Statistical Theory and Related Fields

ISSN: 2475-4269 (Print) 2475-4277 (Online) Journal homepage: https://www.tandfonline.com/loi/tstf20

Small area estimation with subgroup analysis

Xin Wang & Zhengyuan Zhu

To cite this article: Xin Wang & Zhengyuan Zhu (2019) Small area estimation with subgroup analysis, Statistical Theory and Related Fields, 3:2, 129-135, DOI: 10.1080/24754269.2019.1659097

To link to this article: https://doi.org/10.1080/24754269.2019.1659097

Published online: 31 Aug 2019.

STATISTICAL THEORY AND RELATED FIELDS
2019, VOL. 3, NO. 2, 129–135
https://doi.org/10.1080/24754269.2019.1659097

Taylor & Francis
Taylor & Francis Group

Small area estimation with subgroup analysis

Xin Wang[a] and Zhengyuan Zhu [b]

[a]Department of Statistics, Miami University, Oxford, OH, USA; [b]Department of Statistics, Iowa State University, Ames, IA, USA

ABSTRACT
In this article, a new unit level model based on a pairwise penalised regression approach is proposed for problems in small area estimation (SAE). Instead of assuming common regression coefficients for all small domains in the traditional model, the new estimator is based on a subgroup regression model which allows different regression coefficients in different groups. The alternating direction method of multipliers (ADMM) algorithm is used to find subgroups with different regression coefficients. We also consider pairwise spatial weights for spatial areal data. In the simulation study, we compare the performances of the new estimator with the traditional small area estimator. We also apply the new estimator to urban area estimation using data from the National Resources Inventory survey in Iowa.

ARTICLE HISTORY
Received 31 December 2018
Accepted 20 August 2019

KEYWORDS
Linear mixed models; penalty regressions; small area estimation; spatial areal data; subgroup analysis

1. Introduction

Small area estimation (SAE) is an important problem in survey sampling when the sample sizes are not large enough to provide reliable estimates in small domains or areas. See Rao and Molina (2015) and Pfeffermann (2013) for overviews and recent developments in SAE. One of the model-based approaches for SAE is the unit level model, which was first proposed by Battese, Harter, and Fuller (1988). Unit level models are specified for the individual elements of the population and require the availability of unit level auxiliary information.

Traditional unit level models typically assume a linear relationship between the variable of interest and the auxiliary information, and all the areas share the same regression coefficients to borrow information. Random effects are also considered for each small area. However, different relationships can exist in different areas. That is, subgroups could exist for different areas such that areas in one group have the same regression coefficient and areas in different groups have different regression coefficients.

In the linear regression setting, Ma and Huang (2017), Ma, Huang, and Zhang (2016) developed a method to obtain homogeneous groups based on regression coefficients through the alternating direction method of multiplier algorithm (ADMM, Boyd, Parikh, Chu, Peleato, & Eckstein, 2011). In the algorithm, they used pairwise concave penalties based on the smoothly clipped absolute deviation (SCAD) penalty (Fan & Li, 2001) and the minimax concave penalty (MCP) (Zhang, 2010). Wang, Zhu, and Zhang (2019) extended the problem to a regression setting with repeated measures. They also considered spatial weights in the pairwise penalties and showed that spatial weights perform better than equal weights. However, the model cannot be applied to the SAE problems directly, since random effects are not considered.

In this article, we propose a new SAE estimator that allows different regression coefficients in different subgroups under a linear mixed model framework at the unit level. The ADMM algorithm is applied and the variance parameters are also estimated in the algorithm. As in Wang et al. (2019), we use spatial pairwise weights in the pairwise penalties based on the SCAD penalty. In this algorithm, the number of groups and the group structure are also determined.

The article is organised as follows. In Section 2, we introduce the unit level model with areal regression coefficients and the algorithm to find subgroups. In Section 3, we conduct several simulation studies to compare the performance of the proposed estimator with the traditional estimators. In Section 4, we apply the proposed method to a real data set. Finally, Section 5 contains some conclusion and discussion.

2. The model and the algorithm

In this section, the unit level model with area level regression coefficients and the corresponding algorithm to estimate parameters are introduced.

2.1. The unit level model

Suppose there are M areas with known population size N_i and n_i is the sample size in area i for $i = 1, \ldots, M$. Let y_{ih} be the observation of unit h in area i

CONTACT Xin Wang wangx172@miamioh.edu

for $h = 1, \ldots, n_i$, $i = 1, \ldots, M$. Let $\boldsymbol{x}_{ih}$ be the p dimension auxiliary information vector with area population mean $\bar{\boldsymbol{X}}_i = 1/N_i \sum_{h=1}^{N_i} \boldsymbol{x}_{ih}$ known. In the traditional unit level model, that is Battese–Harter–Fuller (BHF) model (Battese et al., 1988), different areas share the same regression coefficient as in (1),

$$y_{ih} = \boldsymbol{x}_{ih}{}^{\mathrm{T}}\boldsymbol{\beta} + v_i + \epsilon_{ih}, \tag{1}$$

where $\boldsymbol{\beta}$ is the unknown regression coefficient vector, v_i's are i.i.d areal random effects with mean zero and variance σ_v^2, and ϵ_{ih}'s are i.i.d random errors with mean zero and variance σ_ϵ^2. Let A_i be the set of observed units and C_i be the set of unobserved units in area i. The predictor for the finite population mean $\bar{Y}_i = 1/N_i \sum_{h=1}^{N_i} y_{ih}$ in area i under model (1) for SAE given in Battese et al. (1988) and the sae package (Molina & Marhuenda, 2015) is

$$\hat{\bar{Y}}_i^{\mathrm{BHF}} = \frac{1}{N_i}\left(\sum_{h\in A_i} y_{ih} + \sum_{h\in C_i}\left(\boldsymbol{x}_{ih}{}^{\mathrm{T}}\hat{\boldsymbol{\beta}} + \hat{v}_i\right)\right), \tag{2}$$

where $\hat{\boldsymbol{\beta}}$ is the estimate of $\boldsymbol{\beta}$ and $\hat{v}_i$ is the empirical best linear unbiased prediction of v_i. In the simulation study, we use the R package *sae* (Molina & Marhuenda, 2015) to obtain the predictions.

Instead of assuming all the areas have the same regression coefficients $\boldsymbol{\beta}$, we assume that there are K mutually exclusive subgroups $\mathcal{G} = \{\mathcal{G}_1, \ldots, \mathcal{G}_K\}$, which is a partition of areas $\{1, 2, \ldots, M\}$. First we assume that each area has its own regression coefficient,

$$y_{ih} = \boldsymbol{x}_{ih}{}^{\mathrm{T}}\boldsymbol{\beta}_i + v_i + \epsilon_{ih}, \tag{3}$$

where $\boldsymbol{\beta}_i$ is the unknown regression coefficient vector for area i. Let $\boldsymbol{y}_i = (y_{i1}, \ldots, y_{in_i})^{\mathrm{T}}$, $\boldsymbol{x}_i = (\boldsymbol{x}_{i1}, \ldots, \boldsymbol{x}_{in_i})^{\mathrm{T}}$ and $\boldsymbol{\beta} = (\boldsymbol{\beta}_1^{\mathrm{T}}, \ldots, \boldsymbol{\beta}_M^{\mathrm{T}})^{\mathrm{T}}$. The weighted log likelihood function is

$$\begin{aligned} l\left(\boldsymbol{\beta}, \sigma_v^2, \sigma_\epsilon^2\right) = &-\frac{1}{2}\sum_{i=1}^{M}\frac{1}{n_i}\log|\Sigma_i| \\ &-\frac{1}{2}\sum_{i=1}^{M}\frac{1}{n_i}\left(\boldsymbol{y}_i - \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}_i\right)^{\mathrm{T}} \\ &\times \Sigma_i^{-1}\left(\boldsymbol{y}_i - \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}_i\right), \end{aligned} \tag{4}$$

where Σ_i is the covariance matrix based on the random effect structure which has the following form:

$$\Sigma_i = \mathbf{1}_{n_i}\mathbf{1}_{n_i}^{\mathrm{T}}\sigma_v^2 + \boldsymbol{I}_{n_i}\sigma_\epsilon^2$$

and

$$\begin{aligned} \Sigma_i^{-1} &= \left(\mathbf{1}_{n_i}\mathbf{1}_{n_i}^{\mathrm{T}}\sigma_v^2 + \boldsymbol{I}_{n_i}\sigma_\epsilon^2\right)^{-1} \\ &= \frac{1}{\sigma_\epsilon^2}\left(\boldsymbol{I}_{n_i} - \mathbf{1}_{n_i}\mathbf{1}_{n_i}^{\mathrm{T}}\frac{\sigma_v^2}{\sigma_\epsilon^2 + n_i\sigma_v^2}\right), \end{aligned}$$

where $\mathbf{1}_{n_i}$ is an $n_i \times 1$ vector with elements 1 and $\boldsymbol{I}_{n_i}$ is an $n_i \times n_i$ identity matrix.

If area i and area j are in the same group, then $\boldsymbol{\beta}_i = \boldsymbol{\beta}_j$. In order to find the estimated partition $\hat{\mathcal{G}} = \{\hat{\mathcal{G}}_1, \ldots, \hat{\mathcal{G}}_{\hat{K}}\}$ with the estimated number of groups $\hat{K}$, the following objective function is considered

$$\begin{aligned} &Q\left(\boldsymbol{\beta}, \sigma_v^2, \sigma_\epsilon^2; \lambda, \psi\right) \\ &= \frac{1}{2}\sum_{i=1}^{M}\frac{1}{n_i}\log|\Sigma_i| + \frac{1}{2}\sum_{i=1}^{M}\frac{1}{n_i}\left(\boldsymbol{y}_i - \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}_i\right)^{\mathrm{T}} \\ &\quad\times \Sigma_i^{-1}\left(\boldsymbol{y}_i - \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}_i\right) \\ &\quad+ \sum_{1\le i<j\le M} p_\gamma\left(\left\|\boldsymbol{\beta}_i - \boldsymbol{\beta}_j\right\|, c_{ij}\lambda\right), \end{aligned} \tag{5}$$

where $\|\cdot\|$ denotes the Euclidean norm, $p_\gamma(\cdot, \lambda)$ is a penalty function with a fixed value γ and a tuning parameter $\lambda \ge 0$. In the penalty function, pairwise weights are considered associated with area i and area j. In this paper, we use the SCAD penalty. In the context of spatial SAE, we define c_{ij} as

$$c_{ij} = \exp\left(\psi(1 - a_{ij})\right), \tag{6}$$

where ψ is a tuning parameter and a_{ij} is the neighbour order between area i and area j. As shown in Wang et al. (2019), pairwise spatial weights can help in spatial areal data.

2.2. The ADMM algorithm

For given λ and ψ, the solution of (5) is

$$\begin{aligned} \left(\hat{\boldsymbol{\beta}}, \hat{\sigma_v^2}, \hat{\sigma_\epsilon^2}\right) = &\underset{\boldsymbol{\beta}\in\mathbb{R}^{Mp}, \sigma_v^2\in\mathbb{R}_+, \sigma_\epsilon^2\in\mathbb{R}_+}{\arg\min} \\ &Q\left(\boldsymbol{\beta}, \sigma_v^2, \sigma_\epsilon^2; \lambda, \psi\right). \end{aligned} \tag{7}$$

The ADMM algorithm is applied to solve (7). Let $\boldsymbol{\delta}_{ij} = \boldsymbol{\beta}_i - \boldsymbol{\beta}_j$, the objective function becomes

$$\begin{aligned} &L_0\left(\boldsymbol{\beta}, \sigma_v^2, \sigma_\epsilon^2, \boldsymbol{\delta}\right) \\ &= \frac{1}{2}\sum_{i=1}^{M}\frac{1}{n_i}\log|\Sigma_i| \\ &\quad+ \frac{1}{2}\sum_{i=1}^{M}\frac{1}{n_i}\left(\boldsymbol{y}_i - \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}_i\right)^{\mathrm{T}}\Sigma_i^{-1}\left(\boldsymbol{y}_i - \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}_i\right) \\ &\quad+ \sum_{1\le i<j\le M} p_\gamma\left(\left\|\boldsymbol{\delta}_{ij}\right\|, c_{ij}\lambda\right) \\ &\quad\text{subject to } \boldsymbol{\beta}_i - \boldsymbol{\beta}_j - \boldsymbol{\delta}_{ij} = \mathbf{0}, \end{aligned}$$

where $\boldsymbol{\delta} = (\boldsymbol{\delta}_{ij}^{\mathrm{T}}, i < j)^{\mathrm{T}}$. The augmented Lagrangian is

$$\begin{aligned} L\left(\boldsymbol{\beta}, \sigma_v^2, \sigma_\epsilon^2, \boldsymbol{\delta}, \boldsymbol{v}\right) = &L_0\left(\boldsymbol{\beta}, \sigma_v^2, \sigma_\epsilon^2, \boldsymbol{\delta}\right) \\ &+ \sum_{i<j}\left\langle \boldsymbol{v}_{ij}, \boldsymbol{\beta}_i - \boldsymbol{\beta}_j - \boldsymbol{\delta}_{ij}\right\rangle \\ &+ \frac{\vartheta}{2}\sum_{i<j}\|\boldsymbol{\beta}_i - \boldsymbol{\beta}_j - \boldsymbol{\delta}_{ij}\|^2, \end{aligned}$$

where $\boldsymbol{v} = (\boldsymbol{v}_{ij}^{\mathrm{T}}, i < j)^{\mathrm{T}}$ are Lagrange multipliers and ϑ is the penalty parameter. Let $\boldsymbol{\tau} = (\sigma_v^2, \sigma_\epsilon^2)$. Given $\boldsymbol{\tau}^m$, $\boldsymbol{\delta}^m$ and $\boldsymbol{v}^m$, $\boldsymbol{\beta}, \boldsymbol{\tau}, \boldsymbol{\delta}$ and $\boldsymbol{v}$ are updated as follows:

$$\begin{aligned}
\boldsymbol{\beta}^{m+1} &= \arg\min L\left(\boldsymbol{\beta}, \boldsymbol{\tau}^m, \boldsymbol{\delta}^m, \boldsymbol{v}^m\right), \\
\boldsymbol{\tau}^{m+1} &= \boldsymbol{\tau}^m + \left[\mathcal{I}\left(\boldsymbol{\tau}^m\right)\right]^{-1} s\left(\boldsymbol{\beta}^{m+1}, \boldsymbol{\tau}^m\right), \\
\boldsymbol{\delta}^{m+1} &= \arg\min L\left(\boldsymbol{\beta}^{m+1}, \boldsymbol{\tau}^{m+1}, \boldsymbol{\delta}, \boldsymbol{v}^m\right), \\
\boldsymbol{v}_{ij}^{m+1} &= \boldsymbol{v}_{ij}^m + \vartheta\left(\boldsymbol{\beta}_i^{m+1} - \boldsymbol{\beta}_j^{m+1} - \boldsymbol{\delta}_{ij}^{m+1}\right).
\end{aligned}$$

Let $\boldsymbol{y} = (\boldsymbol{y}_1^{\mathrm{T}}, \ldots, \boldsymbol{y}_M^{\mathrm{T}})^{\mathrm{T}}$, $\boldsymbol{X} = \mathrm{diag}(\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_M)$ and $\boldsymbol{\Omega} = \mathrm{diag}(1/n_1\Sigma_1^{-1}, \ldots, 1/n_M\Sigma_M^{-1})$. The update of $\boldsymbol{\beta}$ is

$$\begin{aligned}
\boldsymbol{\beta}^{m+1} = &\left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{\Omega}^m\boldsymbol{X} + \vartheta\boldsymbol{A}^{\mathrm{T}}\boldsymbol{A}\right)^{-1} \\
&\times\left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{\Omega}^m\boldsymbol{y} + \vartheta\mathrm{vec}\left(\left(\boldsymbol{\Delta}^m - \vartheta^{-1}\boldsymbol{\Upsilon}^m\right)\boldsymbol{D}\right)\right),
\end{aligned}$$

where $\boldsymbol{A} = \boldsymbol{D} \otimes \boldsymbol{I}_p$, $\otimes$ is the Kronecker product, $\boldsymbol{D} = \{(\boldsymbol{e}_i - \boldsymbol{e}_j)\}^{\mathrm{T}}$ with $\boldsymbol{e}_i$ an $M \times 1$ vector with ith element 1 and other elements 0, $\boldsymbol{\Delta}^m = (\boldsymbol{\delta}_{ij}^m, i < j)_{p\times M(M-1)/2}$ and $\boldsymbol{\Upsilon}^m = (\boldsymbol{v}_{ij}^m, i < j)_{p\times M(M-1)/2}$. When updating $\boldsymbol{\tau}$, $\mathcal{I}(\boldsymbol{\tau})$ is the expected second-order derivative of $-l$ in (4) and

$$s\left(\boldsymbol{\beta}^{m+1}, \boldsymbol{\tau}^m\right) = \left(\frac{\partial l}{\partial\sigma_v^2}, \frac{\partial l}{\partial\sigma_\epsilon^2}\right)^{\mathrm{T}}\Big|_{\boldsymbol{\beta}=\boldsymbol{\beta}^{m+1}, \boldsymbol{\tau}=\boldsymbol{\tau}^m}.$$

The details of $s(\cdot,\cdot)$ and $\mathcal{I}$ are in the appendix. In this step, $\boldsymbol{\tau}$ can be updated several times within one iteration.

Updating $\boldsymbol{\delta}_{ij}$ is based on the result of SCAD penalty. Let $\boldsymbol{\varsigma}_{ij}^m = (\boldsymbol{\beta}_i^{m+1} - \boldsymbol{\beta}_j^{m+1}) + \vartheta^{-1}\boldsymbol{v}_{ij}^m$, then the solution is

$$\boldsymbol{\delta}_{ij}^{m+1} = \begin{cases} S\left(\boldsymbol{\varsigma}_{ij}^m, \lambda c_{ij}/\vartheta\right) & \text{if } \left\|\boldsymbol{\varsigma}_{ij}^m\right\| \le \lambda c_{ij} + \lambda c_{ij}/\vartheta, \\ \frac{S\left(\boldsymbol{\varsigma}_{ij}^m, \gamma\lambda c_{ij}/((\gamma-1)\vartheta)\right)}{1-1/((\gamma-1)\vartheta)} & \text{if } \lambda c_{ij} + \lambda c_{ij}/\vartheta < \left\|\boldsymbol{\varsigma}_{ij}^m\right\| \le \gamma\lambda c_{ij}, \\ \boldsymbol{\varsigma}_{ij}^m & \text{if } \left\|\boldsymbol{\varsigma}_{ij}^m\right\| > \gamma\lambda c_{ij}, \end{cases}$$

where $\gamma > c_{ij} + c_{ij}/\vartheta$ and $S(\boldsymbol{w}, t) = (1 - t/\|\boldsymbol{w}\|)_+\boldsymbol{w}$ and $(t)_+ = t$ if $t > 0$, 0 otherwise.

Remark 2.1: The convergence criteria is based on that given in Boyd et al. (2011). The primal residual and dual residual are defined as $\boldsymbol{r}^{m+1} = \boldsymbol{A}\boldsymbol{\beta}^{m+1} - \boldsymbol{\delta}^{m+1}$ and $\boldsymbol{s}^{m+1} = \vartheta\boldsymbol{A}^{\mathrm{T}}(\boldsymbol{\delta}^{m+1} - \boldsymbol{\delta}^m)$. The stopping criterion is

$$\left\|\boldsymbol{r}^m\right\|_2 \le \epsilon^{\mathrm{pri}}, \quad \left\|\boldsymbol{s}^m\right\|_2 \le \epsilon^{\mathrm{dual}},$$

where

$$\begin{aligned}
\epsilon^{\mathrm{pri}} &= \sqrt{\frac{M(M-1)}{2}p}\epsilon^{\mathrm{abs}} \\
&\quad + \epsilon^{\mathrm{rel}}\max\left\{\left\|\boldsymbol{A}\boldsymbol{\beta}^m\right\|, \left\|\boldsymbol{\delta}^m\right\|\right\}, \\
\epsilon^{\mathrm{dual}} &= \sqrt{Mp}\epsilon^{\mathrm{abs}} + \epsilon^{\mathrm{rel}}\left\|\boldsymbol{A}^{\mathrm{T}}\boldsymbol{v}^m\right\|,
\end{aligned}$$

where ϵ^{abs} is an absolute tolerance and ϵ^{rel} is a relative tolerance. In the simulation study and the application, we use $\epsilon^{\mathrm{abs}} = 10^{-4}$ and $\epsilon^{\mathrm{rel}} = 10^{-2}$.

2.3. The proposed small area estimator

As in Zhu, Zou, Liang, and Zhu (2016), two small area estimators can be defined. Let $\bar{y}_i = 1/n_i\sum_{h=1}^{n_i} y_{ih}$ and $\bar{\boldsymbol{x}}_i = 1/n_i\sum_{h=1}^{n_i}\boldsymbol{x}_{ih}$ be the sample mean of the variable of interest and auxiliary information, respectively. The first one is based on the predictions of random effects, which is defined as

$$\hat{\bar{Y}}_i^{(1)} = \bar{\boldsymbol{X}}_i^{\mathrm{T}}\hat{\boldsymbol{\beta}}_i + \hat{v}_i, \tag{8}$$

where $\hat{\boldsymbol{\beta}}_i$ is the estimate of $\boldsymbol{\beta}_i$ from the proposed algorithm,

$$\hat{v}_i = \hat{\gamma}_i\left(\bar{y}_i - \bar{\boldsymbol{x}}_i^{\mathrm{T}}\hat{\boldsymbol{\beta}}_i\right),$$

and $\hat{\gamma}_i = \hat{\sigma}_v^2/(\hat{\sigma}_v^2 + \hat{\sigma}_\epsilon^2/n_i)$.

In the second estimator, the unobserved values in each area are predicted based on the model, which is given by

$$\begin{aligned}
\hat{\bar{Y}}_i^{(2)} &= \frac{1}{N_i}\left(\sum_{h\in A_i} y_{ih} + \sum_{h\in C_i}\hat{y}_{ih}\right) \\
&= f_i\bar{y}_i + \left(\bar{\boldsymbol{X}}_i - f_i\bar{\boldsymbol{x}}_i\right)^{\mathrm{T}}\hat{\boldsymbol{\beta}}_i + \left(1 - f_i\right)\hat{v}_i,
\end{aligned} \tag{9}$$

where $\hat{y}_{ih} = \boldsymbol{x}_{ih}^{\mathrm{T}}\hat{\boldsymbol{\beta}}_i + \hat{v}_i$ and $f_i = n_i/N_i$. If f_i is small, then the predictor in (9) is nearly identical to the predictor in (8).

3. Simulation study

The simulation setup is designed based on the features of the National Resources Inventory (NRI) survey, which monitors status and trend of natural resources characteristics. One of the characteristic is the area of land uses, such as cropland, pastureland and urban (Nusser & Goebel, 1997). Each state is divided into 'segments' with size of 160 acres. From 1982 to 1997, the full NRI sample was observed in 5-year intervals (1982, 1987, 1992 and 1997) with 300,000 segments. In 2000, the NRI transitioned to an annual sample design with about 70,000 segments.

For the simulation, we construct an artificial population composed of 300,000 segments in 99 counties. The number of counties in the simulated population is the same as the number of counties in Iowa. We treat counties as areas and segments as unit level observations. In the population for the simulation, the number of segments in each county for the 99 counties is between 2210 and 5412. These numbers are the population sizes of segments in counties used in the simulation study. This simulated population maintains features of the NRI data for Iowa. There are around 6000 segments selected in the full sample and around 1500 segments selected in the annual sample in the original NRI design. In the annual sample, fewer segments are sampled, so the accuracy of the estimates is reduced. Thus

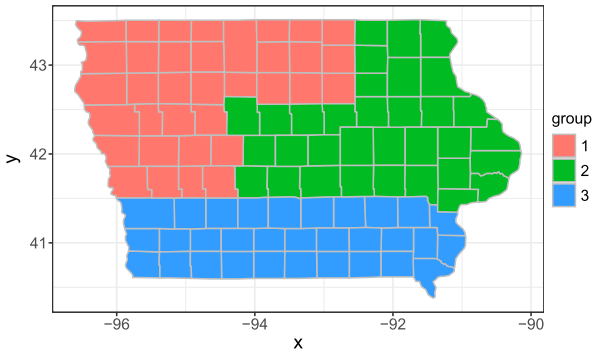

Figure 1. Group information.

auxiliary information should be considered to improve the estimator.

We compare the performances of the proposed estimators to the BHF estimator based on 100 simulations. Tuning parameters are selected based on the following modified BIC (Wang, Li, & Tsai, 2007):

$$BIC = -2l + C_M \log(M)(\hat{K}p), \tag{10}$$

where l is defined in (4) and C_M is a positive number which can depend on M. Here we use $C_M = c_0 \log(\log(Mp + 2))$ with $c_0 = 0.2$ as in Wang et al. (2019).

In the simulation study, simple random sampling is used in each county to select segments. As mentioned before, the population size of segments in each county is between 2210 and 5412. Two sampling rates are considered in each area, 1% and 0.5%. When sampling rate is 1%, there are 3067 selected segments in the whole state and the number of segments in each county is between 22 and 54. When sampling rate is 0.5%, there are 1537 selected segments in the whole state and the range of the number of segments in each county is from 11 to 27. $\boldsymbol{x}_{ih} = (1, x_{ih})^{\mathrm{T}}$ with x_{ih}'s simulated from a normal distribution with mean 1 and standard deviation 1 and v_i's are simulated from a standard normal distribution, that is $\sigma_v^2 = 1$. The assumed group structure in Iowa is shown in Figure 1 with three groups. The three groups are aggregated based on the districts available on https://www.nass.usda.gov/Charts_and_Maps/Crops_County/boundary_maps/indexpdf.php

We consider three different sets of parameters.

- Case I: $\boldsymbol{\beta}_i = (0.5, 0.5)^{\mathrm{T}}$ if $i \in \mathcal{G}_1$, $\boldsymbol{\beta}_i = (2, 2)^{\mathrm{T}}$ if $i \in \mathcal{G}_2$ and $\boldsymbol{\beta}_i = (3.5, 3.5)^{\mathrm{T}}$ if $i \in \mathcal{G}_3$.
- Case II: $\boldsymbol{\beta}_i = (0.5, 0.5)^{\mathrm{T}}$ if $i \in \mathcal{G}_1$, $\boldsymbol{\beta}_i = (1.5, 1.5)^{\mathrm{T}}$ if $i \in \mathcal{G}_2$ and $\boldsymbol{\beta}_i = (2.5, 2.5)^{\mathrm{T}}$ if $i \in \mathcal{G}_3$.
- Case III: $\boldsymbol{\beta}_i = (0.5, 0.5)^{\mathrm{T}}$ if $i \in \mathcal{G}_1$, $\boldsymbol{\beta}_i = (1, 1)^{\mathrm{T}}$ if $i \in \mathcal{G}_2$ and $\boldsymbol{\beta}_i = (1.5, 1.5)^{\mathrm{T}}$ if $i \in \mathcal{G}_3$.

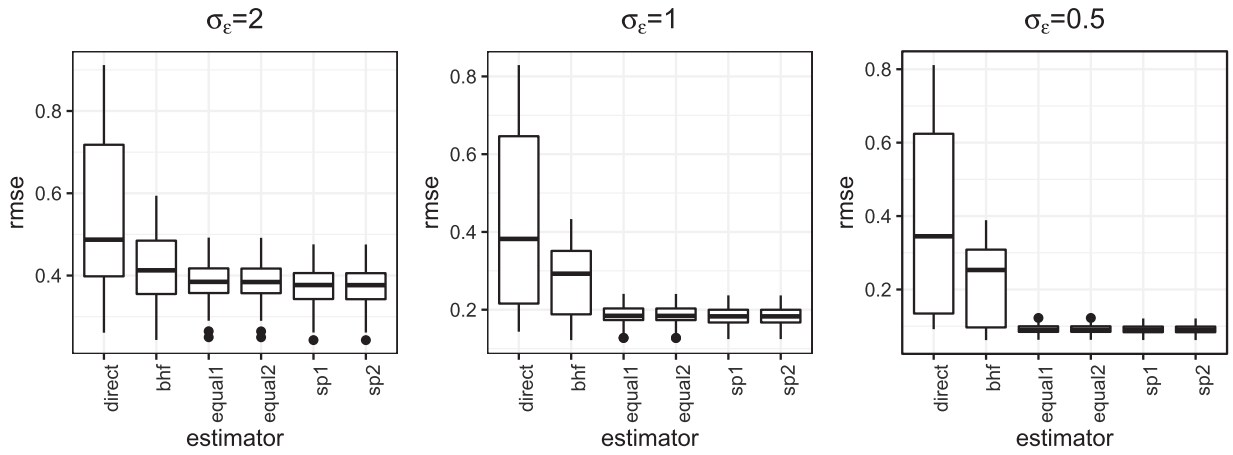

Figure 2. RMSE under Case I.

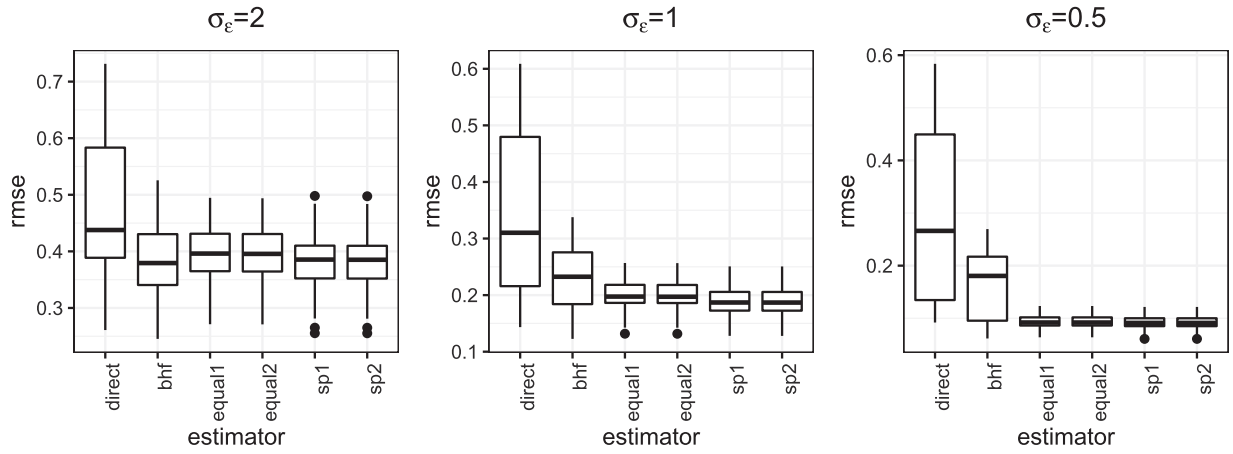

Figure 3. RMSE under Case II.

For each set of parameters, $\sigma_\epsilon = 0.5, 1, 2$ are considered and $\epsilon_{ih} \overset{iid}{\sim} N(0, \sigma_\epsilon^2)$. For the proposed estimator, we consider both the equal weight ($c_{ij} = 1$) and the spatial weight selected based on the modified BIC. Different estimators are compared by

$$\text{RMSE}\left(\hat{\bar{Y}}_i^E\right) = \sqrt{\frac{1}{B}\sum_{b=1}^{B}\left(\hat{\bar{Y}}_{i(b)}{}^{\text{E}} - \bar{Y}_{i(b)}\right)^2},$$

where $\hat{\bar{Y}}_{i(b)}^{\text{E}}$ is the estimated population mean in area i and $\bar{Y}_{i(b)}$ is the population mean in the bth simulation, 'E' is the index of estimators which can be 1 or 2, and $B = 100$. All the simulations are implemented in the Owens clusters of Ohio supercomputer centre (Ohio Supercomputer Center, 2016).

Figures 2, 3 and 4 show the results of the three sets of parameters when the sampling rate is 1% for 99 areas. 'direct' represents the direct estimator, which is

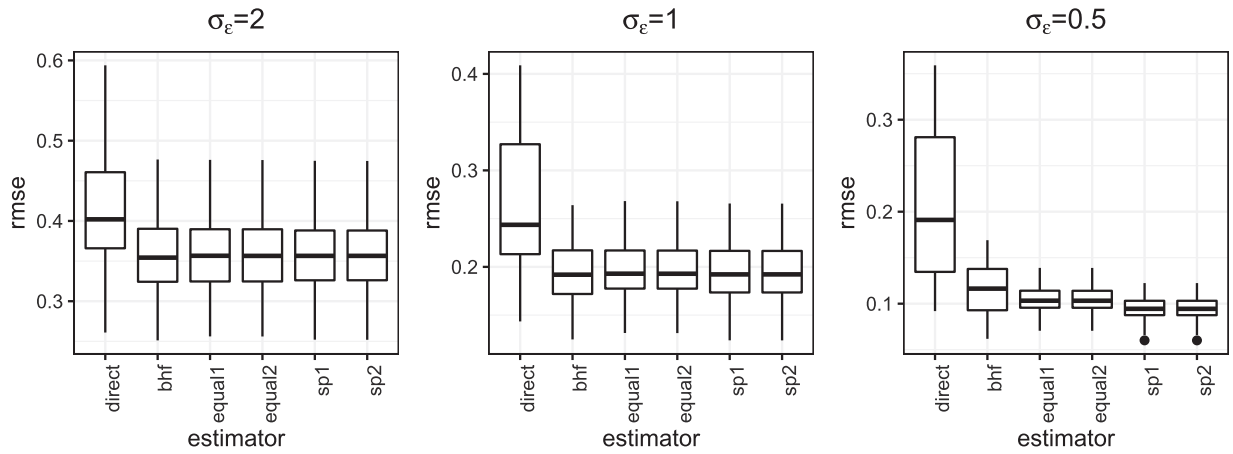

Figure 4. RMSE under Case III.

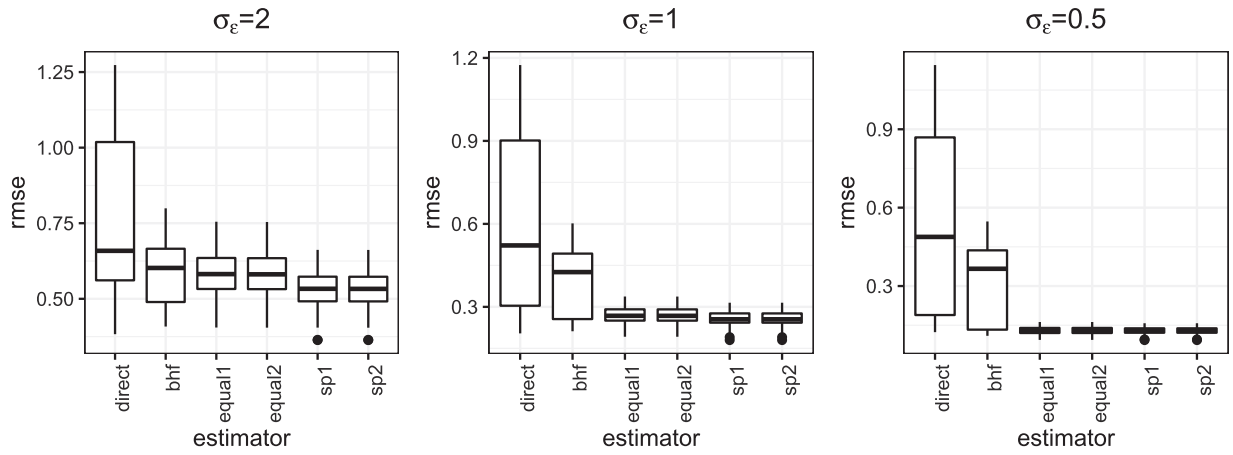

Figure 5. RMSE under Case I.

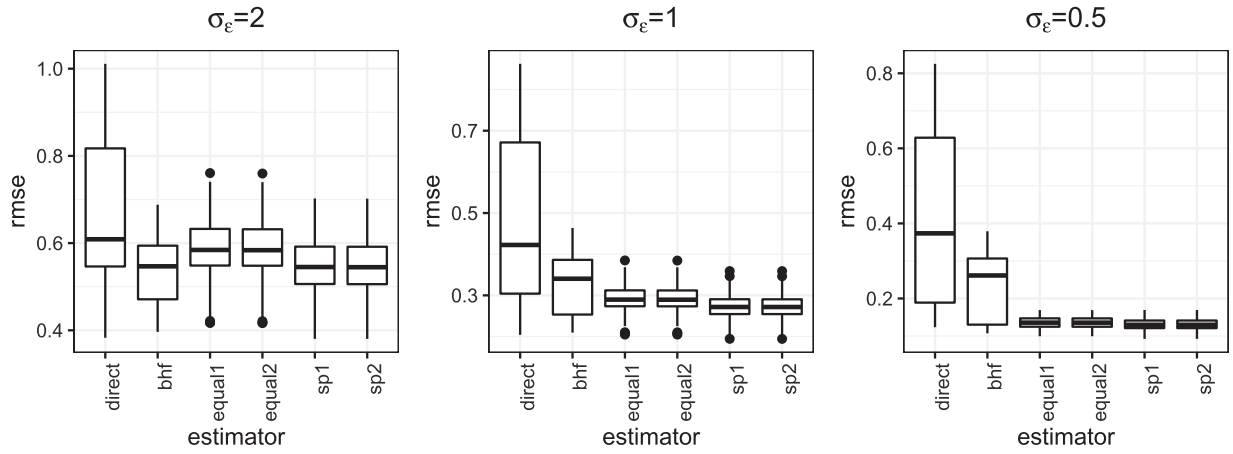

Figure 6. RMSE under Case II.

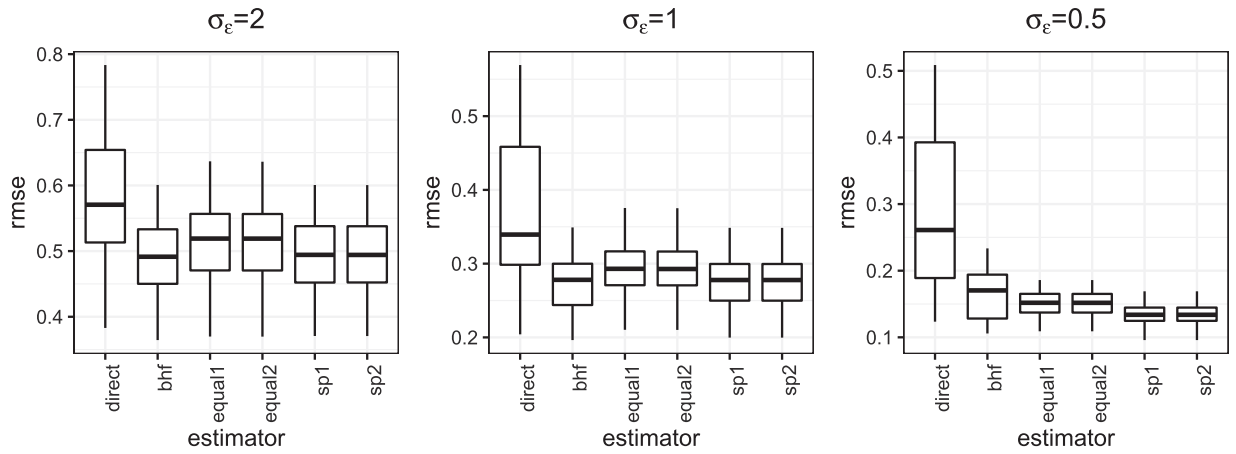

Figure 7. RMSE under Case III.

the sample mean for simple random sampling. 'BHF' is calculated using the *sae* package provided in (1). Under two different weights, we consider two small area estimators described in Section 2.3. 'equal1' and 'sp1' represent the estimator in (8) with equal weights and spatial weights, respectively. 'equal2' and 'sp2' represent the estimator in (9) with equal weights and spatial weights, respectively.

When σ_ϵ is large, the proposed new estimator has the similar performance to the BHF estimator when the group difference is small. As σ_ϵ becomes smaller, the performance gain of the proposed new estimator is better than the classical BHF estimator. Besides that, the estimator with spatial weights performs better than the estimator with equal weights. Since the sampling rate is small, thus f_i is small, there is not much difference between the two estimators in (8) and (9).

Figures 5, 6 and 7 show the results when the sampling rate is 0.5%. When the sample sizes become smaller and σ_ϵ is large, the proposed estimator with equal weights can be worse than the BHF estimator. But the estimator based on spatial weights is still comparable. Similarly to the case with sampling rate 1%, the proposed estimator performs better when the group difference is large or the value of σ_ϵ is small.

4. Real data analysis

In this section, we apply the proposed method to the NRI Iowa urban data in 2015. The auxiliary information used is based on the Landsat data (Li et al., 2018). The Landsat data is matched to the segment level data in the NRI based on segments' locations. The number of segments per county is from 7 to 56. We try different starting values and select the best one with spatial weights. After finding the estimated group structure, we refit the model with known group structure and find the regression coefficients in all groups and then obtain the estimates in each county. Figure 8 shows the estimated group map. And Figure 9 shows the estimated population mean of urban in each county with a comparison with the sample mean and the BHF estimator. The

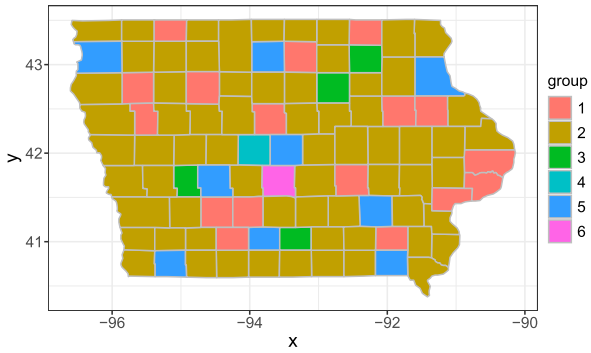

Figure 8. Estimated group structure.

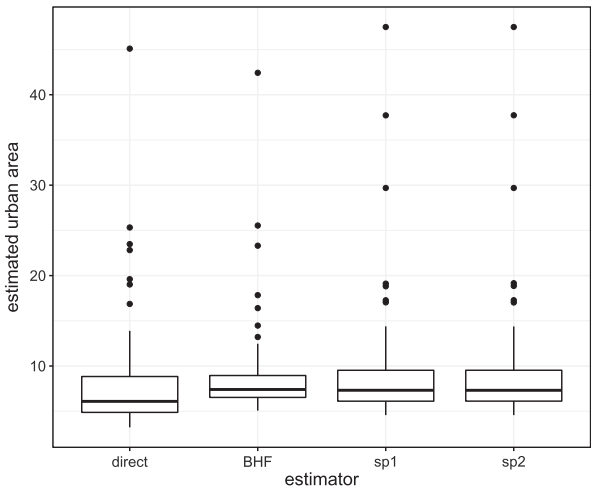

Figure 9. Estimated population mean of urban in each county.

proposed estimates are close to the estimates based on BHF, but with larger variations among different counties due to the fact that more than one groups are used in the estimates.

5. Summary and conclusion

In this article, we propose a new unit level small area estimator based on a penalised regression approach. In the new estimator, we can find subgroups of areas and also borrow information from both auxiliary

information and areas. Besides that, spatial information is also used in the algorithm. We use simulation studies to compare the performance of the new estimator to traditional estimators under several simulation settings, which show that the proposed estimator can improve the estimates.

Variance estimator is also important in survey sampling. A future work is to develop the variance estimator for the proposed new estimator. Another potential future work is to find subgroups for both regression coefficients and random effects together.

Disclosure statement

No potential conflict of interest was reported by the authors.

Funding

This research was supported in part by the Natural Resources Conservation Service of the U.S. Department of Agriculture.

Notes on contributors

Xin Wang is currently an Assistant professor in Department of Statistics at Miami University. Her research interests are spatial data analysis, Bayesian statistics, clustering, convergence rates of MCMC algorithms and survey sampling.

Zhengyuan Zhu is currently a Professor in Department of Statistics at Iowa State University, director of Center for Survey Statistics & Methodology. His research interests include spatial statistics, survey statistics, time series analysis, and multivariate analysis.

References

Battese, G. E., Harter, R. M., & Fuller, W. A. (1988). An error-components model for prediction of county crop areas using survey and satellite data. *Journal of the American Statistical Association*, *83*(401), 28–36.

Boyd, S., Parikh, N., Chu, E., Peleato, B., & Eckstein, J. (2011). Distributed optimization and statistical learning via the alternating direction method of multipliers. *Foundations and Trends in Machine Learning*, *3*(1), 1–122.

Fan, J., & Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American statistical Association*, *96*(456), 1348–1360.

Li, X., Zhou, Y., Zhu, Z., Liang, L., Yu, B., & Cao, W. (2018). Mapping annual urban dynamics (1985–2015) using time series of Landsat data. *Remote Sensing of Environment*, *216*, 674–683.

Ma, S., & Huang, J. (2017). A concave pairwise fusion approach to subgroup analysis. *Journal of the American Statistical Association*, *112*(517), 410–423.

Ma, S., Huang, J., & Zhang, Z (2016). Exploration of heterogeneous treatment effects via concave fusion. arXiv preprint arXiv:1607.03717.

Molina, I., & Marhuenda, Y. (2015). sae: An R package for small area estimation. *The R Journal*, *7*(1), 81–98.

Nusser, S. M., & Goebel, J. J. (1997). The national resources inventory: A long-term multi-resource monitoring programme. *Environmental and Ecological Statistics*, *4*(3), 181–204.

Ohio Supercomputer Center (2016). Owens supercomputer. http://osc.edu/ark:/19495/hpc6h5b1.

Pfeffermann, D. (2013). New important developments in small area estimation. *Statistical Science*, *28*(1), 40–68.

Rao, J. N., & Molina, I. (2015). *Small area estimation*. Hoboken, New Jersey: John Wiley & Sons.

Wang, H., Li, R., & Tsai, C.-L. (2007). Tuning parameter selectors for the smoothly clipped absolute deviation method. *Biometrika*, *94*(3), 553–568.

Wang, X., Zhu, Z., & Zhang, H. H (2019). Spatial automatic subgroup analysis for areal data with repeated measures. arXiv preprint arXiv:11906.01853.

Zhang, C.-H. (2010). Nearly unbiased variable selection under minimax concave penalty. *The Annals of statistics*, *38*(2), 894–942.

Zhu, R., Zou, G., Liang, H., & Zhu, L. (2016). Penalized weighted least squares to small area estimation. *Scandinavian Journal of Statistics*, *43*(3), 736–756.

Appendix

In this appendix, details of partial derivative are provided.

$$\frac{\partial l}{\partial \sigma_v^2} = -\frac{1}{2}\sum_{i=1}^{m}\frac{1}{n_i}tr\left(\Sigma_i^{-1}\frac{\partial \Sigma_i}{\partial \sigma_v^2}\right) - \frac{1}{2}\sum_{i=1}^{m}\frac{1}{n_i}\left(\boldsymbol{y}_i - \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}_i\right)^{\mathrm{T}}\frac{\partial \Sigma_i^{-1}}{\partial \sigma_v^2}\left(\boldsymbol{y}_i - \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}_i\right),$$

$$\frac{\partial l}{\partial \sigma_\epsilon^2} = -\frac{1}{2}\sum_{i=1}^{m}\frac{1}{n_i}tr\left(\Sigma_i^{-1}\frac{\partial \Sigma_i}{\partial \sigma_\epsilon^2}\right) - \frac{1}{2}\sum_{i=1}^{m}\frac{1}{n_i}\left(\boldsymbol{y}_i - \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}_i\right)^{\mathrm{T}}\frac{\partial \Sigma_i^{-1}}{\partial \sigma_\epsilon^2}\left(\boldsymbol{y}_i - \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}_i\right),$$

where

$$\frac{\partial \Sigma_i}{\partial \sigma_v^2} = \mathbf{1}_{n_i}\mathbf{1}_{n_i}^{\mathrm{T}}, \quad \frac{\partial \Sigma_i}{\partial \sigma_\epsilon^2} = \boldsymbol{I}_{n_i},$$

$$\frac{\partial \Sigma_i^{-1}}{\partial \sigma_v^2} = -\frac{1}{\left(\sigma_\epsilon^2 + n_i\sigma_v^2\right)^2}\mathbf{1}_{n_i}\mathbf{1}_{n_i}^{\mathrm{T}}$$

$$\frac{\Sigma_i^{-1}}{\partial \sigma_\epsilon^2} = \frac{1}{\left(\sigma_\epsilon^2\right)^2}\left[\frac{\sigma_v^2\left(2\sigma_\epsilon^2 + n_i\sigma_v^2\right)}{\left(\sigma_\epsilon^2 + n_i\sigma_v^2\right)^2}\mathbf{1}_{n_i}\mathbf{1}_{n_i}^{\mathrm{T}} - \boldsymbol{I}_{n_i}\right].$$

$\mathcal{I}$ can be written as

$$\mathcal{I} = \begin{pmatrix}\mathcal{I}_{11} & \mathcal{I}_{12}\\ \mathcal{I}_{21} & \mathcal{I}_{22}\end{pmatrix},$$

where

$$\mathcal{I}_{11} = \frac{1}{2}\sum_{i=1}^{m}\frac{1}{n_i}tr\left(\Sigma_i^{-1}\frac{\partial \Sigma_i}{\partial \sigma_v^2}\Sigma_i^{-1}\frac{\partial \Sigma_i}{\partial \sigma_v^2}\right) = \frac{1}{2}\sum_{i=1}^{m}\frac{n_i}{\left(\sigma_\epsilon^2 + n_i\sigma_v^2\right)^2},$$

$$\mathcal{I}_{12} = \mathcal{I}_{21} = \frac{1}{2}\sum_{i=1}^{m}\frac{1}{n_i}tr\left(\Sigma_i^{-1}\frac{\partial \Sigma_i}{\partial \sigma_v^2}\Sigma_i^{-1}\frac{\partial \Sigma_i}{\partial \sigma_\epsilon^2}\right) = \frac{1}{2}\sum_{i=1}^{m}\frac{1}{\left(\sigma_\epsilon^2 + n_i\sigma_v^2\right)^2},$$

$$\mathcal{I}_{22} = \frac{1}{2}\sum_{i=1}^{m}\frac{1}{n_i}tr\left(\Sigma_i^{-1}\frac{\partial \Sigma_i}{\partial \sigma_\epsilon^2}\Sigma_i^{-1}\frac{\partial \Sigma_i}{\partial \sigma_\epsilon^2}\right) = \frac{1}{2\left(\sigma_\epsilon^2\right)^2}\sum_{i=1}^{m}\left[1 - \frac{\sigma_v^2\left(2\sigma_\epsilon^2 + n_i\sigma_v^2\right)}{\left(\sigma_\epsilon^2 + n_i\sigma_v^2\right)^2}\right].$$