The Annals of Applied Statistics
2011, Vol. 5, No. 1, 83–87
DOI: 10.1214/10-AOAS398E
Main article DOI: 10.1214/10-AOAS398

SPURIOUS PREDICTIONS WITH RANDOM TIME SERIES: THE LASSO IN THE CONTEXT OF PALEOCLIMATIC RECONSTRUCTIONS. DISCUSSION OF: A STATISTICAL ANALYSIS OF MULTIPLE TEMPERATURE PROXIES: ARE RECONSTRUCTIONS OF SURFACE TEMPERATURES OVER THE LAST 1000 YEARS RELIABLE?

BY MARTIN P. TINGLEY[1]

NCAR and Harvard University

Blakeley B. McShane and Abraham J. Wyner (hereafter, MW2011) find that, under certain scenarios and using the LASSO to fit regression models, randomly generated series are as predictive of past climate as the commonly used proxies (MW2011, Figure 9). They conclude that "the proxies do not predict temperature significantly better than random series generated independently of temperature," a claim that has already been reproduced in the popular press [The Wall Street Journal (2010)]. If this assertion is correct, then MW2011 have undermined all efforts to reconstruct past climate, which are based on the fundamental assumption that natural proxies are predictive of past climate. I disagree with MW2011's conclusion and provide an alternative explanation: the LASSO, as applied in MW2011, is simply not an appropriate tool for reconstructing paleoclimate.

To shed light on the MW2011 results, I turn to an experiment with surrogate data [Tingley (2011)]. The "target" time series, analogous to the Northern Hemisphere mean temperature time series in MW2011, is the sum of a simple linear trend and an AR(1) process, $y(t) = 0.25 \cdot t + \varepsilon(t), t = 1, \ldots, 149$. The AR(1) coefficient in the ε process is 0.4, and the variance of the innovations is 1. I then generate 1138 "pseudo-proxy" time series by adding white noise to this target series. The signal to noise ratio (SNR) of these pseudo-proxies, expressed as the ratio of the standard deviation of the target time series to that of the additive white noise, will take on a range of values $(4, 2, 1, 1/2, 1/4, 1/8)$. In order to compare the performance of these pseudo-proxies to random series, I generate 1138 independent AR(1) time series, each of length 149; the common AR(1) coefficient, α, for these random series will take on a range of values (0, 0.2, 0.4, 0.6, 0.8, 1.0). Two regression models are then fit using 119 of the 149 observations.

The first model, referred to as "composite regression," involves averaging across all predictor series and then using this composite series to predict the target via ordinary least squares regression. The second model applies the LASSO to all

Received September 2010.

[1]A more detailed version of this discussion is available at people.fas.harvard.edu/~tingley/.