Facing spatial massive data in science and society: Variable selection for spatial models

Romina Gonella, Mathias Bourel, Liliane Bel

Facing spatial massive data in science and society : variable selection for spatial models

Romina Gonella[a], Mathias Bourel[a], Liliane Bel[b,*]

[a] *IMERL, Facultad Ingeniería, Universidad de la República, Montevideo*
[b] *UMR MIA-Paris, AgroParisTech, INRAE, Université Paris-Saclay*

Abstract

This work focuses on variable selection for spatial regression models, with locations on irregular lattices and errors according to Conditional or Simultaneous Auto-Regressive (CAR or SAR) models. The strategy is to whiten the residuals by estimating their spatial covariance matrix and then proceed by performing the standard L1-penalized regression LASSO for independent data on the transformed model. A result is stated that proves the sign consistency for general dependent errors provided that the transformed design matrix fulfills standard assumptions for the LASSO procedure and that the estimate of the residual covariance matrix is consistent. Then sufficient conditions on the weight matrix of the SAR or CAR model are given that ensure those conditions hold. A simulation study is driven that shows this method gives good result in terms of variables selection, while some underestimation of the coefficients is noted. It is compared to a strategy that estimates both the regression and the covariance parameters in a LARS procedure. Coefficient are better estimated with the Least Angle Regression (LARS) procedure but it gives in some cases much more false positive in the variable selection. The application is on the regression of income data in rural area of Uruguay on a set of covariates describing socio-economic characteristics of the households.

Keywords: LASSO, Variable Selection, Spatial Statistics

*Corresponding author
Email address: liliane.bel@agroparistech.fr (Liliane Bel)

1. Introduction

Thanks to new technologies, data acquisition is nowadays much easier and the challenge is now to process this mass of data. Variable selection is one of the aspects of the question, the number of covariates is often so large that it must be reduced in order to avoid numerical problems and gaining interpretability. The LASSO (Least Absolute Shrinkage and Selection Operator, [1]) method which cancels the small coefficients of the regression thanks to a ℓ_1 penalty, has become very popular and has experienced many theoretical and practical developments : [2], [3] state the basics of the method, [4] establish the asymptotics, [5] the oracle properties, [6] a consistency result, [7] extend to group-LASSO amongst many others.

In many domains, environment, climate, econometric, agronomy, data are georeferenced and their modeling includes an error term with a spatial dependence. It is well known that this dependence must be taken into account to avoid making errors, whether in the estimation of parameters or their significance. In these domains, the increase in the size of the data is also a reality, even if it does not reach the magnitudes that can be observed in biology for example.

Variable selection by LASSO in spatial models has been studied by several authors : [8] develop an additive model, [9], [10], [11] consider the geostatistical framework, while [12], [13] and [14] consider lattice models. Depending on the case different contexts either on the form of the spatial dependence, or on the approach considered, by maximum likelihood or by minimizing a least squares criterion are handled. We show in this work a sign consistency result, which guarantees to recover asymptotically the true support of the regression parameters, for dependent data. We give necessary conditions on the weight matrix when the spatial model incorporates CAR or SAR errors for the conditions of the theorem to be verified. These two results are presented in section 2 along with the procedure for implementing the selection. In section 3 we compare this method with the method in [13] which regularizes the full likelihood on the same models, on simulated data in several situations. Knowledge of household

income is essential for setting public policy. In developping countries such as Uruguay it may be difficult to collect direct data on income especially for some specific population such as in rural areas and it may be convenient estimating it from covariates easier to collect describing the socio-economic characteristics of the household as well as the comfort level and the satisfaction of their basic needs. In section 4 we compare the two approaches on a survey data in order to explain the per capita income of households from Uruguayan rural areas by covariates A discussion and conclusions are given in section 5.

2. Variable selection for dependent errors, the spatial case

Let us consider a linear model

$$Y = \mathbf{X}\beta + \varepsilon \tag{1}$$

where Y is a $n \times 1$ vector of outputs, X is the $n \times p$ design matrix of covariates, with p that can be very large, β is a $p \times 1$ vector of unknown parameters and ε is a $n \times 1$ random vector with $E(\varepsilon) = \mathbb{0}$ and $Var(\varepsilon) = \Sigma$.

Without loss of generality, it can be assumed that $\mathbf{X}$ and Y are standardized with 0 mean.

One is willing that the fitted model fulfills two properties : it has good predictive ability, it is interpretable that is if the number of covariate is large only a few of the associated coefficients are significant. The LASSO (Least Absolute Shrinkage and Selection Operator, [1]) procedure minimizes the following penalized criterion

$$\mathscr{L}(\beta) = \|Y - \mathbf{X}\beta\|_2^2 + \lambda\|\beta\|_1 \text{ for some } \lambda > 0 \tag{2}$$

The penality $\lambda\|\beta\|_1$ as being a L_γ norm with $\gamma \leq 1$ has the nice feature to cancel parameters when λ is growing, and selects by this means variables. When $\gamma = 1$, the minimization problem is convex, hence it is computationally tractable. Regarding the predictive performance of LASSO, it has been shown that it is equivalent to other penalized methods such as ridge or bridge regression ([2])

and these two features good performance prediction and providing sparse models make the LASSO procedure very popular and efficient for fitting models especially for large number of covariates, even larger than the observation number.

We assume that the true support of the vector β, denoted $A^* = \{j : \beta_j \neq 0,\ 1 \leq j \leq p\}$ is size $|A^*| = p^*$, $0 < p^* \leq p$, and we denote A^{*c} the A^* complementary, $|A^{*c}| = p - p^*$.

The aim of variable selection is to determine the set A^* and estimate the corresponding parameters $\{\beta_j, j \in A^*\}$.

Theoretical results for the LASSO procedure are mainly stated for models with independent errors ($\mathbf{\Sigma} = \sigma^2 Id$), a way to deal with correlated errors is to transform the model in order to remove the correlation :

$$\underbrace{\mathbf{\Sigma}^{-1/2}Y}_{\widetilde{\boldsymbol{Y}}} = \underbrace{\mathbf{\Sigma}^{-1/2}\boldsymbol{X}}_{\widetilde{\boldsymbol{X}}}\beta + \underbrace{\mathbf{\Sigma}^{-1/2}\boldsymbol{\varepsilon}}_{\widetilde{\boldsymbol{\varepsilon}}} \tag{3}$$

with $\widetilde{\boldsymbol{\varepsilon}} \sim N(\mathbf{0}, \boldsymbol{Id})$ and the procedure may be driven on the transformed model. When $\lambda = 0$ the estimator of β is then the GLS estimator $\widehat{\beta} = (\widetilde{\boldsymbol{X}}'\widetilde{\boldsymbol{X}})^{-1}\widetilde{\boldsymbol{X}}'Y = (\boldsymbol{X}'\mathbf{\Sigma}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}'\mathbf{\Sigma}^{-1}Y$. When $\mathbf{\Sigma}$ is not known but may be estimated by $\widehat{\mathbf{\Sigma}}$, a plug-in approach leads to consider the model

$$\underbrace{\widehat{\mathbf{\Sigma}}^{-1/2}Y}_{\widetilde{\boldsymbol{Y}}} = \underbrace{\widehat{\mathbf{\Sigma}}^{-1/2}\boldsymbol{X}}_{\widetilde{\boldsymbol{X}}}\beta + \underbrace{\widehat{\mathbf{\Sigma}}^{-1/2}\boldsymbol{\varepsilon}}_{\widetilde{\boldsymbol{\varepsilon}}} \tag{4}$$

In the following we state conditions under which the estimator of β is sign consistent as defined in [6], that is:

$$\lim_{n\to\infty} P(sign(\hat{\beta}) = sign(\beta)) = 1$$

Sign consistency insures the support of the β estimate is asymptotically the true support.

Without loss of generality , model (1) can written:

$$Y = [\mathbf{X}_{A^*}\ \mathbf{X}_{A^{*c}}] \begin{bmatrix} \beta_{A^*} \\ \beta_{A^{*c}} \end{bmatrix} + \varepsilon \tag{5}$$

where $\beta_{A^*} = \{\beta_j, j \in A^*\}$, $\mathbf{X}_{A^*} = ((X_{.j}))_{j \in A^*}$, where $X_{.j}$ is the jth column of $\mathbf{X}$.

The following theorem is an adaptation to model (1) with correlated errors of the results stated in [6] for models with independent errors.

Theorem 1 (Condition for strong sign consistency when errors are dependent)**.** *Considering the model (4), and $\hat{\beta}$ the LASSO estimator for this model, if there exist positive constants M_1, M_2, M_3, M_4, M_5, M_6 and M_7 and c_1, c_2 with $0 \leq c_1 \leq c_2 \leq 1/2$ such that the following conditions are:*

1. $\frac{1}{n}\mathbf{X}'_{.j}\mathbf{\Sigma}^{-1}\mathbf{X}_{.j} \leq M_1 \ \forall \ j$
2. $\rho_{min}(\frac{1}{n}\mathbf{X}'_{A^*}\mathbf{\Sigma}^{-1}\mathbf{X}_{A^*}) \geq M_2$
3. $p^* = O(n^{c_1/2})$
4. $n^{(1-c_2)/2} \min_{1 \leq j \leq p^*} (|\beta_j|) \geq M_3$
5. *it exists a constant positive vector δ such that* $|\frac{1}{n}\mathbf{X}'_{A^{*c}}\mathbf{\Sigma}^{-1}\mathbf{X}_{A^*}(\frac{1}{n}\mathbf{X}'_{A^*}\mathbf{\Sigma}^{-1}\mathbf{X}_{A^*})^{-1}sign(\beta_{A^*})| \leq \mathbb{1} - \delta$ *where $\mathbb{1}$ is a $p - p^*$ vector of 1, and the inequality holds element-wise.*
6. $\frac{\lambda}{\sqrt{n}} \underset{n \to \infty}{\to} \infty$ *and* $\frac{\lambda}{\sqrt{n}} = o(n^{(c_2 - c_1)/2})$
7. $\left|\left|\left|(\boldsymbol{X}'\boldsymbol{X})/n\right|\right|\right|_\infty \leq M_4$
8. $\rho_{max}(\mathbf{\Sigma}^{-1}) \leq M_6$
9. $\rho_{min}(\mathbf{\Sigma}^{-1}) \geq M_7$
10. $\left|\left|\left|\mathbf{\Sigma}^{-1} - \hat{\mathbf{\Sigma}}^{-1}\right|\right|\right|_\infty = O_P(\frac{1}{\sqrt{n}})$ *when* $n \to \infty$

where $\rho(\mathbf{X})$ is the vector of eigenvalues of $\mathbf{X}$, $\rho_{min}(\mathbf{X})$ and $\rho_{max}(\mathbf{X})$ are respectively the minimum and maximum eigenvalues, $\left|\left|\left|\boldsymbol{X}\right|\right|\right|_\infty = \max_{1 \leq i \leq n} \left(\sum_{j=1}^{n} |X_{ij}| \right)$. Then the strong sign consistency property is achieved for the LASSO estimator $\hat{\beta}$, that is, $\lim_{n \to \infty} P(sign(\hat{\beta}) = sign(\beta)) = 1$.

Condition 5 is the so-called [6] strong irrepresentable condition for matrix $\widetilde{\mathbf{X}}$ in model (3).

Conditions 3, 4 and 6 are present in the result for the independent errors case, while conditions 1, 2 and 5 are analog but for $\widetilde{\mathbf{X}}$ of model (3) instead of $\mathbf{X}$.

Conditions 8 and 9 indicate that eigenvalues of $\mathbf{\Sigma}$ are positive and bounded and condition 10 states that $\widehat{\mathbf{\Sigma}}$ is a good estimator for $\mathbf{\Sigma}$.

The proof of this Theorem is highly related to that of the one presented in [15] established for the multivariate responses case, when the error vector for one observation are dependent, but the errors for different observations are independent. The proof is available in the supplementary material.

2.1. Spatial case

We focus now on spatial models and investigate conditions on the error covariance matrix in order to fulfill the conditions for strong sign-consistency in Theorem 1 . Specifically we deal with models on lattice, focusing on autoregressive CAR or SAR models, with weights matrix W. More precisely the CAR (Conditional Auto Regressive) model is written

$$E(Z_i|Z_j : j \neq i) = \sum_{j=1}^{n} c_{ij} Z_j \qquad Var(Z_i|Z_j : j \neq i) = \sigma^2$$

The Z covariance matrix is

$$\mathbf{\Sigma}_{\text{CAR}} = (\mathbf{Id} - \mathbf{C})^{-1}\mathbf{V}$$

$\mathbf{C} = \theta W$, $|\theta| < 1$, $\mathbf{V} = diag(\sigma^2)$.

The SAR (Simultaneaous Auto Regressive model) model is written

$$Z = \mathbf{C}Z + \nu$$

where ν is a iid random vector with 0 mean and variance σ^2.

The Z covariance matrix is

$$\mathbf{\Sigma}_{\text{SAR}} = (\mathbf{Id} - \mathbf{C})^{-1}\mathbf{V}(\mathbf{Id} - \mathbf{C}')^{-1}$$

with same as for the CAR model.

We seek for conditions on W that guarantee conditions 8, 9 and 10 of Theorem 1 .

To bound $\rho_{max}(\mathbf{\Sigma}^{-1})$ using inequalities results in ([16]) for positive definite matrices

$$\rho_{max}(\mathbf{\Sigma}^{-1}) \leq \gamma(\mathbf{\Sigma}^{-1}) \leq \left|\left|\left|\mathbf{\Sigma}^{-1}\right|\right|\right|_{\infty} \tag{6}$$

where $\gamma(\mathbf{X})$ is the spectral radius of $\mathbf{X}$ and decomposing $\mathbf{\Sigma}^{-1}$ in a product of matrices and bounding each term we obtain for both $\mathbf{\Sigma}_{\text{CAR}}$ and $\mathbf{\Sigma}_{\text{SAR}}$

$$\left|\left|\left|\mathbf{\Sigma}^{-1}\right|\right|\right|_\infty \leq \frac{1}{\sigma^2}\left(1+|\theta|\max_{1\leq i\leq n}\left(\sum_{j=1}^{n}|w_{ij}|\right)\right)^2$$

and the condition 8 holds if $\max\limits_{1\leq i\leq n}\left(\sum\limits_{j=1}^{n}|w_{ij}|\right)$ does no depend on n.
To bound $\rho_{min}(\mathbf{\Sigma}^{-1})$ using the same arguments we obtain for both $\mathbf{\Sigma}_{\text{CAR}}$ and $\mathbf{\Sigma}_{\text{SAR}}$

$$\rho_{min}(\mathbf{\Sigma}^{-1}) \geq \frac{\left(1-|\theta|\max\limits_{1\leq i\leq n}\left(\sum\limits_{j=1}^{n}|w_{ij}|\right)\right)^2}{\sigma^2}$$

and condition 9 holds if $\max\limits_{1\leq i\leq n}\left(\sum\limits_{j=1}^{n}|w_{ij}|\right)$ does not depend on n and $|\theta|\max\limits_{1\leq i\leq n}\left(\sum\limits_{j=1}^{n}|w_{ij}|\right) < 1$. Let $\widehat{C}=\widehat{\theta}W$, and $\widehat{V}=diag(\widehat{\sigma}^2)$ for the CAR model we have

$$\left|\left|\left|\mathbf{\Sigma}_{\text{CAR}}^{-1}-\hat{\mathbf{\Sigma}}_{\text{CAR}}^{-1}\right|\right|\right|_\infty = \left|\frac{1}{\sigma^2}-\frac{1}{\hat{\sigma}^2}\right| + \left|\frac{\hat{\theta}}{\hat{\sigma}^2}-\frac{\theta}{\sigma^2}\right|\max_{1\leq i\leq n}\left(\sum_{j=1}^{n}|w_{ij}|\right)$$

If $\sqrt{n}(\hat{\theta}-\theta)=O_P(1)$ and $\sqrt{n}\left(\frac{1}{\hat{\sigma}^2}-\frac{1}{\sigma^2}\right)=O_P(1)$, and $\max\limits_{1\leq i\leq n}\left(\sum\limits_{j=1}^{n}|w_{ij}|\right)$ is bounded, condition 10 holds for the case CAR.

80

For the SAR model

$$\begin{aligned}\left|\left|\left|\mathbf{\Sigma}_{\text{SAR}}^{-1}-\hat{\mathbf{\Sigma}}_{\text{SAR}}^{-1}\right|\right|\right|_\infty &= \max_{1\leq i\leq n}\left(\left|\sum_{k=1}^{n}w_{ik}^2\left(\frac{\theta^2}{\sigma^2}-\frac{\hat{\theta}^2}{\hat{\sigma}^2}\right)+\frac{1}{\sigma^2}-\frac{1}{\hat{\sigma}^2}\right|\right.\\ &\leq \left|\frac{1}{\sigma^2}-\frac{1}{\hat{\sigma}^2}\right| + \left|\frac{\theta^2}{\sigma^2}-\frac{\hat{\theta}^2}{\hat{\sigma}^2}\right|\max_{1\leq i\leq n}\left(\sum_{\substack{j=1\\ j\neq i}}^{n}\sum_{k=1}^{n}|w_{ik}w_{kj}|\right)+\\ &\quad \left|\frac{2\hat{\theta}}{\hat{\sigma}^2}-\frac{2\theta}{\sigma^2}\right|\max_{1\leq i\leq n}\left(\sum_{j=1}^{n}|w_{ij}|\right)\end{aligned}$$

If $\sqrt{n}(\hat{\theta}-\theta)=O_P(1)$, $\sqrt{n}\left(\frac{1}{\hat{\sigma}^2}-\frac{1}{\sigma^2}\right)=O_P(1)$, and both $\max\limits_{1\leq i\leq n}\left(\sum\limits_{j=1}^{n}|w_{ij}|\right)$,

$\max_{1<\leq i\leq n}\left(\sum_{\substack{j=1\\ j\neq i}}^{n}\sum_{k=1}^{n}|w_{ik}w_{kj}|\right)$ are bounded, hypothesis 10 holds for the case SAR. Table 1 displays the conditions that are to be fulfilled on the weights w_{ij} in order to guarantee hypotheses 8 to 10 of Theorem 1 for covariance matrices of type CAR and SAR.

Identifier	Condition	H.8	H.9	H.10	Apply to
1	$\max_{1\leq i\leq n}\left(\sum_{j=1}^{n}\lvert w_{ij}\rvert\right)<c$	yes	yes	yes	CAR-SAR
2	$\lvert\theta\rvert\max_{1\leq i\leq n}\left(\sum_{j=1}^{n}\lvert w_{ij}\rvert\right)<1$	no	yes	no	CAR-SAR
3	$\max_{1\leq i\leq n}\left(\sum_{\substack{j=1\\ j\neq i}}^{n}\sum_{k=1}^{n}\lvert w_{ik}w_{kj}\rvert\right)<c$	no	no	yes	only SAR

Table 1: Conditions on weights w_{ij} of matrix $\mathbf{W}$.

Let B a binary weight matrix, such that each location has a bounded number of neighbours (for instance stemming from a triangulation neighbour structure), then the following W weight matrix fulfills the 3 conditions of Table 1:

$$w_{ij}=\begin{cases}\min\left(\frac{1}{|\mathscr{N}(i)|},\frac{1}{|\mathscr{N}(j)|}\right) & \text{if } j\in\mathscr{N}(i)\ (equiv.\ i\in\mathscr{N}(j))\\ 0 & \text{in other case}\end{cases}\tag{7}$$

where $\mathscr{N}(i)=\{j:B_{ij}=1\}$ This can be verified bounding $\max_{1\leq i\leq n}\left(\sum_{j=1}^{n}|w_{ij}|\right)$ and $\max_{1\leq i\leq n}\left(\sum_{\substack{j=1\\ j\neq i}}^{n}\sum_{k=1}^{n}|w_{ik}w_{kj}|\right)$. Considering different situations according to the relative sizes of $\mathscr{N}(i)$ and $\mathscr{N}(j)$, we obtain

$$\max_{1\leq i\leq n}\left(\sum_{j=1}^{n}|w_{ij}|\right)=\max_{1\leq i\leq n}\left(\sum_{j=1}^{n}\frac{1}{\max\left(|\mathscr{N}(i)|,|\mathscr{N}(j)|\right)}\right)\leq 1$$

and condition 1 is verified. Condition 2 is verified because $|\theta| < 1$.

We denote $\mathscr{N}_2(i)$ the neighbours of order two of i, and $|\mathscr{N}_2'(i)| = \sum_{j \in \mathscr{N}(i)} |\mathscr{N}(j) - \{i\}|$. We have

$$\max_{1 \leq i \leq n} \left(\sum_{\substack{j=1 \\ j \neq i}}^{n} \sum_{k=1}^{n} |w_{ik} w_{kj}| \right) \leq \max_{1 \leq i \leq n} \left(\frac{|\mathscr{N}_2'(i)|}{|\mathscr{N}(i)|} \right)$$

$\frac{|\mathscr{N}_2'(i)|}{|\mathscr{N}(i)|}$ can be interpreted as the mean quantity of neighbours that have a i-neighbour (without i), and we expect that $\max_{1 \leq i \leq n} \left(\frac{|\mathscr{N}_2'(i)|}{|\mathscr{N}(i)|} \right)$ is bounded for a variety of neighbourhood structures, such as the triangulation one for instance. A simulation study showed empirically that for $n \in \{100, \ldots, 40000\}$, 100 replicates for each n, and a triangulation neighbourhood structure the quantity $\max_{1 \leq i \leq n} \left(\frac{|\mathscr{N}_2'(i)|}{|\mathscr{N}(i)|} \right)$ although increasing slowly with n remains less than 9 and we may consider that condition 3 holds. Parameters θ and σ are estimated from the residuals of model without correlation. They can be estimated by maximum likelihood [17], or by the generalized moments procedure. In the latter case [18] establish the consistency of the estimator so the condition 10 of Theorem 1 holds. To summarize, the procedure to select and estimate the parameters for model 5 is the following:

- Step 1: Fit a LASSO model (Standard LASSO) as if there was no spatial dependence.
- Step 2: Test the spatial autocorrelation of the residuals, using for instance the Moran or Geary index.
- Step 3: If the null hypothesis of no correlation is rejected, consider CAR and SAR models, estimate parameters θ and σ^2 by maximum likelihood or generalized moments and plug them in $\hat{\mathbf{\Sigma}}_{\text{CAR}}$ or $\hat{\mathbf{\Sigma}}_{\text{SAR}}$
- Step 4: Eliminate the spatial dependence using the estimated matrix of the previous step applying 4.

- Step 5: Estimate the LASSO model to the transformed data, selecting the parameter λ by cross validation. This model is named Spatial LASSO.

3. Simulations

In order to investigate how performs the procedure designed in the previous section we proceed to simulations according to several scenarios.

n locations irregularly spaced are randomly drawn, $p = 20$ random variables $N(0,1)$ are simulated forming the $n \times p$ design matrix $\mathbf{X}$. The output Y is set as $Y = X_1 + X_2 + X_3 + X_4 + \varepsilon$, with $\varepsilon \sim N(\mathbb{0}, \mathbf{\Sigma})$, that is $p^* = 4$.

We consider 2 distinct situations, namely

- *Problem 1*: random variables (X_j , $1 \leq j \leq 20$) are independent
- *Problem 2*: random variables (X_j , $1 \leq j \leq 20$) are correlated with $cor(X_i, X_j) = 0.9^{|i-j|}$.

In each case we check that the conditions on the matrix $\widetilde{\mathbf{X}}$ in Theorem 1, that is conditions 1, 2, 5 and 7 hold, and to investigate the asymptotic side of the consistency result we vary n to large numbers. The distinct scenarios correspond to a specific value of n, σ^2, model error (CAR or SAR), the parameter θ is fixed. Values for n, σ^2 and θ are $n \in \{100, 200, 400, 800\}$, $\sigma^2 \in \{0.5, 1, 1.5, 2, 5, 10\}$, $\theta = 0.9$. For each scenario 100 replicates are drawn.

The procedure described in the previous section named "Spatial LASSO" is compared to the "Standard LASSO" procedure for independent errors and to a procedure named "LARS$_m$" described in [13] which selects the significant variables by regularizing the complete likelihood of the model, that is including parameters β, θ and σ^2.

Simulations were driven in R using packages glmnet, lars, MASS, spdep, expm and TruncatedNormal ([19], [20], [21], [22], [23], [24] and [25]). The neighbourhood structure was designed following reference [26].

Table 2 displays the average number of selected variables (Vars), the True Positive and False Positive for each method, according to different parameters of the scenarios. All the detailed results are available in the supplementary material.

scenario	Vars			True Positive			False Positive		
	St LS	Sp LS	LARS$_m$	St LS	Sp LS	LARS$_m$	St LS	Sp LS	LARS$_m$
CAR	4.26	4.28	4.83	3.87	3.89	3.89	0.40	0.39	0.94
SAR	4.13	4.28	4.87	3.72	3.91	3.87	0.41	0.37	0.99
Problem 1	4.31	4.38	5.21	3.79	3.90	3.92	0.52	0.48	1.29
Problem 2	4.09	4.17	4.49	3.80	3.89	3.85	0.29	0.28	0.64
$n = 100$	4.40	4.38	5.17	3.88	3.89	3.94	0.52	0.49	1.24
$n = 200$	4.21	4.38	5.24	3.70	3.92	3.90	0.52	0.47	1.34
$\sigma^2 = 1$	4.55	4.52	5.20	4.00	4.00	4.00	0.55	0.52	1.20
$\sigma^2 = 2$	4.52	4.44	5.11	3.98	4.00	4.00	0.54	0.44	1.11
All simulations	4.20	4.28	4.85	3.80	3.90	3.88	0.40	0.38	0.97

Table 2: Average number selected variables(Vars), True positive, False Positive for Standard LASSO (StLS), Spatial LASSO (SpLS) and LARS$_m$, according to different scenarios

The average number of selected variables is slightly overestimated for the three methods, but more for the LARS$_m$ one. The True Positive are well recovered for Spatial LASSO and LARS$_m$, Standard LASSO is lower. False Positive are much higher for LARS$_m$, almost twice than Standard LASSO and Spatial LASSO, the latter being slightly lower.

Results for CAR and SAR error models are not very different and none give systematically better results. Results are systematically better for *Problem 2* than for *Problem 1* regardless the method, this kind of decreasing correlation in the covariates help selecting the true support. As expected when n increases or σ^2 decreases the three indicators improve.

Let us consider the euclidian distance between the pair (TP,FP) and the optimal situation (4,0) for each scenario in order to rank the three procedures. According to this criterion Spatial LASSO ranks first with **46%** scenarios, then LARS$_\mathbf{m}$ with **33%** and standard LASSO with the remaining **21%**. These results are not really different according to the problem or the n value. When the error model is CAR, Spatial LASSO and LARS$_m$ perform similarly, when the error model is SAR, Spatial LASSO is much better. Spatial LASSO performs better for intermediate values of σ^2, LARS$_m$ for high values of σ^2.

Estimations of the parameters β_j are as expected underestimated by the Stan-

dard LASSO and Spatial LASSO, the bias decreases when n increases. LARS$_m$ gives better estimations for the first 4 coefficients with value 1, but not for the null coefficients as it produces much more False Positive. This is illustrated in Figure 1 for *Problem 2* and SAR error model. Regarding the spatial parame-

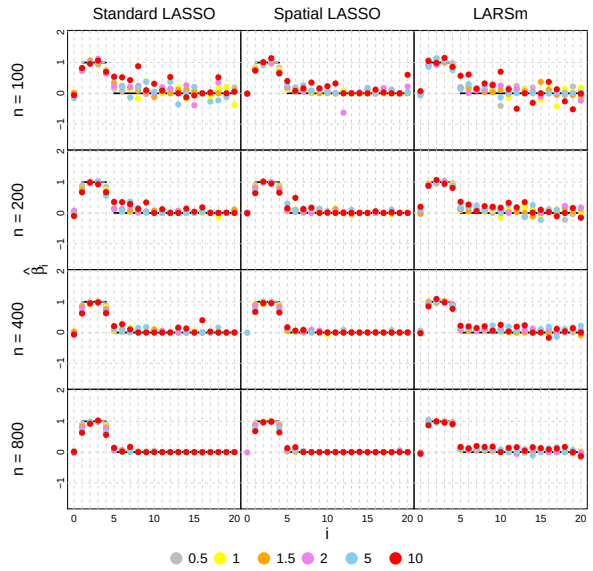

Figure 1: β_j, $j = 0 \dots 20$ estimations for Problem 2, SAR error model. Colours are for different σ^2 values.

ters θ and σ^2 they are better estimated by the LARS$_m$ procedure than by the Spatial LASSO one.

As a conclusion on this simulation study we can assert that the procedure Spatial LASSO recovers better the parameters β support while LARS$_m$ gives better estimations of all the parameters at the cost of a greater number of False Positive. More figures and results are available in the supplementary material.

4. Real case : income in rural areas in Uruguay

Knowing the income level of a household is fundamental to determine when it can apply for the different public policies in a country, such as access to housing, food allowances, etc. In the simplest situation, the household's income is

determined by taking into account all the regular income of each member of the household, through the documentation requested. However, in developing countries, as is the case in Uruguay, and particularly for some specific populations such as rural areas, it is not at all simple to collect this data, since many households receive informal and/or very irregular income, so that in many cases there is no formal documentation to prove the declared income. In these cases, indirect methods can be used to approximate income, such as the level of consumption or the degree of comfort of the household. The characteristics of the territory, the political and economic reality and the moment also influence the determination of income, and identifying the most important variables in the modeling carried out in a spatial analysis framework makes sense, because of their correlations.

We apply the three methods described in the previous sections to an Uruguayan socio-economic dataset including per capita income in rural areas, variables describing labour market (activity, employment, unemployement) and comfort items. The data are from the survey ECH (Encuesta Continua de Hogares) in Uruguay conducted by INE (Instituto Nacional de Estadística).

We use a subset of the survey corresponding to year 2018, including 933 households in rural areas which are not owning their housing and have at most one basic need[1] not satisfied. The microdata of the ECH are freely accessible and are available on the INE website; however, for confidentiality reasons, the georeferencing of the surveyed households is not available. In order to proceed a spatial analysis, and for academic purposes, it was decided to impute to each household a random location drawn within the census tract area to which each household belongs. The spatial location of the households considered, identifying the quartile of per capita income to which they belong, is presented in Figure 2.

[1] measure the lack of access of the population to certain goods and services considered critical for human development: access to decent housing, electricity, potable water, sanitary services, comfort items and access to education

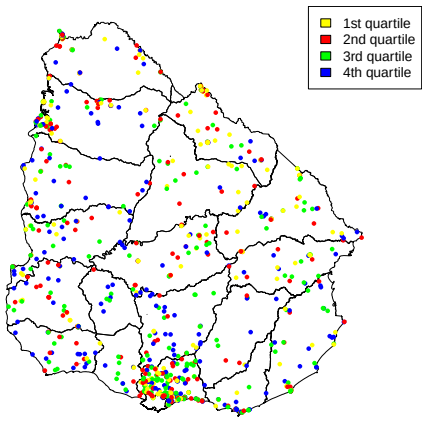

Figure 2: Uruguay map, with households spatial locations and quartiles per capita they belong to.

With the information available in the survey we seek to explain the per capita household income (without rental value and without domestic service), expressed in RU (Readjustable Units[2]), according to a set of quantitative variables, reflecting the main socio-economic characteristics of the household as well as its level of comfort. The 29 covariates are listed in Table 3, they include information on the household members (number, gender, education, occupation) and comfort elements (cars, laptop, mobile phone ...). On real data it is not possible to check that the conditions in Theorem 1 hold as the real covariance matrix Σ is unknown. In particular in our case as the covariates are highly correlated it is doubtful that condition 5. (the irrepresentability condition) which roughly states that variables in the support are weakly correlated to variables out of the support, is fulfilled. Nevertheless it is interesting using this variable selection procedure having in mind that likely we are not in the good framework

[2]It is a unit of measurement, its value is periodically adjusted according to the Average Salary Index, quantifying the variations in the previous twelve months

to investigate what could be the relevant covariates. The weight matrix W is driven from a triangulation neighborhood structure and weights are set according to 7. In addition to the packages already mentioned, we use shapefiles, sp and rgdal ([27], [28] and [29]).

According to this structure each household has a number of neighbours in the range (3, 12) the average number is 5.94.

Table 3 gives the β_j selected and estimated for Standard LASSO, Spatial LASSO and LARS$_m$.

Variable	Parameter	Standard LASSO	CAR		SAR	
			Spatial LASSO	LARS$_m$	Spatial LASSO	LARS$_m$
intercept	$\hat{\beta}_0$	6.14	0.64	-0.91	0.65	5.80
number of Household Members (HM)	$\hat{\beta}_1$	-0.90	-0.80	-	-0.81	-0.31
proportion of Income Earners (IE)	$\hat{\beta}_2$	8.57	9.00	10.15	8.99	4.33
proportion of male IE	$\hat{\beta}_3$	-	-	-	-	0.17
average age of not IE	$\hat{\beta}_4$	-	-	-	-	-
average age of male IE	$\hat{\beta}_5$	-	-	-	-	0.01
average age of female IE	$\hat{\beta}_6$	-	-	-	-	-
average years formal education IE	$\hat{\beta}_7$	0.01	0.005	-	0.01	-
average years formal education male IE	$\hat{\beta}_8$	-	-	-	-	-
average years formal education female IE	$\hat{\beta}_9$	-	-	-	-	-
proportion HM receiving social benefits	$\hat{\beta}_{10}$	-0.21	-0.19	-	-0.19	-
average age HM receiving social benefits	$\hat{\beta}_{11}$	-	-	-	-	-
proportion of employed (E)	$\hat{\beta}_{12}$	0.51	0.36	6.10	0.37	5.46
average age male E	$\hat{\beta}_{13}$	-	-	-	-	-
average age female E	$\hat{\beta}_{14}$	-	-	-	-	-
proportion male E	$\hat{\beta}_{15}$	-	-	-	-	-
average years formal education E	$\hat{\beta}_{16}$	-	-	-	-	-
average years formal education male E	$\hat{\beta}_{17}$	-	-	-	-	-
average years formal education female E	$\hat{\beta}_{18}$	-	-	-	-	-
average number hours worked among E	$\hat{\beta}_{19}$	0.09	0.09	-	0.09	-
average hours worked by male E	$\hat{\beta}_{20}$	-	-	-	-	-
average hours worked by female E	$\hat{\beta}_{21}$	-	-	-	-	-
number unsatisfied basic needs	$\hat{\beta}_{22}$	-0.39	-0.35	-	-0.35	-
number comfort elements	$\hat{\beta}_{23}$	0.06	0.06	-	0.06	-
number cars	$\hat{\beta}_{24}$	2.18	2.08	2.02	2.08	0.43
number motorcycles	$\hat{\beta}_{25}$	-	-	-	-	-
number laptops	$\hat{\beta}_{26}$	2.06	1.96	1.30	1.96	1.50
number air conditioners	$\hat{\beta}_{27}$	-	-	-	-	-
number colour televisions	$\hat{\beta}_{28}$	-	-	-	-	-
proportion HM with mobile phone	$\hat{\beta}_{29}$	1.20	1.32	5.86	1.32	3.65
	$\hat{\theta}$	-	0.41	0.36	0.21	0.22
	$\hat{\sigma}^2$	-	48.02	55.05	48.29	57.17

Table 3: Estimated parameters by Standard LASSO, Spatial LASSO and LARS$_m$, CAR and SAR errors.

The optimal parameter λ calculated by leave one out procedure is equal to 0.575 for Standard LASSO and 0.082 for Spatial LASSO.

Standard LASSO and Spatial LASSO (both CAR and SAR errors model) select the same 11 variables from the 29, they are mostly related to the income

earners, their education, their employment and some comfort elements. LARS$_m$ method, selects 5 with the CAR model and 8 variables with the SAR model. The variables selected included in the set of 11 variables selected by standard and Spatial LASSO.

Both Spatial LASSO and LARS$_m$ results depend on initial values of the parameters. These values are drawn at random and introduce randomness in the results. To evaluate the sensitivity to the initial parameters we draw for each method 100 initial values and examine how vary the estimates. While β estimates obtained with Spatial LASSO do not vary much, those obtained with LARS$_m$ vary greatly, some of them even change signs. Even the number of selected variables may change either with the CAR or the SAR error model. On the other hand θ and σ^2 estimates don not vary much for either Spatial LASSO and LARS$_m$.

To compare the different models we calculate, the Nagelkerke pseudo R^2 ([30]), the AIC and the BIC for each method and type of error model. The models are re-estimated according to the following criteria: a lineaTable 4 shows that Spatial LASSO (both CAR and SAR models) obtain the highest pseudo R^2, and the lowest value of both AIC and BIC.

		CAR		SAR	
Indicator	Standard LASSO	Spatial LASSO	LARS$_m$	Spatial LASSO	LARS$_m$
pseudo R^2	0.442	0.447	0.373	0.447	0.392
AIC	6277	6270	6375	6270	6352
BIC	6340	6338	6413	6338	6405

Table 4: Goodness-of-fit indicators of the estimated models.

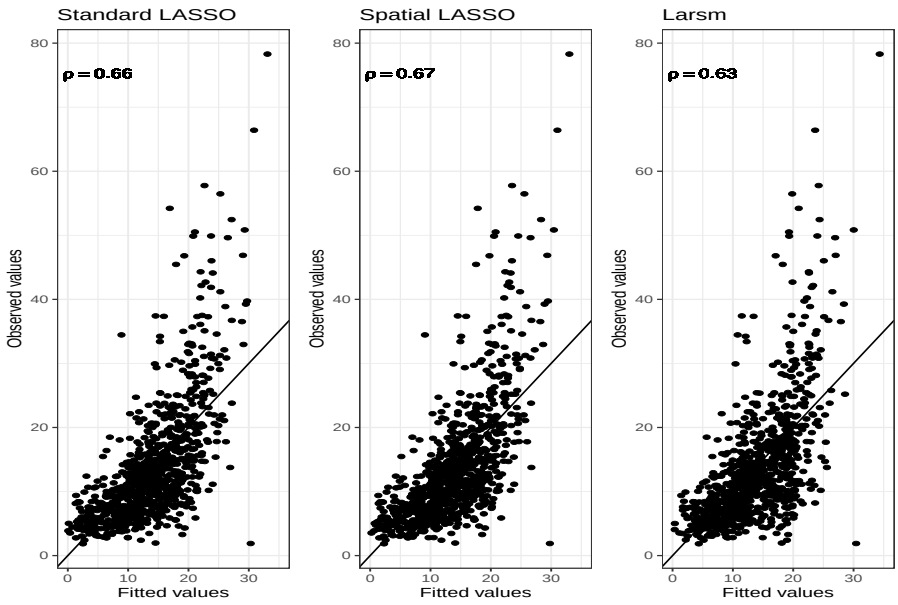

Figure 3: Fitted values versus observed values according to the estimated models.

The LASSO method (both Standard and Spatial by both error models) selected one third of the available variables, while the LARS$_m$ method selected 5 variables with the SAR model and 8 variables with the CAR model. Figure 3 shows the fitted values versus the observed values according to the 3 models. Except for specific cases, in general there is no substantial change in the point estimation of the parameters by the three methods. Comparison according to indicators such as the pseudo R^2, AIC and BIC suggests that the Spatial LASSO has a better fit than the others. While all variables a priori could be eligible to explain per capita household income in the extended rural environment, it is understood that those selected by LASSO form a reasonable subset of those with the greatest explanatory power, even in a context of variables that are not independent of each other. The LARS$_m$ method further narrows the set of variables selected by LASSO, but the drawback is that the results are sensitive to the initial values, the tolerance and the maximum number of steps considered.

5. Discussion and conclusion

In this work we established a result giving conditions to ensure the sign consistency of the LASSO estimator for dependent data. This result guarantees to recover asymptotically the true support of the regression parameters. The conditions involved in this result relate to the design matrix and the error covariance matrix and may be difficult to check in some cases. We focus on the particular case of spatially dependent data on irregular lattice according to a CAR or a SAR model and we give sufficient conditions on the weight matrix for some of these conditions to be satisfied. These conditions are obtained by straightforward calculations, some lighter conditions might be obtained by more precise methods. In a geostatistical framework, properties of the most popular covariance functions could be used to derive similar conditions to the error covariance matrix satisfy the theorem assumptions.

A simulation study shows for the situations investigated that the support is well recovered for sufficient sample sizes and reasonable noise levels. This simulation study also shows that the recommended method: estimate first the parameters in an independent setting and identification of the covariance structure on the residuals and then implementation of the procedure for dependent data using the estimated covariance matrix is more efficient than the method consisting in regularizing the likelihood involving both the regression parameters and the covariance parameters. It would be worthwhile to study more extensively various situations of dependence in errors to determine to what extent the exact support of the parameters can be recovered. Other error models can be considered with higher autoregressive order or for data in a continuous domain instead of irregular lattice.

The application of these methods to a socio-economic dataset, with the objective of explaining the per capita income of rural households in Uruguay, showed that these two approaches give significantly different results nevertheless the usual goodness-of-fit criteria give the advantage to the plug-in method which selects a larger number of variables than the maximum likelihood method. It should be

noted that this dataset, as the covariates are highly correlated, likely does not fulfill the conditions of the theorem, which may explain this discrepancy. Methods taking into account correlation or other kind of proximity in the covariates such as Elastic-Net ([31]), Fused-LASSO ([32]) or Group-LASSO ([33]), should be investigated considering models for data with spatial dependence.

Supplementary material can be download at
https://www6.inrae.fr/mia-paris/Equipes/Membres/Liliane-Bel/SM

References

[1] R. Tibshirani, Regression shrinkage and selection via the lasso, Journal of the Royal Statistical Society. Series B 58 (1) (1996) 267–288.

[2] R. Tibshirani, Regression shrinkage and selection via the lasso: a retrospective, Journal of the Royal Statistical Society. Series B 73 (3) (2011) 273–282.

[3] T. Hastie, R. Tibshirani, M. Wainwright, Statistical Learning with Sparsity. The Lasso and Generalizations, 1st Edition, CRC Press, 2016.

[4] K. Knight, W. Fu, Asymptotics for Lasso-Type Estimators, The Annals of Statistics 28 (5) (2000) 1356–1378.

[5] H. Zou, The Adaptive Lasso and Its Oracle Properties, Journal of the American Statistical Association 101 (2006) 1418–1429.

[6] P. Zhao, B. Yu, On Model Selection Consistency of Lasso, Journal of Machine Learning Research 7 (90) (2006) 2541–2563.

[7] N. Simon, J. Friedman, T. Hastie, R. Tibshirani, A Sparse-Group Lasso, Journal of Computational and Graphical Statistics 22 (2) (2013) 231–245.

[8] S. Nandy, C. Y. Lim, T. Maiti, Additive model building for spatial regression, Journal of the Royal Statistical Society. Statistical Methodology. Series B. 79 (3) (2017) 779–800.

[9] T. Chu, J. Zhu, H. Wang, Penalized maximum likelihood estimation and variable selection in Geostatistics, The Annals of Statistics 39 (5) (2011) 2607–2625.

[10] H. Huang, C. Chen, Optimal Geostatistical Model Selection, Journal of the American Statistical Association 102 (479) (2007) 1009–1024.

[11] H. Wang, J. Zhu, Variable selection in spatial regression via penalized least squares, The Canadian Journal of Statistics 37 (4) (2009) 607–624.

[12] P. Reyes, J. Zhu, B. Aukema, Selection of spatial-temporal lattice models: assessing the impact of climate conditions on a mountain pine beetle outbreak, Journal of Agricultural, Biological and Environmental Statistics 17 (3) (2012) 508–525.

[13] J. Zhu, H. Huang, P. Reyes, On selection of spatial linear models for lattice data, Journal of the Royal Statistical Society Series B 72 (3) (2010) 389–402.

[14] L. Cai, T. Maiti, Variable selection and estimation for high-dimensional spatial autoregressive models, Scandinavian Journal of Statistics 47 (2) (2020) 587–607.

[15] M. Perrot-Dockès, C. Lévy-Leduc, L. Sansonet, J. Chiquet, Variable selection in multivariate linear models with high-dimensional covariance matrix estimation, Journal of Multivariate Analysis 166 (2018) 78–97.

[16] R. A. Horn, C. R. Johnson, Matrix Analysis, 2nd Edition, Cambridge University Press, 2013.

[17] C. Gaetan, X. Guyon, Spatial Statistics and Modeling, Springer, 2010.

[18] H. Kelejian, I. Prucha, A generalized spatial two-stage least squares procedure for estimating a spatial autoregressive model with autoregressive disturbances, The Journal of Real Estate Finance and Economics 17 (2) (1998) 99–121.

[19] R Core Team, R: A Language and Environment for Statistical Computing, R Foundation for Statistical Computing, Vienna, Austria, https://www.R-project.org/ (2020).

[20] J. Friedman, T. Hastie, R. Tibshirani, Regularization Paths for Generalized Linear Models via Coordinate Descent, Journal of Statistical Software 33 (1) (2010) 1–22, http://www.jstatsoft.org/v33/i01/.

[21] T. Hastie, B. Efron, lars: Least Angle Regression, Lasso and Forward Stagewise, R package version 1.2. https://CRAN.R-project.org/package=lars (2013).

[22] W. N. Venables, B. D. Ripley, Modern Applied Statistics with S, 4th Edition, Springer, New York, 2002, iSBN 0-387-95457-0. http://www.stats.ox.ac.uk/pub/MASS4.

[23] R. S. Bivand, E. Pebesma, V. Gomez-Rubio, Applied spatial data analysis with R, 2nd Edition, Springer, New York, 2013, http://www.asdar-book.org/.

[24] V. Goulet, C. Dutang, M. Maechler, D. Firth, M. Shapira, M. Stadelmann, expm: Matrix Exponential, Log, "etc", R package version 0.999-4. https://CRAN.R-project.org/package=expm (2019).

[25] Z. Botev, L. Belzile, TruncatedNormal: Truncated Multivariate Normal and Student Distributions, R package version 2.2. https://CRAN.R-project.org/package=TruncatedNormal (2020).

[26] R. Bivand, E. Pebesma, V. Gómez-Rubio, Applied Spatial Data Analysis with R, 1st Edition, Springer-Verlag, 2008, Ch. Creating Neighbours, pp. 239—251.

[27] B. Stabler, shapefiles: Read and Write ESRI Shapefiles, r package version 0.7. https://CRAN.R-project.org/package=shapefiles (2013).

[28] E. Pebesma, R. Bivand, Classes and methods for spatial data in R, R News 5 (2) (2005) 9–13, https://CRAN.R-project.org/doc/Rnews/.

[29] R. Bivand, T. Keitt, B. Rowlingson, rgdal: Bindings for the "Geospatial" Data Abstraction Library, r package version 1.4-8. https://CRAN.R-project.org/package=rgdal (2019).

[30] N. Nagelkerke, A note on a General Definition of the Coefficient of Determination, Biometrika 78 (3) (1991) 691–692.

[31] H. Zou, T. Hastie, Regularization and variable selection via the elastic net, Journal of the Royal Statistical Society, Series B 67 (2) (2005) 301–320.

[32] R. Tibshirani, M. Saunders, S. Rosset, J. Zhu, K. Knight, parsity and smoothness via the fused lasso, Journal of the Royal Statistical Society: Series B (Statistical Methodology) 67 (1) (2005) 91–108.

[33] M. Yuan, Y. Lin, Model selection and estimation in regression with grouped variables, Journal of the Royal Statistical Society Series B 68 (1) (2006) 49–67.