The Third International Workshop on Climate Informatics
National Center for Atmospheric Research, Boulder, Colorado, September 26-27, 2013
https://www2.image.ucar.edu/event/ci2013

Climate Prediction via Matrix Completion

Mahsa Ghafarianzadeh and Claire Monteleoni
masa@gwu.edu, cmontel@gwu.edu
Department of Computer Science, The George Washington University

1 Introduction

Recently, machine learning has been applied to the problem of predicting future climates, informed by the multi-model ensemble of physics-based climate models that inform the Intergovernmental Panel on Climate Change (IPCC). There are over 20 laboratories, worldwide, running climate model simulations that inform the IPCC. Yet these climate models differ significantly, and there is high variance among their predictions. Climate scientists are currently interested in methods to combine the predictions of this multi-model ensemble of GCMs, in order to better predict future climates. Past work [3,4] demonstrated the promise of (supervised) online learning algorithms applied to this problem. Here we propose a novel approach, in the batch, unsupervised setting, using sparse matrix completion. Consistent with previous work, our method takes the climate models' predictions into account, including their projections into the future, in addition to the past observation data, however our approach to prediction is markedly different. We create a sparse (incomplete) matrix from climate model predictions and observed temperature data, and apply a matrix completion algorithm to recover it, yielding predictions of unobserved temperatures.

2 Technical Approach

Recently, several efficient algorithms have been proposed for sparse matrix completion. Applications of matrix completion have found success in ecology [11], and have also been proposed for *paleo*-climate reconstruction problems [8,9,10]. The algorithm we apply in this paper is OptSpace, a combination of spectral techniques and manifold optimization [5], which minimizes the nuclear norm (sum of the singular values) of the reconstructed matrix, and works well in practice.

We construct an incomplete matrix from the climate model predictions and the true temperature observations as illustrated in Figure 1. The first row of the matrix has the observed temperature data over time (e.g. one value per year), and the rest of the rows are historic and future temperature predictions of the climate models. The missing part of the matrix represents the unknown future temperature observations, and the future predictions of a subset of the model runs. We set all the unknown entries of the matrix to zero, in order to recover them using the OptSpace algorithm.

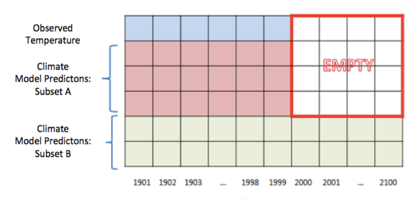


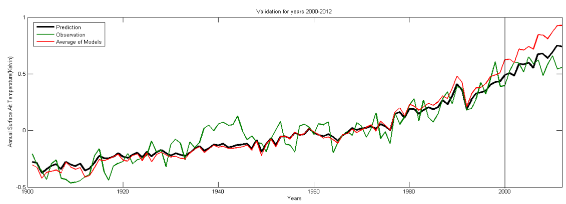
Figure 1: Schematic of Matrix M

We used global mean temperature anomaly data, since it is considered an indicator of climate change and was also studied in previous machine learning applications [3,4]. We used two sets of GCM hindcasts (predictions of years in the past), as well as historical temperature observations. The first set of GCM hindcasts has 7 models obtained from the IPCC Phase 3 Coupled Model Intercomparison Project (CMIP3) archive [2]; the second set has 9. These are all distinct, yielding an ensemble size of 16. We used the Climate of the 20th Century Experiment (20C3M) historic scenario (years 1901-1999), and the SRESB1 experiment future scenario (years 1901-2100). We obtained historical global temperature anomalies from the NASA GISTEMP archive for 1980-2012 [1].

To evaluate matrix completion on the climate prediction task, we compute its error on the predicted (missing) temperature data, with respect to the true observations, on historical validation periods. We compare this error to that of the average prediction over the multi-model ensemble of GCMs (also computed with respect to the true observations). Predicting with the ensemble average is currently the standard method, in climate science, of harnessing the predictions of the multi-model ensemble [6,7].

		8 years (2005-12)	13 years (2000-12)	23 years (1990-12)	33 years (1980-12)	43 years (1970-12)
Prediction	RMSE	0.667	0.620	0.512	0.280	0.237
	σ^2	0.007	0.012	0.022	0.006	0.005
Avg. of the models	RMSE	0.838	0.774	0.648	0.563	0.496
	σ^2	0.014	0.028	0.059	0.066	0.067

Table 1: Comparison between the algorithm's prediction error and that of the average prediction over climate models, on annual temperature anomalies, for 5 different values of T.



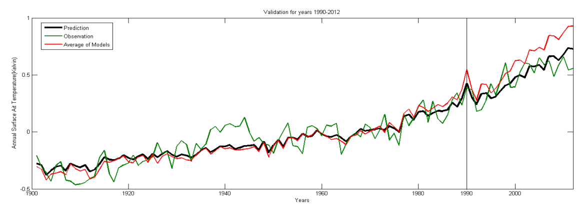


Figure 2: Comparison between the algorithm's prediction and that of the average over climate models, for annual temperature anomalies with validation periods of the past 8, 13 years. (Red: Average over climate models, Green: Observation and Black: Prediction)

In the first experiment (Table 1, Figure 2), we construct a matrix of annual temperature anomalies that has 112 columns for the years 1901-2012. Then, for several values of T, we set to zero all the entries of the past T years in the observation row and assume they are missing. Table 1 shows that the prediction of the matrix completion algorithm has consistently lower root-mean-square error (RMSE) compared to that of the average prediction over all the climate models. This result holds for each of the five experiments, which differ on the number of years to predict.

		5 years (1995-99)	10 years (1990-99)	15 years (1980-99)	20 years (1970-99)	30 years (1960-99)
Prediction	RMSE	0.012	0.010	0.010	0.008	0.007
	σ^2	1.47e-08	1.11e-08	1.14e-08	6.92e-09	5.61e-09
Avg. of the Models	RMSE	0.018	0.018	0.016	0.016	0.014
	σ^2	8.76e-08	1.02e-07	8.48e-08	7.42e-08	5.96e-08

Table 2: Comparison between the algorithm's prediction error and that of the average prediction over climate models, on monthly temperature anomalies, for 5 different values of T

In order to use much larger data sets than in the annual experiment, we also ran experiments on global monthly temperature anomalies by creating a matrix with column size 1188 (12 months × 99 years. Following standard practice [3], anomalies are computed separately per month, to remove seasonal effects. Again in this experiment (Table 2), the matrix completion algorithm's prediction outperforms the average prediction over all the climate models. Notably, the variances of both methods are driven down, versus the annual experiment. This is likely due to the 12-fold increase in the amount of input data, and the similarly increased number of values per validation period, over which the results are averaged.

While this problem may bear some (superficial) similarity to regression, it is markedly different. In particular, to meaningfully fit a regression model, the data should include values for the target variable corresponding to the relevant regions of feature space. In contrast, for all times in the validation interval, although our method is trained on the feature values (a subset of climate model projections), the corresponding (target) observation variable is missing. Nevertheless it would be interesting to compare our results to the results of regression using various techniques and evaluation scenarios.

Since sparse matrix completion typically works in the presence of low intrinsic dimensionality, these encouraging results suggest that perhaps there is a low-dimensional projection along which the climate models cluster, or a lower dimensional subspace in which the most predictive information lies. This is consistent with recent evidence [12] that only a small number of climatological factors (known as predictive components [13]) have a significant affect on climate model "skill."

References

[1] GISTEMP. 2012. NASA GISS Surface Temperature Analysis. http://data.giss.nasa.gov/gistemp/.

[2] CMIP. 2007. The World Climate Research Programme’s (WCRP’s) Coupled Model Intercomparison Project phase 3 (CMIP3) multi-model dataset. http://www-pcmdi.llnl.gov/ipcc/about ipcc.php

[3] Monteleoni, C., Schmidt, G., Saroha, S., and Asplund, E. 2011. Tracking Climate Models. *Statistical Analysis and Data Mining: Special Issue on Best of CIDU* 4(4):72–392.

[4] McQuade, S., and Monteleoni, C. 2012. Global Climate Model Tracking using Geospatial Neighborhoods. *In AAAI Conference on Artificial Intelligence.*

[5] Raghunandan H. Keshavan, Andrea Montanari, and Sewoong Oh. Matrix completion from a few entries. *IEEE Trans. Inf. Theor.,*56(6):2980–2998, June 2010.

[6] Reichler, T., and Kim, J. 2008. How well do coupled models simulate today’s climate? *Bull. Amer. Meteor. Soc.,* 89:303–311.

[7] Reifen, C., and Toumi, R. 2009. Climate projections: Past performance no guarantee of future skill? *Geophys. Res.Lett., 36.*

[8] Schneider, T. 2001. Analysis of incomplete climate data: Estimation of mean values and covariance matrices and imputation of missing values. *J. Climate, 14, 853-887.*

[9] Smerdon, J.E., and A. Kaplan. 2007. Comment on “Testing the fidelity of methods used in proxy-based reconstructions of past climate”: The role of the standardization interval*, Journal of Climate, 20(22), 5666-5670.*

[10] Tingley, M. P., Craigmile, P. F., Haran, M., Li, B., Mannshardt- Shamseldin E., and Rajaratnam, B. 2012. Piecing together the past: Statistical insights into paleoclimatic reconstructions*, Quaternary Science Reviews, 35, 1-22.*

[11] Shan, H., Kattge, J., Reich, P. B., Banerjee, A., Schrodt, F. and Reichstein, M. 2012. Gap Filling in the Plant Kingdom: Trait Prediction Using Hierarchical Probabilistic Matrix Factorization*, In International Conference on Machine Learning (ICML).*

[12]L Jia, T. DelSole, and M. K. Tippett. Can optimal projection improve dynamical model forecasts? *Submitted to J. Climate,*2013.

[13]Timothy DelSole and Michael K. Tippett. Predictability: Recent insights from information theory. *Reviews of Geophysics,*45(4), 2007.