LECTURE 9: Machine learning for paleoclimate

ML-4430: Machine learning approaches in climate science

22 June 2021
Climate Field Reconstructions

1. Guillot, Rajaratnam & Emile-Geay, 2015

Paleoclimate Networks

2. Rehfeld, Molkenthin & Kurths, 2014
STATISTICAL PALEOCLIMATE RECONSTRUCTIONS VIA MARKOV RANDOM FIELDS

BY DOMINIQUE GUILLOT*,1,2, BALA RAJARATNAM*,2 AND JULIEN EMILE-GEAY†,3

Stanford University* and University of Southern California†

1. Guillot, Rajaratnam & Emile-Geay, 2015
1. Guillot, Rajaratnam & Emile-Geay, 2015 → The challenge
Temperature and proxy as multivariate normal random variates

\[(X_1, X_2, \ldots, X_p) \sim N_p(\mu, \Sigma)\]

\[\mu = (\mu_1, \mu_2, \ldots, \mu_p)\]

\[\Sigma = (\sigma_{ij})_{pxp}\]

\[n = n_a + n_m\]

\[p \approx 3000, n \approx 2000, n_a = 150\]

where \(p\) → number of locations, and
\(n\) → number of years
\(n_a\) → number of years with available data
\(n_m\) → number of years with missing data
For large $p$, small $n$, sample covariance matrix is a poor estimator of $\Sigma$

$$p \approx 3000, \, n \approx 2000, \, n_a = 150$$
**Expectation step:** linear regression

(1.1) \[ (x_m - \mu_m^{(l)})^\top = B^{(l)}(x_a - \mu_a^{(l)})^\top, \]

where

\[ B^{(l)} = \Sigma_{ma}^{(l)} \Sigma_{aa}^{(l)} \Sigma_{mm}^{(l)} = \left( \begin{array}{cc} \Sigma_{aa}^{(l)} & \Sigma_{am}^{(l)} \\ \Sigma_{ma}^{(l)} & \Sigma_{mm}^{(l)} \end{array} \right), \]

and

(1.2) \[ \mu^{(l)} = (\mu_a^{(l)}, \mu_m^{(l)}). \]

Here, \( l \) is the index for the current iteration number.

Initial guesses:
- \( \mu^{(0)} \) : sample mean, replacing all missing values:
- \( \Sigma^{(0)} \) : sample covariance
Maximisation step: update of $\mu$ and $\Sigma$

\[
\mu_i^{(l+1)} = \frac{1}{n} \sum_{k=1}^{n} X_{ki}^{(l+1)},
\]

\[
\Sigma_{ij}^{(l+1)} = \frac{1}{n} \sum_{k=1}^{n} [(X_{ki}^{(l+1)} - \mu_i^{(l+1)})(X_{kj}^{(l+1)} - \mu_j^{(l+1)})] + C_{ij}^{(l+1)},
\]

where $C_{ij}^{(l+1)}$ is the covariance of the residuals. Using the same block decomposition as in (1.2), we have

\[
C^{(l+1)} = \begin{pmatrix}
0 & 0 \\
\Sigma_{mm}^{(l)} - \Sigma_{ma}^{(l)}(\Sigma_{aa}^{(l)})^{-1}\Sigma_{am}^{(l)} & \Sigma_{ma}^{(l)} - \Sigma_{ma}^{(l)}(\Sigma_{aa}^{(l)})^{-1}\Sigma_{am}^{(l)}
\end{pmatrix}.
\]
But we do not have a good estimate for $\Sigma$

- Use regularised regression (RegEM & RegEM-TTLS)
- Here, we use Gaussian Markov Random Fields to estimate $\Sigma$
- First, consider the precision matrix,
  $\Omega = (\omega_{ij}) = \Sigma^{-1}$
- Next, note that the partial correlation coefficient $\rho_{ij}$ between $i$ and $j$ given all the other observables is given by
  $\rho_{ij | \text{rest}} = -\omega_{ij} / (\omega_{ii} \omega_{jj})^{-1/2}$
  i.e. $X_i$ and $X_j$ are independent given the rest of the data if the corresponding entry in $\Omega$ is zero, and vice versa!
The graph-based estimator for $\Sigma$

$$\hat{\Sigma}_G = \arg\max_{\Omega > 0} \log \det \Omega - \text{tr}(S\Omega),$$

where $S$ is the sample covariance matrix of $x_1, \ldots, x_n$, given by

$$S = \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})(x_i - \bar{x})^T,$$

But we need to estimate $\Omega$ first.
Estimating $\Omega$

\[
(2.1) \quad \max_{\Omega > 0} l(\Omega) - \rho \|\Omega\|_1,
\]

where $\Omega = \Sigma^{-1}$ denotes the precision matrix of the data, $l(\Omega)$ is the normal log-likelihood of $\Omega$, $\rho > 0$ is a regularization parameter, and $\|\Omega\|_1$ is the 1-norm of $\Omega$:

\[
(2.2) \quad \|\Omega\|_1 = \sum_{i=1}^{P} \sum_{j=1}^{P} |\omega_{ij}|.
\]

(2.4) \quad \max_{\Omega > 0} l(\Omega) - \rho_{TT} \|\Omega_{TT}\|_1 - 2\rho_{TP} \|\Omega_{TP}\|_1 - \rho_{PP} \|\Omega_{PP}\|_1,

\[
\Omega = \begin{pmatrix}
\Omega_{TT} & \Omega_{TP} \\
\Omega_{PT} & \Omega_{PP}
\end{pmatrix},
\]

1. Guillot, Rajaratnam & Emile-Geay, 2015 → The GraphEM algorithm
Evaluating GraphEM

- Pseudoproxy measurements
  - NCAR CSM 1.4 model experiment
  - Last millennium (850–1980 AD)
  - 5° × 5° grid

- Most recent 150 years used for calibration
- The 981 years before that are reconstructed using GraphEM
- Benchmark comparison: RegEM-TTLS

- Performance metrics
  - Mean Squared Error (MSE)
  - Reduction of Error (RE)
  - Coefficient of Efficiency (CE)

\[ P(l, t) = T(l, t) + \frac{1}{\text{SNR}} \cdot \xi(l, t), \]

\[ \text{RE}(l) = 1 - \frac{\text{MSE}(\hat{T})(l)}{\text{MSE}(T_c)(l)}. \]

\[ \text{CE}(l) = 1 - \frac{\text{MSE}(\hat{T})(l)}{\text{MSE}(T_v)(l)}. \]
<table>
<thead>
<tr>
<th>Method</th>
<th>MSE</th>
<th>RE</th>
<th>CE</th>
<th>Bias</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \ell_1 ) method</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>GraphEM (0.3% target sparsity)</td>
<td>0.44 (0.01)</td>
<td>0.33 (0.01)</td>
<td>0.11 (0.02)</td>
<td>0.09 (0.01)</td>
</tr>
<tr>
<td>GraphEM (0.5% target sparsity)</td>
<td>0.42 (0.01)</td>
<td>0.36 (0.01)</td>
<td>0.15 (0.01)</td>
<td>0.08 (0.01)</td>
</tr>
<tr>
<td>GraphEM (0.7% target sparsity)</td>
<td>0.41 (0.01)</td>
<td>0.36 (0.01)</td>
<td>0.16 (0.01)</td>
<td>0.08 (0.01)</td>
</tr>
<tr>
<td>GraphEM (0.9% target sparsity)</td>
<td>0.41 (0.01)</td>
<td>0.36 (0.01)</td>
<td>0.15 (0.01)</td>
<td>0.08 (0.01)</td>
</tr>
<tr>
<td>Neigh</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>GraphEM (600 km radius)</td>
<td>0.42 (0.01)</td>
<td>0.35 (0.01)</td>
<td>0.14 (0.01)</td>
<td>0.06 (0.01)</td>
</tr>
<tr>
<td>GraphEM (800 km radius)</td>
<td>0.39 (0.01)</td>
<td>0.39 (0.01)</td>
<td>0.19 (0.01)</td>
<td>0.06 (0.01)</td>
</tr>
<tr>
<td>GraphEM (1000 km radius)</td>
<td>0.40 (0.01)</td>
<td>0.38 (0.01)</td>
<td>0.18 (0.01)</td>
<td>0.06 (0.01)</td>
</tr>
<tr>
<td>GraphEM (1200 km radius)</td>
<td>0.41 (0.01)</td>
<td>0.36 (0.01)</td>
<td>0.16 (0.01)</td>
<td>0.06 (0.01)</td>
</tr>
<tr>
<td>RegEM-TTLS</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>RegEM-TTLS</td>
<td>0.84 (0.10)</td>
<td>−0.24 (0.14)</td>
<td>−0.61 (0.19)</td>
<td>0.01 (0.02)</td>
</tr>
</tbody>
</table>
1. Guillot, Rajaratnam & Emile-Geay, 2015 → Results
Global Average Temperature

- Guillot, Rajaratnam & Emile-Geay, 2015 → Results
Climate Field Reconstructions

1. Guillot, Rajaratnam & Emile-Geay, 2015

Paleoclimate Networks

2. Rehfeld, Molkenthin & Kurths, 2014
Testing the detectability of spatio–temporal climate transitions from paleoclimate networks with the START model

K. Rehfeld$^{1,2,3}$, N. Molkenthin$^{1,2}$, and J. Kurths$^{1,2}$

$^1$Potsdam Institute for Climate Impact Research, P.O. Box 601203, 14412 Potsdam, Germany
$^2$Department of Physics, Humboldt-Universität zu Berlin, Newtonstr. 15, 12489 Berlin, Germany
$^3$Alfred-Wegener-Institut Helmholtz-Zentrum für Polar- und Meeresforschung, Telegrafenberg A43, 14473 Potsdam, Germany
**Challenge ...**

- Paleoclimate networks can be constructed from sparsely distributed proxy time series
- How well do they capture the dynamics of the system?
- If we had all the data from the region, and constructed a climate network, would the results be the same?
**START model**

- Spatiotemporal autocorrelated time series (START) model

- From expert knowledge, consider 3 sources of wind (perturbations) in the Asian monsoon domain:
  - Westerlies (X)
  - ISM (Y)
  - EASM (Z)

- At each source, place a Gaussian wave front propagating through the domain using the advection-diffusion equation

- Innately, model internal dynamics at each location as a autoregressive process
unidirectional front with a velocity at position $p$ and time point $t$

$$v_X(t, p, m_X, W) = m_X e^{-(p_x-p_{0,x})^2/2W}, \quad (14)$$

with a full width at a half maximum of $2 \sqrt{W \log 2}$. The maximal amplitude of the velocity, $m_X$,

$$m_X(t, F) = m_X = B_X + \alpha F, \quad (15)$$

is found in the center of the Gaussian front, as in Fig. 3. Here, $B_X$ is the baseline strength of the component’s flow, and $\alpha$ is its amplitude, or susceptibility to the external forcing, repre-
Signal at point $i$ ITCZ component

$$R_i = f_X(F, i) R_X + f_Y(F, i) R_Y + f_Z(F, i) R_Z + R_{\text{noise}} \quad (16)$$

Westerlies component EASM component

2. Rehfeld, Molkenthin & Kurths, 2014 → START model
Experimental setup

- Model runs are sampled in two different ways:
  - In the form of a regular grid
  - Locations of paleoclimate records from previous studies

- Three different experiments:
  - ISM off: Westerlies source forcing is removed => $F = -1$
  - Coexistence: All three sources are kept => $F = 0$
  - ISM on: Sources Y and Z are removed => $F = 1$

- Climate networks are constructed:
  - From the gridded samples
  - From the paleoclimate record locations samples

2. Rehfeld, Molkenthin & Kurths, 2014 → Model setup
2. Rehfeld, Molkenthin & Kurths, 2014 → Results
Rehfeld, Molkenthin & Kurths, 2014 → Results
Climate Field Reconstructions

1. Guillot, Rajaratnam & Emile-Geay, 2015

Paleoclimate Networks

2. Rehfeld, Molkenthin & Kurths, 2014