Combining information from different sources: A resampling based approach

S.N. Lahiri

Department of Statistics
North Carolina State University

May 17, 2013
Overview

- Background
- Examples/Potential applications
- Theoretical Framework
- Combining information
- Uncertainty quantification by the Bootstrap
EPA runs computer models to generate hourly ozone estimates (cf. Community Multiscale Air Quality System (CMAQ)) with a resolution of 10mi square.
There also exist a network of ground monitoring stations that also report the O3 levels.
Introduction

- There are many other examples of spatially indexed datasets that report measurements on an atmospheric variable at different spatial supports.
- Our goal is to combine the information from different sources to come up with a better estimate of the true spatial surface.
Consider a function $m(\cdot)$ on a **bounded** domain $D \subset \mathbb{R}^d$ that we want to estimate using data from two different sources.

**Data Source 1:**
- The resolution of Data Source 1 is **coarse**;
- It gives only an averaged version of $m(\cdot)$ over a grid up to an additive noise.

Thus, Data Source 1 corresponds to data generated by Satellite or by computer models at a given level of resolution.
Data Source 2:

- Data Source 2, on the other hand, gives **point-wise** measurements on $m(\cdot)$;
- Has an additive noise that is different from the noise variables for Data Source 1.

Thus, Data Source 2 corresponds to data generated by ground stations or monitoring stations.
Error Structure:

- We suppose that each set of noise variables are **correlated**.
- Further, the variables from the two sources are possibly **cross-correlated**.
- But, we do NOT want to impose any specific distributional structure on the error variables or on their joint distributions.

Goals:

- Combine the data from the two sources to estimate the function $m(\cdot)$ at a given resolution (that is finer than that of Source 1);
- Quantify the associated uncertainty.
Theoretical Formulation

- For simplicity, suppose that $d = 2$ and $\mathcal{D} = [0, 1]^2$.

**Data Source 1:**
The underlying random process is given by:

$$Y(i) = m(i; \Delta) + \epsilon(i), \quad i \in \mathbb{Z}^d$$

where $m(i; \Delta) = \Delta^{-d} \int_{\Delta(i+[0,1]^d)} m(s) ds$, $\Delta \in (0, \infty)$, and

where $\{\epsilon(i), \ i \in \mathbb{Z}^d\}$ is a zero mean second order stationary process.

- The observed variables are

$$\{Y(i) : \Delta(i + [0, 1]^d) \cap [0, 1]^d \neq \emptyset\} \equiv \{Y(i_k) : k = 1, \ldots, N\}.$$
Data Source 1: Coarse grid data (spacings $= \Delta$)
Data Source 2: Point-support measurements

- **Data Source 2:**
  The underlying random process is given by:

  \[ Z(s) = m(s) + \eta(s), \quad s \in \mathbb{R}^d \]

  where \( \{\eta(s), \quad s \in \mathbb{R}^d\} \) is a zero mean second order stationary process on \( \mathbb{R}^d \).

- The observed variables are

  \[ \{Z(s_i) : i = 1, \ldots, n\}. \]

  where \( s_1, \ldots, s_n \) are generated by iid uniform random vectors over \([0, 1]^d\).
Data Source 2: Point-support data

![Plot of point-support data](image-url)
Theoretical Formulation

- Let \( \{\varphi_j : j \geq 1 \} \) be an O.N.B. of \( L^2[0, 1]^d \) and let \( m(\cdot) \in L^2[0, 1]^d \).

- Then,

\[
m(s) = \sum_{j \geq 1} \beta_j \varphi_j(s)
\]

where \( \sum_{j \in \mathbb{Z}} \beta_j^2 < \infty \).

- We consider a finite approximation

\[
m(s) \approx \sum_{j=1}^{J} \beta_j \varphi_j(s) \equiv m_J(s).
\]

- Our goal is to combine the data from the two sources to estimate the parameters \( \{\beta_j : j = 1, \ldots, J\} \).
Estimation on Fine grid

The finite approximation to \( m(\cdot) \) may be thought of as a finer resolution approximation with grid spacings \( \delta \ll \Delta \):
Estimation of the $\beta_j$’s

From Data set 1: \( \{ Y(\mathbf{i}_k) : k = 1, \ldots, N \} \), we have

$$
\hat{\beta}_j^{(1)} = N^{-1} \sum_{k=1}^{N} Y(\mathbf{i}_k) \varphi_j(\mathbf{i}_k \Delta).
$$

It is easy to check that for $\Delta$ small:

$$
E \hat{\beta}_j^{(1)} = N^{-1} \sum_{k=1}^{N} m(\mathbf{i}_k; \Delta) \varphi_j(\mathbf{i}_k \Delta)
\approx N^{-1} \sum_{k=1}^{N} \Delta^{-d} \int_{(\mathbf{i}_k + [0,1]^d)\Delta} m(\mathbf{s}) \varphi_j(\mathbf{s}) d\mathbf{s}
\approx \int_{[0,1]^d} m(\mathbf{s}) \varphi_j(\mathbf{s}) d\mathbf{s}/[N\Delta^d] \approx \beta_j.
$$
Estimation of the $\beta_j$’s

- From Data set 2: $\{Z(s_i) : i = 1, \ldots, n\}$, we have
  \[
  \hat{\beta}_j^{(2)} = n^{-1} \sum_{i=1}^{n} Z(s_i)\varphi_j(s_i).
  \]

- It is easy to check that as $n \to \infty$:
  \[
  E[\hat{\beta}_j^{(2)} | S] = n^{-1} \sum_{i=1}^{n} m(s_i)\varphi_j(s_i)
  \]
  \[
  \to \int_{[0,1]^d} m(s)\varphi_j(s)ds = \beta_j \quad \text{a.s.}
  \]

where $S$ is the $\sigma$-field of the random vectors generating the data locations.
The estimator from Data Set $k \in \{1, 2\}$ is

$$\hat{m}^{(k)}(\cdot) = \sum_{j=1}^{J} \hat{\beta}_j^{(k)} \varphi_j(\cdot).$$

We shall consider a combined estimator of $m(\cdot)$ of the form:

$$\hat{m}(\cdot) = a_1 \hat{m}^{(1)}(\cdot) + a_2 \hat{m}^{(2)}(\cdot)$$

where $a_1, a_2 \in \mathbb{R}$ and $a_1 + a_2 = 1$. 
Many choices of $a_1 \in \mathbb{R}$ (with $a_2 = 1 - a_1$) is possible.

Here we seek an optimal choice of $a_1$ that minimizes the MISE:

$$\int E\left( \hat{m}(\cdot) - m_J(\cdot) \right)^2.$$

Evidently, this depends on the joint correlation structure of the error processes from Data sources 1 and 2.
More precisely, it can be shown that the optimal choice of $a_1$ is given by

$$a_1^0 = \frac{\sum_{j=1}^{J} E\left\{ (\hat{\beta}_j^{(1)} - \hat{\beta}_j^{(2)}) [\hat{\beta}_j^{(2)} - \beta_j] \right\}}{\sum_{j=1}^{J} E[\hat{\beta}_j^{(1)} - \hat{\beta}_j^{(2)}]^2}$$

Since each $\hat{\beta}_j^{(K)}$ is a linear function of the observations from Data set $k \in \{1, 2\}$, the numerator and the denominator of the optimal $a_1$ depends on the joint covariance structure of the processes $\{\epsilon(i) : i \in \mathbb{Z}^d\}$ and $\{\eta(s) : s \in \mathbb{R}^d\}$.

Note that the $\varphi_j$’s drop out from the formula for the MISE optimal $a_1^0$ due to the ONB property of $\{\varphi_j : j \geq 1\}$. 
Joint-Correlation structure

We shall suppose that

- \( \{\epsilon(i) : i \in \mathbb{Z}^d\} \) is SOS with covariogram
  \[ \sigma(k) = \text{Cov}(\epsilon(i), \epsilon(i + k)) \quad \text{for all} \quad i, k \in \mathbb{Z}^d; \]

- \( \{\eta(s) : s \in \mathbb{R}^d\} \) is SOS with covariogram
  \[ \tau(h) = \text{Cov}(\eta(s), \eta(s + h)) \quad \text{for all} \quad s, h \in \mathbb{R}^d; \]

and the cross-correlation function between the \( \epsilon(\cdot)'s \) and \( \eta(\cdot)'s \) is given by

\[ \text{Cov}(\epsilon(i), \eta(s)) = \gamma(i - s) \quad \text{for all} \quad i \in \mathbb{Z}^d, s \in \mathbb{R}^d; \]

for some function \( \gamma : \mathbb{R}^d \to \mathbb{R} \).
This formulation is somewhat non-standard, as the two component spatial processes have different supports.

**Example:** Consider a zero mean SOS bivariate process 
\[ \{(\eta_1(s), \eta_2(s)) : s \in \mathbb{R}^d\} \] with autocovariance matrix \( \Sigma(\cdot) = ((\sigma_{ij}(\cdot))) \). Let \( \eta(s) = \eta_1(s) \) and

\[ \epsilon(i) = \Delta^{-d} \int_{[i+[0,1)^d] \Delta} \eta_2(s) ds, \quad i \in \mathbb{Z}^d. \]

Then, \( \text{Cov}(\epsilon(i), \epsilon(i + k)) \) depends only on \( k \) for all \( i, k \in \mathbb{Z}^d \); (given by an integral of \( \sigma_{11}(\cdot) \)) and \( \text{Cov}(\epsilon(i), \eta(s)) \) depends only on \( i - s \) for all \( i \in \mathbb{Z}^d, s \in \mathbb{R}^d \) (given by an integral of \( \sigma_{12}(\cdot) \)).
Recall that the optimal

$$a_1^0 = \frac{\sum_{j=1}^{J} E \left\{ \left[ \hat{\beta}_j^{(1)} - \hat{\beta}_j^{(2)} \right] \left[ \hat{\beta}_j^{(2)} - \beta_j \right] \right\}}{\sum_{j=1}^{J} E \left[ \hat{\beta}_j^{(1)} - \hat{\beta}_j^{(2)} \right]^2}$$

depends on the population joint covariogram of the error processes that are typically unknown.

It is possible to derive an asymptotic approximation to $a_1^0$ that involves only some summary characteristics of these functions (such as $\int \tau(h) \, dh$ and $\sum_{k \in \mathbb{Z}^d} \sigma(k)$), and use plug-in estimates.
However, the limiting formulae depends on the asymptotic regimes one employs (relative growth rates of $n$ and $N$, and the strength of dependence).

The accuracy of these approximations are not very good even for $d = 2$ due to edge-effects.

These issues with the asymptotic approximations suggest that we may want to use a data-based method, such as the spatial block bootstrap/subsampling that more closely mimic the behavior in finite samples.
Here we shall use a version of the subsampling for estimating $a_1^0$.

The Subsampling method is known to be computationally simpler.

Further, it has the same level of accuracy as the bootstrap for estimating the variance of a linear function of the data.

We shall use the bootstrap for uncertainty quantification of the resulting estimator, as it is more accurate for distributional approximation.
A Spatial Block Resampling Scheme

- We now give a brief description of a spatial version of the Moving Block Bootstrap of Künsch (1989) and Liu and Singh (1992) in the present setup.

- Recall that we have;

  Data Set 1: (Coarse grid) \{ Y(\mathbf{i}_k) : k = 1, \ldots, N \}
  Data Set 2: (Point support) \{ Z(\mathbf{s}_i) : i = 1, \ldots, n \}

- For each data set, we also have an estimate of its mean structure.

- First, form the residuals and center them! Denote these by \{ \hat{\epsilon}(\mathbf{i}_k) : k = 1, \ldots, N \} and \{ \hat{\eta}(\mathbf{s}_i) : i = 1, \ldots, n \}.

- We will resample blocks of \( \hat{\epsilon}() \)'s and \( \hat{\eta}() \)'s.
Next fix an integer $\ell$ such that

$$1 \ll \ell \ll L, \quad (0.1)$$

where $L = N^{1/d} = 1/\Delta$ denotes the number of $\Delta$-intervals along a given co-ordinate.

Here $\ell$ determines the size (volume) of the spatial blocks.

Let $\{B(k) : k \in \mathcal{K}\}$ denote the collection of overlapping blocks of volume $\ell^d \Delta^d$ contained in $[0, 1]^d$.

Note that under (0.1), $K = |\mathcal{K}| =$ the total number of overlapping blocks satisfies

$$K = ([L - \ell + 1])^d \sim N.$$
Spatial Bootstrap

- Resample randomly with replacement from \( \{B_k : k = 1, \ldots, K\} \)
  a sample of size \( b \geq 1 \).
- This yields resampled error variables for both data source 1 and 2, which are used to fill up \([0, 1]^d\).
- For \( b = N/\ell^d \), there are \( N \)-many Data Source 1 error variables \( \{\epsilon^*(i_k) : k = 1, \ldots, N\} \).
- For Data Source 2, this yields a random number \( n_1 \) of error variables \( \{\eta^*(s_i^*) : i = 1, \ldots, n_1\} \).
- It is evident that \( n_1 \sim n \).
Next use the model equations to define the “bootstrap observations”

\[ Y^*(i_k) = \hat{m}^{(1)}(i_k; \Delta) + \epsilon^*(i_k), \quad k = 1, \ldots, N \]

\[ Z^*(s^*_i) = \hat{m}^{(2)}(s^*_i) + \eta^*(s^*_i), \quad i = 1, \ldots, n_1 \]

The reconstruction step is referred to as the **residual bootstrap** (Efron (1979), Freedman (1981)).

For \( b = 1 \), one gets **spatial subsampling**.

Note that for \( b = 1 \), the corresponding bootstrap moments (e.g., the variances/covariances) can be evaluated without any resampling.
The combined estimator

Recall that

\[ a_{1}^{0} = \frac{\sum_{j=1}^{J} E \left\{ [\hat{\beta}_{j}^{(1)} - \hat{\beta}_{j}^{(2)}][\hat{\beta}_{j}^{(2)} - \beta_{j}] \right\}}{\sum_{j=1}^{J} E [\hat{\beta}_{j}^{(1)} - \hat{\beta}_{j}^{(2)}]^{2}} \]

We use the spatial subsampling to estimate \( a_{1}^{0} \); Call this \( \hat{a}_{1}^{0} \).

Then define the **combined estimator** of \( m(\cdot) \):

\[ \hat{m}^{0}(\cdot) = \hat{a}_{1}^{0} \hat{m}^{(1)}(\cdot) + [1 - \hat{a}_{1}^{0}] \hat{m}^{(2)}(\cdot). \]
We can estimate the MISE of our combined estimator by using spatial bootstrap!

Specifically, let $m^{(1)}^* (\cdot)$ be the bootstrap version of $\hat{m}^{(1)} (\cdot)$ that is obtained by replacing $\{ Y_i(k) : k = 1, \ldots, N \}$ with the Bootstrap data set 1: $\{ Y^*_i(k) : k = 1, \ldots, N \}$.

Similarly, define $m^{(2)}^* (\cdot)$ and $a_1^0$, the bootstrap versions of $\hat{m}^{(2)} (\cdot)$ and $\hat{a}_1^0$.

Let $m^0^* (\cdot) = a_1^0 m^{(1)}^* (\cdot) + [1 - a_1^0] m^{(2)}^* (\cdot)$.

Then, the Bootstrap estimator of the MISE of $\hat{m}^0 (\cdot)$ is given by

\[
\hat{\text{MISE}} = \int E^* \left( m^0^* (\cdot) - \hat{m}^0 (\cdot) \right)^2.
\]
Theorem

Suppose that $\Delta = o(1)$, $N = O(n)$, $\ell^{-1} + \ell/L = o(1)$ and that the error random fields satisfy certain moment and weak dependence conditions. Then,

$$\hat{\text{MISE}} / \text{MISE} \to_p 1.$$
Thank You!!!