Mining Extreme Values: Climate and Natural Hazards

Debasish Das[1,2], Evan Kodra[1], Auroop R. Ganguly[1], Zoran Obradovic[2]
[1]Department of Civil and Environmental Engineering, Northeastern University, Boston, MA, USA
[2]Center for Data Analytics and Biomedical Informatics, Temple University, Philadelphia, PA, USA
d.das@neu.edu, a.ganguly@neu.edu, kodra.e@husky.neu.edu, zoran.obradovic@temple.edu.

ABSTRACT
The analysis and modeling of extreme values have traditionally relied on extreme value theory (EVT), which in turn has tended to focus on limiting or asymptotic cases and assumptions of independence. However, disciplines from climate science and digital mapping to infrastructure security and transportation, have been generating massive volumes of data with multidimensional and multivariable dependence, long-memory and long-range associations, and nonlinear interactions, from remote or in-situ sensors and computational models. This motivates the need for automated descriptive and predictive analysis of extremes. The size and complexity of data does not preclude the rarity of the extreme events, but presents the possibility of information extraction from related ancillary variables, or covariates. An empirical analysis of spatiotemporal auto- and cross-correlation structures of extremes, statistical behavior of EVT on finite and noisy data, as well as the impact of noise, nonlinearity and variability, may lead to novel formulations for understanding extremes and their correlations. Based on the results of preliminary data analysis, we proposed using graphical model based on tail dependence for description and analysis of spatial and covariate dependence structure among extremes time-series. Besides providing insights on climate change or natural hazards and the consequences for climate change science or the re-insurance industry, the methods can be generalized to multiple domains ranging from water resources planning and critical infrastructures security to finance, telecommunications, cyber-security and mapping technologies.

Keywords
Precipitation, Climate Change, Extremes Regression, Elastic Net, Sparse Modeling.

1. INTRODUCTION
While most traditional developments in statistics or machine learning have tended to focus on an understanding of usual patterns and frequent occurrences, the ability to understand and predict rare events remains a challenge. Nonetheless, extreme events are growing in importance across disciplines like finance, insurance, hydrology [1] and climate [2-3]. Here we consider the climate change context, where changes in the intensity, frequency or duration of temperature and precipitation extremes are of interest for adaptation and policy. Specifically, attribution to human contributions, credible future projections, emergency preparedness or resource allocations, and insurance or re-insurance risk assessments, may all depend on the ability to understand relations among extreme values and generate predictive insights. Here we are concerned with rare events at the tails of the distributions, or extremely high or low values. Rare events mining in artificial intelligence (AI), which includes classification of imbalanced datasets through synthetic over-sampling [4], deal with situations when rare events do occur in the observations. Methods like skyline [19] or top-K query [20] processing database mining techniques may be useful in sampling large values from available data. However, none of these approaches deal with situations where the rare events may not occur in the observed or model-simulated "training" data, or may not occur in sufficient volumes for direct statistical analysis. The analysis of such events relies on extrapolation beyond what is normally observed. Extreme value theory (EVT) is among the few statistical methods doing true extrapolation; parametric relations are developed to infer about tails of the distribution (e.g., a 100-year, or a one in a thousand, event) with values that are adequately large but not necessarily at the extreme tails [5; 21]. A recent work on EVT [23] is an example of EVT method developed in the machine learning (ML) community with applications to data from climate and the social media.

Permission to make digital or hard copies of portions of this work for personal or classroom use is granted without fee provided that the copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.

ACM SustKDD'12, August 12, 2012, Beijing, China Copyright 2012 ACM 978-1-4503-1558-6/12/08 ....$10.00

2. CHALLENGES & OPPORTUNITIES
Despite decades of development, EVT remains an area with open challenges, many of which may be resolved through statistics, data mining and AI. The growing importance of extremes, for example in the context of climate change and severe rainfall, motivates urgent solutions. Although there are many open challenges [6, 21] related to extreme value analysis, in this paper we have mainly focused on the problem of estimation of spatial dependence structure among extremes and the importance of uncertainty quantification. Being able to accurately estimate this dependence structure will greatly improve the effectiveness of decision making process in multiple sectors including insurance industry and water resource management. For example if two locations are found to be perfectly dependent in terms of extremes events, simultaneous 100-year precipitation events at those locations becomes a 100-year event itself; whereas if they are found to be independent, the same event will be a 10,000 year event. Insurance companies can greatly improve their risk portfolio with this kind of information and set premiums accordingly. On the other hand, a major change in the dependence structure may suggest large-scale change in climate patterns due to urbanization and other anthropogenic activities. However, we have to be careful while interpreting these estimates since they involve extrapolation beyond normally observed events, which emphasizes the importance of supplying appropriate uncertainty estimates.

We have investigated the role of multi-source data and other

covariates in providing additional information content about the dependence structure and reducing the associated uncertainty. We have described the problem with a few motivating experiments, discussed where data guided techniques may help and proposed a potential solution approach, with particular emphasis on uncertainty quantification. Climate change is selected as an exemplar both because of the societal importance [7] and to validate the methods with massive data from sensors and models.

3. BACKGROUND

Rainfall extremes are typically characterized by their intensity, duration and frequency (IDF) for applications from water resources management, flood hazards, and dam design [8]. Recent research has explored changes in the IDF curves under climate change [9].

The *n-year return level,* (RL_n), defined as the level that is reached or exceeded once every *n-years* on the average (alternatively, the probability of exceedances on any given year is *1/n*). The three [5, 8] ways to describe extreme values are the Generalized Extreme Value (GEV) distribution fitted to block maxima (BM) or blocks of time windows like an annual maxima time series, the Poisson arrival of extremes followed by the Generalized Pareto distribution (GPD) fitted to the excesses above a threshold, leading to the Peak-over-Threshold (PoT) as well as the Point Process (PP) approach. The PDF of GEV [5] is given by

$$f(x|\mu,\sigma,\xi) = \frac{1}{\sigma}\left[1+\xi\left(\frac{x-\mu}{\sigma}\right)\right]^{\left(-\frac{1}{\xi}\right)-1} . exp\left\{-\left[1+\xi\left(\frac{x-\mu}{\sigma}\right)\right]^{\left(-\frac{1}{\xi}\right)}\right\} \quad (1)$$

where parameters μ, σ and ξ are called location, scale and shape parameters respectively and *exp*(.) is exponential function. The PDF of GPD [5] is given by

$$f(x;\mu,\sigma,\xi) = \frac{1}{\sigma}\left[1+\xi\left(\frac{x-\mu}{\sigma}\right)\right]^{\left(-\frac{1}{\xi}\right)-1} \quad (2)$$

where parameters μ, σ and ξ are called location, scale and threshold parameters respectively.

From a pragmatic standpoint, the approaches generate estimates of the return levels along with associated uncertainties per time series, but require either the selection of a block size or a threshold. The distributions (GEV or GPD) arise from limiting cases for large sample sizes as well as when the maxima or excess data are independent and identically distributed. Thus, the tradeoffs during the choice of a block size or a threshold may be expressed as a bias versus variance issue: larger block sizes or higher thresholds may imply lower bias but larger variance while smaller block sizes and lower thresholds may imply larger bias. One may refer to [5] for more discussion on the typical choice of methods for practical applications.

Linear Correlation measures (e.g. Pearson's correlation) are useful in estimating the pair-wise correlation between multiple time-series. While computing pair-wise correlation among precipitation time-series at different location may give us some preliminary idea about average spatial decorrelation length-scale among these time-series and dependence structure between mean precipitation events, this is certainly not suitable for extreme events. Firstly, using a method like block-maxima will greatly reduce the number of samples on which the correlation is being computed and introduce large uncertainty. Secondly, the linear correlation measures fail to consider the co-occurrence patterns of the extreme events which are of prime importance for estimating dependence structure among extremes. So, in order to estimate a spatial dependence structure between extremes occurrence, we may need to use more informative measures of dependence like tail dependence, which is loosely given by limiting proportion that one variable exceeds a certain threshold given that the other variable has already exceeded that threshold [22]. Here the threshold may be defined as a percentile. Copula-based methods may also be appropriate in estimating the dependence among extremes time series.

4. PRELIMINARY RESULTS

4.1 Dataset

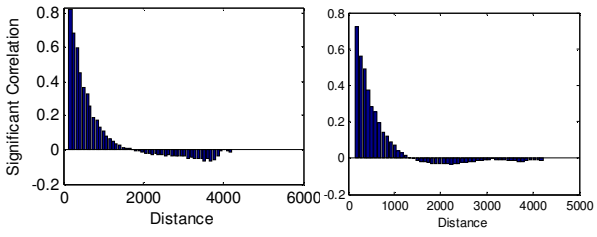
In our experiments we used the precipitation observations from Climate Prediction Center (CPC) which are available at 0.25° x 0.25° grids [22] over entire US. A second type of dataset that we used here falls in the category of "reanalysis datasets" that are generated from physics models that are forced to match with available observations. So, they are mostly uniform in terms of quality. However, they can inherit errors associated with the observations. A summary of the datasets we used are provided in Table 1.

Table 1: Description of the datasets used in the experiments

	Temporal Resolution	Spatial Resolution	Region Used	Variables Used
Observa-tion	Daily	.25° x .25°	US	Precipitation
Reanalysis	Daily	2.5° x 2.5°	US	Precipitation, Temperature

4.2 Experiments

We started by plotting the distribution of the correlation as a function of distance for both the observed and reanalysis precipitation data in Figure 1. We binned all possible distances among grid-points into 50 intervals and for each interval we computed the pairwise correlation between location pairs whose distance falls within that interval and averaged them. However, we only considered the correlations that are significant at 95% significance level.



(a) (b)

Figure 1: Distribution of correlation as a function of distance for (a) Reanalysis and (b) Observation.

Next in Figure 2, we plotted the same distribution for annual maxima instead of the daily time-series. We can see that

correlation decreases significantly for annual maxima time-series. Furthermore, the difference between two types of dataset is more evident among extremes than in regular time-series.

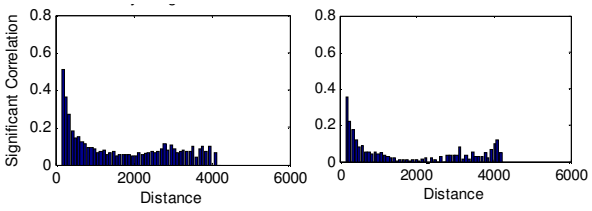


Figure 2: Distribution of correlation among annual maxima as a function of distance for (a) Reanalysis and (b) Observation.

In Figure 3, we plotted the significance of the null hypothesis (that no correlation exists) as a function of distance in a similar way we plotted the correlations. A lower value of significance of null hypothesis means higher significance and lower uncertainty of the correlation. Most of the pairwise correlations computed on annual maxima time-series turned out to be highly uncertain and insignificant. So, the correlation not only became smaller when we used annual maxima instead of the regular time-series, they became insignificant. Moreover, the correlations seem to increase slightly at longer distances when we used maxima which may just be spurious. If not, they may need further analysis. So, simple linear correlation may not be the correct measure to estimate the dependence between extreme events. This shows the need for using other methods such as tail dependence or copula-based methods to quantify the dependence among extremes time-series instead of linear correlation measure.

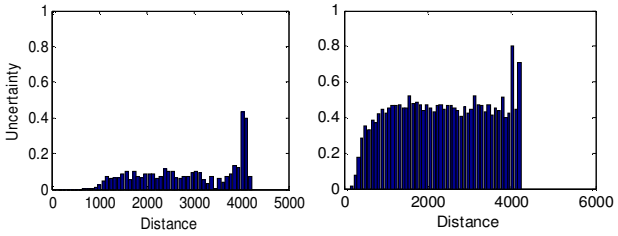
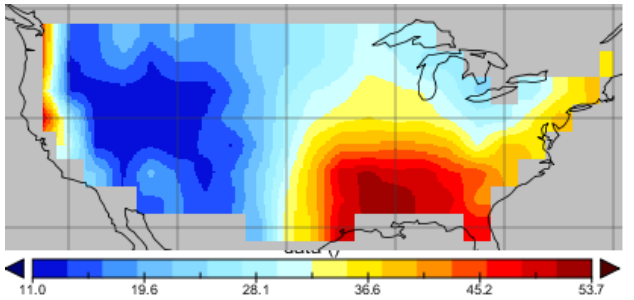


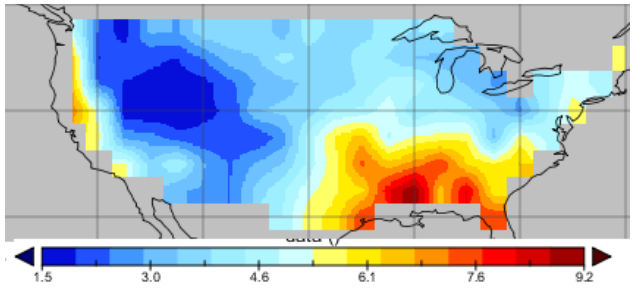
Figure 3: Distribution of significance of null hypothesis (that no correlation exist) as a function of distance for (a) Reanalysis and (b) Observation. (lower value means lower uncertainty)

Now, we fitted GEV distribution (and obtained the maximum likelihood estimates of GEV parameters) at each individual grid-point using the annual maxima and plotted the spatial distribution of the location parameter μ along with corresponding confidence interval in Figure 4. The distribution of parameter clearly shows spatial coherence which we intend to exploit.

In our next experiment, we computed the average temperature over each location and plotted 30 year return level over each location as a function of mean temperature. Figure 5 shows the plot. The plot shows some correlation among average temperature and 30 year return levels. It is a well-known hypothesis in climate that, at global level, precipitation extremes increase with increase in temperature [18]. However, at regional level, the relation is not so clear. Furthermore there may be non-linear dependence and other covariates like relative humidity, precipitable water, updraft velocity etc. may also influence the extremes.



(a)



(b)

Figure 4: Spatial distribution of (a) location parameter and (b) corresponding confidence interval for GEVs fitted at each grid-point

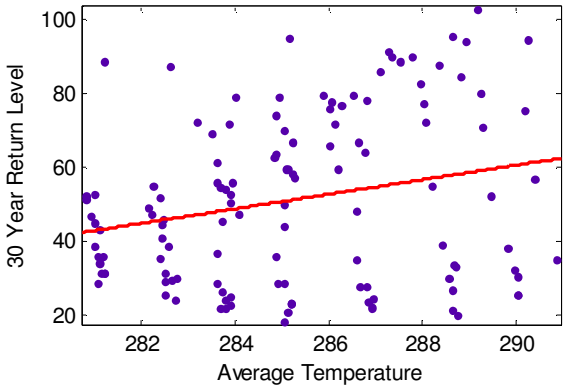


Figure 5: Plot of 30 year return level vs Average temperature at 124 different grid-points within US.

In this section we showed some results from preliminary experiments to give an outline of the key challenges in extremes mining using precipitation extremes as a motivating example. However, more empirical tests are necessary for additional hypothesis before we can effectively design our solutions for addressing these challenges. In the next section we provide some possible approaches for solution to the problem of dependence discovery.

5. FUTURE RESEARCH DIRECTIONS

We propose the development of graphical models for extremes based on tail dependence measures derived from multivariate extensions of EVT. The attributes of the graph are expected to offer descriptive insights about extremes and their space-time correlations. In addition, we propose a graphical model of the

covariate structures based on linear or nonlinear correlations for possible predictive insights on extremes. Based on space-time extensions of time series mining and nonlinear dynamical concepts like motifs and discords [25], we propose to develop methods for predictive insights of extremes and their correlations in space and time.

Estimation of sparse dependence structure between several variables is a well-known problem in machine learning with application across a number of fields. One method which gained popularity recently is called "graphical lasso" which estimates a sparse graphical structure among a large number of variables by inverting the covariance matrix of the variables computed from data under an L1 penalization over components of the inverse covariance matrix in order to force most of them to be zero. Since the elements of the inverse covariance matrix are used to estimate the corresponding elements in the adjacency matrix of the dependency graph; the resulting dependency graph tends to be sparse [24]. The optimization problem for maximizing the log-likelihood of inverse covariance matrix Θ to estimate the graph is given by the following equation:

$$\underset{\Theta}{\operatorname{argmax}}\{\log(\det\Theta) - \operatorname{tr}(S.\Theta) - \rho.||\Theta||_1\}$$

where S is the empirical covariance matrix computed from the data, det(.) and tr(.) are respectively the determinant and the trace of a matrix, $\|.\|_1$ is the L1-norm of the vectorized matrix and ρ is the regularization parameter. However one important constraint that needs to be satisfied for this method to consistently estimate the actual graph is that the variables should be normally distributed [26], which is not the case for extremes. So, one main challenge for adopting this method to estimate the extremes dependence graph is to find a set of constraints under which the tail dependence matrix can be used to estimate the graph instead of the covariance matrix. Additionally, for reasons explained earlier, we need to provide appropriate measure of uncertainty associated with each estimated edge within this graph.

A second challenge involves using the information content from the covariates along with the dependence graph to develop a prediction model for the precipitation extremes. We may have to look for informative patterns in the time-series for covariates that precedes the occurrence of any precipitation extreme after assigning appropriate weights on covariate values that are temporally closer to the time of extreme occurrence. This method can be regarded as supervised "motif discovery" [25] within the covariate time-series where the assumption is that one or more unobserved state variables are causing an extreme to occur and the state variables themselves change their state based on appearance of certain sequence of values within the covariate time-series.

A different approach for utilizing the information content in the covariates may require us to consider this problem as a supervised regression where the precipitation extremes are considered as targets while the covariates at different temporal lag from the time of occurrence of extremes are regarded as features. The process of estimation of information content from covariates can be performed separately from the dependence structure estimation among extremes. However, a better approach may be to merge these methods together since both the process can inform each other. For predictive analysis of precipitation extremes at certain location, time-series from locations that are neighbors in the dependence graph may need to be considered along with the covariate time-series at the location of interest.

A third challenge is to utilize the observation and simulation data for precipitation available from multiple sources in order to improve our estimate of the dependence structure. From the experimental results it can be seen that even though the precipitation time-series from different sources does not differ much in statistical sense, their corresponding extremes time-series might still be different and therefore it is a challenging task to extract and integrate the non-overlapping information available from these time-series and using the overlapping information to reduce the uncertainty in our estimated dependence structure. We may assume the actual underlying distribution of the extremes to be hidden and try to estimate it from data available from multiple sources. In Figure 6, we attempted to cast a simplified version of all three aspects of the problem in a single graphical model. Here g1 to g4 are actual precipitation extremes at different grid locations (there can be more), s1 and s2 are precipitation extremes extracted from observed (or simulated) data by different sources (there can be more) and c1 and c2 are covariates that are known to carry information about the precipitation extremes (there can be more). Shaded circles are observed whereas transparent circles are unobserved variables. Broken line means that the edges need to be estimated. However, this representation is only one of the possible solutions to the problem and other approaches are also possible. Moreover, we need further hypothesis building and data analytics before we can start designing the final solution.

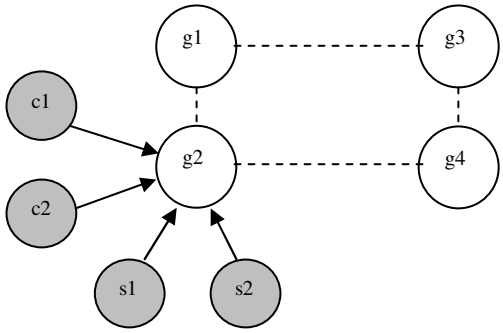


Figure 6: Graphical model showing a simpler version of the dependence structure estimation problem

6. CONCLUSION

In this paper, we described the importance and challenges associated with estimation of spatial dependence structure of precipitation extremes along with appropriate uncertainty estimates. We introduced the challenge of using the information content in covariates and using this information in tandem with the spatial dependence structure for predictive analysis of precipitation extremes. We further described the challenges associated with integration of extremes information from multiple data sources to reduce uncertainty associated with the spatial dependence structure among extremes. Preliminary results show that extremes are significantly different from regular time-series and familiar statistical and/or machine learning tools may not be adequate for analyzing extremes. However, results are still preliminary and more focused investigation is needed before we can make any strong hypothesis and start building solutions. Although we used precipitation extremes as an exemplar, these challenges may generalize across multiple domains where

extremes are important and the solution frameworks are expected to generalize across multiple domains as well.

Acknowledgements

This research is partially supported by NSF grants IIS-1029711, as well as a grant to ARG from the Nuclear Regulatory Commission and ARG's start-up funds from Northeastern University. Please forward any questions related to this publication to *a.ganguly@neu.edu.*

7. REFERENCES

[1] Reiss, R-D., Thomas M: Statistical analysis of extreme values: with applications to insurance, finance, hydrology and other fields, 3rd edition, 2007, Springer, 511 pp.

[2] Min, S.-K., Zhang, X., Zwiers, F.W., Hegerl, G.C., Human contribution to more-intense precipitation extremes. Nature, 470, 2011, 378-381.

[3] Lozano, A.C., Li, H., Niculescu-Mizil, A., Liu, Y., Perlich, C., Hosking, J., Abe, N., Spatio-temporal causal modeling for climate change attribution, Proc. 15th ACM SIGKDD, KDD 2009, 587-596.

[4] Chawla, N.V., Boyer, K.W., Hall, L.O., Kegelmeyer, W.P., SMOTE: Synthetic minority over-sampling technique, Journal of Artificial Intelligence Research, 16, 2002.

[5] Coles, S. G., An introduction to statistical modeling of extreme values, 2001, Springer-Verlag, 208 pp.

[6] Fuentes, M., Reich, B., and Lee, G., Spatial-temporal mesoscale modelling of rainfall intensity using gage and radar data, Annals of Applied Statistics, 2, 2012, 1148–1169.

[7] Field, C.B., Et Al., IPCC, 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, 2012, Cambridge University Press, pp. 1-19.

[8] Katz, R. W., Parlange, M. B., Naveau, P: Statistics of extremes in hydrology. Advances in Water Resources, 25, 1287–1304.

[9] Kao, S. - C., Ganguly A. R: Intensity, duration, and frequency of precipitation extremes under 21st-century warming scenarios, Journal of Geophysical Research, 116(D16), 2011, 14 pp.

[10] Ghosh, S., Das, D., Kao, S.-C., Ganguly A.R., Lack of uniform trends but increasing spatial variability in observed Indian rainfall extremes, *Nature Climate Change* 2012, 86–91.

[11] Khan, S., Kuhn, G., Ganguly, A. R., Erickson III, D. J., and Ostrouchov, G: Spatio-temporal variability of daily and weekly precipitation extremes in South America, Water Resources Research, vol. 43, W11424, 2007, 25 pp.

[12] Ferro, C.A.T., and Segers, J., Inference for clusters of extreme values, Journal of the Royal Statistical Society: Series B (Statistical Methodology), 65(2), 2003, 545-556.

[13] Mannshardt-Shamseldin, E.C., Smith, R.L., Sain, S.R., Mearns, L.O., and Cooley, D., Downscaling extremes: A comparison of extreme value distributions in point-source and gridded precipitation data, Annals of Applied Statistics, 4(1), 2010, 484-502.

[14] Kuhn, G., Khan, S., Ganguly, A.R., and Branstetter, M.L., Geospatial-temporal dependence among weekly precipitation data with applications to observations and climate model simulations in S. America, Advances in Water Resources, 30(12), 2007, 2401-2423.

[15] Martins E.S., Stedinger J.R., Generalized maximum-likelihood generalized extreme-value quantile estimators for hydrologic data, Water Resources Research, 36, 2000, 737-744.

[16] Wehner, M., Sources of uncertainty in the extreme value statistics of climate data. Extremes, 13(2), 2010, 205-217.

[17] Smith R.L., Tebaldi, C., Nychka D., Mearns L.O., Bayesian modeling of uncertainty in ensembles of climate models, Journal of the American Statistical Association, 104 , 2009, 97-116.

[18] O'Gorman, P. A., and Schneider, T: The physical basis for increases in precipitation extremes in simulations of 21st-century climate change. Proc. Natl. Acad. Sci. USA, 106(35), 14773-14777, 2009.

[19] Khalefa, M., Mokbel, M. and Levandoski. J. Skyline query processing for incomplete data. In Proc. 24th Int. Conf. on Data Engineering, 2008.

[20] Ilyas , I. F., Beskales, G. and Soliman, M. A. Survey of top-k query processing techniques in relational database systems. ACM Computing Surveys, 2008.

[21] Das, D., Kodra, E., Obradovic, Z., Ganguly, A. (2012) " Mining Extremes: Severe Rainfall and Climate Change," To appear in *20th European Conf. Artificial Intelligence (ECAI-12)* , Aug. 2012, Montpellier, France.

[22] Joe, H. Multivariate Models and Dependence Concepts, Chapman and Hall, London, 1997.

[23] Liu, Y., Bahadori, M., T., Li, H. Sparse-GEV: Sparse Latent Space Model for Multivariate Extreme Value Time Series Modeling. To appear in *ICML* 2012., Edinburgh, Scotland.

[24] Friedman, J., Hastie, T., and Tibshirani, R. Sparse inverse covariance estimation with the graphical lasso. *Biostatistics*, 2008.

[25] Mueen, A.and Keogh, E. Online Discovery and Maintenance of Time Series Motifs. *SIGKDD* 2010

[26] Ravikumar, P., Raskutti, G., Wainwright, M. and Yu, B. Model selection in Gaussian graphical models: High dimensional consistency of L1-regularized MLE. *NIPS* 2008.