A STATISTICAL ANALYSIS OF MULTIPLE TEMPERATURE PROXIES: ARE RECONSTRUCTION OF SURFACE TEMPERATURE OVER THE LAST 1000 YEARS RELIABLE? : DISCUSSION OF THE PAPER BY MCSHANE AND WYNER

BY PETER CRAIGMILE AND BALA RAJARATNAM∗

The Ohio State University and Stanford University

Professors McShane and Wyner have written a thought provoking paper that intends to challenge some of the conventional wisdom in the paleoclimate literature. Rather than commenting on the merits of the entire methodology we focus on one topic. Namely, we discuss theoretical and practical aspects of the use of the least absolute shrinkage and selection operator (5), more popularly known as the “Lasso”, in the context of paleoclimate reconstruction.

It is important to acknowledge at first sight that the Lasso seems like a natural candidate in the paleoclimate context, since one is immediately faced with a larger number of proxies, compared to the number of data points (For example, in McShane and Wyner (4) (hereafter MW), Section 3.2, the response variable is of length 149 whereas there are 1138 predictors). It is clear that standard regression-based variable selection techniques will not work. The sheer number of predictors does indeed warrant a need for regularization. Many techniques are available for such problems, including popular methods such as ridge regression and principal component regression.

As pointed out by MW the “Lasso tends to choose sparse \( \hat{\beta}_{\text{Lasso}} \) thus serving as a variable selection methodology and alleviating the \( p \gg n \) problem”. This point is very well taken. The model selection capability of the Lasso has made it very relevant in this era of high throughput data and rapidly changing information technology. Consequently the Lasso has been useful in biomedical and genomic applications where genes are often in the tens of thousands, compared to much fewer subjects. Bio-medical scientists often wish to isolate a few, but important genes that are related to disease conditions. The Lasso “zeros out” smaller co-efficients and

∗To whom correspondence may be addressed.
†Bala Rajaratnam was supported in part by NSF grants DMS 0906392, AGS-1003823 and grants SU-WI-EVP10, SUFSC08-SUFSC10-SMSCVISG0906. Peter Craigmile was supported in part by the National Science Foundation under award numbers DMS-0604963 and DMS-0906864

AMS 2000 subject classifications: Primary, 60K35, 60K35; secondary, 60K35

Keywords and phrases: Paleoclimate reconstruction, LASSO, Multiproxy data, Time series dependence
In a more abstract setting, consider a statistical model such as a regression model which has a low signal-to-noise ratio where the co-efficient vector is not sparse. It is quite easy to see that the Lasso can exclude many predictors which have small but non-zero co-efficients. This exclusion will occur with a higher degree of severity, a feature that is not available in ridge regression or principal components regression. The Lasso works well if the signal is sparse; i.e., there are few large non-zero co-efficients and many true zero coefficients. Just as with many other estimators, the signal needs to be also bounded away from zero for the Lasso to be able to recover the non-zero coefficients accurately. Thus, it is not clear immediately whether the very model selection feature that makes the Lasso attractive in so many settings is as equally desirable in the paleoclimate context. As the authors state correctly, the relationship between the predictors and response variable is weak. Hence the co-efficient estimates are very small in magnitude. In this instance the Lasso could potentially zero out many of those co-efficients. It is quite feasible that the proxies could collectively have some predictive power though the contribution of each of the individual proxy time-series may be rather small.

There are also other reasons why the Lasso may not always be appropriate in the paleoclimate context. First, the Lasso can choose at most \( n \) non-zero co-efficients (see (1)). Hence by design any other additional set of proxies which may have almost the same predictive power, but slightly less than the first \( n \) predictors, will have no chance of being selected in the model. This problem comes back to our original point that in the paleoclimate context there may be more than a few sparse signals, but rather a large number of weak signals instead. This is simply due to the irregular, dependent noise in the data and the structural relationships between instrumental records and paleoclimate proxies (see, e.g., Tingley et al. (6)). Furthermore, the standard Lasso does not yield any ridge or Steinian-type shrinkage, a feature that can potentially lead to better RMSE. In this regard, the elastic net proposed by Zou and Hastie (7) might be more suitable. Second while the Lasso has the capability to do model selection by zeroing out certain variables, it also has the adverse effect of shrinking even the larger non-zero co-efficients via soft-thresholding. This indiscriminate shrinking of the co-efficients leads to biased estimates. Third, the Lasso is not an oracle procedure, and there are scenarios in which Lasso variable selection is inconsistent (8). Such theoretical safeguards, under broad assumptions, could be very useful, especially in a hotly debated topic such as climate change. Zou (8) proposes the adaptive Lasso as a possible remedy for these problems and may be worthy of exploration in this context.

A further issue involves the fact that “proxy series contain very complicated and highly autocorrelated time series structures” (MW). The standard Lasso assumes...
that the errors in the regression model are uncorrelated, which is clearly not the case here. Indeed, a review of statistical models appearing in Section 5.2 of MW points to significant autocorrelations that have to be accounted for. Further research is needed, especially for paleoclimatic variable selection (See Gelper and Croux (2) for a time series version of the Lasso applied to economics data). This is further complicated by the “problem of spatial correlation” which MW choose to ignore in their article.

In closing, we have indicated some of our concerns of using the Lasso for paleoclimatic reconstruction. This does not mean that variable selection is unimportant – it is. For example with tree proxies an argument can be made (e.g., Tingley et al. (6), Section 3) that only trees in certain areas contain climate signatures. In defense of MW we note that they do consider techniques other than the Lasso in Section 4 (Space limitations excludes us from commenting on these methods). We conclude by commending the authors on a thought provoking paper, and by referring the reader to a recent manuscript (Tingley et al. (6)) that sheds further light and gives detailed statistical insights into some of the important issues in paleoclimatic reconstruction.

References.


