Statistica Sinica **28** (2018), 921-940
doi:https://doi.org/10.5705/ss.202016.0162

STATISTICAL-PHYSICAL ESTIMATION OF POLLUTION EMISSION

Youngdeok Hwang, Emre Barut and Kyongmin Yeo

Sungkyunkwan University, George Washington University and IBM Thomas J. Watson Research Center

Abstract: Air pollution is driven by non-local dynamics, in which air quality at a site is determined by transport of pollutants from distant pollution emission sources to the site by atmospheric processes. To understand the underlying nature of pollution generation, it is crucial to employ physical knowledge to account for pollution transport by wind. However, in most cases, it is not possible to utilize physics models to obtain useful information; this would require massive calibration and computation. In this paper, we propose a method to estimate the pollution emission from the domain of interest by using the physical knowledge and observed data. The proposed method uses an efficient optimization algorithm to estimate the emission from each of the spatial locations, while incorporating physics knowledge. We demonstrate the effectiveness of the new method through a simulation study.

Key words and phrases: Alternating direction method of multipliers, dispersion, inverse model, penalized regression.

1. Introduction

Air pollution is produced by natural and anthropogenic emissions transported via physical processes driven by wind. The pollutant can be from a variety of sources, including traffic, fossil fuel uses, or burning natural biomass (World Health Organization (2005)). There has been substantial progress in environmental science and engineering in developing computational models to forecast the evolution of physical processes. Notable examples include Weather Research and Forecasting coupled with Chemistry (WRF-Chem, Fast et al. (2006)) and Community Multi-scale Air Quality Model (CMAQ, Byun and Schere (2006)) among many others. These models use knowledge from the physical and chemical processes to construct a system of partial differential equations that compute the transport of pollutant particles from where they are emitted to nearby areas under given weather conditions.

The quality of the physical model prediction depends heavily on the accuracy of the information as to how much pollution is emitted from each geographic location, among many model input parameters. Such pollution emission information is often inaccurate, incomplete, or even unavailable. For example, Fu et al. (2012) compared smoke emission rates from wild fire estimated by two most widely used models, finding that the difference can be a factor of five to eight. It is critical to reduce such a large uncertainty in emission information. Conceptually, this is an 'inverse' problem, as opposed to a 'forward' problem that generates data using given model parameters; interest is estimating physical model parameters using model outputs and observational data. In applied mathematics and computational physics, such are often referred to as *Inverse problems* (e.g., Biegler et al. (2011)).

From a statistical point of view, this is closely related to the *calibration* of a physics model. In this direction, Kennedy and O'Hagan (2001) solved the calibration problem using a Bayesian framework with Gaussian process. Higdon et al. (2008) proposed an extension of Kennedy and O'Hagan (2001) to incorporate the high dimensionality of the computer model outputs. Liu and West (2009) proposed combining Bayesian multivariate dynamic linear models with Gaussian processes for a model with time-dependent functional outputs.

Another effort solves the air pollution problem by building statistical air quality models using measurement data. Carroll et al. (1997) built a spatial-temporal model for hourly ozone level using a Gaussian random field. Paciorek et al. (2009) developed a practical modeling approach to solve the epidemiological problem. Ghosh et al. (2010) studied formation and deformation of atmospheric concentrations of total nitrate using empirical chemical relationships and dynamic statistical models. Williams, Christensen and Reese (2011) estimated the pollution source direction using a dispersion model, treating the computer model outputs as given. Although useful, these efforts do not incorporate enough physical knowledge as to the pollution generation process.

In utilizing a computational model to solve the air quality problem, one makes use of observed data from a monitoring network located over the spatial domain to estimate emission source information. It is crucial to include the knowledge from physics into the modeling framework to incorporate the continuously changing dynamic nature of weather conditions and pollution emission. Physics can introduce such components into the modeling framework, but it is only possible when it is paired with the appropriate statistical modeling.

We propose a statistical framework to exploit physical assumptions of the model linking the computational physics model to the prediction, while obtaining the critical information about the nature of pollution generation. This interdisciplinary framework can also be used for solving the challenging problem of using both physical and statistical knowledge. The major concern of our approach is practicality in computational and operational complexity. At a conceptual level, our work is aligned with Malmberg et al. (2008) in that we combine the statistical model and physical knowledge.

We introduce a reduced order physics model that pairs well with a statistical method to estimate the emission intensity surface of the area of interest. We then describe how our physics model can be reformulated as a regression problem. We develop as well an efficient algorithm to estimate the detailed emission information at each location as it changes over time. Our model imposes sparsity on the estimated coefficients so that the dimensionality of the inverse problem, as well as prior knowledge on the emission, can be properly incorporated. Our method brings in how much pollution emission is produced so that this can be used for policy or administrative decision purposes.

The remainder of the paper is organized as follows. Section 2 describes the fundamental physical process regarding the transport of a pollution on a spatial domain. Section 3 gives our statistical model. Section 4 describes an optimization methodology to estimate the parameters. Section 5 presents the application of the proposed method to synthetic air-quality monitoring data, in which the results and interpretations are presented together. We conclude with some remarks and discussion in Section 6.

2. Physical Process

In this section, we provide a brief introduction to our physical model, and describe how it is connected to the statistical model in Section 3.

The building block of our model is the *dispersion* process. It has two major ingredients, *advection* and *diffusion.* Advection is the transfer of pollutants from one location to another, following the streamline of wind. Diffusion characterizes the movement of pollution from a region of high concentration to one of low concentration due to mixing by atmospheric turbulence.

Let $\phi(\boldsymbol{s}, t)$ denote the pollution concentration at location $\boldsymbol{s}$ at time t. The dispersion process can be expressed as

$$\frac{\partial \phi(\boldsymbol{s}, t)}{\partial t} = -\nabla \cdot (\boldsymbol{u}(\boldsymbol{s}, t)\phi(\boldsymbol{s}, t)) + \nabla \cdot \{\boldsymbol{K}(\boldsymbol{s}, t; \boldsymbol{u}) \cdot \nabla \phi(\boldsymbol{s}, t)\} + Q(\boldsymbol{s}, t), \tag{2.1}$$

which implies that the temporal change of pollution concentration at a location is determined by the advection (the first term on the right hand side) and the diffusion (the second term). Here, $\boldsymbol{u}(\boldsymbol{s},t)$ is the velocity of wind, which can be obtained from a numerical weather prediction model, and $\boldsymbol{K}$ is a diffusion coefficient matrix, defining the rate of mixing of the pollutant. The term $Q(\boldsymbol{s},t)$ represents the rate of newly added pollution at location $\boldsymbol{s}$ and time t.

We illustrate the model in (2.1) using a simple example. Consider a one-dimensional computational domain of three grids, $\boldsymbol{s} = (s_1, s_2, s_3)$. We assume that these grids are the mid-subset of a much larger domain, so that boundary conditions do not affect the calculation. The model output over these grids at each time point t can be expressed as a vector of length three. Now, suppose that the initial concentration at $t = 0$ is zero for all $\boldsymbol{s}$, the wind $\boldsymbol{u} = (1, 1, 1)$, uniform diffusion $K = 1/4$, the spatial grid $\delta_x = 1$, and time step $\delta_t = 1$. The emissions at these three grids are the same for all t, $\boldsymbol{\beta}$, so $Q(\boldsymbol{s},t) = Q(\boldsymbol{s}) = \boldsymbol{\beta} = (\beta_1, \beta_2, \beta_3)$. Using a first order Taylor expansion with finite-difference discretization (Moin (2010)), (2.1) at the j-th grid s_j at time $t+1$ can be approximated by

$$\begin{aligned}\phi(s_j, t+1) \approxeq\, & \phi(s_j,t) + \phi(s_{j-1},t) - \phi(s_j,t) \\ & + \frac{1}{4}\left\{\phi(s_{j-1},t) - 2\phi(s_j,t) + \phi(s_{j+1},t)\right\} + \beta_j,\end{aligned}$$

for $j = 1, 2, 3$. Then for $t = 1, 2$,

$$\phi(\boldsymbol{s},1) = \begin{bmatrix} \frac{1}{2} & \frac{1}{4} & 0 \\ \frac{5}{4} & \frac{1}{2} & \frac{1}{4} \\ 0 & \frac{5}{4} & \frac{1}{2} \end{bmatrix} \begin{bmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \end{bmatrix}; \ \phi(\boldsymbol{s},2) = \begin{bmatrix} \frac{22}{16} & 0 & \frac{1}{16} \\ 0 & \frac{22}{16} & 0 \\ \frac{25}{16} & 0 & \frac{22}{16} \end{bmatrix} \begin{bmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \end{bmatrix}, \tag{2.2}$$

and further calculations proceed similarly.

Now we link the calculated dispersion to monitoring observations using this formulation. We take $\tilde{X}_{t,ij}$ as the dispersion at ith monitoring location, $\boldsymbol{s}_i$, at time t, originating from the jth grid computed by (2.1). In (2.2), for example, $(\tilde{X}_{1,11}, \tilde{X}_{1,12}, \tilde{X}_{1,13}) = (1/2, 1/4, 0)$; $(\tilde{X}_{1,21}, \tilde{X}_{1,22}, \tilde{X}_{1,23}) = (5/4, 1/2, 1/4)$. The pollution concentration observation at $\boldsymbol{s}_i$ at time t and the associated noise are denoted by $y_{t,i}$ and $\epsilon_{t,i}$, respectively, for $t = 1, \ldots, T$ and $i = 1, \ldots, n$. Denoting the emission from grid j by β_j and assuming there are $\tilde{p}$ grids, we have the data generating process

$$y_{t,i} = \sum_{j=1}^{\tilde{p}} \tilde{X}_{t,ij}\beta_j + \epsilon_{t,i}, \ \text{ for } \ i = 1, \ldots, n, \ \text{ and } t = 1, \ldots, T. \tag{2.3}$$

Measurements are available as the observed concentration levels similar to $\phi(\boldsymbol{s}_i, t)$, but with added noise. The physical dispersion process in (2.1) is incorporated in $\tilde{X}_{t,ij}\beta_j$, and the inherent randomness of the measurements are modeled with $\epsilon_{t,i}$.

The convenience here comes from the linearity of the emission intensity $\boldsymbol{\beta}$. When calculating the dispersion, emissions need not be known. As seen above, the dispersion from a source at s_1 can be calculated with a unit vector $Q(\boldsymbol{s}) = (1,0,0)$ for the desired time interval. Once the dispersion field is obtained, the intensity can be scaled by multiplying it by β_1. Similarly, dispersion from the other two coordinates can be calculated with $(0,1,0)$ and $(0,0,1)$ and scaled with β_2 and β_3, which makes $Q(\boldsymbol{s})$ linear in $(\beta_1, \beta_2, \beta_3)$.

In applications, the goal is to estimate the pollution emission surface $Q(\boldsymbol{s})$ over the spatial domain. Without further simplifications, estimation of the pollution emission over all of the spatial domain requires solving an infinite dimensional problem. Accordingly, (2.1) has been used with strong assumptions on the sources. For example, Keats, Yee and Lien (2007) considered the problem of finding the location and emission intensity of a source, while assuming there is only one.

Alternatively, we approximate the emission surface by using a basis representation

$$Q(\boldsymbol{s}) = \sum_{j=1}^{p} \beta_j \Phi(\|\boldsymbol{s} - \boldsymbol{v}_j\|; \tau) = \sum_{j=1}^{p} \beta_j \Phi_j, \tag{2.4}$$

where β_j is the emission intensity associated with jth component, and $\Phi(\|\boldsymbol{s} - \boldsymbol{v}_j\|; \tau) = \exp(-\|\boldsymbol{s} - \boldsymbol{v}_j\|^2/2\tau^2)/2\pi\tau^2$ with $\tau > 0$, a Gaussian density kernel centered at location $\boldsymbol{v}_j$. For a given τ, the $\Phi(\|\boldsymbol{s} - \boldsymbol{v}_j\|; \tau)$ serve as pre-determined basis functions. Then, the emission surface is approximated by a sum of p smooth kernels, centered at the fixed grid points covering the domain, $\boldsymbol{v}_1, \ldots, \boldsymbol{v}_p$. The number of kernels is chosen to be much smaller than the number of grids, $p \ll \tilde{p}$, so it makes the problem tractable. The amount of emission contribution from the area around $\boldsymbol{v}_j$ is estimated by β_j. The computation is still conducted over $\tilde{p}$ grids, but the emissions from the kernels are assumed to behave jointly. Figure 1 shows two basis functions at two locations, where the entire domain is divided into 64 cells, and each cell is represented by the basis centered at the grid in dashed-lines.

As an illustration, Figure 2 shows a series of dispersion model outputs from two sources in the form of (2.4) marked by the red circles, for three consecutive hours. The arrows represent the wind field changing over time, where the color contours depict $\phi(\boldsymbol{s}, t)$ due to the emission concentration. In the following section,

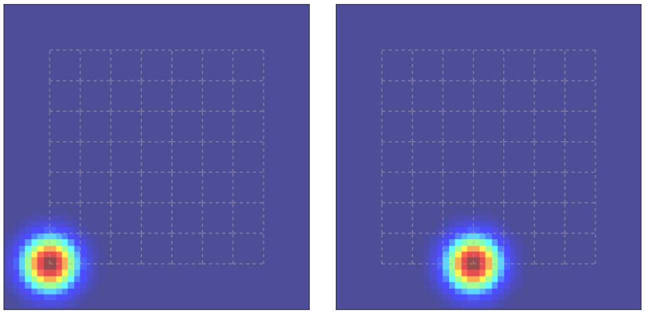


Figure 1. Illustration of our model approach. Dashed grid show the location of centers of basis functions, eight in both horizontal and vertical direction, where two panels show the two sites located at the two grid points of the domain.

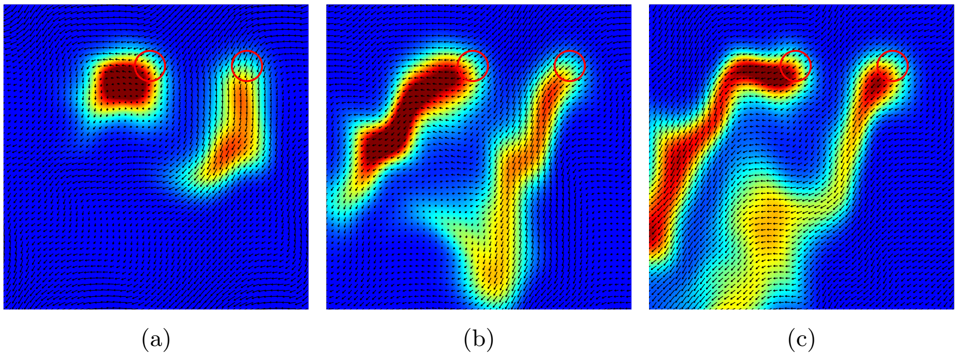


Figure 2. A series of model outputs depicting the dispersion from two sources marked by the red circles, where the snapshots are taken at hour 1 to 3 (a-c). The arrows in the background indicate the changing wind field.

we illustrate how (2.3) and (2.4) are paired to solve our problem.

3. Model

In this section, we present our model extended from equations (2.3) and (2.4). As illustrated in the discussion following equation (2.4), we divide the domain of interest into p sources. We further assume that the emission rate of each source has a diurnal pattern according to the 24-hour cycle of the day. Hence, for $j = 1, \ldots, p$, there is a vector of intensity parameter $\boldsymbol{\beta}(j) = (\beta_{j,1}, \ldots, \beta_{j,24})$. The dispersion emitted from source j and located at station i at time t, calculated using (2.1), is denoted by $X_{t,ij}$. Following these assumptions, we have

$$y_{t,i} = \sum_{j=1}^{p} X_{t,ij}\beta_{j,h(t)} + \epsilon_{t,i}, \tag{3.1}$$

where $h(t) = (t \mod 24) + 1$. This implies that each measurement is represented as a sum of p source location contributions, each of which is further decomposed into sum of 24 components. A similar idea was adopted in the classical chemical mass balance model (Christensen and Gunst (2004)), but in a simpler form. The error terms $\epsilon_{t,i}$ are assumed to be mutually independent.

We stack the hourly emission coefficients for each location, $\boldsymbol{\beta}(j)$, in a $p \times 24$ matrix $\boldsymbol{\beta}$ such that jth row of $\boldsymbol{\beta}$ is $\boldsymbol{\beta}(j)$ $\boldsymbol{\beta} = [\boldsymbol{\beta}(1)^\top, \ldots, \boldsymbol{\beta}(24)^\top]^\top = (\beta_{j,k})$. Although this problem can be solved with multivariate, multi-response linear regression, there are issues due to high dimensionality. In most practical applications, the number of possible emission locations, p, is high, and estimation of hourly emissions requires us to fit $p \times 24$ variables. On the other hand, we expect to have a small number of samples, $n \times T$. Simulating pollution dispersion at a high resolution requires substantial computational effort, and pollution sensors are often costly, which leads to small T and n.

In a setting where the number of variables is of the same order of magnitude as the number of samples, we have to utilize prior knowledge and enforce a special structure on the estimated coefficients. This study was done with emphasis on urban spaces, where two patterns stand out (Gurjar et al. (2004); Morawska (2006); Saarikoski et al. (2008)): most locations emit negligible amounts of pollution; major pollution sources are traffic (motor vehicles, airports and seaports) and industrial areas. They produce pollution in different hourly patterns, and we formulate our model based on them.

Based on the first pattern, $\boldsymbol{\beta}(j)$ should be zero for many j since most locations to do not generate significant pollution. The rows of $\boldsymbol{\beta}$ should be sparse. We enforce this assumption with the group lasso penalty (Yuan and Lin (2006); Simon and Tibshirani (2012)), by adding a penalty term proportional to $\sum_{j=1}^{p} \|\boldsymbol{\beta}(j)\|_2$, where $\|\cdot\|_2$ is the Euclidean norm. The sub-level sets of the group lasso penalty function contain coefficients for which only a few of the rows are non-zero, and hence this penalty ‘sparsifies’ the solution with respect to the rows of $\boldsymbol{\beta}$. The group lasso is commonly used in high-dimensional problems where coefficients are expected to be active (i.e. non-zero) in groups of variables.

According to the second pattern, there are different sources of pollution, but the same sources tend to have similar hourly patterns. Traffic pollution occurs commonly around the main highways, usually spikes in the morning and evening during the rush hours, and is generally constant otherwise. Industrial areas often have emissions that peak around noon, and there are certain industrial areas (e.g. facilities that are not shutdown during the night) that emit a constant level of pollution throughout the day. Ideally, locations' hourly pollution patterns are known in advance, and these can be enforced as conditions on the estimated coefficients, $\boldsymbol{\beta}$. In practice, it is usually difficult to know which regions might have which patterns, especially when the number of potential emission sources is high. We will assume that there are a few unknown daily patterns, and each location follows one of these patterns or a combination of them (e.g. emissions from locations that have power plants with nearby highways might be given by the sum of two daily patterns). Unknown daily patterns can also be thought as latent factors which generate the emission coefficients. In this setting, we expect the rows of $\boldsymbol{\beta}$ to be linearly dependent and that rank($\boldsymbol{\beta}$) should be small. We enforce this assumption with a nuclear norm penalty, which encourages sparsity in singular values (Candès and Recht (2009)). For this, first observe that for an $p \times m$ matrix $\boldsymbol{A}$ with $(p \geq m)$, its nuclear norm $\| \cdot \|_*$ is given by

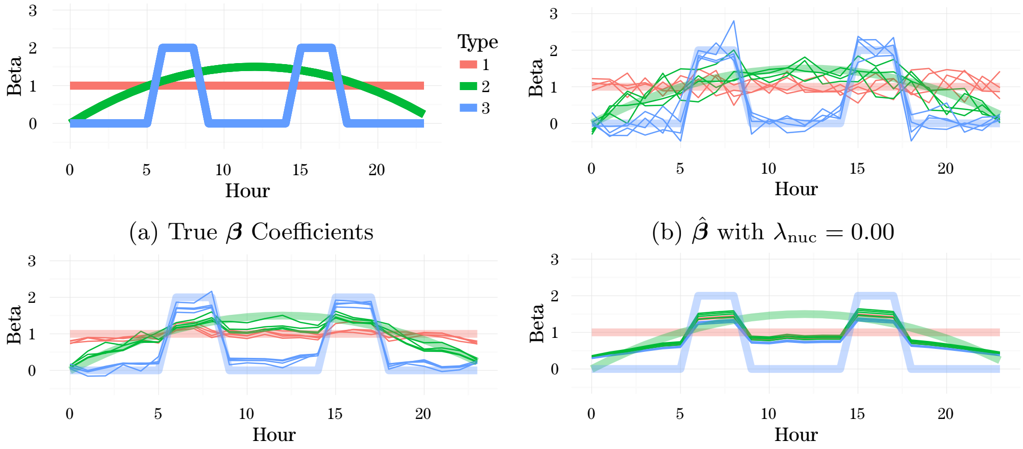
$$\|\boldsymbol{A}\|_* = \sum_{l=1}^{m} \sigma_l,$$

where σ_l is the l^{th} singular value of $\boldsymbol{A}$, l^{th} element of diagonal matrix of $\boldsymbol{\Sigma}$, defined by the singular value decomposition (SVD) of $\boldsymbol{A} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^\top$. Hence this penalty controls linear dependencies among the rows of $\boldsymbol{\beta}$, $\boldsymbol{\beta}(j)$. Example 1 demonstrates how the penalty on singular values works in practice.

Example 1. We display sets of coefficients obtained by solving the nuclear norm penalized regression problem:

$$\min_{\boldsymbol{\beta}} n^{-1}\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_F^2 + \lambda_{\text{nuc}}\|\boldsymbol{\beta}\|_*,$$

where $\| \cdot \|_F$ is the matrix Frobenius norm, $\mathbf{y} \in \mathbb{R}^{T \times 24}, \mathbf{X} \in \mathbb{R}^{T \times p}$, and $\boldsymbol{\beta} \in \mathbb{R}^{p \times 24}$ are the response, predictor, and coefficient matrices, respectively. These variables bear no relationship to other $\mathbf{y}, \mathbf{X}$ and $\boldsymbol{\beta}$ used in the remainder of the paper. The term $\lambda_{\text{nuc}} \geq 0$ is a tuning parameter that controls the effect of the nuclear norm penalty; as it increases, the minimizer is forced to have more linear dependency. We created a toy example with $p = 12$ and $T = 200$ and generated observations from (3.1). Other details, such as the choice of $\mathbf{X}$ and ϵ, are relegated to the Supplementary Material. The true coefficients, $\boldsymbol{\beta}$, and the estimates for three different choices of λ_{nuc} are given in Figure 3. The true coefficients are given by the set of three vectors: $\beta_{\text{Type1}} = (1, \ldots, 1)^\top$, β_{Type2} is a concave quadratic function that peaks at the 12^{th} hour, and β_{Type3} is a vector that contains zeroes except for hours 6-8 and hours 15-17. For each type, four β's were generated which gives $\boldsymbol{\beta}$ with $p = 12$ when stacked.



(a) True $\boldsymbol{\beta}$ Coefficients

(b) $\hat{\boldsymbol{\beta}}$ with $\lambda_{\text{nuc}} = 0.00$

(c) $\hat{\boldsymbol{\beta}}$ with $\lambda_{\text{nuc}} = 0.25$

(d) $\hat{\boldsymbol{\beta}}$ with $\lambda_{\text{nuc}} = 1.00$

Figure 3. Effect of the nuclear norm penalty. Figure 3a displays the values of $\boldsymbol{\beta}$ coefficients for each type. Figures 3b, 3c, and 3d display the fitted $\boldsymbol{\beta}$ obtained using different λ_{nuc}, where each line is one row of $\hat{\boldsymbol{\beta}}$ and the colors display the type of the coefficient. As the weight of the nuclear norm penalty increases, so does the linear dependence between the estimates.

The effects of the nuclear norm penalty on the fitted coefficients are depicted in Figure 3. Coefficients obtained with penalization tend to be more similar, as we force the rows of $\boldsymbol{\beta}$ to be more linearly dependent. The statistical benefits of the penalty are obvious: the estimate with no penalization has a high variance, and a high estimation error as in Figure 3b. When we introduce the nuclear norm penalty term, the estimates have lower matrix rank and since the true coefficients also have a low-rank, this results in estimates with low variance and estimation error, as in Figure 3c. We also introduce some bias: with $\lambda_{\text{nuc}} = 0.25$, Type 2 locations' coefficients are underestimated, but the estimates are much closer to the truth than the results obtained without any penalties. When the penalty term is very large, $\lambda_{\text{nuc}} = 1.00$, $\hat{\boldsymbol{\beta}}$ has rank one, and all of the coefficients are given by the same vector multiplied by a scalar, as seen in Figure 3d, where rows of $\hat{\boldsymbol{\beta}}$ have the same shape. This is the case where we introduce massive bias by significantly shrinking the variance of the estimates.

Remaining assumption in our model is that each source can only *add* emission to the ambient air. Any decay or deposit is assumed to be negligible and hence is absorbed in the error term. Thus we force all of the emission coefficients to be non-negative.

We combine these regularization penalties with a least-squares loss to obtain the objective function

$$\min_{\beta_{jk}\geq 0} \frac{1}{nT}\sum_{t=1}^{T}\sum_{i=1}^{n}\left(y_{t,i}-\sum_{j=1}^{p}X_{t,ij}\beta_{j,h(t)}\right)^2+\lambda_{\text{gl}}\sum_{j=1}^{p}\|\boldsymbol{\beta}(j)\|_2+\lambda_{\text{nuc}}\|\boldsymbol{\beta}\|_*, \quad (3.2)$$

where $\lambda_{\text{gl}} \geq 0$ and $\lambda_{\text{nuc}} \geq 0$ are tuning parameters for the group lasso and nuclear norm penalties, respectively. As before, $h(t) = (t \mod 24) + 1$. The objective function is a sum of convex functions, and hence is convex. The non-negativity constraint is also convex, resulting in a convex problem.

4. Estimation

Multiple non-differentiable components in (3.2) make our problem challenging. For efficient optimization, we propose an *Alternating Direction Method of Multipliers* (ADMM) based approach (Boyd et al. (2010); Parikh and Boyd (2014)). ADMM is built to minimize objective functions with separable components. The algorithm works in a distributed and iterative manner.

There are four main components in (3.2): sum of squared errors, group Lasso penalty, nuclear norm penalty, nonnegative constraints. In this regard, (3.2) is rewritten as

$$\text{minimize } f_{\text{mse}}(\boldsymbol{\beta}) + f_{\text{gl}}(\boldsymbol{\beta}) + f_{\text{nuc}}(\boldsymbol{\beta}) + f_{\text{nn}}(\boldsymbol{\beta}), \quad (4.1)$$

where

$$f_{\text{mse}}(\boldsymbol{\beta}) = (nT)^{-1}\sum_{t=1}^{T}\sum_{i=1}^{n}\left(y_{t,i}-\sum_{j=1}^{p}X_{t,ij}\beta_{j,h(t)}\right)^2,$$

$$f_{\text{gl}}(\boldsymbol{\beta}) = \lambda_{\text{gl}}\sum_{j=1}^{p}\|\boldsymbol{\beta}(j)\|_2,$$

$$f_{\text{nuc}}(\boldsymbol{\beta}) = \lambda_{\text{nuc}}\|\boldsymbol{\beta}\|_*,$$

and $f_{\text{nn}}(\boldsymbol{\beta}) = \sum_{j=1}^{p}\sum_{k=1}^{24}\delta_{\mathbb{R}_+}(\beta_{j,k})$, where

$$\delta_{\mathbb{R}_+}(z) = \begin{cases} 0 & \text{if } z \geq 0, \\ +\infty & \text{otherwise.} \end{cases}$$

Intuitively, our algorithm separately updates the solution to improve in $f_{\text{mse}}, \ldots, f_{\text{nn}}$, and then gradually pulls the solutions toward their average while improving in each direction. The optimization is summarized in Algorithm 1. The detailed effect of each step in Algorithm 1 is deferred to the Supplementary

Algorithm 1 ADMM algorithm to obtain the estimator.

```
Set the initial estimates $\boldsymbol{\beta}^{(1)}_{\mathrm{mse}}, \boldsymbol{\beta}^{(1)}_{\mathrm{gl}}, \boldsymbol{\beta}^{(1)}_{\mathrm{nuc}}, \boldsymbol{\beta}^{(1)}_{\mathrm{nn}}, \bar{\boldsymbol{\beta}}^{(1)}$;
Set the initial differences $\boldsymbol{u}^{(1)}_{\mathrm{mse}}, \boldsymbol{u}^{(1)}_{\mathrm{gl}}, \boldsymbol{u}^{(1)}_{\mathrm{nuc}}, \boldsymbol{u}^{(1)}_{\mathrm{nn}}$.
for $k = 1, \dots, 24$ do                                                ▷ Reshape data
    for $i = 1, \dots, n$ do
        $\mathbf{X}(i,k) \leftarrow \{X_{t,ij} : (t \mod 24) = k-1\}$;
        $\mathbf{y}(i,k) \leftarrow \{y_{t,i} : (t \mod 24) = k-1\}$;
    end for
    $\mathbf{X}(k)^\top \leftarrow [\mathbf{X}(1,k)^\top, \dots, \mathbf{X}(n,k)^\top]$;
    $\mathbf{y}(k)^\top \leftarrow [\mathbf{y}(1,k)^\top, \dots, \mathbf{y}(n,k)^\top]$ ;
end for
for $m = 1, \dots, M$ do                                                 ▷ ADMM Iterations
    for $k = 1, \dots, 24$ do
        $\boldsymbol{\beta}^{(m+1)}_{\mathrm{mse},(:,k)} \leftarrow \left(\mathbf{X}(k)^\top\mathbf{X}(k) + \frac{1}{2\rho}\mathbb{I}_{p\times p}\right)^{-1} \mathbf{X}(k)^\top\mathbf{y}(k) + \frac{1}{2\rho}\left(\bar{\boldsymbol{\beta}}^{(m)}_{:,k} - \rho\boldsymbol{u}^{(m)}_{\mathrm{mse}}(k)\right)$;
                                                                         ▷ MSE
    end for
    $\boldsymbol{\beta}^{(m+1)}_{\mathrm{gl}} \leftarrow \mathrm{sign}(\bar{\boldsymbol{\beta}}^{(m)} - \boldsymbol{u}^{(m)}_{\mathrm{gl}})\left(1 - \frac{\lambda_{\mathrm{gl}}}{\|\bar{\boldsymbol{\beta}}^{(m)} - \boldsymbol{u}^{(m)}_{\mathrm{gl}}\|_2}\right)\mathbf{1}(\bar{\boldsymbol{\beta}}^{(m)} - \boldsymbol{u}^{(m)}_{\mathrm{gl}} > \lambda_{\mathrm{gl}}\rho)$;
                                                                         ▷ Group Lasso
    $\boldsymbol{U}^{(m)}, \boldsymbol{\Sigma}^{(m)}, \boldsymbol{V}^{(m)} \leftarrow \mathrm{SVD}(\bar{\boldsymbol{\beta}}^{(m)} - \boldsymbol{u}^{(m)}_{\mathrm{nuc}})$;
    $\tilde{\boldsymbol{\Sigma}}^{(m)} \leftarrow (\boldsymbol{\Sigma}^{(m)} - \rho\lambda_{\mathrm{nuc}}\mathbb{I}_{p\times p})_+$;
    $\boldsymbol{\beta}^{(m+1)}_{\mathrm{nuc}} \leftarrow \boldsymbol{U}^{(m)}\tilde{\boldsymbol{\Sigma}}^{(m)}{\boldsymbol{V}^{(m)}}^\top$;          ▷ Nuclear Norm
    $\boldsymbol{\beta}^{(m+1)}_{\mathrm{nn}} \leftarrow (\bar{\boldsymbol{\beta}}^{(m)} - \boldsymbol{u}^{(m)}_{\mathrm{nn}})_+$;                   ▷ Nonnegative Projection
    $\bar{\boldsymbol{\beta}}^{(m+1)} \leftarrow \left(\boldsymbol{\beta}^{(m+1)}_{\mathrm{mse}} + \boldsymbol{\beta}^{(m+1)}_{\mathrm{gl}} + \boldsymbol{\beta}^{(m+1)}_{\mathrm{nuc}} + \boldsymbol{\beta}^{(m+1)}_{\mathrm{nn}}\right)\frac{1}{4}$;   ▷ Consensus
    for $g = \{\mathrm{mse}, \mathrm{gl}, \mathrm{nuc}, \mathrm{nn}\}$ do
        $\boldsymbol{u}^{(m+1)}_g \leftarrow \boldsymbol{u}^{(m)}_g + \left(\boldsymbol{\beta}^{(m+1)}_g - \bar{\boldsymbol{\beta}}^{(m+1)}\right)$          ▷ Dual Variables
    end for
end for
```

Material.

The original problem in (3.2) is complex and requires a semidefinite program. Decomposition by (4.1), however, makes the minimization trivial. All of the proximal steps can be calculated in an expeditious manner, which leads to a very efficient and fast algorithm. This is a considerable benefit because the entire algorithm must be executed repeatedly for finding the appropriate tuning parameters. Furthermore, the memory requirement is only linear in the number of variables as the algorithm only tracks the parameters themselves.

ADMM algorithms do not generally have convergence guarantees (Tran-Dinh and Cevher (2015)). By choosing a proper step size, however, those issues can be mitigated. For our problem, ρ needs to be chosen proportional to the minimum eigenvalue of the Hessian of f_{mse}. We can achieve that by setting $\rho \geq (nT)^{-1} \sum_{k=1}^{24} \lambda_{\min}(\mathbf{X}(k)^\top \mathbf{X}(k))$ where $\lambda_{\min}(\cdot)$ is the minimum eigenvalue and $\mathbf{X}(k)$ is the observations in $\mathbf{X}$ that correspond to hour k, whose definition can be found in Algorithm 1.

This condition is required for convergence. However, the speed of convergence with this choice of ρ can be excruciatingly slow. In our analysis, we have observed that by adding a step-size selection procedure, we can empirically ensure convergence and a decent speed of convergence. To avoid an extra computational burden during the step-size selection, we first create a list of step-size candidates from a geometric sequence. In the first 20 steps of the algorithm, all candidates for ρ are tried and the one that gives the largest reduction for the cost function is chosen. In the following steps, only the ρ that was chosen in the last round, and four other ρ candidates that are the closest in value to the previously chosen ρ are tested in the step-size search. By limiting the search, we avoid extensive computation, which results in faster convergence.

There are two tuning parameters in the objective function, λ_{gl} and λ_{nuc}. We employ a brute force search over a grid of possible values, and use cross validation to estimate the sample error for each parameter choice. Choosing large values for λ_{gl} or λ_{nuc} forces all the variables to be zero. Using Karush-Kuhn-Tucker (KKT) conditions, it can be shown that the variables are reduced to 0 when λ_{gl} is larger than $\max_k \bar{\lambda}_{\mathrm{gl},k}$, where $\bar{\lambda}_{\mathrm{gl},k} = \|\mathbf{X}(k)^\top \mathbf{y}(k)\|_2$. Similarly, the variables are shrunk to 0 when λ_{nuc} is larger than $\sqrt{\lambda_{\max}(\boldsymbol{Z}^\top \boldsymbol{Z})}$ where $\boldsymbol{Z}_{:,k} = \mathbf{X}(k)^\top \mathbf{y}(k)$.

Using this, we choose sequences $\boldsymbol{d}_{\lambda_{\mathrm{gl}}}, \boldsymbol{d}_{\lambda_{\mathrm{nuc}}}$ for λ_{gl} and λ_{nuc}, whose ranges are $\exp\left(-5, \log\left(\max_k \bar{\lambda}_{\mathrm{gl},k}\right)\right)$ and $\exp(-5, \log(\sqrt{\lambda_{\max}(\boldsymbol{Z}^\top \boldsymbol{Z})}))$, respectively. Then, the grid for $\lambda_{\mathrm{gl}}, \lambda_{\mathrm{nuc}}$ candidates is given by $\boldsymbol{d}_{\lambda_{\mathrm{gl}}} \times \boldsymbol{d}_{\lambda_{\mathrm{nuc}}}$. When a brute force grid search cannot be afforded due to time constraints, sequential Kriging optimization or global Bayesian optimization methods can be used to find the best tuning parameters (Huang et al. (2006); Snoek et al. (2012)).

5. Case Study

In this section, we illustrate our methodology and evaluate the performance of our proposed estimator with synthetic data. We considered a spatial domain of size 40 km by 40 km, divided into 64 potential source locations (8 × 8), which gives each grid a 5 km × 5 km resolution. We generated 14 day-long meteorological conditions by using a stochastic Fourier series to simulate atmospheric

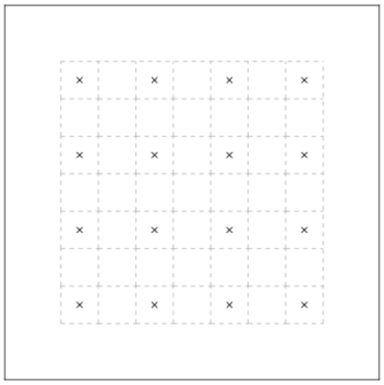


Figure 4. Station locations. The gridded lines indicate the location of the 64 source locations, where × signs indicate the 16 monitoring locations.

turbulence. In a data setting, one can use a numerical weather prediction model (e.g., Skamarock et al. (2008)). The details on simulated atmospheric turbulence are deferred to the Supplementary Material.

The dispersion model calculates the transport from each pollution source to monitoring stations by using the simulated wind condition. We considered 16 monitoring stations, whose locations are depicted in Figure 4, and considered 24 levels for diurnal cycle, based on the time of the day. This setup gives $n \times T = 16 \times 336 = 5376$ observations, each of which is associated with $1536 = 64 \times 24 = p \times 24$ variables and hence $\boldsymbol{\beta} \in \mathbb{R}^{64\times 24}$. The number of variables is comparable to the sample size, and any method that does not employ appropriate regularization is expected to overfit the data.

We assumed that 16 sources are active among the 64 candidates, which resulted in the simulated $\boldsymbol{\beta}^* \in \mathbb{R}^{64\times 24}$, in which only 16 rows of $\boldsymbol{\beta}^*$ have non-zero elements. We divided the 64 source candidates into groups $\mathcal{A}, \mathcal{B}$, and $\mathcal{C}$. Half of the active pollution sources were classified as group $\mathcal{A}$, and were assumed to be constantly active for all 24 hours, $\boldsymbol{\beta}^*(j)^\top = \{1, \ldots, 1\}$ for all $j \in \mathcal{A}$. The remaining eight active sources were assigned to group $\mathcal{B}$, and were fixed to have a diurnal cycle which steadily increases from the morning (6 AM) to noon (12 PM), then steadily decreases until 6 PM. Sites in group $\mathcal{C}$, produced no pollutants, $\boldsymbol{\beta}^*(j) = 0$ for $j \in \mathcal{C}$. The locations of active 16 pollution sources were fixed for all simulations. They were randomly placed, and their daily average emission is depicted in Figure 5. We also tested the robustness of our simulation studies by changing the pollution sources, and the results were very similar.

The design matrix $\mathbf{X}$ and coefficients $\boldsymbol{\beta}$ were determined through the cal-

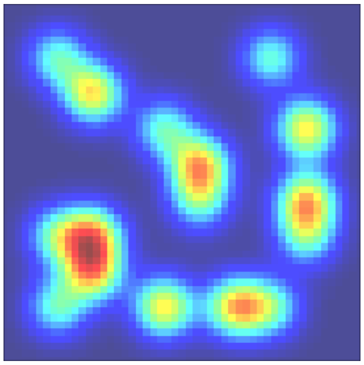


Figure 5. Surface of daily average emission in the simulation.

culated weather conditions and the fixed pollution sources. We then simulated data from the linear regression model

$$\boldsymbol{y}_i = \mathbf{x}_i^\top \boldsymbol{\beta}^* + \boldsymbol{\varepsilon}_i, \qquad \boldsymbol{\varepsilon}_i \sim \mathcal{N}\left(0, \boldsymbol{I}_T\right),$$

where $\mathbf{x}_i^\top$ are computed from the dispersion model in (2.1). The design matrix was also scaled to have average variance of 1 for its columns, this gives a signal to noise ratio close to 1.

We compared three methods: *Non-negative Group Lasso (GL)*, the penalized least-squares estimator with a group L_2-penalty and a non-negativity condition on the coefficients; *Non-negative Nuclear-Norm (NN)*, the penalized least-squares estimator with a nuclear norm penalty and a non-negativity condition on the coefficients; *Our method (GL+NN).*

The GL method is simply our estimator with $\lambda_{\text{nuc}} = 0$; similarly the estimator for the NN method can be obtained by setting $\lambda_{\text{gl}} = 0$. All estimators can be fit using ADMM. To determine the tuning parameters for $\lambda = \{\lambda_{\text{gl}}, \lambda_{\text{nuc}}\}$, we ran 100 separate simulations and stored the λ chosen by 10-fold cross validation. The sequence of candidate λ were obtained from the suggested grid given in Section 3. In the simulations, λ_{gl} was fixed as the average of the 100 λ_{gl} that were chosen in the separate simulations; we repeated the same procedure for choosing λ_{nuc}.

Two performance metrics were compared: L_2 loss, $\|\boldsymbol{\beta}^* - \widehat{\boldsymbol{\beta}}\|_F^2$; L_1 loss, $\sum_j \|\boldsymbol{\beta}^*(j) - \widehat{\boldsymbol{\beta}}(j)\|_1$. We also calculated these performance measures for subsets of $\boldsymbol{\beta}$ with respect to their groups.

The design matrix used in the simulations, $\mathbf{X}$, was heavily correlated. This is common in our methodology, because dispersion from two nearby sites to one monitoring location are driven by the similar wind field, and hence behave similarly. Possibly as a result of this, in all 100 simulations, all estimators were

Table 1. Performances of the Estimators Based on L_1 Loss. Standard errors are given in parentheses.

Method	$\sum_j \lvert\Delta(j)\rvert$	$\sum_{j\in\mathcal{A}} \lvert\Delta(j)\rvert$	$\sum_{j\in\mathcal{B}} \lvert\Delta(j)\rvert$	$\sum_{j\in\mathcal{C}} \lvert\Delta(j)\rvert$
GL	36.89 (0.19)	16.41 (0.09)	9.61 (0.07)	10.86 (0.12)
NN	41.51 (0.25)	13.39 (0.11)	11.74 (0.09)	16.37 (0.14)
GL+NN	**30.65** (0.19)	**12.43** (0.08)	**8.42** (0.06)	**9.79** (0.12)

Table 2. Performances of the Estimators Based on L_2 Loss. Standard errors are given in parentheses.

Method	$\sum_j \Delta(j)^2$	$\sum_{j\in\mathcal{A}} \Delta(j)^2$	$\sum_{j\in\mathcal{B}} \Delta(j)^2$	$\sum_{j\in\mathcal{C}} \Delta(j)^2$
GL	3.94 (0.03)	2.20 (0.02)	1.17 (0.01)	0.57 (0.01)
NN	3.27 (0.04)	**1.21 (0.02)**	1.27 (0.02)	0.79 (0.01)
GL+NN	**2.43** (0.02)	**1.23 (0.02)**	**0.81 (0.01)**	**0.39 (0.01)**

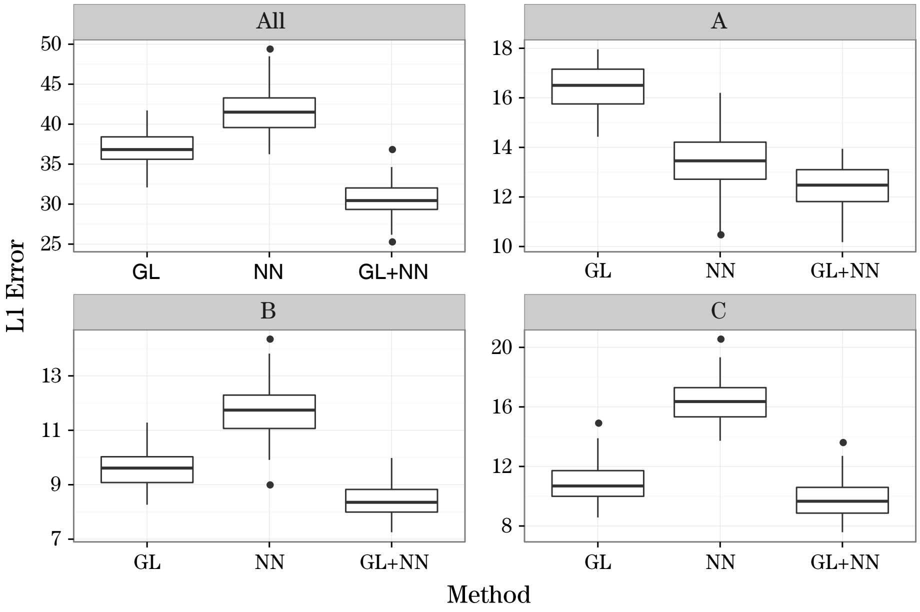


Figure 6. Boxplot of L_1 estimation errors of various estimators.

more conservative, smallerλ values, as these fits tended to have lower prediction error.

For each setting, we present the average of the performance measures based on 100 simulations in Tables 1 and 2. We denote the estimation error for source j as $\Delta(j) = \boldsymbol{\beta}^*(j) - \widehat{\boldsymbol{\beta}}(j)$. A boxplot of the L_1 and L_2 losses for different estimators is given in Figures 6 and 7.

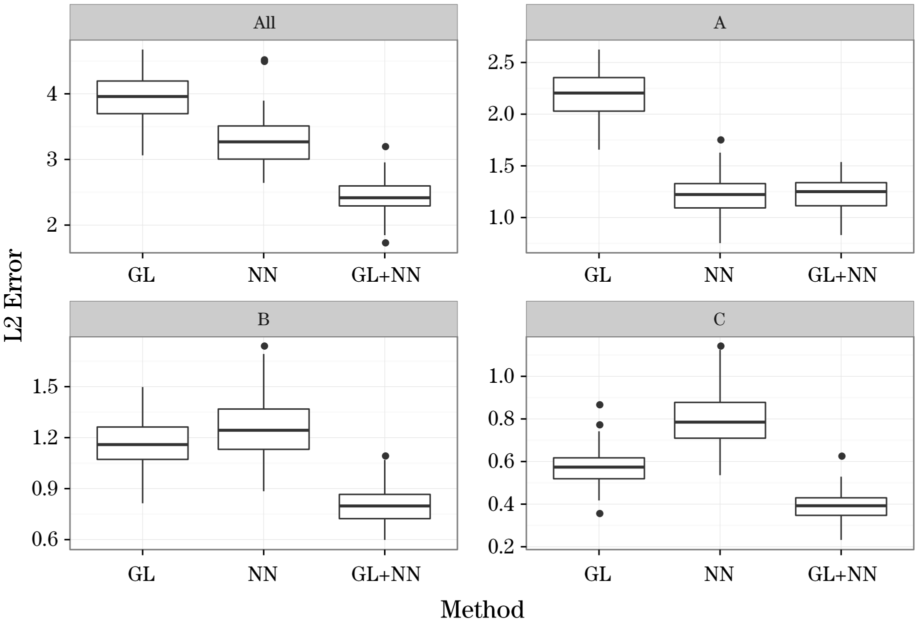


Figure 7. Boxplot of L_2 estimation errors of various estimators.

The simulation results show that the proposed estimator using both nuclear norm and group Lasso penalties overperformed other estimators that made use of only one of them. In Tables 1 and 2, the proposed estimator had lower L_1 and L_2 loss uniformly across all groups of coefficients. This is an expected result because the true coefficients were sparse and had low rank.

Considering the performances across each coefficient group, our estimators had comparable errors in group $\mathcal{C}$ to that of GL. Upon further analysis, we found that ours tended to pick extra sources since coefficients were forced to be similar via the nuclear norm. Accordingly, it was common to see cases where a polluted location's coefficients "bleed onto" nearby grids. From the observation stations' point of view, two nearby sources had similar levels of dispersion. As a result, it became difficult to detect which one of nearby sources was the real polluter.

Comparing the nuclear norm-based estimator to GL, NN produced less-sparse estimates and wrongly selected some of the inactive regions. This can be seen by contrasting the proportion of estimation errors in each group over the total error; 24% of NN's total L_2 loss is from the inactive variables (group $\mathcal{C}$), whereas that number is 14% for GL. As a side note, although these small errors caused NN to have larger L_1 loss compared to GL, since the magnitudes of these estimates were very small, NN had less L_2 loss in total.

With regard to groups with active coefficients, GL had lower error in group $\mathcal{B}$ but higher in group $\mathcal{A}$, since NN can generalize what it learns from one region to another, unlike GL.

As a consequence of the non-negativity constraint on the coefficients, the NN estimator returned sparse solutions (Slawski and Hein (2013)). The median number of non-zero coefficients was 56, compared to 51 for the group Lasso and 53 for our estimator. This also explains why the NN errors are not significantly worse compared to that of GL or our estimator.

All of the active emission sources were estimated to be active; hence the number of false negatives was zero for all methods across all simulations.

6. Discussion

We have proposed a hybrid method to integrate the physical knowledge and statistical methodology to solve a challenging estimation problem. Rather than incorporating more complex physics equations and parameters, we simplified the physical assumptions to utilize available resources and data. This simplification leads to a customized physics model to pair better pair with statistical methods. A statistical model to incorporate the prior domain knowledge and information was proposed, and an efficient algorithm was proposed to solve the resulting optimization problem.

As to future research, when interest is in exploring the uncertainties associated with the attendant physics, a natural extension of our work would be a Bayesian methodology incorporating the prior physical knowledge and previous emission inventory data. There exist more uncertainties related to the weather model output and other physical parameters. Although it is known that such uncertainties can affect the accuracy of the final inference, it is challenging to quantify their impact. A thorough analysis to incorporate such uncertainties can be useful and informative. An extension of our method to the case where the independence assumption is violated can be considered. Research effort is needed to incorporate the dependency structure through spatio-temporal methods. Such methodology needs to address the challenges arising due to estimation of the covariance matrix of the residuals.

Supplementary Materials

The details regarding the atmospheric turbulence simulation, the simulation setting for Figure 3, and the derivations for the ADMM algorithm can be found in the Supplementary Material.

Acknowledgment

We are grateful to an associate editor and two anonymous reviewers for their insightful comments, which significantly improved our manuscript.

References

Biegler, L., Biros, G., Ghattas, O., Heinkenschloss, M., Keyes, D., Mallick, B., Marzouk, Y., Tenorio, L., van Bloemen Waanders, B. and Willcox, K. (2011). *Large-Scale Inverse Problems and Quantification of Uncertainty.* West Sussex, United Kingdon: John Wiley & Sons.

Boyd, S., Parikh, N., Chu, E., Peleato, B. and Eckstein, J. (2010). Distributed optimization and statistical learning via the alternating direction method of multipliers. *Foundations and Trends in Machine Learning* **3**, 1–122.

Byun, D. and Schere, K. L. (2006). Review of the governing equations, computational algorithms, and other components of the models-3 Community Multiscale Air Quality (CMAQ) modeling system. *Applied Mechanics Reviews* **59**, 51–77.

Candès, E. and Recht, B. (2009). Exact matrix completion via convex optimization. *Foundations of Computational Mathematics* **9**, 717–772.

Carroll, R., Chen, E., Li, T., Newton, H., Schmiediche, H. and Wang, N. (1997). Ozone exposure and population density in harris county, texas. *Journal of the American Statistical Association* **92**, 392–404.

Christensen, W. F. and Gunst, R. F. (2004). Measurement error models in chemical mass balance analysis of air quality data. *Atmospheric Environment* **38**, 733–744.

Fast, J. D., Gustafson, W. I., Easter, R. C., Zaveri, R. A., Barnard, J. C., Chapman, E. G., Grell, G. A. and Peckham, S. E. (2006). Evolution of ozone, particulates, and aerosol direct radiative forcing in the vicinity of houston using a fully coupled meteorology-chemistry-aerosol model. *Journal of Geophysical Research: Atmospheres* **111**.

Fu, J. S., Hsu, N. C., Gao, Y., Huang, K., Li, C., Lin, N.-H. and Tsay, S.-C. (2012). Evaluating the influences of biomass burning during 2006 base-asia: a regional chemical transport modeling. *Atmospheric Chemistry and Physics* **12**, 3837–3855.

Ghosh, S., Bhave, P., Davis, J. and Lee, H. (2010). Spatio-temporal analysis of total nitrate concentrations using dynamic statistical models. *Journal of the American Statistical Association* **105**, 538–551.

Gurjar, B., Van Aardenne, J., Lelieveld, J. and Mohan, M. (2004). Emission estimates and trends (1990–2000) for megacity delhi and implications. *Atmospheric Environment* **38**, 5663–5681.

Higdon, D., Gattiker, J., Williams, B. and Rightley, M. (2008). Computer model calibration using high-dimensional output. *Journal of the American Statistical Association* **103**, 570–583.

Huang, D., Allen, T., Notz, W. and Zeng, N. (2006). Global optimization of stochastic black-box systems via sequential kriging meta-models. *Journal of Global Optimization* **34**, 441–466.

Keats, A., Yee, E. and Lien, F.-S. (2007). Bayesian inference for source determination with applications to a complex urban environment. *Atmospheric Environment* **41**, 465–479.

Kennedy, M. C. and O'Hagan, A. (2001). Bayesian calibration of computer models. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)* **63**, 425–464.

Liu, F. and West, M. (2009). A dynamic modelling strategy for bayesian computer model emulation. *Bayesian Analalysis* **4**, 393–411.

Malmberg, A., Arellano, A., Edwards, D. P., Flyer, N., Nychka, D. and Wikle, C. (2008). Interpolating fields of carbon monoxide data using a hybrid statistical-physical model. *Annals of Applied Statistics* **2**, 1143–1580.

Moin, P. (2010). *Fundamentals of Engineering Numerical Analysis*. New York, NY: Cambridge University Press, 2nd ed.

Morawska, L. (2006). Motor vehicle emissions as a source ofindoor particles. *Indoor Environment: Airborne Particles and Settled Dust*.

Paciorek, C. J., Yanosky, J. D., Puett, R. C., Laden, F. and Suh, H. H. (2009). Practical large-scale spatio-temporal modeling of particulate matter concentrations. *The Annals of Applied Statistics* **3**, 370–397.

Parikh, N. and Boyd, S. (2014). Proximal algorithms. *Foundations and Trends in Optimization* **1**, 123–231.

Saarikoski, S., Timonen, H., Saarnio, K., Aurela, M., Järvi, L., Keronen, P., Kerminen, V. and Hillamo, R. (2008). Sources of organic carbon in fine particulate matter in northern european urban air. *Atmos. Chem. Phys* **8**, 6281–6295.

Simon, N. and Tibshirani, R. (2012). Standardization and the group lasso penalty. *Statistica Sinica* **22**, 983–1001.

Skamarock, W. C., Klemp, J. B., Dudhia, J., Gill, D. O., Barker, D. M., Duda, M. G., Huang, X.-Y., Wang, W. and Powers, J. G. (2008). *A Description of the Advanced Research WRF Version 3*. Boulder, Colorado, USA: National Center for Atmospheric Research.

Slawski, M. and Hein, M. (2013). Non-negative least squares for high-dimensional linear models: Consistency and sparse recovery without regularization. *Electronic Journal of Statistics* **7**, 3004–3056.

Snoek, J., Larochelle, H. and Adams, R. P. (2012). Practical bayesian optimization of machine learning algorithms. In *Advances in Neural Information Processing Systems*.

Tran-Dinh, Q. and Cevher, V. (2015). Splitting the smoothed primal-dual gap: Optimal alternating direction methods. *arXiv preprint arXiv:1507.03734* .

Williams, B., Christensen, W. F. and Reese, C. S. (2011). Pollution source direction identification: Embedding dispersion models to solve an inverse problem. *Environmetrics* **22**, 962–974.

World Health Organization (2005). Who air quality guidelines for paticularte matter, ozone, nitrogen dioxide and sulfur dioxide–global update (who/sde/phe/oeh/06.02), world health organization. In *Air Quality Guidelines - Global Update 2005*.

Yuan, M. and Lin, Y. (2006). Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society. Series B (Methodological)* **68**, 49–67.

Department of Statistics, Sungkyunkwan University, Seoul, 03063, Korea
E-mail: yhwang@skku.edu
Department of Statistics, George Washington University, Washington, DC 20052, U.S.A.
E-mail: barut@gwu.edu
IBM Thomas J. Watson Research Center, Yorktown Heights, NY 10598, U.S.A.
E-mail: kyeo@us.ibm.com

(Received March 2016; accepted April 2017)