Introducing High dimensional statistics and the properties of the LASSO estimator : a tool for recovering the true spatial weight matrix : First Draft subject to changes

Barra

9 septembre 2016

Abstract

In this paper we want to estimate a spatial weight matrix out of the data. We introduce problematic from the High Dimensional Statistics and more particularly the LASSO from Tibshirani. We introduce some algorithms allowing to solve this problem in case where $K > N$. We estimate W using one of those algorithm in a pure spatial autoregressive framework. Although a lot remains to do, notably establishing a clear link between these optimization techniques and the economical interpretation of the outcome and the computation of standard error , we believe in the orientation used to attempt to make W endogenous. We apply this techniques in order to estimate the influence of the number of suicide in one region on the other by data taken over 22 French regions.

1 Introduction

Spatial Econometrics is nowadays a well known tool for estimator bias or inefficiency correction when cross sectional dependence or heterogeneity is suspected in a data set. These two latter effect can be seen as a matter of model misspecification. The very nature of the data collected by regional economists suffers from two major drawback. The spatial unit of observation, say a district, is not necessarily the good layer to analyse a spatial phenomenon, hence there may be spill overs between the boundaries of two units of observation leading to non spherical variance/covariance matrix of the error term. But more importantly, what is observed as a point in space is likely to influence what happens in an other. This phenomenon of diffusion is very important in the sense that it causes a type of endogeneity that classic econometrics fails to handle.
This has led regional scientists and econometricians to develop techniques based on the neighbourhood notion to tackle these estimation problems. Anselin in its pioneer work, developed estimations techniques based on the use of spatial weight matrix W. This matrix is a way to formalize the neighbourhood notion, and can be set up in various ways. The most common are the contiguity matrices, where if two observations i and j share a common border, the entry w_{ij} will take value 1, 0 if not, as proposed by Moran(1948) and Geary(1954). Another popular way to model this structure is to postulate that the effect of a phenomenon will decrease as a function of the distance, in this framework, the entry w_{ij} will take value $w_{ij} = \frac{1}{d_{ij}^n}$ as proposed by Cliff

and Ord and reflecting Tobler's first law of geography "everything is related to everything else, but near things are more related than distant things".
However all the definitions proposed for the weight matrix entry rely on the knowledge, the expertise of researcher about the phenomenon under study. It is in no case data driven. In other word, the spatial dependence structure is exogenous and fixed a priori by the researcher. One interesting field of research in spatial statistics is to consider that this dependence should be endogenous and estimated. The determination of the w_{ij} entry has been widely discussed and its choice is often subject to critics. Why should a researcher use such W matrix and not an other ? Which power of the distance should be chosen when using inverse distance matrices ? The controversies are often endless. The main critic is formulated in Anselin(1988) :

"An important problem results from the incorporation of parameters in the weights. Typically these weights are taken to be exogenous and the parameter values are determined a priori, or in a step separate from the rest of the spatial analysis[] In particular it could potentially lead to the inference of spurious relationships, since the validity of the estimates is pre-conditioned by the extent to which the spatial structure is correctly reflected in the weights"

As the validity of the parameters is influenced by the choice upon the weight matrix, one should be interested in jointly estimating the slope parameters and the weight matrix in a unified procedure. Moreover it is conceptually twisted to fix a priori a quantity we wish to estimate. Indeed our goal must not be only to retrieve good properties for the $\beta's$ but also to recover the "true" data generating process. Hence, to infer the weight matrix entry from the data.
Here we face a major problem, that the pioneers of the discipline avoided in choosing a priori W, the number of parameter to estimate is far more important than the number of observations available, and this by construction. Indeed W relating each observation to the others, it is of dimension N.N, with the main diagonal being 0 if we consider that a location cannot be its own neighbour. All these is well known by anyone interested in spatial econometrics. Considering each entry as a parameter to estimate we have $N^2 - N$ parameters to estimate with N observation. Increasing N will not provide suitable asymptotic and more over an OLS estimation would not be unique as the number of unknown is greater than the number of equation to determine them, this is known as the incidental parameter problem.
Therefore, the spatial econometricians must shift their paradigm and slip into a world where K, the number of parameters to estimate is greater than N and moreover where it is a growing function of N. High dimensional statistics can provide the econometricians with such tools. Tibshirani in 1996 proposed a interesting approach by minimizing the model Residual Sum of Square(RSS) under the constraint that the parameter's L1 norm is inferior to some constant. The Least Absolute Shrinkage and Selection Operator (LASSO) was born. The LASSO achieve subset selection by the geometry of its constraint and ensure sparsity (parsimony) to the model. Hence we face a classic convex objective function but here the constraint although convex is not differentiable in 0. This procedure will be the core of this paper and we aim at explaining it in a comprehensive manner to spatial econometricians and will present various ways to solve it. Tibshirani (1996) proposed an algorithm to solve this optimization problem when N>K.
Most of the following researches were focusing on finding less greedy algorithm to solve this problem. A major contribution comes from Osborne, Prescott and Turlach (2000) using the dual of the optimization problem to allow K>N. Johnston and Donoho (1996) posed the same problem under the name of "basis pursuit". Sardy et al. (2000) proposed a block relaxation algorithm to solve it when K>N as well but where the design matrix must be "orthonormal-union complete". An algorithm, close to the forward stagewise for choosing most relevant variables into a linear model, the Least Angle Regressions (LARS). The latter allows to incorporate the variable most

correlated with the residuals one by one in a path dependent manner. Here at each step the algorithm add one more regressor to the model. If we have K regressors and that we implement the Algorithm K times, the result will be the same than an OLS regression. This is not really helpful to overcome our dimension problem, however the LARS algorithm can be modified to achieve variable selection and to reduce the dimension of the parameters at play in the model. Efron, Hastie, Johnstone and Tibshirani (2004) show the links between the LARS and the LASSO.
Our goal being to estimate the entries of the W matrix, we want some properties to be preserved. The main goal is to recover the true data generating process. So we must have a procedure which avoid false positive and false negative, which is, any $w_{ij} \neq 0$ must be in the active set of parameter and any $w_{ij} = 0$ must be outside the active set. We will formally define later this notion of active set and show that the LASSO respect this property. Of course the trade off will be at the expense of a bias in the point estimate.
Some modifications must be made in order to apply the LASSO to spatial econometrics. First of all our parameter space is here a matrix which will need to be vectorized to exhibit a suitable form for estimation. Secondly we have two type of parameters in our problem. One is formed of the $N^2 - N$ "spatial parameters" and another is formed by the K parameter slope related to the problem under study. The objective function being convex in the spatial parameters conditionally to the parameter slope and vice versa but not jointly. This will require a modification of the classics algorithm to fit the procedure to our problem.
Lam and Souza (2014) implement elegantly a block descent coordinate algorithm for solving the LASSO in case of a Spatial Durbin Model with stationary error term. Bahttacharjee (2015) implement a two step LASSO using instrumental variable to estimate the W entries.
As far as I know these are the only two attempt to overcome Tobler's first law of geography by econometricians. If not providing substantial advances to these two authors contribution, this essay aims at democratizing the comprehension of the LASSO procedure and to show its powerful application to spatial econometrics.
Using the specification designed in Sardy et al. We aim to estimate the W matrix entries in a cross sectional framework, which has as far as we know, never been done.
A sparse group LASSO strategy proposed by Friedman, Hastie and Tibshirani (2010) allowing to set some blocks to 0, in our context, meaning that some observations interacts with no others. But also to achieve sparsity within the blocks which allow to avoid some strong condition could be applied.
The rest of the essay is organized as follow : section 2 will sketch a brief historical survey of LASSO. Section 3 will describe the LASSO, its important properties and some ways to solve it, which are according to us key, either because they represent an historical break-through or because they are widely used. This is nothing but exhaustive as the literature in optimization of non differentiable function is wide. Section 4 will introduce the properties of such L1 penalized models in term of model selection and bound for the prediction error.
Section 5 describe the spatial setting and show how the LASSO can be used to estimate the W matrix and will present an example using flue (or suicide) data over 22 French metropolitan regions

2 Historical context

The development of high dimensional models (the number of covariate in the design matrix are high), by the access to larger and larger dataset has risen the problem of very volatile OLS estimators. Indeed if K is large and/or very correlated with each others the variance/covariance matrix of the estimators is large leading to a large prediction error.

Two ways to overcome this problem was subset selection (that we will skip for the moment) and the ridge regression. We will first focus on the ridge regression because it is a closed parent of the LASSO. The initial objective was to constraint the vector parameter to a given length. Hence the RSS minimization had to be constrained. The euclidean norm of the vector was a good constraint as it is convex and differentiable. The Ridge solution was given by :

$$\hat{\beta}_{ridge} = Argmin \sum_{N} (Y_n - \sum_{K} \beta_k . x_{nk})^2 \tag{1}$$

$$\text{S.T} \sum_k \parallel \beta_k \parallel_2^2 \leq cst$$

Geometrically, the form of the constraint is spherical as shown in fig.1. This allow to shrink parameters toward the origin fulfilling the objective to constrain their variation. But no parameter can be exactly set to 0 for a finite constant. From fig.1, one can intuitively see that the smaller the constant, the closer to origin is the sphere. Hence when cst tend to 0 the parameters tend to 0 as well. On the other hand, if the minimal RSS is inside the feasible region (i.e the constant is large), then the constraint is not binding and $\hat{\beta}_{ridge} = \hat{\beta}_{OLS}$
Facing a convex optimization problem where both the loss function and the constraint are differentiable, a closed form solution exists. It is given by :

$$\hat{\beta}_{ridge} = (X'X + I_K . \lambda) X'Y$$

Breiman(1993) introduce a new technique for subset selection which is shrinking some parameters and setting others to 0 called the non negative garrote.
Consider a data set $(y_i, X_{i1}, ..., X_{iK})$ for $i = 1...N$
Suppose K large, so that for a better model interpretation we want to reduce this dimension. Suppose also that we can reasonably think that few of the predictors play a role in predicting the response vector Y. **This assumption of sparseness is key in all that follow**. We face a classical problem of trade off between variance and bias. Although adding new variables increase the regression equation variance. Indeed, considering Gaussian predictors, say :
$X \sim N(\mu, \sigma^2)$
The prediction error will be given by $\frac{E[(\hat{\beta}-\beta)' X' X (\hat{\beta}-\beta)]}{n} = \frac{\sigma^2}{n} . K$
On the other hand, adding new regressors reduces the estimation bias. Suppose that a proportion $\frac{K}{c}$ with $c > 1$ a constant of the covariate actually help in predicting Y, then we can both reduce the variance without affecting the bias. Breiman proposed an estimation procedure that both select the relevant variables and reduces the variance, the non negative garrote (NNG).
Consider a linear model, the NNG goes as follow : take c_k to minimize

$$\sum_{N} (Y_n - \sum_{K} c_k . \hat{\beta}_k . x_{nk})^2 \tag{2}$$

$$\text{S.T } c_k > 0 \text{ and } \sum_K c_k < s$$

$\hat{\beta}$ being the initial OLS estimate. The new estimated parameter will therefore be $\tilde{\beta}_k = c_k \hat{\beta}_k$
The argument is that by reducing s more of the c_k become 0 and their associated regressors are set to 0.
Tibshirani(1996) critics the non negative garrote procedure in that its solution is sensitive to both the sign and magnitude of the first step OLS estimates. In a framework where OLS behave poorly, for example when the Gram matrix $X'X$ exhibit high level of correlation between the covariates, the garrote solution will suffer the initial non stability of the OLS.
Tibshirani introduces the LASSO which is not sensitive to the OLS parameter sign. The motivation for such techniques was driven by the will to reduce a large set of genes potentially

responsible for prostate cancer with a low number of patients. The idea is to minimizes the model RSS under a L1 norm constraint on the parameter vector. The geometry of the L1 constraint which is for $K = 2$ a rotated square. The LASSO solution is reached when the RSS meet the rotated square, according to the data it can happen that this constraint is met on a corner of the constrained space. In this case, one of the parameter will be set to 0 and considered as non active.
More formally :

$$\hat{\beta}_{lasso} = argmin \sum_{i=1}^{N} (y_i - \sum_k \beta_k . x_{ki})^2 \tag{3}$$

$$\text{S.T} \sum_k \mid \beta_k \mid \leq cst$$

As both the objective function and the constraint are convex in β an equivalence can be made between the constant and a Lagrangian. Hence the above equation can be written as follow :

$$\hat{\beta}_{lasso} = argmin \sum_{i=1}^{N} (y_i - \sum_k \beta_k . x_{ki})^2 + \lambda \mid \beta \mid_1$$

It is clear from this equation that the bigger the constant is, the less binding is the constraint exactly as in the ridge regression framework. There is a direct link in between the constant and λ when the former increase the latter tend to 0. It is also straightforward to see that if $\lambda = 0$ the LASSO solution is equivalent to the OLS. The choice of λ will therefore be of key interest to respect the non false negative and non false positive property that we will now refer to as *Oracle property.*
The choice of the L1 norm is motivated by the geometry of its constraint and is a modification of the ridge regression where the penalization is done with respect to the parameter L2 norm, which shrinks the parameter but cannot set any of them to 0. The figure below show a geometrical example with K=2 covariates.

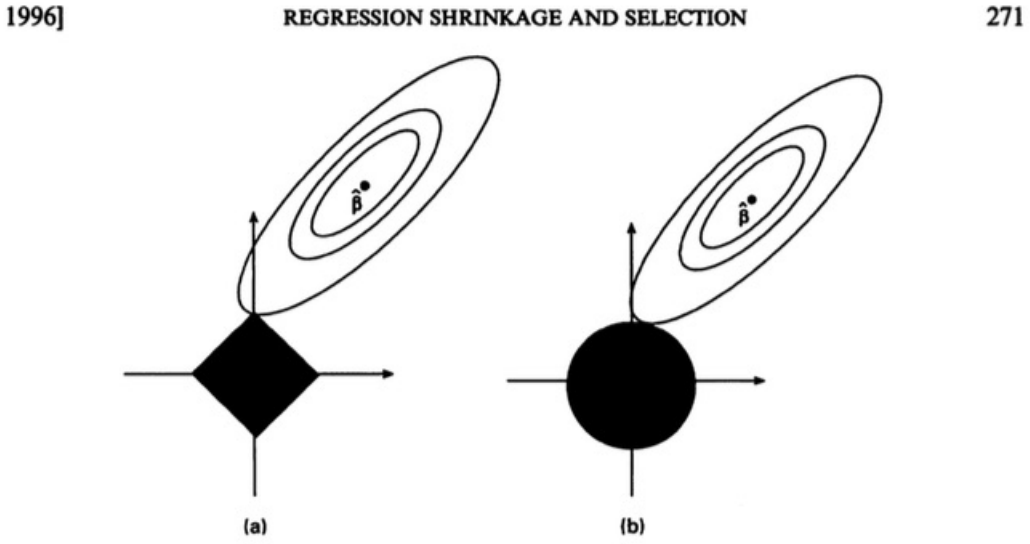

Fig. 2. Estimation picture for (a) the lasso and (b) ridge regression

source : Tibshirani (1996)

The objective function being convex and differentiable, the constraint being the sum of convex functions, hence convex, the solution to (3) is given by solving a quadratic programming problem under linear inequality constraint. The fact that the absolute value function is not differentiable in 0 leads to no closed form solution. Many ways to represent this constraints have been proposed.
The most straightforward way is to represent it as combination of linear constraints on the sign of β. The constraint is made on the sign of the parameters because of the nature of the absolute value function.

for $\beta \in \Re$
$\mid \beta \mid = \beta$ if $\beta > 0$
$\mid \beta \mid = -\beta$ if $\beta < 0$
$\mid \beta \mid = 0$ if $\beta = 0$

Suppose K=3, the set of linear constraints will be :

$+\beta_1 + \beta_2 + \beta_3 \leq cst$
$+\beta_1 - \beta_2 + \beta_3 \leq cst$
$+\beta_1 - \beta_2 - \beta_3 \leq cst$
$+\beta_1 + \beta_2 - \beta_3 \leq cst$
$-\beta_1 + \beta_2 + \beta_3 \leq cst$
$-\beta_1 + \beta_2 - \beta_3 \leq cst$
$-\beta_1 - \beta_2 + \beta_3 \leq cst$
$-\beta_1 - \beta_2 - \beta_3 \leq cst$

Any vector β minimizing the model RSS and satisfying these constraints will be a solution to the LASSO. However one can see that the number of constraint is of order 2^K which can quickly become un-tractable when K is large.

3 Algorithms

The number of constraints being large, the solution to such a system of equation can become very long to solve. In this section we give a closer look to the various algorithms used to solve this constrained (easily transformed into unconstrained) problem.

3.1 Case where $N \leq K$

first algorithm proposed by tibshirani Consider γ_i a 1.K vector of the form (+/-1,...,+/-1)' with i= 1...2^p , the number of constraint.
Consider two sets of indices $E = [i \mid \gamma_i'\beta = cst]$ the equality set, and a set for which the constraint is slack $S = [i \mid \gamma_i'\beta < cst]$.
Consider G_E a matrix whose rows are γ_i for $i \in E$

The Algorithm add sequentially the constraints and goes as follow :
a) Start with $E = i_0$ where $\gamma_{i0} = \text{sign}(\hat{\beta}^0)$, with $\hat{\beta}^0$ the over all OLS estimate.
b) Find $\hat{\beta}$ minimizing the RSS under constraint $G_{E0}.\beta \leq cst$
c) while $\sum_k \mid \beta_k \mid \leq cst$

d) add i to E where $\gamma_i = \text{sign}(\hat{\beta})$

To get a better understanding let us take our example with 3 parameters.
$\hat{\beta}^0 = \text{Argmin}(Y - X\beta)'()Y - X\beta)$

imagine $\hat{\beta}_1^0 < 0 \quad \hat{\beta}_2^0 > 0 \quad \hat{\beta}_3^0 > 0$ and $\mid \hat{\beta}_1^0 \mid + \mid \hat{\beta}_2^0 \mid + \mid \hat{\beta}_3^0 \mid > cst$

Min $(Y - X\beta)'(Y - X\beta)$
S.T $\begin{bmatrix} -1 & 1 & 1 \end{bmatrix} . \begin{bmatrix} \hat{\beta}_1^0 \\ \hat{\beta}_2^0 \\ \hat{\beta}_3^0 \end{bmatrix} \leq cst$
And so on until $\mid \hat{\beta}_1 \mid + \mid \hat{\beta}_2 \mid + \mid \hat{\beta}_3 \mid \leq cst$

3.2 Case where $K > N$

Dual and active set method
Osborne, Presnell and Turlach (2000) using a local linearisation with respect to $\hat{\beta}$ and the dual properties of the optimization problem propose a procedure to solve the LASSO even in case K>N.
This is a Major advantage remembering that our ultimate objective is to estimate the entries of a weight matrix, where in a cross sectional framework, the number of parameter is far greater than the number of observation
Without entering into the proof of weak and strong duality theorem (we refer to the classic convex programming theory), the algorithm is based on the minimisation of the local linearised RSS with respect to the current parameter vector. An active set strategy is adopted, which means that the covariates are of two kinds, if $\beta_k \neq 0$ then $\beta_k \in \sigma$ with σ being the active set. On the other hand if $\beta_k = 0$ then $\beta_k \in \sigma^c$
Hence a convex optimization problem is transformed the minimization of a linear function over a convex set spanned by the constraint.
To understand what is happening in this optimization problem, one must keep in mind the primal form of the Lagrangian function.
Recall the initial minimization problem (3). It can be reformulated as :

$$\hat{\beta}_{LASSO} = Argmin(Y - X\beta)'(Y - X\beta) - \lambda(cst - \sum_{j=1}^{K} \beta_j)$$

Let us consider the upper bound of this function in case $\lambda > 0$

$$\begin{cases} Argmin(Y - X\beta)'(Y - X\beta) & if cst - \sum_{j=1}^{K} \beta_K \geq 0 \\ \infty & if cst - \sum_{j=1}^{K} \beta_K \leq 0 \end{cases} \tag{4}$$

Minimizing the upper bound of the Lagrangian is equivalent to solve (3). Moreover, this upper bound function is discontinuous and impose an infinite penalty for violating the constraint. And the Upper bound of the Lagrangian plays the role of a barrier function modulated by a barrier parameter μ
the algorithm goes as follow :

$$MinF(\beta + h) \tag{5}$$

$$\text{s.t } \theta'_\sigma(\beta_\sigma + h_\sigma) \leq cst \text{ and h} = (h_\sigma \ 0)\text{'}$$

h being the local linearisation of $(Y - X\beta)'(Y - X\beta)$
Let us develop it to see clearer what we are minimizing.

h $= RSS^c + \sum_{i=1}^{K} \sum_{j \neq i} (x'_i y + 2\beta_i^c x'_i x_i + \beta_j x'_i xj \beta_j x'_j x_i)(\beta_i - \beta_i^c)$; Where the upper script c stands for "current"
Stacked in matrix form we find back the notation as in Osborne Presnell and Turlach. We hence search for a β_σ minimizing the first order expanded RSS. subject to the fact that :
$sign(\beta_\sigma^c).\beta_\sigma^c + h_{sigma}$ is feasible.

If the constraint is binding, the Karush Kuhn Tucker conditions are written as in Nash and Sofer (1996) :

$$(X'_\sigma X_\sigma).h_\sigma + \mu.\theta_{sigma} = X'_\sigma(Y - X_\sigma\beta_\sigma)$$
$$\theta'_\sigma.h_\sigma = cst - \theta'_\sigma\beta_\sigma$$

Solving this system of equation for h_σ, we obtain :

$$\begin{cases} (X'_\sigma X_\sigma)^{-1}[X'_\sigma(Y - X_\sigma\beta_\sigma)] + 0 & if \mu < 0 \\ (X'_\sigma X_\sigma)^{-1}[X'_\sigma(Y - X_\sigma\beta_\sigma)] + \dfrac{\theta'_\sigma(X'_\sigma X_\sigma)^{-1}Y - cst}{\theta'_\sigma(X'_\sigma X_\sigma)^{-1}\theta_\sigma}\theta_\sigma & if \mu > 0 \end{cases} \tag{6}$$

update $\beta^c = \{\beta_\sigma^c \quad 0\}$' by $\beta^+ = \{\beta_\sigma^+ \quad 0\}$' = $\{\beta_\sigma^c \quad 0\}$' + $\{h_\sigma^c \quad 0\}$'

$$\begin{cases} sign(\beta_\sigma^+) = \theta_\sigma & \beta^+\text{is sign feasible} \\ sign(\beta_\sigma^+) \neq \theta_\sigma & \beta^+\text{is not sign feasible} \end{cases} \tag{7}$$

If β^+ is not sign feasible
Move to the new zero component in the direction of the gradient h. That is find the smallest step $0<\gamma<1$ and the corresponding element $\beta_k^c \in \sigma$ such that $\beta_k^c + \gamma h_k = 0$ and set $\beta^c = \beta^c + \gamma.h$
Change the element θ_k for $-\theta_k$
Minimize (4) with the new vector β and sign(β) and solve for h, call its solution $\hat{h}$
If $f(\beta + \hat{h}) < f(\beta + h)$ and compatible with the new vector sign(β). Then h is a feasible descent direction.
else delete k from the active set obtain β_σ^- and θ_σ^-
Minimize (4) with the new vector β_σ^- and θ_σ^- and solve for h^-, and check if it is sign feasible.
Iterate until a sign feasible combination is found.
this latter step is actually equivalent to testing if the solutions to the minimization problem are inside the feasible region, the sign restriction is due to the nature of the Absolute norm constrained problem.

Once a sign feasible β^+ is found we know the problem is feasible, now we are searching for optimality.
We want to verify the condition $-X'(Y - X\beta^+) + \lambda\theta = 0$

Compute $v^+ = \frac{X'(Y - X\beta^+)}{\|X'(Y - X\beta^+)\|_\infty} = \begin{bmatrix} v_1^+ \\ v_2^+ \end{bmatrix}$
With v_1^+ the ratio of the correlation between the design matrix and the residuals on the maximum correlation taken among the regressors in the active set.

With v_2^+ the same ratio taken among the regressors outside the active set.
If -1< v_2^+ <1 The solution is reached
else pick the parameter s associated with maximum | v_2^+ |
Append a 0 to β_σ and expend θ_σ of the sign of $(v_2^+)_s$
Minimize (4) with this new parameter and iterate until -1< v_2^+ <1

The advantage of this procedure is that one can start the algorithm with $\sigma = \emptyset$ and hence is not dependant of the OLS solution. When all parameters are set to 0. To determine which variable should enter first, we refer to the second part of the algorithm. The variable outside the active set with maximum correlation with the residual will enter the set. In the specific case where $\beta = 0_K$ this correlation can be written as :
$corr(X, Y - X\beta) = corr(X, Y)$ Hence it sounds logical that the first variable entering the active set be the most correlated with the response vector.
The Lars of Efron and Tibshirani uses the same starting strategy. Although more recent we will not develop it here as it is not implementable if K>N. The full procedure and its links to LASSO and stagewise forward regression are developed in Efron, Hastie, Johnstone and Tibshirani (2004).
We now describe a Generalized block relaxation procedure for solving basis pursuit, the other named for LASSO developed contemporaneously in engineering. This technique allow for K>N by optimizing blocks of parameters holding the other constant.

Sardy, Bruce and Tseng (2000) : Block coordinate relaxation method for solving LASSO
The block coordinate relaxation method has been used in many optimization problem, and is a generalization of the Gauss-Seidel algorithm for solving a linear system of equation, to blocks of parameters and not to a single one.
The idea of this method is fairly elegant and allows to divide the problem in M sub-problem.
In this case, the authors propose to apply it to blocks of parameters. Where we optimize one block letting the others fixed, and iterate cyclically through the blocks (for the most simple version).
Suppose that you can decompose the design matrix X in M subgroups by columns. such that :

$$\text{columns(X)} = \bigcup_{M=1}^{M} \text{Columns } (X^{(M)})$$

Hence each sub-matrix is of size N. $(\frac{K}{M})$ and the parameters vector is also divided by M, and blocks of $(\frac{K}{M})$ parameters are estimated with respect to corresponding design matrices of size N. $(\frac{K}{M})$.

The algorithm goes as follow :

1. Choose an initial guess for the vector parameter β without loss of generality it can be 0. If an OLS or a ridge estimate is possible, it might be better to use it as it is closer to the solution.
2. select the block $m \in [1, M]$
3. Partition the column of X and β in two sets : $X^{(m)}$and$\beta^{(m)}$ and $X^{(-m)}$and$\beta^{(-m)}$
4. Define a residual vector $e^{\hat{(m)}} = Y - X^{(-m)}\beta^{(-m)}$
5. $\beta^{\hat{(m)}} = Argmin\frac{1}{2}(e^{\hat{(m)}} - X^{(m)}\hat{\beta})'(e^{\hat{(m)}} - X^{(m)}\hat{\beta}) + \lambda \parallel \hat{\beta} \parallel$
6. Solve this LASSO sub-problem by an adapted method, as an example by the dual, active set method of Osborne et al. described above.

7. iterate through the blocks until a stopping criterion for $\hat{\beta}$ is met

Friedman, Hastie and Tibshirani (2010) : The sparsistent grouped LASSO
As in Sardy, Bruce and Tseng the regressors are divided in L groups of dimension K/M . 1.
Derived from the group LASSO of Yuan and Lin (2007) who minimize the RSS penalized by a non squared L2 norm

$$Argmin \sum_N (Y_n - \sum_L \beta_L . X_{nL})^2 \text{ S.T } \sum_k \parallel \beta_k \parallel_2 \leq cst$$

Yuan and Lin show that such a specification achieves block sparsistency, which is, a block of parameter L is $0_{K/M}$
Friedman, Hastie and Thibshirani (2010) develop an elegant improvement of the group LASSO allowing for either block and intra blocks sparsistence. Moreover it allows relaxation of the orthogonality conditions for intra-block variables.
They propose to minimize the RSS like in Yuan and Lin but to add a L1 norm penalty on the overall parameter vector

$$\text{Min } \parallel Y - \sum_{l=1}^{L} X_l \beta_l \parallel_2^2 + \lambda_1 \sum_{l=1}^{L} \parallel \beta_l \parallel_2 + \lambda_2 \parallel \beta \parallel_1$$

The first part is differentiable for any value of β except in 0, hence, the first part subgradient equations are :

$$-X_l'(Y - \sum_l^L X_l \beta_l) + \lambda_1 \frac{\beta_l}{\|\beta_l\|} \text{ If } \beta_l \neq 0$$

$$-X_l'(Y - \sum_l^L X_l \beta_l) \leq \lambda_1 \text{ If } \beta_l = 0$$

A solution for the non zero β_l can be found :

$$\hat{\beta}_l = (X_l' X_l + \frac{\lambda}{\|\hat{\beta}_l\|})^{-1} X_l' (Y - \sum_{k \neq l} X_k' \hat{\beta}_k)$$

It is clear that sparsistency within blocks l where parameters are non zero is not achieve as there is no L1 norm expression in this setting.
If we consider the full model (sparse group LASSO)of the authors, this problem is tackled.
Let us remember the a group X_l is composed of M/K covariates that we will denote Z as in the original paper. As well β_l that we will denote θ_i
As in Sardy et al. Denote $\hat{e}_l = Y - \sum_{s \neq l} X_s \hat{\beta}_s$ and adopt a block coordinate descent strategy, and the sub-gradient equations are given by :

$$-Z_j(\hat{e}_l - \sum_j Z_j \theta_j) + \lambda_1 \eta_j + \lambda_2 \varsigma_j = 0$$

$$Z_j \hat{e}_l = \lambda_1 \eta_j + \lambda_2 \varsigma_j - Z_j \sum_j Z_j \theta_j$$

A system of M/K equations where $\eta_j = \frac{\theta_j}{\|\beta_l\|}$ if $\beta_l \neq 0$ and $\parallel \eta \parallel_2 \leq 1$ if $\beta_l \neq 0$
$\varsigma_j = \text{sign}(\theta_j)$ if $\theta_j \neq 0$ and $\varsigma_j \in [-1; 1]$ if $\theta_j \neq 0$ as it is the sub differential of the absolute value function in 0 (as we see as well in **4.2**).
From the above system equations we see that the only way for β_l to be zero is that this system has a solution respecting the conditions $\parallel \eta \parallel_2 \leq 1$ and $\varsigma_j \in [-1; 1]$.
To do so :

$$\frac{\partial(\frac{1}{\lambda_1^2}) \sum_{j=1}^{K/M} (Z_j \hat{e}_l - \lambda_2 \varsigma_j)}{\partial \varsigma_j}$$

for $\varsigma_j \in [-1;1]$
One can verify that the numerator is indeed $\sum_{j=1}^{K/M} \eta_j^2$
Minimizing this quadratic form and solving for ς_j and obtaining $\hat{\varsigma_j}$ and plugging it into the numerator must obtain a value for the latter ≤ 1 as $\| \eta \|_2 \leq 1$ must be satisfied.
If it is not the case, the whole vector can not be 0 and we go to another step of the procedure where we try to achieve sparsistency inside the non "full 0" blocks.
To do so, minimize the criterion :

$$\frac{1}{2}\sum_{i=1}^{N}(\hat{e}_i - \sum_{j=1}^{K/M} Z_{ij}\theta_j)^2 + \lambda_1 \| \beta_l \|_2 + \lambda_2 \sum_{j=1}^{K/M} | \theta_j |$$

Which is equivalent to a LASSO specification within one block.

4 Properties of the LASSO, sparse models

We have two major objectives in searching for estimating W, and more broadly in statistical learning. As we are in a framework where K is a growing function of N and that we have by construction constrained the RSS we must abandon the notion of consistency for an estimator, and the notion of asymptotic as well. Our first objective will be to find an upper bound for the prediction error.
More importantly the main objective is to isolate the signal from the noise out of a large number of potentially active covariates. In other words we seek to discover the relevant predictive variables, the "true" data generating process.
Hence a good procedure must fulfil oracles properties as defined in Fan and Lee (2011).

1. Identify the right subset model
2. Having the maximum estimation rate when the true subset is known

4.1 The prediction error upper bound and its probability

Following the notations as in Bülhman and Van de Geer, for indices $j = 1, ..., K$
Let $S_0 = [j \mid \beta_j \neq 0]$ be the active set, the set of indices such that β is non zero. And let $s_0 = \mid S_0 \mid$ be its cardinality, also called the sparsity index of the active set.
Recall from section 2 that each variable is estimated with a precision $\frac{\sigma^2}{n}$ If we would know s_0 the prediction error would be given by $\frac{E[(\hat{\beta}-\beta)' X' X (\hat{\beta}-\beta)]}{n} = \frac{\sigma^2}{n}.s_0$; with $s_0 <$ K by sparsity. But as this set is unknown having an exact value for the prediction error is impossible. The best we can do is derive an upper-bond that limits its value.
For simplicity sake, let us derive it first in a non stochastic framework $Y = X\beta^o$ the extension to the stochastic case which is our concern will be easier.

$$\beta^* = Argmin \frac{1}{2} \| X\beta - X\beta^o \|_2^2 + \lambda \| \beta \|_1$$

As β^* is a minimizer, one can write $\frac{1}{2} \| X\beta^* - X\beta^o \|^2 + \lambda \| \beta^* \|_1 \leq \lambda \| \beta^o \|_1$; as $X\beta^o - X\beta^o$ cancels in the first part of the right of the inequality.
As the L1 norm is a linear operator, I can decompose $\| \beta^* \|$ into $\| \beta^*_{s_0} \| + \| \beta^*_{s_0{}^c} \|$ and obtain :

$$\frac{1}{2} \| X\beta^* - X\beta^o \|^2 + \lambda \| \beta^*_{s_0{}^c} \|_1 \leq \lambda \| \beta^o \|_1 - \lambda \| \beta^*_{s_0} \|_1 \leq \lambda \| \beta^*_{s_0} - \beta^o \|_1 \tag{8}$$

Where the second inequality holds by triangular inequality.

Let us now introduce an assumption on the $X'X$ matrix of moment. For the theory to work we need a certain compatibility between the L1 and L2 norm.

Hp1 : Let L > 0 a constant, the compatibility constant is defined as follow :
$\phi^2_{X'X}(S_0, L) = min[\sqrt{s_0}.(X\beta_S - X\beta_{S^c}) \mid\mid \beta_{s_0} \parallel_1 \leq 1; \parallel \beta^*_{s_0^c} \parallel \leq L]$
$\phi^2_{X'X}(S_0, L) \neq 0$

When applied to our case, by (7), as $\parallel X\beta^* - X\beta^o \parallel^2$ is positive we know that the L1 norm outside the active set is bounded by the L1 norm inside. Hence according to the definition in **Hp1**, L=1, more over as all β^o outside the active set are zero, I can rewrite $\parallel \beta^*_{S_0} - \beta^o \parallel_1 = \parallel \beta^*_{S_0} - \beta^o_{S_o} \parallel_1$
By definition of the compatibility constant

$$\parallel \beta^*_{S_0} - \beta^o_{S_o} \parallel_1 \leq \frac{\|X\beta^* - X\beta^o\|}{\sqrt{s_0}\|X\beta_{S_0} - X\beta_{S_0^c}\|} \cdot \sqrt{s_0}$$

Hence we eventually have :

$$\parallel X\beta^* - X\beta^o \parallel^2 + 2\lambda \parallel \beta^*_{s_0^c} \parallel_1 \leq 2\lambda \frac{\|X\beta^* - X\beta^o\|}{\phi^2_{X'X}(S_0,1)} \cdot \sqrt{s_0} \leq \frac{1}{2} \parallel X\beta^* - X\beta^o \parallel^2 + \frac{2\lambda^2\sqrt{s_0}}{\phi^2_{X'X}(S_0,1)^2}$$

Where the last inequality holds by Cauchy-Schwartz inequality.

From this we have that :

1. the prediction error $\parallel X(\beta^* - \beta^o) \parallel^2 \leq \frac{4\lambda^2 s_0}{\phi^2_{X'X}(S_0,1)^2}$
2. $\parallel \beta^*_{S_0} - \beta^o \parallel \leq \frac{2\lambda s_0}{\phi^2_{X'X}(S_0,1)^2}$ as it is also bounded by $\frac{\|X\beta^* - X\beta^o\|}{\sqrt{s_0}\|X\beta_{S_0} - X\beta_{S_0^c}\|} \cdot \sqrt{s_0}$

Now if we are in a stochastic world :

The LASSO problem becomes :

$$\beta^* = Argmin \frac{1}{2} \parallel Y - X\beta - Y - X\beta^o \parallel_2^2 + \lambda \parallel \beta \parallel_1$$

Note that here the L2 norm is at the square (which correspond to a RSS min under L1 constraint)
Consider that I want that the maximum correlation between the error term and the covariate to be less than a constant λ_0, formally

$$\parallel \epsilon' X \parallel_\infty \leq \lambda_0$$

I also want the tuning parameter to kill the noise, otherwise the OLS solution is also the LASSO solution. hence $\lambda \geq \lambda_0$
Take $L = \frac{\lambda + \lambda_0}{\lambda - \lambda_0} > 1$
Doing exactly the same steps than previously except that L is now greater than one, let us proceed :
As $\hat{\beta}$ is a minimizer, one can write

$$\frac{1}{2} \parallel X\hat{\beta} - X\beta^o \parallel^2 + \lambda \parallel \hat{\beta} \parallel_1 \leq \lambda \parallel \beta^o \parallel_1 + \epsilon' X(\hat{\beta} - \beta^0)\,;$$

By dual norm inequality : $\mid \epsilon' X\beta \mid \leq \parallel \epsilon' X \parallel_\infty . \parallel \beta \parallel_1$
Which is logic as the L1 norm of the correlation between the residual and the parameters can not exceed the **maximum** correlation between the residuals and the covariates times the L1 norm of the parameters. So I can bound the last part of the above equation.

$$\frac{1}{2} \parallel X\hat{\beta} - X\beta^o \parallel^2 + \lambda \parallel \hat{\beta^{}_{s_0^c}} \parallel_1 \leq \lambda \parallel \beta^o \parallel_1 - \lambda \parallel \hat{\beta^{}_{s_0}} \parallel_1 + \lambda_0 \parallel \hat{\beta} - \beta \parallel_1 \tag{9}$$

Again $\hat{\beta}$ contains also elements outside the active set and I can decompose it L1 norm and pass the elements outside the active set on the left hand of the inequality.

$$\frac{1}{2} \parallel X\hat{\beta} - X\beta^o \parallel^2 + (\lambda - \lambda_0) \parallel \hat{\beta_{s_0^c}} \parallel_1 \leq \lambda \parallel \beta^o \parallel_1 - \lambda \parallel \hat{\beta_{s_0}} \parallel_1 + \lambda_0 \parallel \hat{\beta_{S_0}} - \beta \parallel_1$$

By the same triangular inequality than previously :

$$\frac{1}{2} \parallel X\hat{\beta} - X\beta^o \parallel^2 + (\lambda - \lambda_0) \parallel \hat{\beta_{s_0^c}} \parallel_1 \leq \lambda \parallel \beta^o \parallel_1 - \lambda \parallel \hat{\beta_{s_0}} \parallel_1 + \lambda_0 \parallel \hat{\beta_{S_0}} - \beta \parallel_1 \leq (\lambda + \lambda_0) \parallel \hat{\beta_{S_0}} - \beta \parallel_1$$

Applying the compatibility constant definition with the constant L set as above in order to kill the noise, I have

$$\frac{1}{2} \parallel X\hat{\beta} - X\beta^o \parallel^2 + (\lambda - \lambda_0) \parallel \hat{\beta_{s_0^c}} \parallel_1 \leq (\lambda + \lambda_0) \frac{\parallel X\hat{\beta} - X\beta^o \parallel . \sqrt{s_0}}{\phi^2_{X'X}(S_0, L)}$$

Again by Cauchy-Schwartz

$$\parallel X\hat{\beta} - X\beta^o \parallel^2 + (\lambda - \lambda_0) \parallel \hat{\beta_{s_0^c}} \parallel_1 \leq \frac{1}{2} \parallel X\hat{\beta} - X\beta^o \parallel^2 + \frac{2(\lambda + \lambda_0)^2 . s_0}{\phi^2_{X'X}(S_0, L)^2}$$

Hence the error bound in the stochastic case is :

$$\parallel X\hat{\beta} - X\beta \parallel^2 \leq \frac{4(\lambda + \lambda_0)^2 . s_o}{\phi^2_{X'X}(S_0, L)^2}$$

And we want this bound to hold with great probability. Say ϵ is drawn out of a Gaussian distribution of mean 0 and variance 1 : $\epsilon \sim N(0; I)$

Consider the set $\tau = [\parallel \epsilon' X \parallel_\infty \leq \lambda_0]$ and $\lambda_0 \asymp \sqrt{\frac{2log(2k)+\delta}{n}} . \underbrace{max_{1 \leq j \leq K} \parallel X_j \parallel}_{=1 if Normalized}$

$P[\tau^c] = P[max_{1 \leq j \leq K} \mid \epsilon' X \mid > \lambda_0] \underbrace{\leq}_{\text{By Union Bond}} K.max_{1 \leq j \leq K} P[\frac{|\epsilon' X|}{\parallel X_j \parallel} > \frac{\lambda_0}{\parallel X_j \parallel}]$

$K.max_{1 \leq j \leq K} P[\frac{|\epsilon' X|}{\parallel X_j \parallel} > \frac{\lambda_0}{\parallel X_j \parallel}] \underbrace{\leq}_{\text{By inequality of concentration for Gaussian variables}} 2Kexp[\frac{2log(2K)+\delta}{n}]$ Hence the probability to respect the prediction error upper bound is

$$1 - 2K.exp[\frac{2log(2K) + \delta}{n}] \tag{10}$$

4.2 The variable selection

We now want to show that this framework allow for good variable selection properties :
let $\hat{S} = [j \mid \hat{\beta}_j \neq 0]$ be the estimated active set.
An optimal variable selection would be $\hat{S} \subset S_0$ and $S_0 \subset \hat{S}$
Like in the previous section we start by analysing the non stochastic case, and then jump to the random case.

let $\hat{\Sigma} = X'X/n$ be decomposed as : $\begin{bmatrix} \hat{\Sigma}_{11} & \hat{\Sigma}_{12} \\ \hat{\Sigma}_{21} & \hat{\Sigma}_{22} \end{bmatrix}$; with 1 corresponding to the active variable and 2, to the non active.
Let's introduce the notion of non-representability condition.
Intuitively, we do not want the variables which are outside the active set to be correlated with the variables inside. Hence regressing $X_2 \notin S_0$ on $X_1 \in S_0$ The maximum absolute value of

parameter for such a regression must not exceed 1. The ideal case being the latter to be 0 (Very unlikely in high dimension).

$$\text{for any } \| \tau \| \leq 1; \| \hat{\Sigma}_{21}\hat{\Sigma}_{11}^{-1}.\tau \|_{\infty} \leq \theta \tag{11}$$

For $0 \leq \theta < 1$, β^*_{LASSO} leads to no false positive : $S^* \subset S_0$
By the Karush Kuhn Tucker for the LASSO problem : $\hat{\Sigma}(\beta^* - \beta^o) = -\lambda\varsigma^*$
With ς^* the constraint sub-gradient. Indeed as the absolute value is not differentiable in 0. The notion of derivative does not apply. Hence we apply the sub-gradient notion.

$$\begin{cases} \varsigma^* = 1 & if \beta^* > 0 \\ \varsigma^* = -1 & if \beta^* < 0 \\ \varsigma^* \in (-1;1) & if \beta^* = 0 \end{cases}$$

Decomposing by active and inactive variables we have

$$\begin{cases} \hat{\Sigma}_{11}(\beta_1^* - \beta^o) + \hat{\Sigma}_{12}\beta_2^* = -\lambda\varsigma_1^* \\ \hat{\Sigma}_{21}(\beta_1^* - \beta^o) + \hat{\Sigma}_{22}\beta_2^* = -\lambda\varsigma_2^* \end{cases} \tag{12}$$

Solving the system of linear equations (11), multiplying the first equation by $\hat{\Sigma}_{21}.\hat{\Sigma}_{11}^{-1}$, solving for $(\beta_1^* - \beta^o)$ and replacing into the second equation, we get :

$$\begin{cases} \hat{\Sigma}_{21}(\beta_1^* - \beta^o) + \hat{\Sigma}_{21}\hat{\Sigma}_{11}^{-1}\hat{\Sigma}_{12}\beta_2^* = -\lambda\hat{\Sigma}_{21}\hat{\Sigma}_{11}^{-1}\varsigma_1^* \\ \hat{\Sigma}_{22} - \hat{\Sigma}_{21}\hat{\Sigma}_{11}^{-1}\hat{\Sigma}_{12}\beta_2^* = -\lambda\varsigma_2^* + \lambda\hat{\Sigma}_{21}\hat{\Sigma}_{11}^{-1}\varsigma_1^* \end{cases}$$

Multiplying both sides of the second equation by β_2^* and keeping in mind that $\beta^*.\varsigma^* = \| \beta^* \|_1$
We solve the system :

$$\underbrace{\beta_2^{*'}(\hat{\Sigma}_{22} - \hat{\Sigma}_{21}\hat{\Sigma}_{11}^{-1}\hat{\Sigma}_{12})\beta_2^*}_{\text{Quadractic form, hence PSD}} = -\lambda \| \beta_2^* \|_1 + \lambda\beta_2^{*'}\hat{\Sigma}_{21}\hat{\Sigma}_{11}^{-1}\varsigma_1^* \underbrace{\leq}_{By(11)} -\lambda \| \beta_2^* + \lambda \| \beta_2^* \| .\theta \leq$$
$$-\lambda(1-\theta) \| \beta_2^* \|_1$$

By hypothesis $\theta < 1$ hence $(1-\theta) > 0$ and $-\lambda(1-\theta)$. $\underbrace{\| \beta_2^* \|_1}_{if \beta \text{is non zero}} < 0$
This is clearly in contradiction with the positive semi definite form of the equation's left hand side. Hence β_2^* of dimension K-s_0 Must be 0.

In the stochastic case the proof is similar except that $\theta < \frac{1}{L}$; with $\mathrm{L} = \frac{\lambda+\lambda_0}{\lambda-\lambda_0}$
The proof is done thanks to the KKT conditions for the problem (3) given by :

$$2\hat{\Sigma}(\hat{\beta} - \beta^o) = -(\lambda - \lambda_0)\hat{\varsigma}$$

The same decomposition as in the noiseless case applies and we solve the system of equation exactly the same way to obtain :

$$\underbrace{\hat{\beta_2}'(\hat{\Sigma}_{22} - \hat{\Sigma}_{21}\hat{\Sigma}_{11}^{-1}\hat{\Sigma}_{12})\hat{\beta_2}}_{\text{Quadractic form, hence PSD}} = -(\lambda - \lambda_0) \| \hat{\beta}_2 \|_1 + (\lambda - \lambda_0)\hat{\beta}_2\hat{\Sigma}_{21}\hat{\Sigma}_{11}^{-1}\varsigma_1^* \leq -(\lambda - \lambda_0)(1-\theta) \| \hat{\beta}_2 \|_1$$

Recall that to kill the noise we must have $\lambda > \lambda_0$ hence L>1. Again if θ <1 and $\| \hat{\beta}_2 \|_1$ would be a violation of the KKT. As $\theta = \frac{1}{L} < 1$ $\| \hat{\beta}_2 \|_1$ can not be non zero.

This irrepresentable condition is strong and nothing says that in reality it is fulfilled. for a weaker conditions one can use adaptative LASSO proposed by Zhou (2006) or refer to the (non exhaustive) discussions of Guyon and Elisseeff (2003), Meinshausen and Bühlmann (2006), Zhao and Yu (2006).

For no false negative, $S_0 \in \hat{S}$ the condition is called the β_{min} The idea is that a true active variable must not have a parameter slope associated too small to be confounded with the noise and be cut off the active set .

$$\| \hat{\beta} - \beta^o \| \asymp \lambda_0$$

5 Application to spatial econometrics

Let us consider for simplicity's sake a purely autoregressive model of the form.

$$Y = WY + \epsilon \tag{13}$$

Where Y is N.1 response vector
W is N.N spatial weight matrix connecting each observation to the others.
Hence each observation is a weighted average of the other observations.

$$\begin{bmatrix} y_1 = w_{11}y_1 + w_{12}y_2 & + & ... & + & w_{1N}y_N & + & \epsilon_1 \\ y_2 = w_{21}y_1 + w_{22}y_2 & + & ... & + & w_{2N}y_N & + & \epsilon_2 \\ ..... & ... & ... & ... & ... & + & ... \\ ..... & ... & ... & ... & ... & ... & ... \\ ..... & ... & ... & ... & ... & ... & ... \\ y_N = w_{N1}y_1 + w_{N2}y_2 & + & ... & + & w_{NN}y_N & + & \epsilon_N \end{bmatrix}$$

But we consider that an observation can not be its own neighbour. Thus, w_{ii} for i =1 ... N ; must be 0.
We want the system to be written as :

$$\begin{bmatrix} y_1 = w_{12}y_2 & + & ... & + & w_{1N}y_N & + & \epsilon_1 \\ y_2 = w_{21}y_1 & + & w_{23}y_3 & ... & w_{2N}y_N & + & \epsilon_2 \\ ..... & ... & ... & ... & ... & + & ... \\ ..... & ... & ... & ... & ... & ... & ... \\ ..... & ... & ... & ... & ... & ... & ... \\ y_N = w_{N1}y_1 & + & ... & + & w_{NN-1}y_{N-1} & + & \epsilon_N \end{bmatrix}$$

So the space of the parameters to estimate is a N.(N-1) matrix. And this is not a convenient form to estimate them. We would like find back a more conventional form to "store" this parameter, hence a way to rewrite the above model with the w_{ij} stacked into a column vector.
To do so, we are going to "panelize" this cross sectional setting as follow :

1. Delete W's diagonal $\equiv W^*$
2. Get W^{*}'
3. Stack W^{*}' as a vector by column which is for column 1 to N-1, take column i+1 and stack it under column i.

4. do $Y \otimes I_N \equiv Z$
5. Transpose Z, Z' being of dimension N by N^2
6. for i= 1 to N^2 delete column i with step = N+1. Doing so the first observation is dropped from the first line, the rest of the first column being constituted of 0. The second observation is deleted from the second line, the rest of the second column being constituted of 0 as well, and so on until the end of the matrix. Call X this N by $N^2 - N$ matrix.

We end up with a model :

$$Y = XW + \epsilon$$

With X being a N by $N^2 - N$ matrix.
W being a $N^2 - N$ by one vector. One can easily verify that this set up brings us back to the situation that we seek and is just a "trick" to rewrite the model in a convenient form.

We are hence in the case where we want to estimate a system of linear equations where the system has more parameters than observations.
Hence we fall in the case described by Sardy et al.(2000), Osborne et al. (2000).

An other solution could be to consider the sparse group lasso proposed by Simon Hastie Friedman Tibshirani (2010) but for matter of time I can not consider it until tuesday

With X being of the form :

$$\begin{bmatrix} Y_2 & ... & Y_N & 0 & ... & 0 & 0 & ... & ... & 0 \\ 0 & ... & 0 & Y_1 & Y_3 & ... & Y_N & 0 & ... & 0 \\ ... & & & & & & & & & ... \\ 0 & ... & 0 & ... & 0 & ... & 0 & Y_1 & ... & Y_{N-1} \end{bmatrix}$$

Following the block coordinate relaxation algorithm we take blocks of size N-1 and will have therefore N blocks of this size.

1. Take block 1 : $\begin{bmatrix} Y_2 & ... & Y_N \\ 0 & ... & 0 \\ ... & ... & ... \\ 0 & ... & 0 \end{bmatrix}$
2. Compute a residual vector $\begin{bmatrix} \hat{e_1^{(m)}} \\ \hat{e_2^{(m)}} \\ . \\ . \\ . \\ \hat{e_N^{(m)}} \end{bmatrix} = \begin{bmatrix} Y_1 - 0 \\ Y_2 - \sum_{j \neq i=1}^{N} w_{2j} Y_j \\ . \\ . \\ . \\ Y_N - \sum_{j \neq i=1}^{N} w_{Nj} Y_j \end{bmatrix}$

3. Take this vector of residuals and search the vector of predictors $\begin{bmatrix} \hat{w_{12}} \\ . \\ . \\ . \\ \hat{w_{1N}} \end{bmatrix}$ such that :

$$\hat{W}_{i=1} = Argmin \sum_{i=1}^{N} (\hat{e}_1^{(m=1)} - X_i^{(m=1)} W_1)^2 + \lambda \parallel W_1 \parallel_1 \tag{14}$$

4. Solve (14) applying Osborne Prescott and Turlach (2000)as in section (3.2)
5. Take m=2, the second block and repeat 2. , 3. and 4. until the last block
6. repeat 1. , 2. , 3. , 4. , 5. until a convergence criterion is found

5.1 Empirical example on suicide data on 22 French Region

We took our data out of the "Sentinelle" network, giving public access to datasets from the National Institute for Health and Medical Research. We took the raw quantity of suicides for each of the 22 French regions.
Applying the above procedure we obtain the following matrix :

0,00	0,08	0,03	0,04	0,06	0,05	0,02	0,04	0,02	0,17	0,01
0,08	0,00	0,05	0,06	0,08	0,07	0,03	0,06	0,03	0,25	0,01
0,04	0,03	0,00	0,02	0,03	0,03	0,01	0,02	0,01	0,10	0,00
0,02	0,04	0,05	0,00	0,04	0,03	0,01	0,03	0,02	0,12	0,01
0,04	0,06	0,08	0,03	0,00	0,05	0,02	0,04	0,02	0,17	0,01
0,05	0,05	0,06	0,03	0,03	0,00	0,01	0,03	0,02	0,14	0,01
0,01	0,02	0,02	0,01	0,01	0,02	0,00	0,01	0,01	0,05	0,00
0,01	0,04	0,06	0,02	0,03	0,04	0,03	0,00	0,02	0,12	0,01
0,02	0,02	0,03	0,01	0,02	0,02	0,02	0,01	0,00	0,07	0,00
0,13	0,32	0,43	0,18	0,22	0,32	0,25	0,10	0,24	0,00	0,04
0,02	0,01	0,01	0,00	0,00	0,01	0,01	0,00	0,01	0,00	0,00
0,01	0,06	0,08	0,03	0,04	0,06	0,05	0,02	0,04	0,02	0,18
0,02	0,02	0,03	0,01	0,02	0,02	0,02	0,01	0,02	0,01	0,06
0,01	0,02	0,03	0,01	0,01	0,02	0,02	0,01	0,01	0,01	0,06
0,01	0,03	0,04	0,02	0,02	0,03	0,02	0,01	0,02	0,01	0,09
0,01	0,03	0,04	0,01	0,02	0,03	0,02	0,01	0,02	0,01	0,08
0,02	0,04	0,06	0,02	0,03	0,04	0,03	0,01	0,03	0,02	0,12
0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00
0,00	0,02	0,03	0,01	0,01	0,02	0,02	0,01	0,01	0,01	0,06
0,00	0,01	0,01	0,01	0,01	0,01	0,01	0,00	0,01	0,00	0,03
0,01	0,03	0,04	0,02	0,02	0,03	0,02	0,01	0,02	0,01	0,09
0,02	0,03	0,05	0,02	0,02	0,03	0,03	0,01	0,03	0,01	0,10

0,06	0,02	0,02	0,03	0,03	0,04	0,00	0,02	0,01	0,03	0,04
0,08	0,03	0,03	0,04	0,04	0,06	0,00	0,03	0,02	0,04	0,05
0,03	0,01	0,01	0,02	0,02	0,02	0,00	0,01	0,01	0,02	0,02
0,04	0,02	0,01	0,02	0,02	0,03	0,00	0,01	0,01	0,02	0,02
0,06	0,02	0,02	0,03	0,03	0,04	0,00	0,02	0,01	0,03	0,04
0,05	0,02	0,02	0,02	0,02	0,03	0,00	0,02	0,01	0,02	0,03
0,02	0,01	0,01	0,01	0,01	0,01	0,00	0,01	0,00	0,01	0,01
0,04	0,02	0,02	0,02	0,02	0,03	0,00	0,02	0,01	0,02	0,03
0,02	0,01	0,01	0,01	0,01	0,02	0,00	0,01	0,00	0,01	0,01
0,33	0,12	0,12	0,17	0,15	0,23	0,00	0,12	0,06	0,17	0,19
0,01	0,00	0,00	0,00	0,00	0,01	0,00	0,00	0,00	0,00	0,00
0,00	0,02	0,02	0,03	0,03	0,04	0,00	0,02	0,01	0,03	0,04
0,00	0,00	0,01	0,01	0,01	0,02	0,00	0,01	0,00	0,01	0,01
0,00	0,02	0,00	0,01	0,01	0,01	0,00	0,01	0,00	0,01	0,01
0,00	0,03	0,01	0,00	0,01	0,02	0,00	0,01	0,01	0,02	0,02
0,00	0,03	0,01	0,01	0,00	0,02	0,00	0,01	0,00	0,01	0,02
0,01	0,04	0,02	0,01	0,02	0,00	0,00	0,01	0,01	0,02	0,02
0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00
0,00	0,02	0,01	0,01	0,01	0,01	0,01	0,00	0,00	0,01	0,01
0,00	0,01	0,00	0,00	0,01	0,00	0,01	0,00	0,00	0,01	0,01
0,00	0,03	0,01	0,01	0,02	0,01	0,02	0,00	0,01	0,00	0,02
0,00	0,03	0,01	0,01	0,02	0,02	0,02	0,00	0,01	0,01	0,00

6 references

— Ahrens, A., & Bhattacharjee, A. (2015). Two-step lasso estimation of the spatial weights matrix. Econometrics, 3(1), 128-155.
— Anselin, L. (2013). Spatial econometrics : methods and models (Vol. 4). Springer Science & Business Media.
— Bhattacharjee, A., & Jensen-Butler, C. (2011). Estimation of the spatial weights matrix under structural constraints.
— Bühlmann, P., & Van De Geer, S. (2011). Statistics for high-dimensional data : methods, theory and applications. Springer Science & Business Media.
— Cliff, A. D., & Ord, J. K. (1981). Spatial processes : models & applications (Vol. 44). London : Pion.
— Efron, B., Hastie, T., Johnstone, I., & Tibshirani, R. (2004). Least angle regression. The Annals of statistics, 32(2), 407-499.
— Friedman, J., Hastie, T., & Tibshirani, R. (2010). A note on the group lasso and a sparse group lasso. arXiv preprint arXiv :1001.0736.
— Geary, R. C. (1954). The contiguity ratio and statistical mapping. The incorporated statistician, 5(3), 115-146.
— Giraud, C. (2014). Introduction to high-dimensional statistics (Vol. 138). CRC Press.
— Griva, I., Nash, S. G., & Sofer, A. (2009). Linear and nonlinear optimization. Siam.
— Guyon, I., & Elisseeff, A. (2003). An introduction to variable and feature selection. The Journal of Machine Learning Research, 3, 1157-1182.
— Knight, K., & Fu, W. (2000). Asymptotics for lasso-type estimators. Annals of statistics, 1356-1378.
— Lam, C., & Souza, P. C. (2015). Detection and estimation of block structure in spatial weight matrix. Econometric Reviews, 1-30.
— Lam, C., & Souza, P. C. (2013). Regularization for spatial panel time series using the adaptive LASSO. Journal of Regional Science.[PubMed].
— Mark Schmidt. Graphical Model Structure Learning with L1-Regularization. Ph.D. Thesis, University of British Columbia, 2010.
— Meinshausen, N., & Bühlmann, P. (2006). High-dimensional graphs and variable selection with the lasso. The annals of statistics, 1436-1462.
— Moran, P. A. (1948). The interpretation of statistical maps. Journal of the Royal Statistical Society. Series B (Methodological), 10(2), 243-251.
— Osborne, M. R., Presnell, B., & Turlach, B. A. (2000). A new approach to variable selection in least squares problems. IMA journal of numerical analysis, 20(3), 389-403.
— Osborne, M. R., Presnell, B., & Turlach, B. A. (2000). On the lasso and its dual. Journal of Computational and Graphical statistics, 9(2), 319-337.
— Qin, Z., Scheinberg, K., & Goldfarb, D. (2013). Efficient block-coordinate descent algorithms for the group lasso. Mathematical Programming Computation, 5(2), 143-169.
— Qu, X., & Lee, L. F. (2015). Estimating a spatial autoregressive model with an endogenous spatial weight matrix. Journal of Econometrics, 184(2), 209-232.
— Sardy, S., Bruce, A. G., & Tseng, P. (2000). Block coordinate relaxation methods for nonparametric wavelet denoising. Journal of computational and graphical statistics, 9(2), 361-379.
— Simon, N., Friedman, J., Hastie, T., & Tibshirani, R. (2013). A sparse-group lasso. Journal of Computational and Graphical Statistics, 22(2), 231-245.
— Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society. Series B (Methodological), 267-288.

— Van Der Vaart, A. W., & Wellner, J. A. (1996). Weak Convergence (pp. 16-28). Springer New York.
— Van De Geer, S. A., & Bühlmann, P. (2009). On the conditions used to prove oracle results for the Lasso. Electronic Journal of Statistics, 3, 1360-1392.
— Yuan, M., & Lin, Y. (2006). Model selection and estimation in regression with grouped variables. Journal of the Royal Statistical Society : Series B (Statistical Methodology), 68(1), 49-67.
— Zhao, P., & Yu, B. (2006). On model selection consistency of Lasso. The Journal of Machine Learning Research, 7, 2541-2563.
— Zou, H. (2006). The adaptive lasso and its oracle properties. Journal of the American statistical association, 101(476), 1418-1429.