \prod_{i=1}^{n} p(N_i|e^{\psi_i}) p(\psi_i|\boldsymbol{\beta}, \tau, \mathbf{x}_i) \qquad (3)$$

The distribution $p(N_i|e^{\psi_i})$ can be modeled by a Poisson distribution and $p(\psi_i|\boldsymbol{\beta}, \tau, \mathbf{x}_i)$ can be modeled by the Gaussian given by $\mathcal{N}(\boldsymbol{\beta}^\top \mathbf{x}_i, \tau^{-1})$. In order to implement sparsity, we can use a Laplace prior over $\boldsymbol{\beta}$ (Park & Casella, 2008). We can see that the dependence structure between the average rates and the feature variables is captured by the values of $\boldsymbol{\beta}$. A zero value of any of the β_j means the extremes frequencies have no dependence on the corresponding feature whereas a non-zero value means otherwise. We can find the parameters $\boldsymbol{\beta}$ by maximizing the log-marginal $p(\{N_i\}|\boldsymbol{\beta}, \tau, \mathbf{X})$ with respect to the parameters. However, directly optimizing the above log-marginal will not be feasible as it involves marginalization of the joint distribution in (3) over the latent variables. We propose to use a variational Bayesian (VB) approximation algorithm (Beal, 2003) to find an approximate posterior distribution over each of the parameters which is described in the following subsection.

2.3. VB Approximation for Bayesian Sparse Regression

We denote the sets $\{\psi_i\}$ and $\{N_i\}$ by vectors $\boldsymbol{\Psi}$ and $\mathbf{N}$ respectively. Let us also denote all the unknown parameters by $\boldsymbol{\Theta}$ and regard them as latent variables as well. A set of latent variables is therefore given by $\mathbf{Z} = \{\Psi, \boldsymbol{\Theta}\}$. Moreover, from now on, we will ignore observed feature variables $\mathbf{X}$ from the list of conditioning variables. From a Bayesian view-point, the joint distribution $p(\mathbf{N}, \mathbf{Z})$ can be factorized in conditionals by assuming priors over each of the parameters in a hierarchical fashion. As mentioned earlier, sparsity over $\boldsymbol{\beta}$ can be implemented by enforcing a Laplace prior on components of $\boldsymbol{\beta}$. We use following hierarchical form of Laplace prior (Kabán, 2007) to make the inference tractable.

$$p\left(\boldsymbol{\beta}; \tau, \gamma\right) = \prod_{j=1}^{p} \frac{\sqrt{\gamma_j \tau}}{2} \exp\left(-\sqrt{\gamma_j \tau}|\beta_j|\right)$$

$$= \prod_{j=1}^{p} \int \mathcal{N}\left(\beta_j; 0, \tau^{-1}\alpha_j^{-1}\right) \text{InvGa}\left(\alpha_j; 1, \frac{\gamma_j}{2}\right) d\alpha_j$$

where τ is the precision parameter of the additive noise in (2), α_j is the latent precision parameter for the coefficient β_j and InvGa(.) is the Inverse gamma distribution. Moreover, Gamma priors are imposed on τ and feature-specific individual parameters γ_j (Park & Casella, 2008; Kabán, 2007). We assume non-informative priors over hyper-parameters a_0, b_0, c_0, d_0 and typically assign a_0=b_0=c_0=d_0=10^{-6}. The final joint distribution $p(\mathbf{N}, \mathbf{Z})$ is provided in the Appendix ((A.1)) We assume a factorized form approximation for the joint distribution and solve for each component using Variational Bayes method details of which is given in the Appendix. The final update equations for variational inference are given here. The moments of the posterior components are given in Appendix.

(1) Distribution of $\boldsymbol{\Psi}$:

$$q_{\boldsymbol{\Psi}}(\boldsymbol{\Psi}) = \prod_{i=1}^{n} q_{\psi}(\psi_i) = \prod_{i=1}^{n} \frac{1}{Z_i} \mathrm{e}^{\left\{\tilde{N}_i \psi_i - \frac{\langle\tau\rangle}{2}\psi_i^2 - \mathrm{e}^{\psi_i}\right\}} \quad (4)$$

$$\textit{with} \qquad \tilde{N}_i = N_i + \langle\tau\rangle X_i^\top \langle\boldsymbol{\beta}\rangle \quad (5)$$

In order to make the inference tractable, we approximated $q_\psi\left(\psi_i\right)$ by a Gaussian using Laplace approximation. The approximation method is briefly described in the Appendix.

(2) Distribution of $\boldsymbol{\beta}$:

$$q_{\boldsymbol{\beta}}(\boldsymbol{\beta}) = \mathcal{N}(\boldsymbol{\beta}; \boldsymbol{\mu}, \boldsymbol{\Sigma}) \quad (6)$$

$$\textit{with} \qquad \boldsymbol{\Sigma} = \left(\langle\tau\rangle\mathbf{X}^\top\mathbf{X} + \langle\tau\rangle\text{diag}(\langle\boldsymbol{\alpha}\rangle)\right)^{-1}; \quad (7)$$

$$\textit{and} \qquad \boldsymbol{\mu} = \boldsymbol{\Sigma}\left(\mathbf{X}^\top\langle\boldsymbol{\Psi}\rangle\right)\langle\tau\rangle \quad (8)$$

Here diag($\langle\boldsymbol{\alpha}\rangle$ corresponds to the LASSO (Tibshirani, 1994) shrinkage.

(3) Distribution of $\boldsymbol{\alpha}$:

$$q_{\boldsymbol{\alpha}}(\boldsymbol{\alpha}) = \prod_{j=1}^{p} q_\alpha(\alpha_j) = \prod_{j=1}^{p} \text{InvGaussian}\left(\alpha_j; g_j, h_j\right) \quad (9)$$

$$\textit{with} \qquad g_j = \sqrt{\frac{\langle\gamma_j\rangle}{\langle\tau\rangle\langle\beta_j^2\rangle}}; \qquad h_j = \langle\gamma_j\rangle \quad (10)$$

(4) Distribution of $\boldsymbol{\gamma}$:

$$q_{\boldsymbol{\gamma}}(\boldsymbol{\gamma}) = \prod_{j=1}^{p} q_\gamma(\gamma_j) = \prod_{j=1}^{p} \text{Ga}\left(\gamma_j; a_j, b_j\right) \quad (11)$$

$$\textit{with} \qquad a_j = a_0 + 1; \qquad b_j = b_0 + \frac{1}{2}\langle\alpha_j^{-1}\rangle \quad (12)$$

(5) Distribution of τ:

$$q_\tau(\tau) = \text{Ga}\left(\tau; c, d\right) \quad (13)$$

$$\textit{with} \qquad c = c_0 + \frac{n+p}{2}; \qquad d = d_0 + \frac{I}{2} + \frac{J}{2} \quad (14)$$

where $I = \sum_{i=1}^{n}\left(\langle\psi_i^2\rangle - 2\langle\psi_i\rangle\mathbf{x}_i^\top\langle\boldsymbol{\beta}\rangle + \langle\boldsymbol{\beta}^\top\mathbf{x}_i\mathbf{x}_i^\top\boldsymbol{\beta}\rangle\right)$; and $J = \sum_{j=1}^{p}\langle\alpha_j\rangle\langle\beta_j^2\rangle$.

The parameters of each of the distributions has dependency on moments of one or more of the other variables. We can only find an optimum solution via iterative updates, starting with random initial values of the relevant moments no closed form solution exists. We can compute the *variational lower bound* J by replacing $q_{\boldsymbol{\Lambda}}(\boldsymbol{\Lambda})$, $q_{\boldsymbol{\beta}}(\boldsymbol{\beta})$, $q_{\boldsymbol{\alpha}}(\boldsymbol{\alpha})$, $q_{\boldsymbol{\gamma}}(\boldsymbol{\gamma})$ and $q_\tau(\tau)$ in (A.3). In each update iteration, we compute the lower bound J

and continue the process as long as the bound keeps increasing. Note that, once the approximate solution is reached, we can compute the marginal distributions over coefficients β_i which is a Gaussian with mean μ_i and variance Σ_{ii}. We can thereby perform a t-test to determine whether the corresponding covariate influences the extremes frequencies.

3. Related Work

We compared the results of our method with the following closely comparable alternatives.

3.0.1. Regular LASSO

We considered regular LASSO (Tibshirani, 1994) to be a baseline since we are interested in obtaining a sparse model. However, as mentioned earlier, LASSO assumes Gaussianity on target variable which does not hold in this case.

3.0.2. Transfer Entropy Method

Transfer entropy is an information-theoretic measure of dependence between two variables which is given by the decrease in entropy in one variable when the other variable is observed. Once the pairwise transfer entropies are computed for each feature with the target variable, a feature selection algorithm such as IAMB (Tsamardinos et al., 2003) can be used to obtain the dependency structure of the target variables. However, the problems with this method are that it does not utilize additional information specific to the distribution of extremes and the accuracy of the method depends on the performance of the feature selection algorithm chosen.

3.0.3. Copula Method

Copula method is used for dependency analysis (Liu et al., 2009; Lafferty & Wasserman, 2012) between random variables having non-Gaussian marginal distributions. In this approach, random variables having non-Gaussian distributions are first transformed into Gaussian copula domain using a transformation $U_i = \Phi^{-1}\left(F_i\left(X_i\right)\right)$ where F_i is the marginal distribution over a non-Gaussian variable X_i. Now a graph learning algorithm such as `glasso` (Friedman et al., 2008) can be used on the transformed variables to find the dependence graph among the variables, from which we can select the edges that indicate dependence relationships between a target variable and the remaining covariates. However the copula-based graph learning methods is also generic and therefore does not use information specific to extremes. Moreover, it is necessary to determine an optimal value of the penalty parameter for the graph learning algorithm which is not required in our Bayesian approach.

4. Experimental Results

In order to evaluate the performance of our algorithm we have applied it on both synthetic datasets and a climate dataset consisting of several observed and derived climatological variables. We evaluated our algorithm in terms of how accurately it recovers the true dependence structure. However, to verify the accuracy of our algorithm we need to have knowledge of the true dependence structure. Therefore we tested our algorithm on a number of different synthetic datasets with varying complexity. On the other hand, results on climate dataset is partially validated with existing knowledge. Additionally, we presented some interesting discoveries to demonstrate the value of using our method on climate data.

4.1. Synthetic Data

We used 8 different $\boldsymbol{\beta}$ values having different sparsity varying from 2 non-zero elements to 30 non-zero elements. The number of non-zero elements were incremented by 4 and noise variance was fixed at 0.1. We generated 500 samples of the features $\mathbf{X}$ (dimensionality p fixed at 40) from normal distributions $\mathcal{N}(0,1)$ and generated λ_i using equation (2), i.e. by exponentiating the linear combination of features $\mathbf{X}$ in the right hand side. We sampled the frequencies N_i from Poisson distributions having average parameters λ_i. In our final dataset, we kept only those data-points for which N_i is less than 20 and larger than 0 to simulate natural scenarios where extremes are rare. We sampled the features multiple times and reported the average F_1-score for each experiment. F_1 score measures how accurately our method recovers the zero/non-zero structure of the coefficients. It is given by $F_1 = \frac{2PR}{P+R}$ where P is the precision and R is the recall.

As mentioned before, we decided whether each β_i is zero or not based on a t-test that rejects the null hypothesis ($\beta_i \neq 0$) at 1% confidence level. Average and standard deviations of F_1-scores from 10 repetitions of the above experiment are shown in Figure 1. We can see that our method is far superior in terms of accurately finding the dependence structure. When we decrease sparsity, the F_1-score of both VB approach and LASSO increases monotonically with almost negligible variance. However, the F_1-score of the Gaussian copula based method decreases consistently with decreasing sparsity. Finally, the transfer entropy method does not show any particular trend, although it's ac-

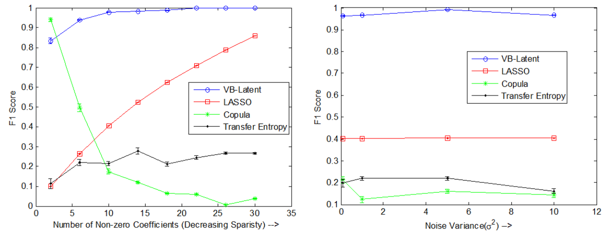

Figure 1. F_1-score achieved by different methods with decreasing sparsity (noise variance fixed at 0.1)

curacy is very low.

We presented a few sample plots from individual experiments in Figure 2 to show the true and estimated coefficients for different methods. We can see that our method (Figure 2a) very accurately captured the underlying dependence structure between the frequency and the feature variables. However, LASSO correctly identifies the non-zero coefficients (Figure 2b), but not the zero coefficients. Although, many estimated values of zero coefficients are close to zero, unlike VB method, we do not have any additional information to infer whether the coefficients are actually zero or not. Therefore, decreasing sparsity or increasing number of non-zero coefficients results in increasing accuracy for LASSO. On the other hand, Copula-based methods always ends up with a very sparse solution (Figure 2c), with most of the coefficients estimated to be zero. Therefore it's accuracy is high when the true solution is very sparse, but the accuracy decreases when sparsity decreases.

4.2. Climate Application

Precipitation extremes are important owing to their consequences for flood hazards and water resources. Unfortunately, the processes leading to regional precipitation extremes are still poorly understood, but are known to have mechanistic dependence both on regional covariates as well as large-scale climate oscillators (Aryal et al., 2009; Gregersen et al., in press). An improved understanding of the dependence relationship between extremes and their drivers may potentially help in better projection of extremes. We applied our algorithm to estimate the dependence of yearly occurrences of precipitation extremes on a number of potential local, regional and large-scale climate variables for one climatologically homogeneous region in US. Note that our method does not automatically model the spatial and temporal dependence among frequency of occurrence of extremes. However a reasonable assumption is that by carefully choosing the potential feature variables, we can account for spatial and temporal correlations among extremes frequencies (Aryal et al., 2009) if there is any.

4.2.1. Dataset

We chose only one climatologically homogeneous region out of 9 for our analysis - north-east US (A map showing the regions can be found at www.ncdc.noaa.gov/img/climate/research/1998/ann/usrgns_pg.gif). We obtained the daily precipitation, minimum and maximum temperature station data (not gridded) from the repository of US Historical Climatology Network (USHCN) (Easterling et al., 1996) for states falling under the north-east region. As mentioned earlier, our target variable is yearly frequency of precipitation extremes over stations within a region. Extremes were defined as events exceeding a threshold set at 99 percentile of all precipitation events at a particular location and we counted the number of such events in each year.

We used a number of covariates for all regions and let our algorithm find the covariates that influence the precipitation extremes frequency. The covariates we used falls in one of three broad categories - local, regional or climate indices (global) (Gregersen et al., in press). Local covariates originate from each station and exhibits both spatial and temporal variability. Regional variables are spatially averaged local variables over the entire region under consideration that exhibits regional dynamics. Climate indices are global variables that represent large-scale signals in climate variables. Usually they are principal components (although they are called Empirical Orthogonal Function (EOF) in climate literature) of space-time anomaly matrix of global climate variables like Sea Surface Temperature (SST), Sea Level Pressure (SLP) etc. Anomalies in climate variables are computed by subtracting the seasonal mean from the daily values. A list of covariates used of each category is given in Table 1 in the Appendix. While all of the local and regional features were derived from the station data obtained from USHCN, the climate indices were obtained from NOAA's Climate Prediction Center data repository (http://www.esrl.noaa.gov/psd/data/climateindices/list/). We could use the covariates between 1950 to 2011 as most of the climate indices are available only for that period. Also, if more than 50% of the daily observations out of a year are found to be missing for any of variables at a specific location, we simply discarded all variables for that year and for that specific location. Global climate indices are available monthly. We averaged monthly indices over a year if at least 2 monthly values are available. Otherwise, we discarded the corresponding year. Finally, there were

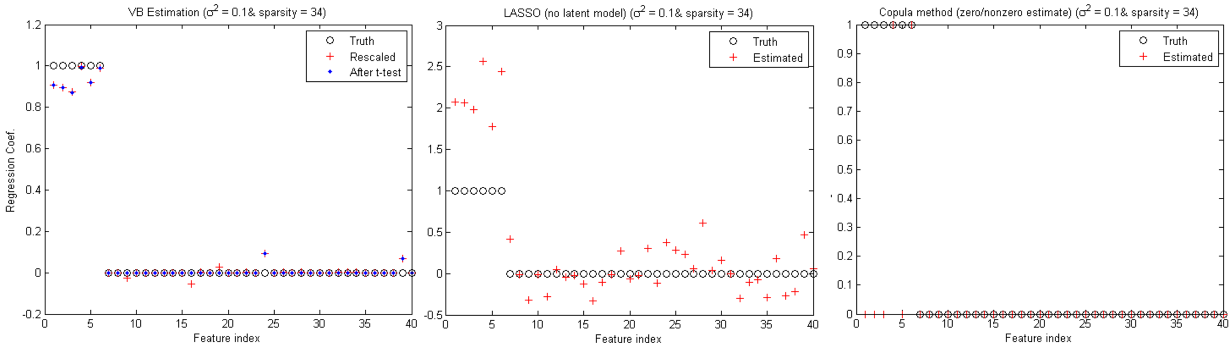

Figure 2. Sample results with 6 non-zero coefficients in β and $\sigma^2 = 0.1$ for a) VB method b) LASSO and c) Copula method.

5888 data-points from 158 stations located in notheast region. We used 33 potential covariates which are listed in Appendix (See Table 1).

4.2.2. Discussion of Results

We can see that for both North-east (Figure 3) region of continental US, frequency of precipitation extremes has statistically significant dependence on Mean Annual Precipitation (`MAP`) and Summer Mean Precipitation (`SMP`). This results may be interpreted in a number of ways. First of all, `MAP` explains the spatial variability of the extreme frequencies and one way to interpret the results is that the spatial variability is the strongest signal in the frequencies and once that is explained, there is only little information within the remaining covariates. A second and more interesting interpretation is that variability of frequencies of rainfall extremes depend both on the regional summer maximum temperature (`SMTRegmax`) anomaly and global temperature anomaly (`GMT`). So, we can expect to see more frequent extreme rainfall events as a result of regional warming and global warming, at least for the region under consideration. Also note that, in NE region, there is some dependence on a few climate indices (`Nino1+2`,`Nino3`,`TNA` etc.). However, an important next step is the translation of the statistical dependencies discovered using our method to causal insights, which needs rigorous physical interpretation.

5. Conclusion and Future Work

In this paper, we presented a statistically rigorous framework for finding covariates that influences the extremes frequencies. We accommodated the Poisson process model for describing extremes occurrences within the Bayesian hierarchical description of sparse regression and came up with an efficient Variational Bayes inference technique to estimate the probability distribution over the regression parameters. Our method shows a high level on accuracy on synthetic dataset and outperforms competing methods by a large margin. We have also shown the value of using our method to find the dependence of precipitation extremes frequency on covariates of different spatial scale

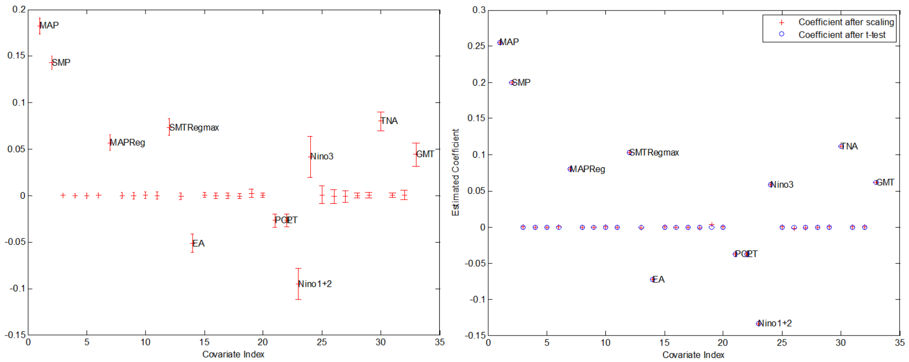

Figure 3. a) Mean and standard deviation of posterior (Gaussian) distributions of coefficients corresponding to each of the potential covariates for north-east region (only the potentially significant covariates are labeled) and b) Statistically significant covariates after t-test for the same region

notwithstanding the fact that these findings are only recommendations awaiting physical interpretation to be qualified as novel insights.

The dependence analysis performed in this paper is mainly exploratory in nature. We plan to extend the model to perform predictive analysis in future. We also plan to develop similar sparse model for dependence analysis of extremes intensity which is significantly more challenging since it involves distributions outside of exponential family. Finally, we plan to incorporate spatio-temporal autoregressive model within our probabilistic framework to leverage the spatio-temporal correlation inherent in climate and other spatio-temporal datasets.

References

Aryal, S. K. et al. Characterizing and Modeling Temporal and Spatial Trends in Rainfall Extremes. *Journal of Hydrometeorology*, 10:241253, February 2009. ISSN 1525-7541.

Beal, Matthew J. *Variational Algorithms for Approximate Bayesian Inference*. PhD thesis, Gatsby Computational Neuroscience Unit, University College London, 2003.

Coles, Stuart. *An Introduction to Statistical Modeling of Extreme Values*. Springer-Verlag New York, Inc., London, UK, 2001. ISBN 1852334592.

Easterling, D. R. et al. United States Historical Climatology Network (USHCN) Monthly Temperature and Precipitation Data. Technical report, Oak Ridge National Laboratory, U.S. Department of Energy, Oak Ridge, Tennessee, 1996. URL http://cdiac.ornl.gov/epubs/ndp/ushcn/ushcn.html.

Field, C.B. et al. Summary for policymakers. In *Managing the Risks of Extreme Events and Disasters to Advance Climate Change Adaptation. Special Report of the Intergovernmental Panel on Climate Change*, pp. 1–19. Cambridge University Press, 2012.

Friedman, Jerome, Hastie, Trevor, and Tibshirani, Robert. Sparse inverse covariance estimation with the graphical lasso. *Biostatistics (Oxford, England)*, 9(3):432–41, July 2008. ISSN 1468-4357.

Friedman, Jerome, Hastie, Trevor, and Tibshirani, Robert. Regularization Paths for Generalized Linear Models via Coordinate Descent. *Journal of statistical software*, 33(1):1–22, January 2010. ISSN 1548-7660.

Gregersen, I. D. et al. A spatial and nonstationary model for the frequency of extreme rainfall events. *Water Resources Research*, 49:127–136, January in press. doi: {10.1029/2012WR012570}.

Kabán, Ata. On bayesian classification with laplace priors. *Pattern Recogn. Lett.*, 28(10):1271–1282, 2007. ISSN 0167-8655.

Lafferty, John and Wasserman, Larry. The Nonparanormal SKEPTIC. In *ICML*, 2012.

Liu, Han, Lafferty, John, and Wasserman, Larry. The nonparanormal: Semiparametric estimation of high dimensional undirected graphs. *J. Mach. Learn. Res.*, 10:2295–2328, December 2009. ISSN 1532-4435.

Liu, Yan, Bahadori, Mohammad Taha, and Li, Hongfei. Sparse-GEV: Sparse latent space model for multivariate extreme value time series modeling. In *ICML*, 2012.

Lozano, Aurelie C. et al. Spatial-temporal causal modeling for climate change attribution. In *KDD*, pp. 587. ACM Press, 2009.

Park, Trevor and Casella, George. The Bayesian LASSO. *Journal of the American Statistical Association*, 103(482):681–686, 2008.

Reiss, Rolf-Dieter and Thomas, Michael. *Statistical Analysis of Extreme Values: with Applications to Insurance, Finance, Hydrology and Other Fields.* Springer, London, UK, 3rd edition, 2007. ISBN 3764372303.

Tibshirani, Robert. Regression shrinkage and selection via the LASSO. *Journal of the Royal Statistical Society, Series B*, 58:267–288, 1994.

Tsamardinos, I., F., Aliferis C., and A., Statnikov. Algorithms for Large Scale Markov Blanket Discovery. In *Proceedings of the Sixteenth International Florida Artificial Intelligence Research Society Conference*, pp. 376–381. AAAI Press, 2003.

Appendix

A. VB Approximation for Bayesian Sparse Regression

The joint distribution $P(N, Z)$ is given by

$$p\left(\mathbf{N}, \mathbf{\Psi}, \boldsymbol{\beta}, \tau, \boldsymbol{\alpha}, \boldsymbol{\gamma}\right) = \prod_{i=1}^{n} \text{Pois}\left(N_i; e^{\psi_i}\right) \mathcal{N}\left(\psi_i; \boldsymbol{\beta}^T \mathbf{x}_i, \tau^{-1}\right)$$
$$\times \text{Ga}\left(\tau; c_0, d_0\right) \times \prod_{j=1}^{p} \Big\{ \mathcal{N}\left(\beta_j; , 0, \tau^{-1}\alpha_j^{-1}\right)$$
$$\times \text{InvGa}\left(\alpha_j; 1, \frac{\gamma_j}{2}\right) \text{Ga}\left(\gamma_j; a_0, b_0\right) \Big\} \tag{A.1}$$

The assumed factorized form of the posterior is given by:

$$q(\mathbf{\Lambda}, \boldsymbol{\beta}, \tau, \boldsymbol{\alpha}, \boldsymbol{\gamma}) = q_{\mathbf{\Lambda}}(\mathbf{\Lambda}) q_{\boldsymbol{\beta}}(\boldsymbol{\beta}) q_{\tau}(\tau) q_{\boldsymbol{\alpha}}(\boldsymbol{\alpha}) q_{\boldsymbol{\gamma}}(\boldsymbol{\gamma}) \tag{A.2}$$

Using Jensen's inequality, it can be shown that the log-marginal log$p(\mathbf{N})$ has a lower bound which is given by

$$J = \int q(\mathbf{\Lambda}, \boldsymbol{\beta}, \tau, \boldsymbol{\alpha}, \boldsymbol{\gamma}) \text{log} \frac{p\left(\mathbf{N}, \mathbf{\Lambda}, \boldsymbol{\beta}, \tau, \boldsymbol{\alpha}, \boldsymbol{\gamma}\right)}{q(\mathbf{\Lambda}, \boldsymbol{\beta}, \tau, \boldsymbol{\alpha}, \boldsymbol{\gamma})} \text{d}\mathbf{\Lambda} \text{d}\boldsymbol{\beta} \text{d}\boldsymbol{\alpha} \text{d}\boldsymbol{\gamma} \text{d}\tau \tag{A.3}$$

Solving for each of the components in (A.2) by maximizing the lower bound J, we obtain the update equations provided in section 2

A.1. Moments of the components in the posterior

(1) Distribution of $\mathbf{\Psi}$: We approximate $q_\psi(\psi_i)$ by a Gaussian whose mean is same as the mode of $q_\psi(\psi_i)$ (i.e. the point $\hat{\psi}_i$ that maximizes $q_\psi(\psi_i)$), or equivalently

$$\left.\frac{df(\psi_i)}{d\psi_i}\right|_{\psi_i=\hat{\psi}_i} = 0$$

where $f(\psi_i) = \text{e}^{\left\{\tilde{N}_i\psi_i - \frac{\langle\tau\rangle}{2}\psi_i^2 - \text{e}^{\psi_i}\right\}}$. This gives $\hat{\psi}_i = \tilde{N}_i/\langle\tau\rangle - \mathcal{W}(\exp(\tilde{N}_i/\langle\tau\rangle))/\langle\tau\rangle$. Now, making a second order Taylor expansion of log$f(\psi_i)$ around $\hat{\psi}_i$ and then exponentiating both sides, we get

$$f(\psi_i) \simeq f(\hat{\psi}_i) \text{exp} \left\{ -\frac{A}{2} \left(\psi_i - \hat{\psi}_i\right)^2 \right\}$$

where $A = -\frac{d^2}{d\psi_i^2} \text{ln} f(\psi_i)\Big|_{\psi_i=\hat{\psi}_i} = \left(\left|\langle\tau\rangle + e_i^{\psi}\right|_{\psi_i=\hat{\psi}_i}\right)^{-1}$. So, the Gaussian approximation of $q_\psi(\psi_i)$ in (4) is given by $q_\psi(\psi_i) \approx \mathcal{N}(\psi_i; m_i, \sigma_i^2)$ where $m_i = \tilde{N}_i/\langle\tau\rangle - \mathcal{W}(\exp(\tilde{N}_i/\langle\tau\rangle))/\langle\tau\rangle$ and $\sigma_i = (|\langle\tau\rangle + e^z|_{z=m_i})^{-1}$ where $\mathcal{W}(.)$ is Lambert's W-function.

(2) Distribution of $\boldsymbol{\beta}$: The moments are given by $\langle\boldsymbol{\beta}\rangle = \boldsymbol{\mu}$; and $\langle\beta_j^2\rangle = \Sigma_{jj} + \mu_j^2$.

(3) Distribution of $\boldsymbol{\alpha}$: InvGaussian $(\alpha_j; g_j, h_j)$ denotes inverse Gaussian distribution with mean g_j and shape parameter h_j having the following density function.

$$\text{InvGaussian}\left(\alpha_j; g_j, h_j\right) = \sqrt{\frac{h_j}{2\pi\alpha_j^3}} \text{exp}\left(-\frac{h_j\left(\alpha_j - g_j\right)^2}{2g_j^2\alpha_j}\right)$$

with $\alpha_j > 0$ The relevant moments are given by $\langle\alpha_j\rangle = g_j$ and $\langle\alpha_j^{-1}\rangle = g_j^{-1} + h_j^{-1}$.

(4) Distribution of $\boldsymbol{\gamma}$: Relevant moment $\langle\gamma_j\rangle = a_j/b_j$.

(5) Distribution of τ: The moment is given by $\langle\tau\rangle = c/d$.

B. List of Potential Covariates for Climate Application

Table 1. Potential covariates used for dependency analysis of frequecy of rainfall extremes

Local: Mean Annual Minimum Temperature (MATmin), Mean Annual Maximum Temperature (MATmax), Mean Summer Minimum Temperature (SMTmin), Mean Summer Maximum Temperature (SMTmax), Mean Annual Precipitation (MAP),Mean Summer Precipitation (SMP)

Regional: Regional Mean Annual Minimum Temperature (MATRegmin), Regional Mean Annual Maximum Temperature (MATRegmax), Regional Mean Summer Minimum Temperature (SMTRegmin), Regional Mean Summer Maximum Temperature (SMTRegmax), Regional Mean Annual Precipitation (MAPReg), Regional Mean Summer Precipitation (SMPReg)

Climate Indices: North Atlantic Oscillation (NAO), East Atlantic Pattern (EA), West Pacific Pattern (WP), East Pacific/North Pacific Pattern (EPNP), Pacific/North American Pattern (PNA), East Atlantic/West Russia Pattern (EAWR), Scandinavia Pattern (SCA), Tropical/Northern Hemisphere Pattern (TNH), Polar/Eurasia Pattern (POL), Pacific Transition Pattern (PT), Nino 1+2, Nino 3, Nino 3.4, Nino 4, Southern Oscillation Index (SOI), Pacific Decadal Oscillation (PDO), Northern Pacific Oscillation (NP), Tropical/Northern Atlantic Index (TNA), Tropical/Southern Atlantic Index (TSA), Western Hemisphere Warm Pool (WHWP), Global Mean Temperature Anomaly (GlobalMeanTemp)