

Spurious predictions with random time series:
The LASSO in the context of paleoclimatic reconstructions.

A Discussion of
“A Statistical Analysis of Multiple Temperature Proxies:
Are Reconstructions of Surface Temperatures
over the Last 1000 Years Reliable?”

by Blakeley B. McShane and Abraham J. Wyner.

Martin P. Tingley*

September 17, 2010

Blakeley B. McShane and Abraham J. Wyner (hereafter, MW2010) 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 (MW2010, Fig. 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 MW2010 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 MW2010’s conclusion and provide an alternative explanation: the LASSO, as applied in MW2010, is simply not an appropriate tool for reconstructing paleoclimate.

To shed the light on the MW2010 results, I turn to an experiment with surrogate data. The “target” time series, analogous to the Northern Hemisphere mean temperature time series in MW2010, is the sum of a simple linear trend and an AR(1) process, $y(t) = .25 \cdot t + \epsilon(t), t = 1 \dots 149$. 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

*NCAR and Harvard University. e-mail address: tingley@fas.harvard.edu. A more detailed version of this discussion is available at people.fas.harvard.edu/~tingley/.

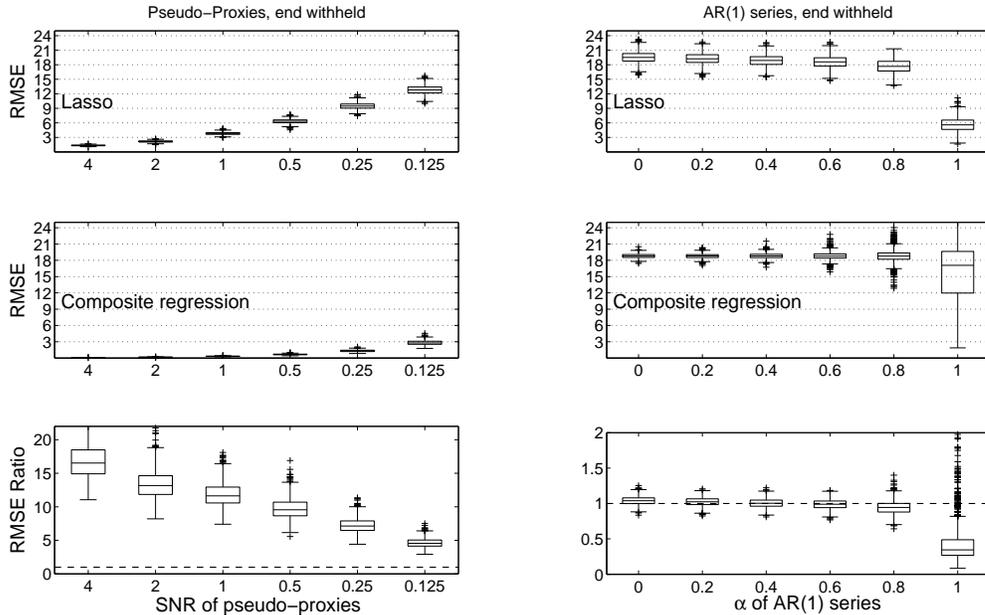


Figure 1: Out-of-sample RMSE calculated using 30 values withheld from the end of each surrogate data set. Left column: using pseudo-proxies as predictors. Right column: using independent AR(1) series as predictors. Top row: regression using the LASSO. Middle row: composite regression. Bottom row: the ratio of the LASSO RMSE value to the composite regression RMSE.

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 predictor series, and is fit using the algorithm described in Friedman et al. [2007, 2010] and the `glmnet` package for Matlab (available at <http://www-stat.stanford.edu/~tibs/glmnet-matlab/>). The LASSO penalization parameter (λ on page 13 of MW2010) is set to be 0.05 times the smallest value of λ for which all coefficients are zero; the LASSO penalization is thus very small.

Box plots of the out-of-sample RMSE are shown in Figure 1 for 1000 experiments that calculate the RMSE using observations withheld from the end of the data set; results are similar when observations are withheld from the interior. Composite regression results in lower RMSE than the LASSO for all values of the pseudo-proxy SNR (Figure 1, left column). For an SNR of 1/4, the LASSO RMSE is about 7.5 times larger than the composite regression RMSE. This is a clear indication that the LASSO is not making effective use of the information contained in the pseudo-proxies.

Applying the LASSO to AR(1) series with sufficiently high α values results in lower out-of-

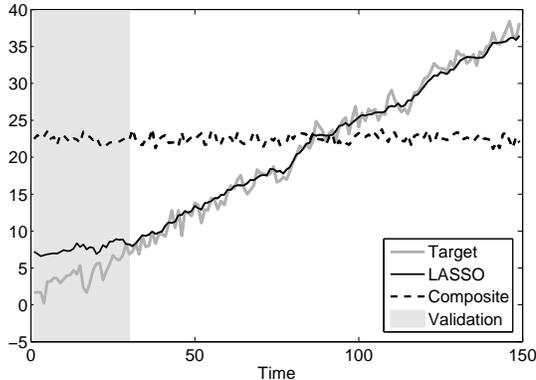


Figure 2: Example fits from applying the LASSO to random walk predictors and composite regression to white noise predictors. Shading indicates the portion of the data set withheld for validation.

sample RMSE values than applying the LASSO to the noisier pseudo-proxies (compare the two top panels of Figure 1). This is the result discussed in MW2010: the LASSO gives better results when applied to highly structured random time series than when applied to noisy predictors that do in fact contain information about the target series. Note in addition that for values of $\alpha \geq 0.8$, the LASSO on the AR(1) series results in lower RMSE than using composite regression on AR(1) series with $\alpha = 0$ (the limiting case of an SNR of zero for the pseudo-proxies). These results can be explained by the structure of the surrogate data experiment, which sets the target series to be linear in time, with additive AR(1) noise. The LASSO applied to AR(1) series with $\alpha = 1$ results in non-zero coefficients for only those predictor series that display strong, linear trends over the calibration interval, and the expected value of a predictor series during the validation interval is then the last value in the calibration interval. In contrast, as the $\text{SNR} \rightarrow 0$, composite regression on the pseudo-proxies approaches (in expectation) the intercept model. These features are illustrated in Figure 2.

MW2010 point out that highly structured random series (large α) are well suited to interpolation, and to a lesser extent extrapolation, on short time scales (page 22). As the variance of the white noise component of the pseudo-proxies increases, these predictors become both less informative of the target series, and less structured in time. At a certain SNR, short term interpolations or extrapolations based on independent, but more temporally structured series, perform better. This threshold SNR is a decreasing function of the length of the extrapolation/interpolation interval. As the goal in a paleoclimate context is extrapolation on long timescales, composite regression on extraordinarily noisy proxies will outperform the LASSO applied to random walks.

The LASSO gives inferior results in situations where each of a large number of predictors is only weakly correlated with the target series, but the mean across all predictors is highly correlated with that target. It is well known that the LASSO is the posterior mode which results from placing a common double exponential prior on the regression coefficients [Park and Casella, 2008]. It is difficult to imagine a scientifically defensible reason for specifying such a prior in the paleoclimate

context. A more scientifically reasonable approach is to modify the LASSO prior to shrink the regression coefficients not towards zero, but towards a common, data determined value. Such a prior reflects the assumptions that 1) the regression coefficients are likely to be similar to one another, and 2) all predictors are informative of the target series. Within the paleoclimate context, where the expectation is that each proxy is weakly correlated to the northern hemisphere mean (for two reasons: proxies generally have a weak correlation with local climate, which in turn is weakly correlated with a hemispheric average) the LASSO as used by MW2010 is simply not an appropriate tool. It throws away too much information.

More generally, MW2010 have perhaps missed a larger point. The presence of a large number of correlated predictors is intrinsic to the paleoclimate reconstruction problem, and has a geophysical basis. MW2010 state that, “it is unavoidable that some type of dimensionality reduction is necessary, even if there is no principled way to achieve this” (page 8–9). This is simply not the case. A more scientifically sound approach recognizes that the proxies are related to the local climate, which in turn displays both spatial and temporal correlation. These ideas can be encoded in hierarchical statistical models, which can combine the specification of a parametric spatiotemporal covariance form for the target climate process (e.g., surface temperature anomalies) with reasonable forward models that describe the conditional distribution of the proxy observations, given the climate process. Such approaches naturally account for the $p \gg n$ problem, and for the strong correlations between the proxies. These models are derived from the rich development of Bayesian statistics over the past twenty years, and are being adapted by the paleoclimate community. See Tingley and Huybers [2010] for a specific example, and Tingley et al. [2010] for a comprehensive discussion.

Acknowledgements

This manuscript benefited from discussions with Peter Huybers.

References

- J. Friedman, T. Hastie, H. Höfling, and R. Tibshirani. Pathwise coordinate optimization. *The Annals of Applied Statistics*, 1(2):302–332, 2007.
- J. Friedman, T. Hastie, and R. Tibshirani. Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, 33(1):1–22, 2010.
- T. Park and G. Casella. The Bayesian Lasso. *Journal of the American Statistical Association*, 103(482):681–686, 2008.
- The Wall Street Journal. Editorial: Climate of uncertainty, 02 September 2010. Downloaded from <http://online.wsj.com/article/SB10001424052748703467004575463433671739148.html> on 7 September 2010.

M.P. Tingley and P. Huybers. A Bayesian Algorithm for Reconstructing Climate Anomalies in Space and Time. Part 1: Development and applications to paleoclimate reconstruction problems. *Journal of Climate*, 23(10):2759–2781, 2010.

M.P. Tingley, P.F. Craigmile, M. Haran, B. Li, E. Mannshardt-Shamseldin, and B. Rajaratnam. Piecing together the past: Statistical insights into paleoclimatic reconstructions. Technical Report 2010–09, Stanford University, Department of Statistics, 2010. http://statistics.stanford.edu/~ckirby/reports/2010_2019/reports2010.html.