Robust Statistical Methods in the Geosciences

Noel Cressie
(ncressie@uow.edu.au)

National Institute for Applied Statistics Research Australia
University of Wollongong

International Conference on Robust Statistics 2017

- **Resistance**: John Tukey liked this term to describe statistics that were insensitive to drastic changes in just a few data.

- **Robustness**: Insensitivity of inference to small changes to a central model or model class.

Geophysicists want resistance, but we should sell them robustness.
1 The primary consumers for this L4 flux product are going to be plant physiologists, ecosystem biologists, folks who study the impact of climate anomalies on CO$_2$ fluxes, etc.

. . . . . . . .

2 By default, the web portal will provide download links to the model median/mean and range. Downloading individual model results will be a few clicks away, but possible.

. . . . . . . .

3 Regarding [#2], it is still not decided whether we will provide the model median or mean by default. We need to plot both and see if outliers screw up one over the other.
Sample medians, $\tilde{x} \equiv \text{med}\{x_1, \ldots, x_n\}$, have distributions, but when it comes time to make scientific inferences with them, I have seen rocket scientists use:

$$\text{var}(\tilde{x}) = S^2 / n,$$

where

$$S^2 = \sum_{i=1}^{n} (x_i - \tilde{x})^2 / (n - 1).$$

Suppose data $\{x_1, \ldots, x_n\}$ are spatially dependent. I have seen variances of sample means, $\bar{x} \equiv \text{ave}\{x_1, \ldots, x_n\}$ (and sample medians), put equal to $S^2$ (in place of $S^2 / n$).

More sophisticated inferences can involve bootstrapping.
Need for Robustness

- Tropical Pacific *Sea Surface Temperatures* (SSTs) and their seasonal signatures

- Remote sensing of *atmospheric CO₂* and its retrieval from an individual sounding

- Calibration/validation of ground-based instruments looking up, with satellite-based instruments looking down [see Bohai Zhang’s talk this afternoon after the coffee break]
Tropical Pacific SSTs


- SST is an important component of many global climate models.

- **El Niño** represents the distribution of warmer-than-average waters in the eastern tropical Pacific Ocean. (La Niña represents a cooler eastern tropical Pacific.)

- **Anomalies** are traditionally used to analyze SSTs: For example, there are 30 “Jan” SST values at pixel \( s \) during the anomaly period, 1971-2000. Average them to give \( \overline{SST}^{Jan}(s) \). Then, if \( y_{2017}^{Jan}(s) \) is the SST value in January 2017 at pixel \( s \), the **SST anomaly** is

\[
a_{2017}^{Jan}(s) = y_{2017}^{Jan}(s) - \overline{SST}^{Jan}(s).
\]
Figure: Niño 3.4 Region is defined by the black box.
Monthly raw SST data in the Niño 3.4 Region are given by \( \{ y_t^M(s_i) : i = 1, \ldots, 500 \} \), where \( M \in \{ \text{Jan}, \text{Feb}, \ldots, \text{Dec} \} \).

Look for a mean-standard deviation (sd) relation in the raw data; it would not be obvious from anomalies.

**Spatial Diagnostic Plot:**

Spatial mean: \( \bar{y}_t^M = \frac{\sum_{i=1}^{500} y_t^M(s_i)}{500} \).

Spatial sd: \( S_t^M = \left\{ \frac{\sum_{i=1}^{500} (y_t^M(s_i) - \bar{y}_t^M)^2}{499} \right\}^{1/2} \).

Make x-y plots, where

\[
\begin{align*}
y &= \text{spatial sd} \\
x &= \text{spatial mean},
\end{align*}
\]

and stratify \( \{ \text{Jan}, \text{Feb}, \ldots, \text{Dec} \} \) by year, seasons, and months.
Spatial Diagnostic Plot: Year

\[ S_z(t) \]

\[ z(t) \]

Cressie (UOW)  Robust Statistical Methods in the Geosciences
Spatial Diagnostic Plot: Seasons

- December
- January
- February
- March
- April
- May
- June
- July
- August
- September
- October
- November

$S_z(t)$ vs $\bar{z}(t)$

Cressie (UOW) Robust Statistical Methods in the Geosciences
Spatial Diagnostic Plots: Month
The negative slope in the spatial diagnostic plot is consistent from year, to seasons, to months. There are enough data available to stratify on months; our analysis was for November 1981-December 2014.

Some months exhibit outliers.

The fitted lines shown on the previous slide were obtained using the robust Thiel-Sen method, discussed here for estimating the slope $\beta$ in the generic linear regression,

$$y_i = \alpha + \beta x_i + \epsilon_i \;; \; i = 1, \ldots, n.$$
Convert the problem to **location estimation**:

\[
\frac{y_i - y_j}{x_i - x_j} = \beta + \frac{\epsilon_i - \epsilon_j}{x_i - x_j} ; \ i = 1, \ldots, n,
\]

and take the weighted average (wt \(\propto (x_i - x_j)^2\)) or the weighed median (wt \(\propto |x_i - x_j|\)) of the “data,” \(\{(y_i - y_j)/(x_i - x_j)\}\).

- The **weighted Thiel-Sen estimator** is:

\[
\hat{\beta}_{WTS} = \text{wt med}\{(y_i - y_j)/(x_i - x_j)\},
\]

with weights \(|x_i - x_j|\).

- Note that the **OLS estimator** is \(\text{wt ave}\{(y_i - y_j)/(x_i - x_j)\}\) with weights \((x_i - x_j)^2\).
From the Diagnostic Plot, we observe that the raw SSTs exhibit heteroskedasticity, where the variance depends on the mean.

What transformation, \( \{ f(y^M_t) \} \), of the raw data would remove the heteroskedasticity? Motivation: On the transformed scale, the SST dynamics are likely to be linear and additive.

Let \( x \) be a r.v. with mean \( \mu_x \) and variance \( \sigma^2_x \), and suppose
\[
\sigma^2_x = h(\mu_x) \quad \text{(i.e., heteroskedasticity)}.
\]
Then the \( \delta \)-method gives,

\[
\text{var}(f(x)) \sim \text{const},
\]

for

\[
f(x) \propto \int^x h(\mu_x)^{-1/2} d\mu_x.
\]
From the Diagnostic Plot, put $h(\mu_x) = (\alpha + \beta \mu_x)^2$. Hence,

$$f(x) \propto \log(\alpha + \beta x).$$

Each $M \in \{Jan, Feb, \ldots, Dec\}$ has a (robust) estimate $\hat{\alpha}^M_{WTS}$ and $\hat{\beta}^M_{WTS}$. Hence transform the month-$M$ data $\{y_t^M(s)\}$ to

$$v_t^M(s) \equiv \log(\hat{\alpha}^M_{WTS} + \hat{\beta}^M_{WTS} y_t^M(s)).$$

On the transformed scale (i.e., the $v$-scale), the dynamics are expected to follow an AR($p$) process with higher forecast accuracy than if the same process “were” fitted on the original $y$-scale.

On the $v$-scale, we expect homoskedasticity. So we can interpret the spatial mean unambiguously (i.e., on the $v$-scale, the ocean is warm or cool independently of the spatial variance).
Figure: Slopes are non-zero; bootstrap confidence intervals are shown.
Figure: Slopes are zero; bootstrap confidence intervals are shown.
Figure: The spring barrier is apparent; bootstrap confidence intervals are shown.
Satellites remotely sense the Earth system.

Greenhouse gases in the atmosphere are of particular interest.

**CO$_2$** is a leading greenhouse gas that can be measured by satellite campaigns, such as Japan’s GOSAT and USA’s OCO-2.

Coverage is global but retrieval of CO$_2$ data is local (at the footprint level).
Figure: The OCO-2 launch (02:56 PDT, July 2, 2014) and the OCO-2 spacecraft (artist concept); images are from https://www.nasa.gov/content/oco-2-launch/
The raw radiance data (photon counts) are $y$ (approximately 3,000 dimensional), and the atmospheric state is $x$ (approximately 100 dimensional). Twenty state variables represent CO$_2$ values (in ppm) at different geopotential heights.

For each footprint, $y$ is observed (called a “sounding”), and an attempt is made to infer $x$ for the atmospheric column above the footprint.

Some attempts fail; those that succeed are called “retrievals.”

This talk will concentrate on the ACOS/OCO-2 retrieval.
Figure: ACOS retrieval locations of GOSAT data: June 5 – July 26, 2009
Nonlinear forward model:

\[ y = F(x) + \varepsilon. \]

Solve for \( x \) by minimizing the penalized weighted sum of squares. That is, minimize

\[
(y - F(x))' S_\varepsilon^{-1} (y - F(x)) + (x - x_\alpha)' S_\alpha^{-1} (x - x_\alpha),
\]

with respect to the state vector \( x \). The diagonal elements of \( S_\varepsilon \) (respectively, \( S_\alpha \)) are \( \{\sigma_\varepsilon^2,i\} \) (respectively, \( \{\sigma_j^2\} \)).

Rodgers (2000) gives two possible iteration schemes. (In practice, a convergence criterion has to be met.) ACOS/OCO-2 uses a Levenberg-Marquardt iteration scheme.
Figure: ABO₂, WCO₂, and SCO₂ data on 6 August, 2014 (first light)
Define the Jacobian, $J_{ij} \equiv \frac{\partial F_i(x)}{\partial x_j}$, and the unit-free Jacobian,

$$\phi_{ij} \equiv J_{ij} \sigma_j / \sigma_{\varepsilon,i}; \ i = 1, \cdots, n_\varepsilon, j = 1, \cdots, n_\alpha.$$ 

Let $\hat{J}_{ij}$ denote the retrieved Jacobian; then $\hat{\phi}_{ij} \equiv \hat{J}_{ij} \sigma_j / \sigma_{\varepsilon,i}$, is unit-free, and $\hat{\Phi} \equiv \{\hat{\phi}_{ij}\}$ is an $n_\varepsilon \times n_\alpha$ array.

For each $j$, we consider $\{\hat{\phi}_{ij} : i \in \text{band}\}$ to be a sample from a distribution with mean $\mu_{j}^{\text{band}}$, where $\text{band} = \text{OA}, \text{WC}, \text{SC}$.

We look for those $j$ for which we cannot reject the null hypothesis, $H_0 : \mu_{j}^{\text{band}} = 0$, versus $H_1 : \mu_{j}^{\text{band}} \neq 0$. This represents a way of classifying the Jacobian values as weakly sensitive or strongly sensitive to changes in the state. To find such $j$, we use the test statistic,

$$\hat{\phi}_{j}^{\text{band}} \equiv \text{med} \left\{ |\hat{\phi}_{ij}|^{1/2} : i \in \text{band} \right\}; \ \text{band} = \text{OA}, \text{WC}, \text{SC}$$ (see the next slide).
A near-zero Jacobian for a given state-vector element means that there is weak sensitivity in the radiances, to changes in that element:

\[ y \approx F(\hat{x}) + J(x - \hat{x}) + \varepsilon \]

\[ = F(\hat{x}) + J_1(x_1 - \hat{x}_1) + J_2(x_2 - \hat{x}_2) + \varepsilon \]

\[ \approx F(\hat{x}) + J_1(x_1 - \hat{x}_1) + \varepsilon, \]

where \( x = (x'_1, x'_2)' \), and \( x_2 \) are the state-vector elements that correspond to near-zero Jacobian entries. Retrieval of \( x_2 \) is difficult.

The elements may still be important in the forward equation, but their lack of sensitivity to changes makes the problem more ill-posed. These elements need strong priors in order for the Levenberg-Marquardt algorithm to converge.
For each fixed $j$, suppose $\{\hat{\phi}_{ij}: i \in \text{band}\}$ is a sample from the random variable, $\Phi_{j}^{\text{band}}$, where band $=$ oxygen A (OA), weak CO$_2$ (WC), and strong CO$_2$ (SC).

Suppose that $\Phi_{j}^{\text{band}}$ has distribution, 

$$\Phi_{j}^{\text{band}} \sim \text{Dist} \left( \mu_{j}^{\text{band}}, (\sigma_{j}^{\text{band}})^2 \right),$$

for all $j = 1, \cdots, n_{\alpha}$. Thus, we have multiple hypotheses:

$$H_0 : \mu_{j}^{\text{band}} = 0 \quad \text{v.} \quad H_1 : \mu_{j}^{\text{band}} \neq 0; \quad j = 1, \cdots, n_{\alpha}$$

and recall that band $=$ OA, WC, SC.

We first make the data more resistant. Define

$$D_{j}^{\text{band}} \equiv \left\{ |\hat{\phi}_{ij}|^{1/2} : i \in \text{band} \right\}.$$
Consider generic iid random variables $W_1, \cdots, W_m$, with distribution $\text{Gau}(\mu_W, \sigma^2_W)$. To test $H_0: \mu_W = 0$ vs. $H_1: \mu_W \neq 0$, use the robust test statistic based on a median,

$$\tilde{X} \equiv \text{med}\{|W_i|^{1/2} : i = 1, \cdots, m\}.$$ 

Under the null hypothesis $H_0: \mu_W = 0$,

$$|W_i|^{1/2} \sim \text{Gau}\left(0.82216\sigma^2_W^{1/2}, 0.12192\sigma_W\right),$$

approximately; see the next slide. 

Then, under $H_0$,

$$\tilde{X} \sim \text{Gau}\left(0.82216\sigma^2_W^{1/2}, \frac{0.12192\pi\sigma_W}{2m}\right),$$

approximately; see the next slide.
Aside: Some Distribution Theory

- See previous slide (second bullet): If $Y \sim \text{Gau}(0, 1)$, then

  $$E(|Y|^{1/2}) = E(|\chi_1^2|^{1/4}) = 0.82216$$
  $$\text{var}(|Y|^{1/2}) = \text{var}(|\chi_1^2|^{1/4}) = 0.12192.$$  

- See previous slide (third bullet): Suppose $X_1, \ldots, X_m \overset{iid}{\sim} \text{Dist}(\mu_X, \sigma_X^2)$, where “Dist” is symmetric. If $\tilde{X} \equiv \text{med}\{X_1, \ldots, X_m\}$, then

  $$E(\tilde{X}) = \mu_X, \quad \text{var}(\tilde{X}) \simeq \frac{1}{4mf^2(\mu_X)},$$

  where $f(\cdot)$ is the pdf of $X_1$. When Dist = Gau, $f^2(\mu_X) = 1/(2\pi)$, and hence

  $$\text{var}(\tilde{X}) \simeq \frac{\pi\sigma_X^2}{2m}.$$
Multiple hypothesis testing is done at the 1% significance level. For each hypothesis test, we conclude $H_1$ (i.e., strong sensitivity) when

$$\tilde{X} > 0.82216\sigma_{\tilde{W}}^{1/2} + \Phi^{-1}(1 - \alpha)\sqrt{\frac{0.12192\pi\sigma_{\tilde{W}}}{2m}}.$$ 

Here, $\alpha = 0.01/(112 \times 3)$, the Bonferroni-adjusted level.

The standard deviation, $\sigma_{\tilde{W}}$, needs to be estimated. To maintain robustness, we can use the relation, $\sigma_{\tilde{W}} = \sigma_{|W|^{1/2}}^2 / 0.12192$, to estimate $\sigma_{\tilde{W}}$ by first estimating $\sigma_{|W|^{1/2}}$.

We use the median absolute deviation (MAD) to estimate $\sigma_{|W|^{1/2}}^2$. Consequently, $\hat{\sigma}_W = \hat{\sigma}_{|W|^{1/2}}^2 / 0.12192$, where

$$\hat{\sigma}_{|W|^{1/2}} = \text{med} \left\{ \left| |W_i|^{1/2} - \tilde{X} \right| : i = 1, \ldots, m \right\} \cdot 1.4826.$$
Statistical Significance Filter, ctd

- We call this collection of robust hypothesis tests the Statistical Significance Filter.

- Use $\tilde{\phi}_j^{\text{band}} \equiv \text{med}\{|\hat{\phi}_{ij}|^{1/2} : i \in \text{band}\}$ as test statistic. If a combination (band, $j$) does not pass the Statistical Significance Filter (i.e., $H_0$ is true), then the associated state element $x_j$ is considered to be only weakly sensitive to changes in the state.

- If a combination (band, $j$) does pass the statistical significance filter (i.e., $H_1$ is true), then the associated state element $x_j$ has unit-free Jacobian values that are estimated by $\tilde{\phi}_j^{\text{band}} \equiv \text{med}\{\hat{\phi}_{ij} : i \in \text{band}\}$, for band = OA, WC, SC.

- This Jacobian analysis is applied to the ACOS B2.8 retrieval, where $n_\varepsilon = 2240$ and $n_\alpha = 112$, but it is applicable to any generic retrieval based on regularization. We now illustrate our methodology at location 1, one of the 30 shown over Australia.
Figure: Unit-free Jacobian estimates that pass the Statistical Significance Filter (location 1)
Figure: Unit-free Jacobian estimates that pass the Statistical Significance Filter (location 1)
Figure: Unit-free Jacobian estimates that pass the Statistical Significance Filter (location 1)
**Figure:** Unit-free Jacobian estimates that pass the Statistical Significance Filter (location 1)
Figure: Unit-free Jacobian estimates that pass the Statistical Significance Filter (location 1)

OA
WC
SC

Green: \( \mu_{\text{band}}^j = 0 \)
Red: \( \mu_{\text{band}}^j \neq 0 \)
Near-Zero Jacobian Entries: ACOS retrievals

- Our analysis for **location 1** yields near-zero Jacobian entries in all three bands (i.e., a continuous vertical green stripe) for:
  
<table>
<thead>
<tr>
<th>Location Numbers</th>
<th>Parameter</th>
</tr>
</thead>
<tbody>
<tr>
<td>1, 2, 3</td>
<td>CO₂ concentration</td>
</tr>
<tr>
<td>21</td>
<td>H₂O scale factor</td>
</tr>
<tr>
<td>23</td>
<td>Temperature offset</td>
</tr>
<tr>
<td>105, 107, 109</td>
<td>Albedo slope for the three bands</td>
</tr>
<tr>
<td>110, 111, 112</td>
<td>Spectral dispersion offset</td>
</tr>
</tbody>
</table>

- For the **30 retrievals** shown over Australia, we consistently see near-zero Jacobian entries in all three bands for:
  
<table>
<thead>
<tr>
<th>Location Numbers</th>
<th>Parameter</th>
</tr>
</thead>
<tbody>
<tr>
<td>21</td>
<td>H₂O scale factor</td>
</tr>
<tr>
<td>23</td>
<td>Temperature offset</td>
</tr>
<tr>
<td>105, 107, 109</td>
<td>Albedo slope for the three bands</td>
</tr>
<tr>
<td>110, 111, 112</td>
<td>Spectral dispersion offset</td>
</tr>
</tbody>
</table>
Conclusions

- Geophysicists want resistance, but inference for anything but OLS estimates is uncommon.

- Statisticians want resistance and valid inferences.

- Both are found in Robust Statistics!

