5th International Workshop on Climate Informatics
September 24-25, 2015

Hosted by the National Center for Atmospheric Research in Boulder, CO

Copula based covariate selection in climate for statistical downscaling

Yi Li[1], Yale Chang[2], Thomas Vandal[3], Debasish Das[3], Adam Ding[1], Auroop Ganguly [3], Jennifer Dy[2]

Abstract—**It is imperative to accurately assess the impacts of climate change at regional scale in order to inform stakeholders to make policy decisions on critical infrastructures, management of natural resources, humanitarian aid, and emergency preparedness. However, Global Climate Models (GCMs) currently provide relatively coarse resolution outputs which preclude their application to accurately assess the effects of climate change on finer regional scale events. Statistical downscaling are methods that use statistical models to infer the regional-scale or local-scale climate information from coarsely resolved climate models. To make accurate predictions, covariate selection must be used to reduce the dimensionality of high dimensional climate data. Covariates in climate data tend to be highly dependent and non-linear in nature requiring advanced covariate selection methods. In this work, we propose a novel copula-based dependence measure that can capture non-linear relationships between variables as a criterion for feature selection. We demonstrate its effectiveness in discovering relevant features important for prediction with a non-parametric Bayesian mixture of sparse regression models applied to statistical downscaling.**

I. Motivation

The relatively coarse resolutions of Global Climate models (GCMs) often preclude their application to accurately assess the effects of climate change on finer regional-scale phenomena. Statistical downscaling uses statistical models to learn empirical statistical relationships between large-scale GCM features (predictors) and regional-scale climate variable(s) (predictands) to be projected. Moreover, these climate models consist of a range of variables that are highly dependent and non-linear, including air temperature, pressure, relative humidity, and others over space and time, as well as different elevations.

Projecting climate variables at a finer resolution from GCM of local-scale and/or large-scale variables is highly non-trivial, challenging traditional statistical models. A wide range of methods have been applied to statistical downscaling including multiple linear regression, bias correction spatial disaggregation, artificial neural networks [1], and a non-parametric Bayesian mixture of sparse regressions [2], each suffering from high dimensionality. The selection of relevant covariates, both generalizable to a changing climate and physically interpretable, is equally as important as the statistical model chosen for projections.

In this paper, *we propose a copula-based dependency for covariate selection which preserves the interpretability of covariates, accounts for non-linear relationships, and is computationally efficient.* This covariate/feature selection algorithm can then be used to efficiently pre-select the relevant features as inputs to downstream complex prediction models, such as [2]. We demonstrate its effectiveness in discovering covariates which may be relevant to statistical downscaling while projecting annual precipitation.

II. Methodology

Key to any feature selection method is the criterion measure of feature importance. A feature X is relevant for predicting target variable Y if it is statistically dependent on X. A traditional measure is the Pearson's correlation coefficient ρ. However, it focuses on linear relationships and is not appropriate for exploring complex relationships that we deal with. There are two deficiencies for ρ: (1) It is not invariant to monotone transformations of variables. (2) It does not treat all deterministic relationships equally, and cannot capture some non-monotone nonlinear relationships. For mixture models such as [2], even though the relationship in each clusters is linear, the overall relationship (mixed linear) is nonlinear. And for some mixed linear relationships ρ can be zero. Consequently, we propose a dependence measure based on copula theory to help us find both linear and non-linear relationships.

The joint distribution for d random variables $X = (X_1, ..., X_d)$ can be decomposed by Sklar's Theorem [3]: $F(x_1, ..., x_d) = C[F_1(x_1), ..., F_d(x_d)]$, where $F_j(x)$ is the marginal distribution of X_j and C is a cop-

[1]Dept. of Mathematics, liyi.1989@gmail.com [2]Dept. of Electrical and Computer Engineering, [3]Sustainability and Data Sciences Lab, all from Northeastern University. This work is supported by NSF CCF-1442728.

ula – a joint distribution on the d-dim unit cube. $C(u_1, ..., u_d) = P(U_1 \leq u_1, ..., U_d \leq u_d)$ is the distribution function of copula-transformed, uniformly distributed variables $U_j = F_j(X_j)$. This decomposition separates the dependence structure in the data from the marginals. All dependence information is contained in the copula C.

Reshef [4] proposed the concept of equitable dependence measure: a measure that captures the relationship based on the noise level regardless of the function types (linear or nonlinear). Kinney [5] formalized mathematically the *self-equitability* definition. We propose to use the Silvey [6]'s Δ or Copula correlation $CCor$ measure, which is not only self-equitable but also *robust-equitable* [7], meaning the dependence measure can capture signals hidden in background noise. These properties enable $Ccor$ to be a suitable tool for feature selection. We apply $CCor$ together with the minimum redundancy maximum relevance feature search strategy [8], which selects the features most related to the response variable while having least correlation among the features themselves. The features selected are then fitted with a non-parametric Bayesian mixture of sparse regression models [2].

III. Numerical Examples

Mean Precipitation in USA. We investigate the performance of $CCor$ as a dependence measure for feature selection on the statistical downscaling of annual mean precipitation data from the US Historical Climatology Network[9]. The predictors include the station-scale local variables and global variables from National Oceanic and Atmospheric Administration (NOAA), including monthly, seasonal, and regional statistics (mean, max, min, standard deviation), resulting in 580 features.

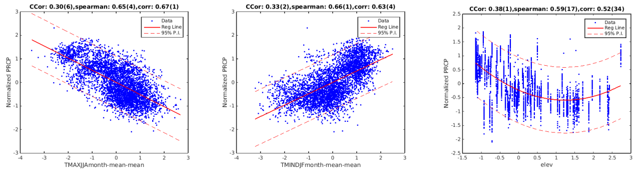

Fig. 1. Features in the northwest region. The rank (shown in parenthesis) of each feature is provided after the dependence score.

In Fig 1, we illustrate the advantage of $CCor$ versus ρ and Spearman's ρ for capturing nonlinear dependencies. Observe that they can all detect linear structures in the first two plots (as shown by their ranks). However, only $CCor$ can detect the non-monotonic nonlinear structure in the last plot.

We compare the performance of feature selection method with CCor and ρ followed by the non-parametric Bayesian mixture of sparse regression models. We also show the results on no feature selection (all features) with linear regression as a baseline. The 5-fold cross validated mean squared errors (MSE) for predicting annual mean precipitation in the west (CA, NV), northwest (WA, OR, ID) regions [2] and the US continent are shown in Fig 2. Note that the prediction performance with selected $CCor$ features performed best (lowest MSE). Besides the traditional temperature mean and extremes, we find higher moments, the standard deviation are also important predictors for precipitation.

IV. Conclusion

Covariates in climate prediction tend to be highly dependent and non-linear in nature requiring advanced covariate selection methods. In this paper, we proposed an equitable dependence measure to detect linear and non-linear relationships and demonstrate its effectiveness in discovering relevant features important for statistical downscaling.

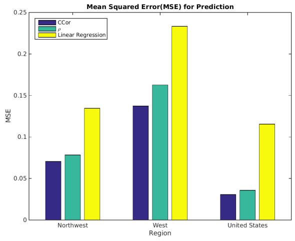

Fig. 2. Predictive MSE.

References

[1] G. Bürger, T. Murdock, A. Werner, S. Sobie, and A. Cannon, "Downscaling extremes-an intercomparison of multiple statistical methods for present climate," *Journal of Climate*, vol. 25, no. 12, pp. 4366–4388, 2012.

[2] D. Das, J. Dy, J. Ross, Z. Obradovic, and A. R. Ganguly, "Non-parametric bayesian mixture of sparse regressions with application towards feature selection for statistical downscaling," *Nonlinear Processes in Geophysics*, vol. 21, no. 6, pp. 1145–1157, 2014.

[3] R. B. Nelsen, *An introduction to copulas*. Secaucus, NJ, USA: Springer-Verlag New York, Inc., 2006.

[4] D. Reshef, Y. Reshef, H. Finucane, S. Grossman, G. McVean, P. Turnbaugh, E. Lander, M. Mitzenmacher, and P. Sabeti, "Detecting novel associations in large data sets," *Science*, vol. 334, no. 6062, pp. 1518–1524, 2011.

[5] J. Kinney and G. Atwal, "Equitability, mutual information, and the maximal information coefficient," *Proceedings of the National Academy of Sciences*, vol. 111, pp. 3354–3359, 2014.

[6] S. D. Silvey, "On a measure of association," *Ann. Math. Statist.*, vol. 35, pp. 1157–1166, 09 1964.

[7] A. Ding and Y. Li, "Copula correlation: An equitable dependence measure and extension of pearson's correlation," 2014.

[8] H. Peng, F. Long, and C. Ding, "Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy," in *IEEE Trans. on Pattern Analysis and Machine Intelligence*, pp. 1226–1238, 2005.

[9] D. R. Easterling, T. C. Peterson, and T. R. Karl, "On the development and use of homogenized climate datasets," *Journal of climate*, vol. 9, no. 6, pp. 1429–1434, 1996.