

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*

The abstract of the article by Blakeley B. McShane and Abraham J. Wyner (hereafter, MW2010) asserts 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. Such a bold claim warrants more investigation than is provided in MW2010.

More specifically, 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). There are (at least) two explanations for this result: 1) the proxies are no better than random series, or 2) some aspect of the LASSO method fails to make efficient use of the information in the proxies. The results of a simple experiment with surrogate data and considerations of the properties of the LASSO shows the latter explanation to be correct: the LASSO, as applied in MW2010, is simply not an appropriate tool for reconstructing paleoclimate.

The LASSO [Tibshirani, 1996] is a well-established technique for fitting regression models in which the number of variables is greater than the number of observations ($n \gg p$). As noted in MW2010, the LASSO generally results in many of the regression coefficients being set to zero, so acts as a variable selection procedure that “helps reduce the problem of spatial correlation amongst the proxies.” (Page 13). For these reasons, MW2010 “believe it should provide predictions which are as good or better than other methods” (page 14), and an investigation of the LASSO’s utility in the paleoclimate context is certainly welcome.

*NCAR and Harvard University. e-mail address: tingley@fas.harvard.edu

To shed the light on the MW2010 result that, using the LASSO, random series are as predictive of past climate as the actual proxies, 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 the standard deviation 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 30 withheld observations, either from one end or the middle of each data set, are used to calculate the out-of-sample root mean squared error (RMSE) for each model.

The first model, referred to as “composite regression,” involves averaging across all predictor series (either the pseudo-proxies or the random series), and then using this single, 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/>). I do not perform the cross-validation procedure used in MW2010 to determine the LASSO penalization parameter (λ on page 13 of MW2010). Instead, I use the default setting of the `glmnet` package, which sets λ to be 0.05 times the smallest value of λ for which all coefficients are zero. The LASSO penalization is thus very small.

I generate 1000 target time series, and run $2 \cdot 2 \cdot (6 + 6) = 48$ experiments on each. There are two levels to both the “method” factor (LASSO and composite regression) and “validation” factor (values withheld from the end or middle of each data set). There are twelve levels to the “predictor” factor, divided into 2 sub-factors, as the predictors are formed by adding white noise to the target series (6 different SNRs), or by generating independent AR(1) series (6 different values of α). Box plots of the out of sample RMSE are shown in Figures 1 and 2, and box plots of the ratio of the LASSO RMSE to the composite regression RMSE are shown in Figure 3.

First and foremost, composite regression results in lower RMSE values than the LASSO for all values of the pseudo-proxy SNR, when either the end or the middle of the data set is withheld. Withholding the end (middle) of the data sets for validation, the LASSO RMSE is about 7.5 (2.5) times larger than the composite regression RMSE for a pseudo-proxy SNR of 1/4. This is a clear indication that the LASSO is not making effective use of the information contained in the pseudo-proxies. If each of the 1,138 proxies from Mann et al. [2008] used by MW2010 is indeed informative of the Northern Hemisphere mean, but only weakly so, then the LASSO is simply not an optimal analysis tool for using them to infer past climate.

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

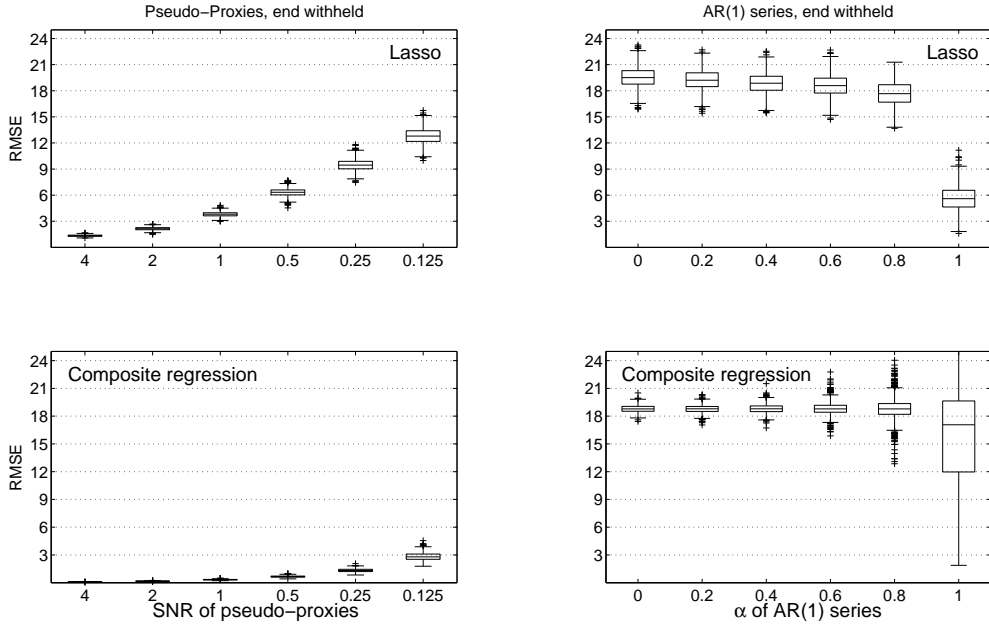


Figure 1: Out-of-sample RMSE calculated using 30 values withheld from the end of each surrogate data set. Left column: using pseudo-proxies as predictors. Right column: using independent AR(1) series as predictors. Top row: regression using the LASSO. Bottom row: composite regression.

sample RMSE values than applying the LASSO to the noisier pseudo-proxies (compare the two top panels in Figures 1 and 2). 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.

When values from the end of the data set are withheld (Figure 1), composite regression on the pseudo-proxies (lower left) results in lower RMSE than regression with the LASSO on the AR(1) series (upper right), for all values of the pseudo-proxy SNR and AR(1) coefficient. Note, however, that the limiting case of an SNR of zero for the pseudo-proxies corresponds to using AR(1) series with $\alpha = 0$. For values of $\alpha \geq 0.8$, the LASSO on the AR(1) series results in lower RMSE than using the composite regression on AR(1) series with $\alpha = 0$ (equivalent to an SNR of 0). Thus, for very noisy proxies and high values of α , the LASSO applied to random series results in lower out of sample RMSE than either the LASSO or composite regression applied to the pseudo-proxies. This result 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. Consider applying the LASSO to AR(1) series with $\alpha = 1$, which results in non-zero coefficients for only those predictor series that display strong, unidirectional trends over the calibration interval. By the nature of a random walk, the expected value of a predictor series during the validation interval is the last value in the calibration interval. In contrast, as the $\text{SNR} \rightarrow 0$, composite regression on the pseudo-proxies approaches (in

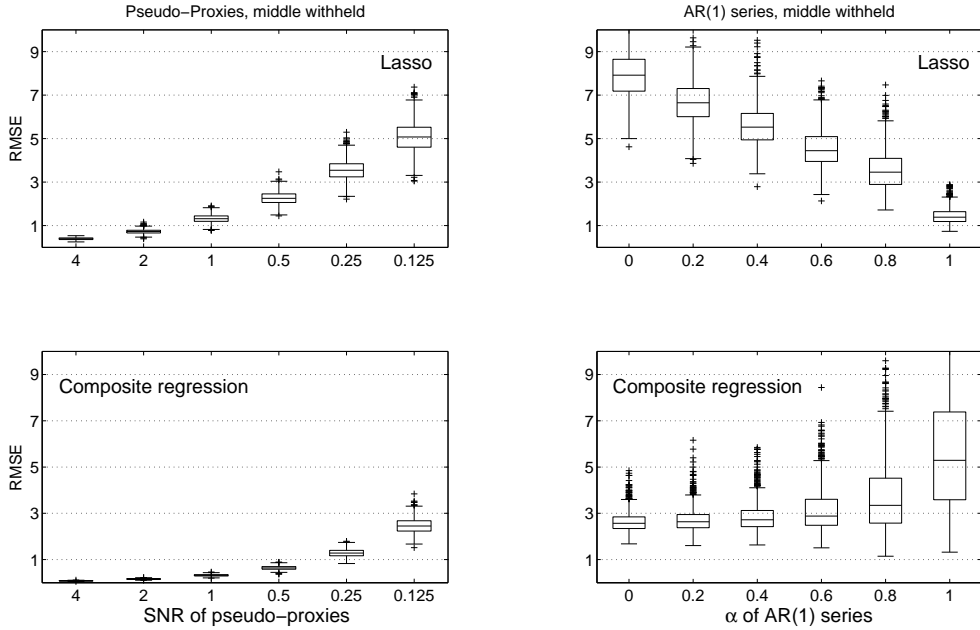


Figure 2: As in Figure 1, but withholding 30 observations from the middle of each surrogate data set. Note the change in vertical scale from Figure 1.

expectation) the intercept model, where the predicted values during the validation interval are given by the mean over the calibration interval. These features are illustrated in Figure 4, which plots representative results of applying the LASSO to random walk predictors and composite regression to white noise predictors. Note that if the target series reverts to the calibration mean during the validation interval, then the performance of the LASSO using random walk predictors will suffer, while that of composite regression on white noise series will improve.

When the withheld values fall in the middle of the data set (Figure 2) the intercept model is actually quite good, as evidenced by the fact that the RMSE box plot for composite regression on white noise series ($\alpha = 0$) is centered at the same value as that for composite regression applied to pseudo-proxies with an SNR of $1/8$. As above, for sufficiently high values of α and sufficiently low values of the pseudo-proxy SNR, applying the LASSO to AR(1) series results in lower RMSE values than applying either composite regression or the LASSO to pseudo-proxies. This result can likewise be explained by the structure of the experiment. The LASSO applied to AR(1) series picks out those series that have strong, monotonic trends for both the early and late parts of the data set. When $\alpha = 1$, the predictor series during the validation interval behave like discrete Brownian bridges, pinned by the two values on either side of the calibration interval. In expectation, then, the predictors picked by the LASSO will display linear trends over the validation interval, a feature illustrated in Figure 4. If the target series during the calibration interval has a (roughly) linear structure, as is the case in these experiments, then the LASSO applied to highly

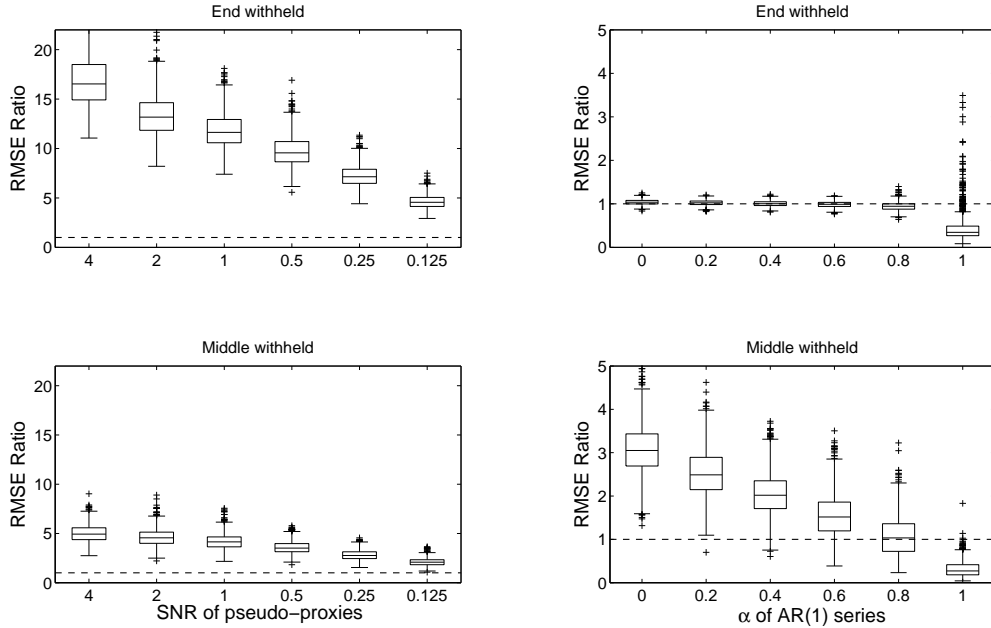


Figure 3: Ratio of the RMSE values from the LASSO to those from composite regression. Left column: using pseudo-proxies as predictors. Right column: using independent AR(1) series as predictors. Top row: RMSE ratio calculated using 30 values withheld from the end of each surrogate data set. Bottom row: RMSE calculated using 30 values withheld from the middle of each surrogate data set. Note the change in vertical scale between the left and right panels.

autocorrelated AR(1) series successfully exploits the time series structure of the predictors. The resulting predictions have generally lower RMSE than either the LASSO or composite regression applied to sufficiently noisy pseudo-proxies.

These results are (in part) a product of the structure of this experiment, which involves predicting a target series with a constant linear trend. Note, however, that many of the 30 year hold-out blocks used in MW2010 (their Figure 8) display strong linear trends, so the experiment presented here is immediately applicable to the interpretation of the MW2010 results. 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.

As evidenced by the modeling and analysis approach adopted in Section 5, MW2010 clearly favor Bayesian methods. It is both apt and revealing to consider the Bayesian interpretation of

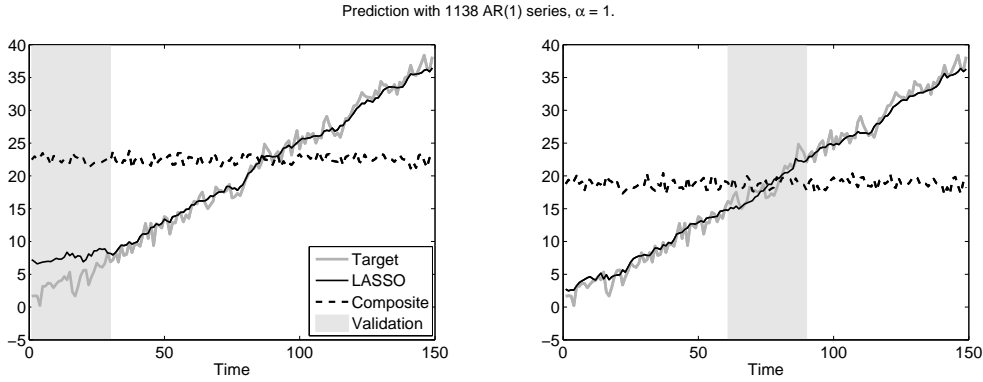


Figure 4: Example fits from applying the the LASSO to random walk predictors and composite regression to white noise predictors. Values are withheld for validation from either one end (left panel) or the middle (right panel) of the surrogate data set.

both the LASSO and the composite regression approaches considered here. Following the notation of MW2010, the basic model for both analyses is,

$$y_i = \beta_0 + \sum_{j=1}^p x_{ij}\beta_j + \epsilon_i \quad (1)$$

$$\epsilon_i \sim N(0, \sigma^2), \text{ iid } \forall i.$$

The LASSO results from placing a flat prior on β_0 and a common double-exponential prior on each $\beta_j, j \geq 1$:

$$p(\beta_0) \propto 1 \quad (2)$$

$$p(\beta_j) = \frac{\lambda}{4\sigma^2} \exp\left(-\frac{\lambda|\beta_j|}{2\sigma^2}\right).$$

Under these priors, the posterior distribution of the regression coefficients (conditional on λ and σ^2) takes the form,

$$p(\beta|\cdot) \propto \exp\left(-\frac{1}{2\sigma^2} \left(\sum_{i=1}^n \left(y_i - \beta_0 - \sum_{j=1}^p x_{ij}\beta_j \right)^2 + \lambda \sum_{j=1}^p |\beta_j| \right)\right). \quad (3)$$

The posterior mode (conditional on λ and σ^2) results from minimizing the argument of the exponent with respect to the $\beta_j, j = 0 \dots p$. This is equivalent to the LASSO optimization problem [Park and Casella, 2008, cf. MW2010, page 13]. It is difficult to imagine a scientifically defensible reason for specifying such a prior in the paleoclimate context. In a Bayesian analysis, draws from the conditional posterior for the β will *never* result in *any* of the coefficients being exactly equal to zero: the posterior mode (i.e., the LASSO) is structurally much different from all draws.

Now consider an alternative parameterization, $\beta_j = \mu + \delta_j, j = 1 \dots p$, which expresses each regression coefficient as the sum of a term common to all and a deviation. Place flat priors on β_0 and μ , and a common double-exponential prior on the δ_j :

$$\begin{aligned} p(\beta_0) &\propto 1 \\ p(\mu) &\propto 1 \\ p(\delta_j) &= \frac{\lambda}{4\sigma^2} \exp\left(-\frac{\lambda|\delta_j|}{2\sigma^2}\right). \end{aligned} \tag{4}$$

The posterior of β_0, μ , and the δ_j is then (conditional on λ and σ^2),

$$p(\beta_0, \mu, \delta|\cdot) \propto \exp\left(-\frac{1}{2\sigma^2} \left(\sum_{i=1}^n \left(y_i - \beta_0 - p\mu \cdot \bar{x}_i - \sum_{j=1}^p x_{ij}\delta_j\right)^2 - \lambda \sum_{j=1}^p |\delta_j|\right)\right). \tag{5}$$

where \bar{x}_i is the average across the p predictors for the i^{th} observation. This hierarchical model shrinks the regression coefficients, $\beta_j = \mu + \delta_j$, towards a common, data determined value. In the limit as all $\delta_j \rightarrow 0$, the δ_j drop out, and the mean across the design matrix is the sole predictor: this is the simple composite regression model used in the surrogate data experiments. Setting μ to zero (i.e. replacing the flat prior with a unit mass at zero) results in the LASSO model used by MW2010.

Both the LASSO and the composite regression approach are shrinkage estimators that can be understood in terms of the priors placed on the regression coefficients. The LASSO shrinks the predictors towards zero, and use of the posterior mode results in many of the coefficients being set to exactly zero. If each of the predictors is informative (even weakly so) of the target, then the LASSO does not use all available information. Composite regression is a limiting case of placing a common double-exponential (or normal, for that matter) prior distribution on the deviations of the regression coefficients from their mean value, and placing a flat prior on this mean value. This prior shrinks the regression coefficients towards a common, data determined value. If all predictors share a similar relationship with the target series, then this is a reasonable assumption. Put crudely, the LASSO is an, “if in doubt, throw it out” estimator, while placing the double exponential prior on the δ_j is an “if and doubt, shrink it to the average” estimator.

As demonstrated here, 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. As an analogy, consider the music produced by a large orchestra. Each individual instrument may be only weakly correlated with the overall sound of the orchestra, but (ignoring differences in volumes) the simple average across all instruments gives the performance heard by the audience. The LASSO, in contrast, picks out only those few instruments that are most correlated with the overall composition during a calibration interval, and these few instruments can be poorly correlated with the performance as a whole outside of this interval. 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 far 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.
- M.E. Mann, Z. Zhang, M.K. Hughes, R.S. Bradley, S.K. Miller, S. Rutherford, and F. Ni. Proxy-based reconstructions of hemispheric and global surface temperature variations over the past two millennia. *Proceedings of the National Academy of Sciences*, 105(36):13252–13257, 2008.
- 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.
- R. Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 58(1):267–288, 1996.

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.