Testbed for Detection and Attribution Methods

Nathan Lenssen

December 2017

0.1 Model and parameters of interest

Throughout this discussion, we represent matrices by capital, bold letters such as $\mathbf{C}$ and $\boldsymbol{X}$, columns of these vectors by subscripts of the corresponding lowercase letter where $\boldsymbol{x}_m$ is the m^{th} column of the matrix $\boldsymbol{X}$. Vector quantities are denoted by lower case bold letters such as $\boldsymbol{y}$ and $\boldsymbol{\beta}$, in agreement with the notation of matrix columns. Scalar elements of vectors are denoted of plain lowercase with subscripts as in β_m.

The historical formation of the detection and attribution problem is that the observed climate is a consequence of responses to external forcings. The signficant of each of the forced responses is found by the ordinary least square solution (OLS) of the linear model

$$\boldsymbol{y} = \boldsymbol{X}^*\boldsymbol{\beta} + \boldsymbol{u} \tag{1}$$

In this relationship, $\boldsymbol{y}$ is the observed climate response to the various forced responses $\boldsymbol{X}^*$. We use the 'star' in $\boldsymbol{X}^*$ as a reminder that the OLS formulation assumes that the forced responses are known, nonrandom quantities (without error).

In general, the climate responses $\boldsymbol{y}$ are some sort of observational data. In contrast, the forced responses that comprise the $\boldsymbol{X}^*$ matrix are typically experiments from climate models where each of the columns $\boldsymbol{x}^*_m$ represent the outcome of an independent forcing experiment run in a climate model. The problem is difficult statistically as we cannot measure either the climate response or the forced responses precisely; there are various uncertainties that need to be accounted for.

In the OLS model (following Allen and Tett), we assume that we have random error according to the mean-zero internal climate variability $\mathbf{C}$.

$$\begin{aligned} \boldsymbol{y} &= \boldsymbol{X}^*\boldsymbol{\beta} + \boldsymbol{u} \\ &= \left(\sum_{m=1}^{M} \boldsymbol{x}^*_m \beta_m\right) + \boldsymbol{u}\,, \qquad \boldsymbol{u} \sim \mathcal{N}(0, \mathbf{C}) \end{aligned} \tag{2}$$

Where the second line is rewriting the The critical step in inference on β is estimation of the climate variability. Since the dimensionality of the problem is generally much greater than the sample of control runs, empirical estimates of the climate variability will likely be non-invertible. The primary technique used is a pseudo-inverse method involving the eigen-decomposition to estimate the precision matrix describing the climate variability.

The OLS model neglects that the forced responses are observed with error by assuming that they are perfectly known. However, climate model output is also subject to internal climate variability. The

class of statistical regression models where only noisy observations of the regressor are available is called error in variable (EIV) models. Making the reasonable assumption that the forced responses follow the same climate variability as the control runs, we have the model

$$\boldsymbol{y} = \sum_{m=1}^{M} (\boldsymbol{x}_m - \boldsymbol{u}_m)\beta_m + \boldsymbol{u}_0 \,, \quad \boldsymbol{u}_0, \boldsymbol{u}_1, \ldots, \boldsymbol{u}_M \overset{\text{iid}}{\sim} \mathcal{N}(0, \mathbf{C}) \tag{3}$$

This is known as the total least squares (TLS) problem and the solution is found algorithmically through optimization of the likelihood function. Like the OLS method, the climate variability $\mathbf{C}$ is assumed fixed during the estimation of $\boldsymbol{\beta}$ and needs to be estimated separately.

The formulation of the TLS model in equation 3 can be cumbersome to work with as the EIV structure is hidden all in one equation. To better see the how the various uncertainties effect the estimation of β, it can be helpful to formulate the problem in two equations: one describing the observed climate response and another describing the model-derived forced responses. Following the notation from above, we can rewrite our EIV model as

$$\begin{aligned} \boldsymbol{y} &= \boldsymbol{X}^*\boldsymbol{\beta} + \boldsymbol{\nu} \\ \boldsymbol{X} &= \boldsymbol{X}^* + \boldsymbol{U} \,, \qquad \boldsymbol{U} = (\boldsymbol{u}_1, \ldots, \boldsymbol{u}_M) \overset{\text{iid}}{\sim} \mathcal{N}(0, \mathbf{C}) \end{aligned} \tag{4}$$

We will use this formulation for the remainder of the discussion as it transparently shows the different variabilities acting upon the observed climate response and the model generated forced responses. Note that we did not specify a distribution for the error on the observed response $\boldsymbol{y}$, denoted $\boldsymbol{\nu}$ in equation 4. In the formulation of TLS presented in equation 3, $\boldsymbol{\nu} = \boldsymbol{u}_0$ and is a random error according to the climate variability $\mathbf{C}$. One of the current areas of interest in the statistical detection and attribution community is thinking more deeply about the form of the error on $\boldsymbol{y}$. For the purpose of the testbed, we have an error on the observed response $\boldsymbol{\nu}$ of the form

$$\text{Var}(\boldsymbol{\nu}) = \mathbf{C} + \mathbf{W}$$

Where $\mathbf{C}$ is the covariance matrix of the climate variability and $\mathbf{W}$ is the covariance matrix of the observational error (It is also proposed to include the climate model error and linearization error).

As a sanity check that we are describing the same system, we want to make our expanded EIV formulation in the form of the TLS problem. Making the statistical models in equation 3 and 4 agree, we write our classical TLS formulation as

$$\begin{aligned} \boldsymbol{y} = \sum_{m=1}^{M} (\boldsymbol{x}_m - \boldsymbol{u}_m)\beta_m + \boldsymbol{\nu}' \,, \quad \boldsymbol{u}_1, \ldots, \boldsymbol{u}_M &\overset{\text{iid}}{\sim} \mathcal{N}(0, \mathbf{C}) \\ \boldsymbol{\nu}' &\sim \mathcal{N}(0, \mathbf{W}) \end{aligned} \tag{5}$$

which results in an error on the observed $\boldsymbol{y}$ of the form

$$\boldsymbol{\nu} = (\boldsymbol{u}_0 + \boldsymbol{\varepsilon}) \sim \mathcal{N}(0, \mathbf{C} + \mathbf{W})$$

where $\boldsymbol{\varepsilon} \sim \mathcal{N}(0, \mathbf{W})$ represents the observational error.

I am almost certain my formulation in equation 5 is incorrect. Would love to discuss to figure out where I am losing track of the variance!

0.2 Estimation of Climate Variability

Classical detection and attribution methods are predicated on the assumption that the climate variability $\mathbf{C}$ is a known quantity.

Control runs...

Finish this section later!!!!

0.3 Model Ensembles

So far, our discussion has not addressed how multi-run ensembles of climate model output are incorporated into the EIV framework. Climate ensembles come into play in the forced response experiments as well as the control runs used to estimate the climate variability. We denote ℓ^{th} run of the m^{th} forced response experiment as $\boldsymbol{x}_m^{(\ell)}$. Likewise, we denote the ℓ^{th} control run as $\boldsymbol{x}_0^{(\ell)}$. Similarly, the total number of runs in the ensemble for the m^{th} forced response is given by L_m and the total number of control runs is L_0. This notation is for a single climate model, but can be extended to multimodel ensembles.

0.4 Testbed Output

The testbed returns two groups of data objects: objects that mimic the variables that are available in real-data problems, and objects that are unknown, or latent, in real-data problems and are used in the testing setting to determine the performance of methods. As mentioned in the previous section, we prefer viewing the problem through the expanded EIV formulation

$$\begin{aligned} \boldsymbol{y} &= \boldsymbol{X}^*\boldsymbol{\beta} + \boldsymbol{\nu} \qquad \boldsymbol{\nu} \sim \mathcal{N}(0, \mathbf{C} + \mathbf{W}) \\ \boldsymbol{X} &= \boldsymbol{X}^* + \boldsymbol{U}\,, \qquad \boldsymbol{U} = (\boldsymbol{u}_1, \ldots, \boldsymbol{u}_M) \overset{\text{iid}}{\sim} \mathcal{N}(0, \mathbf{C}) \end{aligned} \tag{6}$$

as it clearly separates the error processes for the observed climate responses $\boldsymbol{y}$ and the M model generated forced responses $\boldsymbol{X} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_M)$. The formulation in equation 6 is that it clearly separates variables that are known and unknown. We will also see that this formulation serves as pseudocode for the simulations required for our testbed.

- **Observed Data Objects:** These objects simulate the data that we have to work with in real-world (non-simulation) dettection and attribution problems. They are the only data used when running a detection and attribution method.

  $\boldsymbol{y}$: The observed climate response. Generally some sort of observational product derived from station and remote sensed data.

  $\boldsymbol{X}$: The forced climate response according to climate model output. There are K different forced responses where the k^{th} forced response has an ensemble of size L_k

  $\boldsymbol{x}_0$: Control runs of climate models used to estimate the climate variability $\mathbf{C}$. We have L_0 of these runs where

  $$\boldsymbol{x}_0 = (\boldsymbol{x}_0^{(1)}, \ldots, \boldsymbol{x}_0^{(\ell)}, \ldots, \boldsymbol{x}_0^{(L_0)})$$

- **Latent Data Objects:**

$\boldsymbol{y}^*$: The true climate response from $\boldsymbol{y}^* = \boldsymbol{X}^*\boldsymbol{\beta}$

$\boldsymbol{X}^*$: The true forced responses. Again, there are K forced responses

$\mathbf{C}$: The true covariance matrix of climate variability

$\mathbf{W}$: The true observational error covariance on the climate response

0.5 Testbed Modules

The testbed has three major modules that are used to generate artificial data to be used for the testing of detection and attribution methods. They are all independent and individually tunable and replaceable to allow greater flexibility of the testbed.

(M1) **Forced Response $\boldsymbol{X}^*$:**

(M2) **Internal Climate Variability C:**

(M3) **Observational Error W:**

0.6 Testbed Computational Parameters

A list of the parameters that can be tuned in the Matlab code used to generate the testbed data. There are four groups of parameters, each controlling a separate part of the process. The climate variability and forced response modules are interchangeable to allow for more flexibility to the simulated scenarios.

Problem Size/Dimensionality:
The problem is determined on the square domain of $[0,1]^2$*. Code runs in less than 10 seconds with* $n = 2500$*, but slows quickly past due to some large matrix inversions. Setting the number of response and control runs high allows for robustness analysis of models to ensemble size without rerunning the testbed.*

`n` Total number of spatial locations

`q` Number of locations per grid dimension where $\texttt{q} = \sqrt{\texttt{n}}$

`L0` Number of control runs (previously fixed at $L_0 = 300$)

Regression Parameters

`M` Number of response patterns (previously `p`)

`L` Vector with number of simulations per response pattern $L = (L_1, \ldots, L_M)$ (previously `m`)

`betaTrue` True regression response parameters (previously `beta0`) $[M \times 1]$

`sigmaW` Variance of the (iid) observational error on the climate response $[1 \times 1]$

Climate Variability Simulation [C]
We are currently using an exponential kernel parameterized by d where we modify the top `nx` *eigenvalues by multiplicative factors. The code is written so that different climate variability schemes can be dropped in as Matlab functions.*

`dExp` Parameter controlling the scale of the exponential kernel

`nx` Eigenvalue cutoff for the rectangular basis

`lambda` Multiplicative eigenvalue modifications

`delta` Climate variability whitening parameter $\rho \in [0,1]$ with $\rho = 0$ corresponding to spatially independent field. (previously `rho`)

Forced Response Simulation [$\boldsymbol{X}^*$]
We are currently using Matérn random fields to simulate forced responses due to their speed and flexibility. Each response is individually parameterized. As in the climate variability, it is easy to change the method by a single-line edit.

`alphax` Inverse of the range of the correlation $[1 \times M]$

`smoothnessx` Number of derivatives used in Matérn expansion $[1 \times M]$

`xscale` Magnitude of each forced response $[1 \times M]$

`gammaC` Noise magnitude of each forced response $[1 \times M]$

0.7 Parameter Settings

Fixed Parameter Settings:

Dimensionality: $(q, n, L, L_0) = (50, 2500, 25, 1000)$

Regression: $(M, \beta_0) = (2, [1, 1])$

0.8 Features to Show Off

- Variety of possible covariance functions that we can generate. Generate a really strange, non-isotropic one!!!
  - Looks like cranking up the d parameter on the exponential kernel way above 1 and making the deviations of the eigenvalues very high creates some crazy covariance fields.
  - Settings that make a crazy plot are:

    ```
    dExp = 5;
    nx = 100;
    rng(251); lambda = exp(unifrnd(-2.5,2.5,nx,1));
    ```

  - See `code/presentationExperiments/climVarExperimentation.m` for simulation code and plotting
- Show a variety of Matérn fields. End up picking two that are pretty different in terms of correlation range so that we can maybe see the pattern of each when added together.
  - This is quite easy to tweak other than the relative scale of the two fields which still needs to be *ad hoc* tuned
  - Adjusting by the average square value of the fields seems like a reasonable hacky fix.
  - Code for experimentation and plotting can be found in
- Use a variety of ensemble sizes of the forced responses to show how estimates of the forced responses through the ensemble mean alone improve. (whiteness comes into play here)
- Think about how the current figures can be put into a presentation to show how we are simulating the final observed and forced responses.
- Show how estimation of the climate variability improves with use of more control simulations

0.9 Questions to answer

- Should we be whitening **C**, the climate variability? If so, what is the physical meaning behind this whitening?
- Do we still need both the x-scale and γ_C parameters? Looks like probably: x-scale modifies the magnitude of the signal, γ_C modifies the magnitude of the noise. There may be a different way to parameterize this, but we do need two parameters to set the relative scale of the forced responses as well as the signal-to-noise ratio of each of the forced responses.
  - In a similar vein, is there a smarter way to balance the relative scales of the Matérn generated forced responses? How do you take the norm of a matrix? What norm could we use to fix the ratio between the two fields?
  - Could this philosophy be used to fix a signal-to-noise ratio in a more robust way as well?
- I'm starting to get in trouble with language. I want to refer to observed and latent variables/objects, but we are also using the word observed in "observed climate response".

0.10 TODO

(1) Quickly get a simulation to Dorit that can be used for testing of the Bayesian method
   - Ask about sample sizes wanted
   - Fix climate variability and pattern, then manually tweak the relative orders of magnitude (use `norm()` to set relative scales possibly?)

(2) Outline Presentation to determine what Figures I need
   - Write code for the generation of these Figures

(3) **Bonus:** Develop a more robust method for balancing signal/noise(s)

(4) **Super Bonus:** Rewrite code so that we have a L_m setting rather than just a L for all forced response experiments.