Ph.D. Research Proposal
Geographical Applications of modern statistical learning algorithms

Ron Sarafian

Ben Gurion University

The Faculty of Engineering Sciences

Department of Industrial Engineering and Management

Advisor: Dr. Johnathan Rosenblatt

Advisor: Prof. Itali Kloog

Advisor: Prof. Israel Parmet

Contents

1 Abstract in Hebrew

2 Abstract

Estimating air pollution concentrations is valuable for environmental exposure assessment and epidemiological studies. The use of satellite-based data models to estimate particulate matter (PM) has increased substantially over the past few years. These models employ statistical learning algorithms to analyze spatially and temporally resolved datasets to provide an assessment of air pollution levels in locations where no measurement is performed. These assessments are then used as covariates in epidemiological studies; hence, their reliability is of high importance.

However, some complex issues arise when analyzed data have spatial and temporal structure, many of which are neither clearly defined nor completely resolved. Spatio-temporal dependency is a challenge for classical statistical methods, especially those used for prediction and model performance assessment. Therefore, proper statistical techniques that reduce prediction's errors are necessary in order to avoid biased epidemiological results that may lead to erroneous conclusions. scale

Our research focuses on both theoretical and practical aspects of modern statistical learning algorithms implemented in environmental and geographical applications. It is carried out as a research collaboration between the Department of Industrial Engineering and Management and the Department of Geography and Environmental Development under the supervision of Dr. Johnathan Rosenblatt and Prof. Itai kloog.

We study the statistical issues regarding the estimation and model assessment of air pollutants concentrations in the presence of spatio-temporal autocorrelation in the data. In particular, the research aims at: (1) Propose an extension to classical methods that estimate prediction error (such as Cross Validation) for exposure assessment models with spatio-temporal structured data. (2) Improve models prediction performances while reducing its complexity using linear models. (3) Cope with the computational challenges associated with these models, taking advantage of some recent significant achievements in the fields of memory efficiency and parallel computing.

We hope that this research will greatly contribute to the field of environmental exposure and geo-statistics; not only in theoretical perspective, but also in that it would harness scientific knowledge to produce applications that would improve people's life.

3 Introduction

How does air quality affect people health? A quantitative answer for this question is of interest for policy makers, especially regarding to the allocation of pollutants. To provide a general answer, epidemiologists usually employ observational studies, in which the relationship between air pollution expusure level and health indices is being investigated. In many cases the experimental units are spatio-temporal entities, each with its own measurement of pollution concentration. The fine Particular matter (PM) abundance is one of the common indicators for pollution levels [q][1]. However, PM mass concentrations are measured at ground monitoring stations, therefore highly limited in terms of spatial coverage. In order to provid PM assesments in geographical areas where no mesurments are available some statistical methodology is needed.

PM levels show high spatio-temporal variation [Pelletier et al., 2007][2]; therfore, geostatistical interpolation methods in which spatial PM levels are modeled as opposed to some smoothness minimizer polynomial spline may introduces PM exposure error ([Zeger et al., 2000][3]). Over the years, methods which utilize additional data to minimize the exposure error have been developed. The Land-use Regression (LUR) approach take advantage of traffic, topography, and other geographic variables to train models to predict pollution levels at any location ([q][14],[q][15]). The LUR methods provided more accurate exposure assessments at unmonitored locations; However they were significantly limited mostely since the included covariates are generally not time varying, therefore were better suited for long-term exposure assessments. Along with the improvement in the epidemiologists' analysis tools the demand for spatio-temporally resolved datasets of PM concentrations was rising and new approaches were needed.

The increasing in availability of satellite data has enabled environmental scientists to harness remote sensing technologies to PM concentration assesments. Satellite-based aerosol optical depth (AOD) is the measure of particles in air (e.g., haze, smoke, dust) distributed within a column of air from the earth's surface to the top of the atmosphere. AOD retrievals were found to be associated with ground PM measurments in different geographical areas. The use of AOD as a covariate in studies of PM concentration estimation has increased substantially over the past few years (e.g., [Kloog et al., 2014][4],[Chudnovsky et al., 2014][5],[ref][]).

In recent studies, ground measured PM data were integrated with rerpeated AOD measurments along with other spatio-temporal covariates in order to build a model that can be used to assess PM exposure in locations where no measurement is performed ([ref][]; [ref][]). The clustered structure of the data requires a model that is able to account for several sources of variability (mostly, space and time), and a common practice is the *mixed models* statistical framework (see for instance [ref][PM,AOD,MixedModel]; [ref][]). The AOD based models have shown substantially well performances in predicting exposure levels, and their great advantage is in providing exposure assessments within short intervals ([q][good performance]; [q][short intervals])

The generated PM predictions allow to examine the effect of air pollution on health and disease conditions, not only near monitoring stations, but also in remote, usually less populated areas. Studies that have used generated PM predictions which were obtained using the described above procedure found statistically significant correlation between PM and varaity of health related mesurments such as adverse birth outcomes ([Kloog et al., 2012][6]); natural-cause mortality [Wang et al., 2016][7]; Acute Myocardial Infarction ([Madrigano et al., 2013][11]) and more.

In the following we propose . . . , introduce . . . , and present . . . . Since that . . . we also consider . . . and try to . . . .

4 Study Approach

4.1 Challenges

The mission of PM exposure levels assesment is actually not well defined without additional setting. Put differently, the study meta-purpose must be declared. This purpose is of great importance since it derives the appropriate performance evaluation methodology. An old proverb says that "The proof is in the pudding", therefore, a significant decision one should make is which pudding to choose. Should the assessed PM values minimize the mean squared prediction error? Perhaps a more conservative approach is appropriate and the maximum error should be minimized? Obviously, these choices should be considered in light of the epidemiological study goals. For instance, when estimating the effect of air pollution on health condition, should the emphasis of PM assesment be placed on being as accurate as possible in populated areas, rural areas, or evenly across space?

Moreover, even after the study purpose was determined, executing model performance evaluation is not a trivial task. The spatial and temporal structure of the analyzed data raises some complex statistical issues, many of which are neither clearly defined nor completely resolved. In particular, when spatio-temporal dependency exist, assessing how the results of the model will generalize to an independent data set becomes a challenging problem that is less explored. *Cross-validation* (CV) is a widespread strategy that is used for model's predictive performance estimation. The main idea behind CV is to avoid overfitting by splitting the data several times into training and validation samples which are independent. However, when data are spatio-temporaly dependent, classical CV analyses break down and must be modified.

Anoter challenge which arises from the structure of the data is related to statistical model that is being used. Providing PM assesments in geographic locations and at different time points requires the learning from complex dynamic spatial datasets. These datasets have a natural hierarchical or multi-level structure. On one hand, it offers researchers extended modeling possibilities, thus to increase model performance. On the other hand, the analyses of these datasets requires a proper modeling of the dependence between observations. There are several approaches to model the dependency structure, each with its own pros and cons. Obviously, any selected approach has its own influence on the epidemiological resuls.

When databases are also very large, the computational complexity of the model is additional consideration should be taken into account. Different models usually have different estimation procedures with its own properties of computational complexity. For instance, the Mixed Models are usually estimated using *Maximum Likelihood* (ML) estimation method, therefore are based on numerical optimization of nonlinear functions with no closed-form solution. If on the other hand, a linear model is chosen there is a closed-form analytical solution, so the estimates can be expressed in matrix notation terms. Generally these term are easier to compute, depending on the structure of the data. Also, diffrent estimation approaches could affect not only the computation time, but may also lead to significantly different results.

somthing about sparse representation, parallel computing and memory efficiency

4.2 Research Objectives

The research objectives follows the challenges that were described in the previous section. In principle, the order of tackling the objectives is as it appears in the following sections, however, there are plenty of overlaps subjects, so that progress is expected in several channels in parallel. The objectives are:

- Investigate the procedure of PM prediction model performance estimation with CV in conditions of spatio-temporal dependency and in light of the epidemiological goals.
- Propose a generalization to the state of the art models that are used to predict PM levels in order to improve their performances. We will examine the genralized least squares (GLS) model and suggest several methods for estimating its required covariance matrix while modeling the spatio-temporal dependencies. Moreover, a regularization of the GLS model would also be investigated.
- Present an improvement in computation memory use, taking advantage of the fact that linear models consume less compututional resources. Also, we would employ some of the recent achievements from the fields of sparse representation, parallel computing and memory efficiency in our framework.

5 Research Framework

5.1 Performance Estimation

There are many validation techniques for evaluating the predictive performence of a statistical algorithm. Probably the simplest and most widely used is *cross-validation* (CV). The main idea behind CV is as follow: First, split the data into a training set and a validation set. It is crucial that these sets are completely independent, otherwise CV is likely to be overfitting. Second, use the training set to train a statistical algorithm. Third, evaluate the predictive performence of the trained algorithm on the validation set.

In Spatio-temporal autocorrelated settings, it is fundamental to find a strategy that split the data into two samples that would be independent, yet would not unwittingly induce extrapolations by restricting unnecessary range between training and validation samples.

Forthermore, an obvious question is how to evaluate the predictive performence. In a regression setting for example, the performence of a model is measured by the discrepancy between the real and the predicted values, usually in some form of the *sum of squers*. Following the framework of [Arlot and Celisse, 2010][8], the quality of the predicted values as an approximation to the real values is quantified by its loss $\mathcal{L}$. We will later discuss the exact settings of the loss function $\mathcal{L}$, but before that we would like to emphasize the importance of choosing it.

Supervised learning tasks such as regression can be formulated as the minimization of a loss function over a training set. Thus, the chice of the loss function should reflect the purpose of the study. For instance, if the epidemiological objective is to assess the effect of air pollution on health, then PM prediction by its own is not the ultimate goal. In this case, choosing the loss function is not trivial; It is necessary to find the loss function which through its minimization the estimated effect of air pollution on health is closest to the truth.

It seems that when one is willing to perform a CV procedure in this settings, he faces two different hurdles: (1) choosing the right loss function so that it meets the objectives of the study, and (2) avoiding CV overfitting due to spatio-temporal dependency by choosing the folding scheme. We will break down this challeng into two decisions that must be made in accordance with the defined goals and the data structure.

5.1.1 The Loss Function

In order to illustrate the meaning of choosing the loss function and its influence on epidemiological results, let us discuss an example of an extreme situation: Consider the case where a population of 100 people in a particular area live in only two regions: 99 people live in a dense city, while one man lives far away in a small village. Suppose that we know the PM exposure level for each of the people. A reasonable assumption is that the exposure levels of the city's people are highly correlated, so that for every practical purpose their PM values are the same.

Formally, let y be a 1×100 vector of the population PM exposure levels, and $\hat{y} = f(x)$ the predicted PM vector using some model f which can be estimated from the population data x. Let $\mathcal{F}$ denote the set possible values for f. For instance, f can be thought as some linear function of x. The quality of the model f is measured by its loss $\mathcal{L}(f)$, where $\mathcal{L}(f) : \mathcal{F} \to \mathbb{R}$ is called the *loss function.*

We would like to consider two different loss functions that are used in the PM prediction stage, and their effect on epidemiological outcomes.

The (unweighted) *quadratic loss function*:

$$\mathcal{L}_{UW} = \Big(y - f(x)\Big)'\Big(y - f(x)\Big),$$

and the *weighted quadratic loss function*:

$$\mathcal{L}_{W} = \Big(y - f(x)\Big)'\mathbf{W}\Big(y - f(x)\Big).$$

One option is to set the weights matrix $\mathbf{W}$ as the inverse of the data variance-covariane matrix Σ_n^{-1}. In this case, $\mathcal{L}_W$ can be thought as the squared Mahalanobis length of the residuals vector.

Assume that f_{UW} and f_W both minimize $\mathcal{L}_{UW}$ and $\mathcal{L}_W$ over $\mathcal{F}$ respectively.

Note that f_{UW} was chosen in a manner that gives each squared error of y an equal weight, therefore, de facto f_{UW} devotes all of its efforts to predict the values of y in the city, so that the error weight of the the village man is practically zero. Without going to much into details, the estimation with f_{UW} would result in a poor prediction for the village man PM level. Intuitively, f_{UW} does not express the real value of the data.

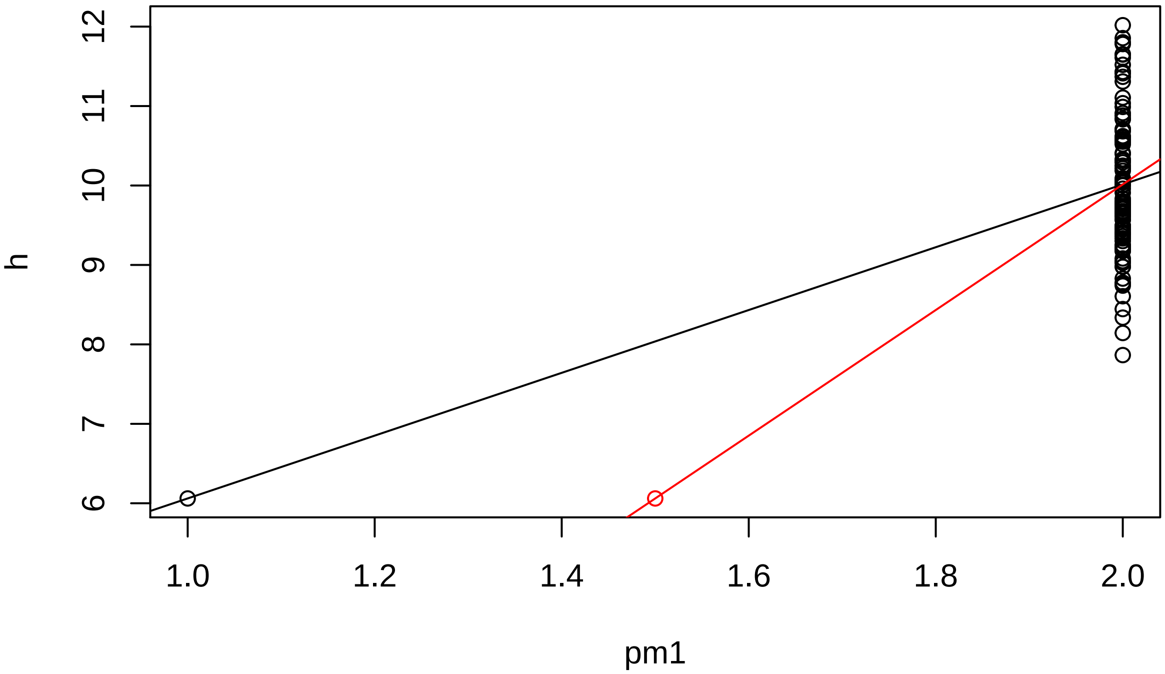
Conversely, f_W recognizes that the numerous PM observations in the city are actually the same observation, hence the weight of the error in the village will be substantially higher, and so the PM prediction in the village would be more accurate.

Easy to see that under $\mathcal{L}_{UW}$ the model tend to be more acurate in the areas where there are many similar observations (which will usually be more populated places) in a price of a higher error in areas with fewer observations (usually unpopulated remote areas). Also, $\mathcal{L}_W$ is forcing the model to utilize the genuine value that data provide (at least in a geographical sense), hence, predictions accuracy for uncorrelated areas would increase.

Epidemiologists who estimate the efect of air pollution on some health indices **would prefer high varianced PM measurments.** and **to be aqurate in the village**

about measuring prediction errors with the loss - too optimistic (w.r.t epidemiological results) and, what is the different between estimation and validation in loss function?

Weighted quadratic loss function requaires a prespecified weights matrix W. A natural option is to estimate in some way the aforementioned variance-covariance inverse matrix Σ_n^{-1}. The estimation of Σ_n^{-1} is substantial for our research and has great influence not only on model performance evaluation but also on model estimation stage. In section 5.2.3 we go into more details and discuss several estimation approaches.



5.1.2 The Folding

Conventional CV techniques assume that experimental units are independent. In data with temporal and spatial observations which are widely common in environmental and geographical studies this assumption is violated. Independency of observations means that randomly splitted CV (i.e. k-fold CV) divides the data into dependent training and validation samples, resulting in overfitting ([Larimore and Mehra, 1985][]; [ref][]). In other words, the estimated errors are downward biassed, so that performance estimates are actually overoptimistic ([Mosteller and Tukey, 1977][]).

The procedure of CV under dependency conditions has been studied extensively over the last few decade in several contexts. Much progress has been made in the field of nonparametric regression. [Hart and Wehrly, 1986][13] for example, proved that when data are positively correlated, using standard CV will overfits for choosing the bandwidth of a kernel estimator in regression. [Chu and Marron, 1991][] proposed a *modified CV* when selecting the nuisance parameter in nonparametric curve estimation with dependent data. [Burman et al., 1994][10] continued this line, introducing the *h-block* CV as a verssion of leav-one-out (LOO) CV method optimized for use with dependent observations. Their idea is simple: Rather than remove a single case in each CV iteration, remove as well a block of h cases from either side of it. they suggest to take h as a fixed function of the number of cases, and to correct for the underuse of the sample by adding a term to the estimates.

However, as a verssion of LOO, *h-block* CV has proven to be asymptotically inconsistent (see [shao, 1993][]). [Racine, 2000][] proposed the *hv-block* as a modification which is also asymptotically optimal. It extends *h-block* by defining the test set to be v-sized observations block instead of being a singletone, while maintaining near-independence of the training and validation data via h-blocking.

Generally, the main CV approach used in the literature to overcome time series dependence, is to choose training set I^t, and validation set I^v such that $\arg\min_{i \in I^t, j \in I^v} |i - j| > h > 0$, where h expresses the distance from which observations i and j are independent. However, these methods were not suitable for data with spatial dependence, and some adaptations were required.

Apperently CV approaches appropriate for spatialy dependent data recive less attention in the statistical literature. However, some progress has been made in recent years, mainly in the fields of geographical and environmental studies.

[Telford and Briks, 2009]][] suggested that the scheme of h-block can be adopted for the spatial case by omitting observations within a radius of h of the test set (also referred as spatial-LOO). They proposed using the range of a variogram model to appropriately define h. [Trachsel and Telford, 2015][] proposed further methods for determining h e.g. by finding the distance at which the root mean squared error (RMSE) of h-block CV and the RMSE of an independent validation set are similar. [Rest et. al., 2014][] considered a variable selection model under conditions of spatial autocorrelated data. They compared a spatial-LOO version with a classical model selection with Akaike information criterion (AIC) while accounting for residual spatial autocorrelation (RSA). Using simulations they found that spatial-LOO is particulary more useful when the range of RSA was small.

[Roberts et. al., 2016][] examine the utility of blocking procedures for CV in a number of dependent settings, and propose several blocking schemes. They also discuss the possibility of extrapolation (i.e. when blocking hold out entire portion of the predictor space) and state that when extrapolation is the modelling goal deliberate blocking in predictor space may should be considered.

Some contribution to blocking approaches comes also from the bootstrap practice. In fact, the procedure of resampling in bootstrap and in CV methods are the same, excepting of with or without replacing. [Davison and Hinkley, 1997][bootstrap methods] review some of the schemes for block resampling proposed for complex dependence such as time series and point processes. Examples are: *Post-blackening*, *Blocks-of-blocks*, and *Stationary bootstrap*. They also provide a detailed instruction for the choice of block length In an iterative fashion under suitable assumptions.

stationarity However, these CV approaches are all suitable only for cases where observation form a general stationary sequence, hence without additional adjustmant, is not applicable for process whose joint probability distribution change when shifted in time or in spatial domain.

spatio-temporal structure. to the best we know… Yet, more advancement is expected with the rapid development of large spatial databases along with the need for proper statistical tools.

still not covered:

- anisotropic analysis, highly unbalanced spatial data

- not stationary data
- Anti-Learning Phenomenon

5.2 Statistical Model

Although dependency among observations is a common phenomenon in observational studies, it violates one of the standard statistical assumptions, and challenging many classical statistical techniques. Geographical and environmental data generally show both spatial and temporal dependency between observations, known as autocorrelation. Standard regression techniques ignor this dependency structure and usually lead to unstable estimates [ref][], bad predictions [Cressie, 1993][book,p.435], and biassed results [ref][] resulting in erroneous conclusions.

In the last decades there has been an explosion of research in dependence data modeling. Logitudial modeling is perhaps the most familiar one. It mostly focuses on the dependence among obsevations over time such as time series, and is commonly used in many fields such as biology, economics and more. A number of approaches to modeling longitudinal data have been proposed in the statistical literature. A very common approach is to use a parameterized covariance model. parameterized models assume an underlying structure of stationary stochastic process, which can be described 'using a with small number of parameters. such as structures are *autoregressive* (AR), or *moving average* (MA) models. For a comprehensive review see [Weiss, 2006][book].

Spatial statistics deals with more complex dependency structure where data represent observations that are associated with points or regions. According to the first law of geograpy ([Tobler (1970)][first_law]), samples in geographical space tend to be more similar, resulting in spatial autocorrelation (SAC). Modeling SAC is not a simple task, and is definitely not a straightforward extension of time series into two dimensions. [Cressie, 1993][book] introdue the *conditionally specified gaussian* model that use the spatial locations of samples to model the probability distribution of the errors under gaussian assumption. The spatial econometrics literature uses the so-called spatial weights matrix W to denote variables lagged in space. W describe the spatial arrangement of the geographical units and can be assumed or estimated. in his seminal book, [Anselin, 1988][book] present several estimation methods for linear regression model with spatially dependent error terms from an econometric perspective. [LeSage and Pace, 2009][book] called this model *spatial error model* (SEM) and reffered it as a special case of the more general *spatial Durbin model* (SDM). Anothe method that have recently shown to be effective is *Bayesian hierarchical* modeling approaches (see [Banerjee et. al., 2014][book]). These models handle complex relationships such as multi-level data by estimating the parameters of the posterior distribution using the bayesian methods.

Spatio-temporal data enable the researcher to take advantage of time-sapce interactions, thus to provide more accurate predictions in a higher resolution. However, the analysis of spatio-temporal data might be quite complicated since both spatial and temporal dependencies need to be accounted. It is an emerging reseach field and modeling approaches are still developing. In the environmental exposure assessment studies, *Mixed models* ([Henderson et. al., 1959][The Estimation of Environmental]; [Robinson, 1991][that BLUP is…]) are probably the most prevalent statistical framework for spatio-temporal data, particularly those analysing satellite based data. Mixed models cope with clustered data by distinguishing between two sources of variation: between clusters, and within clusters. Another suitable approach for modeling complex dependency structures is the *generalized least squares* (GLS). The GLS is a linear regression model which uses the estimatied variance-covariance matrix of the error terms to efficiently estimate model's parameters in the presence of dependency. Since it is a linear model, the GLS has some nice properties as we will discuss later. Altought GLS based models have been known in the statistical literature for decades ([Besag, 1974][gls1], [Cliff and Ord, 1981][gls2]), their application in geographical and environmental studies has been very limited so far.

In the following we discuss about some of the features of the mixed effect model and the GLS, including their estimation and regularization. As we shall see, these two approaches eventually deal with the same challenge: To characterize the observations dependency structure. More specifically, they ask: "how does the residuals variance-covariance matrix should be modeled?".We state that, as far as regressions are concerened, the almost only difference between models is the definition of the dependence structure through the covariance matrix of the residuals terms.

The GLS model estimate the parameters using a prespecified covariance matrix of the error terms. Thus, it can be thought as a more general approach to model complex structured data, since any covariance matrix can be used. Therefore, the mixed model should be considered as one of GLS's special cases. Moreover, we argu that any other regression model fall under the extensive settings of the GLS.

Since it is the errors covariance matrix that practically determine the model, we will make an effort to explore the its modeling. The estimation of this matrix is undoubtedly a gentle art, as the dependence structure consists of temporal and spatial correlations which their patterns are unknown.

Moreover, in choosing the statistical model we should also consider its goal. The purpose of the model in our case is predication rather than inference. For this reason, we might cosider reducing prediction error in cost of biassed model, using regularization. Suitable regularization approaches for regressions are Shrinkage methods such as *Ridge regression* or *Lasso*. It should be noted that while such procedures are relatively easy to adapt for linear models, It might be more complex for mixed effect models. [ref][]

Some subjects that we might consider but not covered here are:

- Bayesian hierarchical modeling
- modifiable areal unit problem (MAUP) [Cressie, 1996][Change of support and the modifiable areal unit problem]

5.2.1 Mixed Models

Mixed models, sometimes referred as Hirarchial models, are a class of statistical models suited for the analysis of structured data. Mixed models are particularly useful when obseving repeated measurements of the same statistical units. The mixed models are widely used in environmental studies due to their ability to genuinely combine the data by intruducing multilevel random effects. In theses studies levels are usually time periods, spatial areas, or their interactions.

For each level, the mixed effect model defines clusters. In time level, typical clusters are days or hours, and in spatial areas level, they might be spatial grid cells. The model assumes that observations between clusters are independent, while obserrvations within cluster are dependent since they belong to the same subpopulation. For instance, when days are the only clusters in a PM spatio-temporal dataset, PM measurments for a specific day of all geographic units are dependent, as they are assumed to be drawn from the same subpopulation. That is to say, that every day is unique in its distribution of PM measurments across different geographic locations. This cluster-specific uniqueness is reflected in a estimated posteriori coefficient and reffered as *random effects*. Other model coefficients are fixed across clusters (usually referred as *fixed effects*) and have the same meaning as in standard regression models.

A very common model in the exposure asseement literatue is the *linear mixed effect* (LME) model that was originally developed by [Laired and Ware (1982)][]. It can be formalated as:

$$y_j = X_j\beta + Z_jb_j + \varepsilon_j \qquad j = 1, ..., T$$

where:

j represent a cluster, s_j is the number of observations in cluster j, and T is the number of clusters; y_j is an $s_j \times 1$ vector of responses of the jth cluster; X_j is a $s_j \times m$ design matrix of fixed effects; β is an $m \times 1$ fixed effects coefficients; Z_j is an $s_j \times k$ design matrix of random effects; b_j is an $k \times 1$ random effects coefficients with mean zero and covariance matrix $\sigma^2 D$; and ε_i is an $s_j \times 1$ independent and identically distributed (*iid*) error terms vector. each component of ε_j (obsevation in the jth cluster) is assumed to have mean zero and variance σ^2.

The compressed form of N equations is:

$$y = X\beta + Zb + \varepsilon$$

or,

$$y = X\beta + \eta$$

where

$$\eta = \begin{bmatrix} \eta_1 \\ \vdots \\ \eta_T \end{bmatrix} = \begin{bmatrix} \varepsilon_1 + Z_1b_1 \\ \vdots \\ \varepsilon_T + Z_Tb_T \end{bmatrix}$$

The model assumes that $E(\eta) = 0$. Note that $\mathrm{Var}(\eta)$ is an $N \times N$ covariance matrix, where $N = \sum_{j=1}^{T} s_j$. Let us define $V = \mathrm{Var}(\eta) = E(\eta\eta')$. V has the following block diagonal form:

$$V = \sigma^2 \begin{bmatrix} I_{s_1} + Z_1 D Z_1' & 0 & 0 & 0 \\ 0 & I_{s_2} + Z_2 D Z_2' & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & I_{s_T} + Z_T D Z_T' \end{bmatrix}$$

where I_{s_i} is the identity matrix in s_i dimension.

In other words, without further assumptions (i.e residuals autocorrelation), the LME can be considered as the familiar linear model, exept that it assume that the residuals covariance matrix V follows a specific structure. In particular, V is forced to have a diagonal block design, where each block represents a cluster.

Note that as more levels are added (i.e. more clusters), the less sparse the covariance matrix would be. However, block design covariance matrix does not allow for correlation between clusters. Lack of correlation between clusters is very unlikely when the clusters are spatial or time units, therefore some adjustment could be useful.

5.2.2 GLS

The GLS ([Atiken, 1934][first gls]), extends the Gauss–Markov theorem to the case where the covariance of the error terms is not a scalar matrix.

To understand GLS estimator, consider the linear regression model in the following matrix notation:

$$(*) \qquad y = X\beta + \varepsilon.$$

According to Gauss-Markov theorem, for a known covariance matrix of the error terms Σ, the best linear unbiased estimator (BLUP) for β is:

$$b(\Sigma) = (X'\Sigma^{-1}X)^{-1}X'\Sigma^{-1}y.$$

However, Σ is usually unknown, and so GLS estimators replace Σ with the its estimator:

$$\hat{\beta}_{GLS} = b(\hat{\Sigma}) = (X'\hat{\Sigma}^{-1}X)^{-1}X'\hat{\Sigma}^{-1}y.$$

Clearly, the quality of the GLS estimator lies in the estimation of Σ. In the proposed study, we will examine and discuss several estimation methods of Σ. Now it is easy to realize that GLS includes as its special cases various specific models that are determined by the estimated error terms covariance matrix.

In addition, we want to emphasize that GLS estimator is also the minimizers of the squared mahalanobis length of the residual vector $y - X\beta$. This is particulary important since we might want to choose an estimator that minimizes the *weighted loss function* which is described above as an option to measure model performance using the estimated Σ:

$$\hat{\beta}_{GLS} = \arg\min_{\beta}\Big((y - X\beta)^{-1}\hat{\Sigma}^{-1}(y - X\beta)\Big).$$

Note that GLS can be considered as an estimation method that de-correlate the scale of the *ordinary least squares* (OLS) errors. This means that as long as we reasonably estimated Σ, strongly dependent observations, wich usually have highly correlated errors, will have less impact on the values of the estimators than independent observations.

5.2.3 The Variance-Covariance Matrix

Whether it's a mixed model, AR, or SEM, it is the covariance matrix that essentially tells the story of data dependency. The decision regarding the covariance modeling is the researcher's statement about the data generating process.

Here we review several models specifications which may be applied. We focus on *parameterized* covariance matrices, where all the components of the covariance matrix are function of $q \in 1, ..., \frac{N(N+1)}{2}$ parameters, where N is the number of error terms. We define the covariance model as $\Sigma(\theta)$, where θ reffers to the distinct unknown parameters that need to be estimated from the data. [Weiss, 2005][modelinglongitudinal] states that the covariance model $\Sigma(\theta)$ should be choosen so that the true covariance matrix is from the type defined be the model (i.e. AR, SEM, etc.), but parsimonious, meaning that q is small as possible.

We point that any parameterized covariance matrix $\Sigma(\theta)$ can be considered as a compromise between two possibilities: The first, is the variance-scalled identity matrix: $\sigma^2 I_N$, somtimes called *spherical error variance* matrix [Hayashi, 2000][econometrics]. This structure assumes *homoscedasticity* and no autocorrelation, and requires estimation of only one parameter. The second, is the most general model, called the *unstructured covariance matrix* and specifies no patterns:

$$\mathrm{Var}(\varepsilon) = \begin{bmatrix} \sigma_1^2 & \sigma_{1,2} & \dots & \sigma_{1,n} \\ \sigma_{2,1} & \sigma_2^2 & & \\ \vdots & & \ddots & \vdots \\ \sigma_{n,1} & & \dots & \sigma_n^2 \end{bmatrix}.$$

Unfortunately, the use of unstructured covariance matrix is not feasible in most cases since it requires fitting $\frac{N(N+1)}{2}$ parameters, which most datasets do not support.

We would like to examine covariance models at an increasing complexity. Firstly, we divide the procedure into temporal and spatial modeling and finally propose an integrated model. We will stick to a spatio-temporal framework in which $i \in 1, ..., S$ indicates a spatial unit and $j \in 1, ..., T$ indicates a time unit.

5.2.3.1 Time perspective: *Autoregressive* Errors

When PM measurments are regressed agisnt environmental covariates, both the response and predictors vary over time j. Thus a case to suspect is *autocorrelation between errors*. To illustrate the errors covariance matrix that describe process like this we consider the $AR(1)$ model of the errors, in which the error term depends on its (1) previous values. This model can be easily extanded to $AR(p)$.

Consider the model in $(*)$, only with the extension:

$$\varepsilon_{ij} = \rho\varepsilon_{i(j-1)} + \delta_{ij},$$

where, ρ is referred as the *autocorrelation* parameter, and δ_{ij} is white noise iid process and follows a normal distribution: $\delta_{ij} \sim \mathcal{N}(0, \sigma_\delta^2)$. Note that $|\rho| < 1$ defines the process as *wide-sence stationary*. In this case the correlation function would be:

$$\text{Corr}(\varepsilon_{ij}, \varepsilon_{il}) = \rho^{|j-l|},$$

We ignore here the spatial pattern and assume that the process is spatially constant. That is, at every spatial unit i the $T \times T$ covariance matrix is:

$$\text{Var}(\varepsilon_j) = \tau^2 \begin{bmatrix} 1 & \rho & \rho^2 & & \rho^{T-1} \\ \rho & 1 & \rho & \dots & \rho^{T-2} \\ \rho^2 & \rho & 1 & & \rho^{T-3} \\ & \vdots & & \ddots & \vdots \\ \rho^{T-1} & \rho^{T-2} & \rho^{T-3} & \dots & 1 \end{bmatrix},$$

where $\tau^2 = \text{Var}(\varepsilon_{ij}) = \frac{\sigma_\delta^2}{1-\rho^2}$. Note that this covariance model requires the estimation of 2 parameters.

5.2.3.2 Spatial perspective:

Econometric approach: model the errors generating process through the weight matrix approach

Consider the model in $(*)$, now with different assumption regarding to ε. Remaining in matrix notation:

$$\varepsilon = \lambda\tilde{\varepsilon} + \omega$$

where $\tilde{\varepsilon} = W\varepsilon$

$$\varepsilon = (I - \lambda W)^{-1}\omega$$
$$\text{Var}(\varepsilon) = \sigma_\omega^2(I - \lambda W)^{-1}(I - \lambda W')^{-1}$$

Some alternatives for choosing the components w_{uv} of W:

k-nearest neighbors:

$$w_{uv} = \begin{cases} 1 & , u \in N_k(v) \\ 0 & , otherwise \end{cases}$$

Radial Distance:

$$w_{uv} = \begin{cases} 1 & , 0 \leq d_{uv} \leq L \\ 0 & , otherwise \end{cases}$$

Power Distance:

$$w_{uv} = \frac{1}{d_{uv}^{\alpha}}$$

Direct Specification of the correlation matrix components K: a functional form for the covariance structure is assumed. parameters might be estimated jointly with regression parameters using ML.

Some alternatives:

- *Negative exponential*:

$$\sigma_{ik} = b_1 \exp - \frac{d_{ik}^{\alpha}}{b_2}$$

when $\alpha = 2$ it is a gaussian

- *Gaussian*:

after constructingthe correlation matrix K takes $\hat{\Sigma} = \sigma_{\varepsilon}^2 K$ (how to get σ_{ε}^2 ?)

5.2.3.3 ***FGLS***??

$\Sigma(\theta)$ can be consistently estimate using the *feasible generalized least squares* (FGLS) estimator. The FGLS is a common method mainly in the econometric field, in which great emphasis is put on data structure. FGLS, modeling proceeds in two stages: (1) the model is estimated ignoring the dependence structure of the data (e.g. with OLS), and the residuals are then used to construct the error covariance matrix estimator $\widehat{\Sigma(\theta)}$. (2) The GLS estimation is performed using the first stage estimated $\widehat{\Sigma(\theta)}$.

5.2.3.4 Spatio-temporal perspective: *something*

toy model

Consider the toy 2-time-units × 2-spatial-units symmetric covariance matrix of the residuals η from the model $y = X\beta + \eta$

$$V_{2\times 2} = V'_{2\times 2} = \begin{bmatrix} v_{1,1} & \rho_{1,\{1,2\}} & \gamma_{\{1,1\},\{2,1\}} & \gamma_{\{1,1\},\{2,2\}} \\ & v_{1,2} & \gamma_{\{1,2\},\{2,1\}} & \gamma_{\{1,2\},\{2,2\}} \\ & & v_{2,1} & \rho_{2,\{1,2\}} \\ & & & v_{2,2} \end{bmatrix}$$

More generally, for time-unit j and spatial-unit i, $j = 1, ..., T$, $i = 1, ..., N$:

- $v_{i,j}$ correspond with the variance of the residuals in spatial-unit i in time-unit j.
- $\rho_{i,\{j,j'\}}$ correspond with the correlation between the residuals in in time-units j and j' in the same spatial-unit i, i.e. *temporal autocorrelation*
- $\gamma_{\{i,j\},\{i',j\}}$ correspond with the correlation between the residuals in in spatial-points i and i' at the same time-unit j, i.e. *spatial autocorrelation*
- $\gamma_{\{i,j\},\{i',j'\}}$ correspond with the correlation between the residuals in in spatial-points i and i' at time-units j and j' respectively, i.e. *spatio-temporal correlation*

Some simplifications are possible, for instance: $\rho_{i,\{j,j+t\}} \to 0$ when $g_t(t) > c_t$, $\gamma_{\{i,j\},\{i+d,j\}} \to 0$ when $g_d(d) > c_d$ and, $\gamma_{\{i,j\},\{i+d,j+t\}} \to 0$ when $g_{td}(t,d) > c_{td}$, where g_t, g_d, g_{td} are functions describing the correlations decay, and c_d ,c_t, c_{td} can be assumed or estimated.

Moreover, in spatial context, $g()$ can be assumed isotropic or anisotropic. [Cressie and Chan (1989)][7] chose $g()$ based on the estimated variogram of the obser vations

isotropy vs. unisotropy

5.2.4 Estimation Perspective

5.2.4.1 Estimation of the Mixed Model

Back to the simple LME model. A typical assumption is that:

$$\varepsilon_i \sim \mathcal{N}(0, \sigma^2 I) \qquad b_i \sim \mathcal{N}(0, \sigma^2 D)$$

and $I = I_{s_j}$.

The multivariate normal distribution of y_j can then be writen as:

$$y_j \sim \mathcal{N}(X_j\beta, \sigma^2(I + Z_j D Z'_j)),$$

and the log likeliyhood function for the linear mixed model is given by:

$$l(\beta, \sigma^2, D) = -\frac{N}{2}\ln 2\pi - \frac{1}{2}\Big(N \ln \sigma^2 + \sum_{j=1}^{N}\Big(\ln|I + Z_j D Z'_j| + \sigma^2 (y_j - X_j\beta)'(I + Z_j D Z'_j)^{-1}(y_j - X_j\beta)\Big)\Big)$$

This log likelihood function involves matrix inverse and determinante which might be a difficult task if the matrix dimension is high. However, some dimension reduction formulation can be employed in order to make calculation easier [ref][]

5.2.4.2 Estimation of the GLS

GLS estimation is essentially applying OLS to the transformed data. To see this, consider Σ's Cholesky's decomposition: $\Sigma = L\Lambda L'$ where L is a unitriangular matrix and Λ is a diagonal matrix. Easy to see that:

$$\Sigma^{-1} = PP',$$

where $P = L^{-1}\Lambda^{-\frac{1}{2}}$ and that $P\Sigma P' = I$.

Note that mltiplying both sides of $(*)$ by P yields:

$$(**) \qquad \tilde{y} = \tilde{X}\beta + \tilde{\varepsilon},$$

where, $\tilde{y} = Py$, $\tilde{X} = PX$ and $\tilde{\varepsilon} = P\varepsilon$. Also note that $E(\tilde{\varepsilon}) = 0$ and $\mathrm{Var}(\tilde{\varepsilon}) = E(P\varepsilon\varepsilon'P') = \sigma^2 P\Sigma P' = \sigma^2 I$, hence β can be estimated using $(**)$ and the OLS:

$$\begin{aligned}\hat{\beta}_{GLS} &= (\tilde{X}'\tilde{X})^{-1}\tilde{X}'\tilde{y} \\ &= (X'P'PX)^{-1}X'P'Py \\ &= (X'\Sigma^{-1}X)^{-1}X'\Sigma^{-1}y\end{aligned}$$

Therefore ....

5.2.5 Regularization

Reducing the variance of the predicted values can be done by *shrinking*, while sacrificing a little bit of bias. When the goal is to improve prediction accuracy this should be considered. Shrinkage methods are continuous selection models (not limited to discrete variable selection) that impose a penalty on the regression coefficients. The most familiar are *ridge regression*, the *least absolute shrinkage and selection operator* (Lasso), and the *elastic net*, (see [Hastie et. al, 2009][ESL ch. 3.4] for an enlightening review). However, these methods are mostly suitable for data with independent observations, and are not straightforward in a spatio-temporal datasets.

5.2.5.1 Regularization in the Mixed Model

Altought regularization in regression models have received considerable attention over the past years, literature on regularized LME models is somewhat scarce. The challenge in regularization of mixed models is to properly select random effects together with the fixed effects. This challenge stems from the fact that as long as the random effects are not determined its covariance matrix is unknown. One option is to perform selection in separate stages, but it may lead to different regularization solutions depending on the order of the stages.

Recently, several procedures have been proposed to identify both the random and fixed effects. [Bondell et. al., 2010][Joint Variable Selection] proposed a simultaneous selection of the fixed and random effects in an LME model using a modified Cholesky decomposition. Their method is based on a penalized joint log-likelihood with an adaptive penalty (*adaptive Lasso*). [Fan and Li, 2012][var sele in LME] proposed to use a proxy matrix in the penalized profile likelihood to overcome the difficulty of unknown covariance matrix of the random effects. One drawback of these kind of methodsis that they usually involve complex numerical optimization, therefore are computational intensive in relation to classical regularizations methods

5.2.5.2 Regularization in GLS

As described, in GLS estimaton the OLS method is implemented on the whitening transformation of the data. Therefore, its regularization is as easy as regularization of the OLS model:

$$\begin{aligned}\hat{\beta}_{RGLS} &= \arg\min_{\beta}\Big\{(y - X\beta)'\hat{\Sigma}^{-1}(y - X\beta) + \lambda g(\beta)\Big\}\\ &= \arg\min_{\beta}\Big\{(\tilde{y} - \tilde{X}\beta)'(\tilde{y} - \tilde{X}\beta) + \lambda g(\beta)\Big\}\end{aligned}$$

where $g(\beta)$ is some penalization on model complexity. For instance by setting: $g(\beta) = \sum_{i=k}^{p} \beta_k^2$ we get the ridge regression estimator:

$$\hat{\beta}_{RGLS} = \beta_{ridge} = (\tilde{X}'\tilde{X} + \lambda I)^{-1}\tilde{X}'\tilde{y}$$

Note that β_{RGLS} does not involve numeric optimization and enjoys a closed form solution.

5.3 Example: Validation and Estimation in Epidemiological Perspective

EIV Rmd file goes here

5.4 Computational Challenges

why gls is easy to compute and to parellelize

Although both formulas superficially require large matrices to be inverted, the calculations can be organized to allow individual rows and columns of both to be computed without the need for matrix inversion. In practice, a few representative rows are usually all that is needed, for it is impractical to completely analyze very large matrices, anyway.

5.4.1 Sparse Representations

5.4.2 Memory Efficiency

5.4.3 Parallel Computing

6 Preliminary Results

7 Appendix

8 References

[1]: [2]: (http://pelletierb.perso.math.cnrs.fr/Publications_files/bp-jgr07.pdf) [3]: (https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1638034/pdf/envhper00306-0069.pdf) [4]: (http://www.sciencedirect.com/science/article/pii/S1352231014005354) [5]: (https://ntrs.nasa.gov/archive/nasa/casi.ntrs.nasa.gov/20150006835.pdf) [6]: (https://www.ncbi.nlm.nih.gov/pubmed/22709681) [7]: (https://ehp.niehs.nih.gov/wp-content/uploads/124/8/ehp.1409671.alt.pdf) [8]: (https://projecteuclid.org/download/pdfview_1/euclid.ssu/1268143839) [9]: (http://onlinelibrary.wiley.com/doi/10.1111/j.1467-9892.1992.tb00102.x/epdf) [10]: (http://www.ssc.wisc.edu/~bhansen/718/BurmanChowNolan1994.pdf) [11]: (https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3569684/) [12]: (https://www.jstor.org/stable/pdf/2676791.pdf?refreqid=excelsior%3Af543fd1414d620015a70259258b347d2) [13]: (https://www.jstor.org/stable/pdf/2289087.pdf?refreqid=excelsior:6c5480d9ab4761bd800086db207ac0a8) [14]: () [15]: () [16]: () [17]: () [18]: () [19]: ()

Weiss, Robert E. Modeling Longitudinal Data. Springer Science & Business Media, 2006.