An Errors-in-Variables Model for Calibration/Validation of Satellite Remote Sensing Data

Bohai Zhang

National Institute for Applied Statistics Research Australia (NIASRA)
University of Wollongong (UOW), Australia

Joint work with Noel Cressie (NIASRA, UOW), and Debra Wunch (University of Toronto)

International Conference on Robust Statistics 2017
Remote sensing data can be collected on a global scale within a short time period. Such datasets help scientists understand the spatial distributions and temporal dynamics of environmental variables.

The Orbiting Carbon Observatory-2 (OCO-2) mission aims to produce precise, time-dependent measurements of atmospheric carbon dioxide (CO$_2$).

For many reasons, the retrieved OCO-2 data come with errors. Bias-correction procedures are required to produce more accurate data products.
The OCO-2 Mission

Figure: The OCO-2 launch and the OCO-2 space craft (artist concept); images are from https://www.nasa.gov/content/oco-2-launch/.
Figure: TCCON stations, produced by Reto Stöckli, NASA Earth Observatory (NASA Goddard Space Flight Center).
Figure: Points represent a TCCON value (ground truth) of column-averaged CO₂ (X) and an OCO-2 value (remote sensing) of column-averaged CO₂ (Y). The solid black line is the fitted regression line (Y = OCO-2 regressed on X = TCCON), and the red-dashed line is the reference line with slope 1.
An Errors-in-Variables (EIV) Model

- Satellite remote sensing data ($Y$) are often calibrated by using ground-station data/aircraft data ($X$). However, measurement errors are usually present in both $X$ and $Y$; cf. version 7 of OCO-2’s data products (Mandrake et al., 2015).

- Let $\{(X_i, Y_i)\}_{i=1}^{N}$ be $N$ pairs of observations from two data sources, with $X_i$ and $Y_i$ as the covariate and the response, respectively. Let

$$X_i = x_i + \epsilon_{x,i}, \quad Y_i = y_i + \epsilon_{y,i}.$$  \hspace{1cm} (1)

- The scalar quantities $x_i$ and $y_i$ are fixed but unknown mean parameters, related by:

$$y_i = a + bx_i.$$  \hspace{1cm} (2)

- For $i = 1, \ldots, N$, $E(\epsilon_{x,i}) = E(\epsilon_{y,i}) = 0$; $\text{var}(X_i) = \sigma_{x,i}^2$ and $\text{var}(Y_i) = \sigma_{y,i}^2$; $\epsilon_{x,i}$ and $\epsilon_{y,i}$ are mutually independent.
York (1966, 1968) proposed least-sum-of-weighted-squares (LWS) estimators of $a$ and $b$ that minimise

$$S(a, b) = \sum_{i=1}^{N} \frac{w_{x,i} w_{y,i}}{b^2 w_{y,i} + w_{x,i}} (Y_i - a - bX_i)^2.$$  

(3)

The regression weights, $w_{x,i}$ and $w_{y,i}$, are pre-specified, and they should be chosen to be the reciprocals of the variances of $X_i$ and $Y_i$, respectively, to ensure consistent estimation of regression coefficients (e.g., see Zhang et al., 2017).

The LWS estimators obtained from minimising (3) are also maximum (profile) likelihood estimators, under certain assumptions (Titterington and Halliday, 1979).
Motivations

- $X_i$ and $Y_i$ are typically aggregated data (summary statistics such as sample means/medians of individual observations).
- Temporal/spatial dependence may need to be accounted for when obtaining their variances.
- Systematic errors may also exist in $X_i$ and $Y_i$.

(a) TCCON time series

(b) OCO-2 observations around target
The Single-Covariate Case

- For $i = 1, \ldots, N,$
  \[ X_i = x_i + \eta_{x,i} + \epsilon_{x,i}, \quad Y_i = y_i + \eta_{y,i} + \epsilon_{y,i}, \quad y_i = a + bx_i. \quad (4) \]

- The systematic errors, $\eta_{x,i}$ and $\eta_{y,i},$ have mean zero and variances $\tau^2_x$ and $\tau^2_y,$ respectively; and $\epsilon_{x,i}$ and $\epsilon_{y,i}$ are mean-zero random errors with variances $\sigma^2_{\epsilon,x,i}$ and $\sigma^2_{\epsilon,y,i},$ respectively.

- The profile log-likelihood function is:
  \[
  \ell(\theta|\{X_i, Y_i\})
  = -\frac{1}{2} \sum_{i=1}^{N} \frac{(Y_i - a - bX_i)^2}{b^2(\sigma^2_{\epsilon,x,i} + \tau_x^2) + \sigma^2_{\epsilon,y,i} + \tau_y^2} - \frac{1}{2} \sum_{i=1}^{N} \log(\sigma^2_{\epsilon,x,i} + \tau_x^2)
  - \frac{1}{2} \sum_{i=1}^{N} \log(\sigma^2_{\epsilon,y,i} + \tau_y^2) + \text{constant}. \quad (5)
  \]
In making inference from the profile log-likelihood function, not all model parameters are identifiable.

\( \tau_x^2 \) and \( \sigma_{\epsilon,x,i}^2 \) are not identifiable, but their sum is identifiable. Similarly, \( \tau_y^2 \) and \( \sigma_{\epsilon,y,i}^2 \) are not identifiable; it can also be shown that \( b, \tau_x^2, \) and \( \tau_y^2 \) are not identifiable.

Estimation strategy:

1. Estimate \( \sigma_{\epsilon,x,i}^2 \) and \( \sigma_{\epsilon,y,i}^2 \) (random-error variances) based on datasets of individual observations from which \( X_i \) and \( Y_i \) are calculated.
2. Estimate \( \tau_x^2 \) from an independent validation dataset that is free of systematic errors.
3. Estimate the remaining parameters, \( \{a, b, \tau_y^2\} \), through the regression analysis of \( \{(X_i, Y_i) : i = 1, \ldots, N\} \).

At each stage, estimated parameters from the previous stages are substituted into the estimating equations as if they are known.
Fix $i$, and let $\mathcal{D}_{x,i} \equiv \{\tilde{X}_{i,1}, \ldots, \tilde{X}_{i,n_x,i}\}$ be the dataset of individual observations for calculating $X_i$. Similarly, let $\mathcal{D}_{y,i} \equiv \{\tilde{Y}_{i,1}, \ldots, \tilde{Y}_{i,n_y,i}\}$ be the dataset of individual observations for calculating $Y_i$.

For $i = 1, \ldots, N$, $k = 1, \ldots, n_{x,i}$, and $\ell = 1, \ldots, n_{y,i}$, assume that

$$
\tilde{X}_{i,k} = x_i + \eta_{x,i} + \tilde{\epsilon}_{x,i,k}, \quad \tilde{Y}_{i,\ell} = y_i + \eta_{y,i} + \tilde{\epsilon}_{y,i,\ell},
$$

where $E(\tilde{\epsilon}_{x,i,k}) = E(\tilde{\epsilon}_{y,i,\ell}) = 0$; and $\text{var}(\tilde{\epsilon}_{x,i,k}) = \tilde{\sigma}^2_{x,i}$ and $\text{var}(\tilde{\epsilon}_{y,i,\ell}) = \tilde{\sigma}^2_{y,i}$ (i.e., homogeneous variances within the datasets of individual observations).
For our remote-sensing application, $\tilde{X}_{i,k}$ and $\tilde{Y}_{i,\ell}$ are typically observed over time or space. In this case, parametric temporal/spatial covariance functions $C_x(\cdot, \cdot; \theta_{x,i,j})$ and $C_y(\cdot, \cdot; \theta_{y,i})$ are assumed for modelling dependence of individual observations.

The TCCON individual observations are free of outliers. For TCCON regression data, $X_i = \sum_{k=1}^{n_{x,i}} \tilde{X}_{i,k} / n_{x,i}$, is the sample mean. By assuming Gaussian measurement errors, REML (Patterson and Thompson, 1971; Harville, 1977) was applied to estimate $\theta_{x,i,j}$.

The OCO-2 individual observations can be contaminated by outliers. Hence, the sample median, $Y_i \equiv \text{med}\{\tilde{Y}_{i,\ell}\}$, was used to obtain the regression datum.
The OCO-2 individual observations \( \{ \tilde{Y}_{i,\ell} \equiv \tilde{Y}(s_{i,\ell}) : \ell = 1, \ldots, n_{y,i} \} \) are modelled as realisations from a spatial Gaussian process:

\[
\tilde{Y}_{i,\ell} = y_i + \eta_{y,i} + \tilde{\epsilon}_{y,i,\ell},
\]

where \( s_{i,\ell} \in \mathbb{R}^2 \) is a spatial location.

Let \( C_y(\cdot, \cdot; \theta_{y,i}) \) be the flexible isotropic Matérn covariance function (e.g., Stein, 1999; Matérn, 2013), which we use to evaluate \( \text{cov}(\tilde{Y}_{i,\ell}, \tilde{Y}_{i,\ell'}) \). That is, the spatial covariances between OCO-2 data are modelled by

\[
C_y(s_{i,\ell}, s_{i,\ell'}; \theta_{y,i}) = \frac{\tilde{\sigma}_{y,i}^2 2^{1-\nu_{y,i}}}{\Gamma(\nu_{y,i})} \left( \frac{\|s_{i,\ell} - s_{i,\ell'}\|}{\phi_{y,i}} \right)^{\nu_{y,i}} K_{\nu_{y,i}} \left( \frac{\|s_{i,\ell} - s_{i,\ell'}\|}{\phi_{y,i}} \right).
\]
The estimates of the covariance-model parameters, $\theta_{y,i}$, need to be robust to outliers. We use robust estimators of the semivariogram (e.g., Cressie and Hawkins, 1980; Cressie, 1993) and fit the covariance-function parameters using weighted least squares (Cressie, 1993).

The Cressie-Hawkins semivariogram estimator is:

$$
\hat{\gamma}(h(k)) = \frac{1}{2} \left( \frac{1}{|N(h(k))|} \sum_{N(h(k))} |\tilde{Y}_{i,j} - \tilde{Y}_{i,\ell}|^{1/2} \right)^4 / \left( 0.457 + \frac{0.494}{|N(h(k))|} \right),
$$

where $N(h(k)) \equiv \{(j, \ell) : s_{i,j} - s_{i,\ell} \in \text{tol}(h(k)), j, \ell = 1, \ldots, n_{y,i}\}$, and $\text{tol}(h(k))$ is a pre-specified tolerance region around the spatial lag $h(k)$. 

Bohai Zhang (UOW)
The covariance parameters $\theta_{y,i}$ are estimated by weighted least squares (Cressie, 1985); that is, minimise with respect to $\theta_{y,i}$,

$$W(\theta_{y,i}) = \sum_{k=1}^{K} |N(h(k))| \left( \frac{\hat{\gamma}(h(k))}{\gamma(h(k); \theta_{y,i})} - 1 \right)^2.$$

The random-error variance in $Y_i$ is $\sigma^2_{\epsilon, y, \ell} \equiv \text{var}(\text{med}\{\tilde{\epsilon}_{y,i,\ell}\})$. Under mild conditions (Sen, 1972; Cressie and Glonek, 1984), the asymptotic variance is,

$$\sigma^2_{\epsilon, y, i} \equiv \text{var}(\text{med}\{\tilde{\epsilon}_{y,i}\}) \simeq \frac{\pi \tilde{\sigma}^2_{y,i}}{2n_{y,i}} + \frac{\tilde{\sigma}^2_{y,i}}{n^2_{y,i}} \sum_{j=1}^{n_{y,i}} \sum_{k \neq j} \text{arcsin}(C_{y,i;j,k}/\tilde{\sigma}^2_{y,i}),$$

where recall that $C_{y,i;j,k} = \text{cov}(\tilde{\epsilon}_{y,i,j}, \tilde{\epsilon}_{y,i,k}) = C_y(s_{i,j}, s_{i,k}; \theta_{y,i})$ is given by the Matérn covariance function.
Random-Error Estimation Results

Figure: Empirical and fitted semivariogram plots for Lamont/3590.

Table: Parameter estimation results, sample variances, and the effective sample sizes, \( \tilde{n}_{x,i} \) and \( \tilde{n}_{y,i} \), for Lamont/3590.

<table>
<thead>
<tr>
<th>TCCON</th>
<th>( X_i )</th>
<th>( \hat{\sigma}^2_{x,i} )</th>
<th>( S^2_{x,i} )</th>
<th>( n_{x,i} )</th>
<th>( \tilde{n}_{x,i} )</th>
<th>( \hat{\sigma}^2_{x,i} )</th>
<th>( \hat{\phi}_{x,i} )</th>
<th>( \hat{\nu}_{x,i} )</th>
</tr>
</thead>
<tbody>
<tr>
<td>OCO-2</td>
<td>( Y_i )</td>
<td>( \hat{\sigma}^2_{y,i} )</td>
<td>( S^2_{y,i} )</td>
<td>( n_{y,i} )</td>
<td>( \tilde{n}_{y,i} )</td>
<td>( \hat{\sigma}^2_{y,i} )</td>
<td>( \hat{\phi}_{y,i} )</td>
<td>( \hat{\nu}_{y,i} )</td>
</tr>
<tr>
<td></td>
<td>401.0840</td>
<td>0.0063</td>
<td>0.3602</td>
<td>65</td>
<td>57.05</td>
<td>0.3607</td>
<td>0.0039</td>
<td>0.5 (fixed)</td>
</tr>
<tr>
<td></td>
<td>400.0395</td>
<td>0.0023</td>
<td>0.3025</td>
<td>2961</td>
<td>202.32</td>
<td>0.2989</td>
<td>0.7117</td>
<td>0.1849</td>
</tr>
</tbody>
</table>
In this application, the systematic-error variance $\tau^2_x$ is estimated by using aircraft profile data (Wunch et al., 2010; Messerschmidt et al., 2011).

For regression coefficients $a$ and $b$, and the systematic-error variance $\tau^2_y$, these parameters are estimated based on adjusted score equations (McCullagh and Tibshirani, 1990).
We carried out an EIV regression on 66 pairs of TCCON \((X)\) and OCO-2 \((Y)\) data. The TCCON datum \(X_i\) is the sample mean of a time series selected in a 2-hour window, and the OCO-2 datum \(Y_i\) is the sample median of individual satellite observations in a small spatial domain \((12\text{km} \times 12\text{km})\) observed over a few minutes.

(a) TCCON time series

(b) OCO-2 observation locations around target
Table: Parameter-estimation results for the regression analysis, with estimated asymptotic standard errors of parameter estimates given in parentheses. The slope $b$ is associated with the TCCON covariate. The sum of squared residuals (SSR) is also reported, where $SSR = \sum_{i=1}^{N} (Y_i - \hat{a} - \hat{b}X_i)^2$. The current validation results are given in the "Version 7" row.

<table>
<thead>
<tr>
<th>Model parameters</th>
<th>$a$</th>
<th>$b$</th>
<th>$\tau_y^2$</th>
<th>SSR</th>
</tr>
</thead>
<tbody>
<tr>
<td>$a$, $b$, $\tau_y^2$</td>
<td>$-6.462$ (17.937)</td>
<td>$1.013331$ (4.51 $\cdot$ 10$^{-2}$)</td>
<td>$0.514$ (0.159)</td>
<td>$66.34$</td>
</tr>
<tr>
<td>$a = 0$, $b$, $\tau_y^2$</td>
<td>0 (fixed)</td>
<td>$0.997100$ (3.00 $\cdot$ 10$^{-4}$)</td>
<td>$0.516$ (0.158)</td>
<td>$65.96$</td>
</tr>
<tr>
<td>$a = 0$, $b$, $\tau_x^2 = \tau_y^2 = 0$</td>
<td>0 (fixed)</td>
<td>$0.996601$ (5.26 $\cdot$ 10$^{-5}$)</td>
<td>$0$ (fixed)</td>
<td>$69.22$</td>
</tr>
<tr>
<td>Version 7</td>
<td>0 (fixed)</td>
<td>$0.996941$ (1.15 $\cdot$ 10$^{-3}$)</td>
<td>$-$</td>
<td>$66.43$</td>
</tr>
</tbody>
</table>
Table: Parameter-estimation results for the proposed EIV model with $p = 2$ and intercept $a = 0$, where the asymptotic standard errors of parameter estimates are given in parentheses.

<table>
<thead>
<tr>
<th>Covariates</th>
<th>$b_1$ (TCCON)</th>
<th>$b_2$</th>
<th>$\tau^2_Y$</th>
<th>SSR</th>
</tr>
</thead>
<tbody>
<tr>
<td>TCCON</td>
<td>0.997100 ($3.00 \cdot 10^{-4}$)</td>
<td>--</td>
<td>0.516 (0.158)</td>
<td>65.96</td>
</tr>
<tr>
<td>TCCON&amp; Lat</td>
<td>0.997473 ($3.41 \cdot 10^{-4}$)</td>
<td>-0.0074 (0.0035)</td>
<td>0.461 (0.148)</td>
<td>62.52</td>
</tr>
<tr>
<td>TCCON&amp; Hem</td>
<td>0.998346 ($5.16 \cdot 10^{-4}$)</td>
<td>-0.7087 (0.2459)</td>
<td>0.416 (0.139)</td>
<td>59.83</td>
</tr>
<tr>
<td>TCCON&amp; Sza</td>
<td>0.997497 ($1.27 \cdot 10^{-3}$)</td>
<td>-0.0035 (0.0110)</td>
<td>0.513 (0.157)</td>
<td>66.12</td>
</tr>
</tbody>
</table>
For the calibration of OCO-2 data using TCCON, quantifying variance components and estimating them is crucial for obtaining consistent estimates of regression coefficients.

Robust semivariogram estimators were applied to estimate spatial dependence structure for the OCO-2 individual observations; the approximate variance of the sample median was obtained to estimate the random-error variances of the OCO-2 regression data.

The proposed method, with systematic errors expressed by random effects, provides a reliable calibration straight line.

When compared to the Version 7 calculation, a slightly different estimate of the slope $b$ (and its asymptotic standard error) is obtained. The Version 7 result assumes that $\tau_X^2 = \tau_Y^2 = 0$, however, this can lead to substantial bias. Hence, systematic errors should always be in the model.


