# A screening procedure for high-dimensional autologistic models

**Rodolfo Metulini**, **Francesco Giordano**

Department of Economics and Statistics (DISES) - University of Salerno

June 25th, 2021

SIS 2021 - 50th Scientific meeting of the Italian Statistical Society

# Table of contents

# Introduction

- Bank failure prediction has been diffusely employed with a statistical modelling approaches:
  1. Discriminant Analysis (**Altman, 1968**)
  2. Logit (and Probit) models (**Ohlson, 1980**)
  3. Support Vector Machines (**Shin et al, 2005**)
- A fundamental aspect is that of correctly measuring and interpreting the effect of a large number of covariates (possibly accounting for non linearity, additivity (**Berg, 2007**) and interaction effects) **Amendola et al (2017)**.
- We deal the problem of selecting the set of relevant features by resorting on **screening procedures** based on selecting important covariates by means of a **marginal approach** for **ultra-high dimensional data** (**Fan and Lv, 2008**)
  - Penalized variable selection methods suffer for noise accumulation (**Fan et al, 2012**).
- In this paper we focus on generalized linear models (GLM) with logit link function (**McCullagh and Nelder, 1989**) and accounting for Spatial Autocorrelation (SAR) by resorting on the **autologistic model** (**Hughes et al, 2011**).

# Objective

- The aim of this paper is to evaluate **sure screening property** (SSP) (**Fan and Lv, 2008**) of our screening procedure:
  - By means of a **simulation experiment**
  - within the framework of **Maximum Pseudo Likelihood Estimation** (MPLE).
- The key point for every screening procedure is the SSP:

  - estimated set of relevant covariates contains the true relevant ones with a probability that tends to 1, when the sample size grows -

- In the case of GLM, **Fan and Song (2010)** demonstrated such a property, but nothing has been done to prove SSP for autologistic.

# Modelling framework - i

- Autologistic models for binary response in regular lattice data set-up (**Cressie, 2015**) has been firstly proposed by **Besag (1975)** by directly imposing a joint Markov random field.
  - They remind the formulation of the logistic regression derived by **McCullagh and Nelder (1989)**.
- Let $Y_i \in \{0, 1\}$, $i = 1, ..., n$, be the $i-$th binary element of the vector $\mathbf{Y}$, $\mathbf{X}_i$ be the vector column corresponding to the $i-$th row of the design matrix $\mathbf{X}$ with $n$ rows and $p$ columns, $\beta$ be the vector containing the $p$ regression parameters to be estimated.
- The full conditional distribution of $\mathbf{Y}$ according to autologistic considering the assumption of **stationary** and **isotropic** processes along with **Cressie's clique n. 2** (**Cressie, 2015**) is given by:

$$log \frac{P(Y_i = 1 \mid \mathbf{X}, \mathbf{Y})}{P(Y_i = 0 \mid \mathbf{X}, \mathbf{Y})} = \mathbf{X}_i^{'} \beta + \eta \sum_{j \neq i} w_{ij} Y_j, \tag{1}$$

where $\eta$ is a scalar and $w_{ij}$ is the $(i, j)$ element of the $n - by - n$ matrix $\mathbf{W}$, with $w_{ij} = 1$ if $i$ is neighbour of $j$, 0 otherwise.

# Modelling framework - ii

- **Caragea and Kaiser (2009)** proposed a centred re-parametrization to provide meaningful interpretations of the parameters.
  - $Y_i$ is replaced by $Y_i - \mu_i$ in eq. 1, where $\mu_i$ is the unconditional expectation of $Y_i$.
- By assuming positivity condition (i.e., if $P(Y_i) > 0,\ i = 1, ..., n$, then $P(Y_1, ..., Y_n) > 0$) and Brook's Lemma (**Besag, 1974**, pag. 195) it is possible to generate the following joint distribution:

$$\pi(\mathbf{Y} \mid \boldsymbol{\theta}) = c(\boldsymbol{\theta})^{-1} exp\left(\mathbf{Y}'\mathbf{X}\beta - \eta\mathbf{Y}'\mathbf{W}\boldsymbol{\mu} + \frac{\eta}{2}\mathbf{Y}'\mathbf{W}\mathbf{Y}\right), \qquad (2)$$

where $\boldsymbol{\mu} = (\mu_1, ..., \mu_n)'$ is the vector of expectations, $\boldsymbol{\theta} = [\beta', \eta]'$ and $c(\boldsymbol{\theta})$ is the normalizing constant.

# Estimation method

- We place in the framework of **MPLE** method (**Besag, 1975**).
  - MPLE circumvents the issue of computational intractability of the normalizing constant $c(\boldsymbol{\theta})$ by maximizing the pseudo-likelihood with respect to the parameters as if it were a standard maximum likelihood.
- Related literature focused on methods for obtaining this normalizing constant (**Ogata and Tanemura, 1984**).
- However, despite MPLE is not efficient, with a loss of efficiency positively related to the absolute value of $\eta$, **asymptotical consistency** and normality are guaranteed (**Besag, 1975**).

# Screening procedure - i

- Methods based on using a penalty for penalization of the model coefficients has been proposed, such as least absolute shrinkage and selection operator (LASSO) (**Tibshirani, 1996**) and its generalizations.
- However, variable selection methods specifically proposed for autologistic model (e.g., **Fu et al, 2013**) have never been proved in high dimensional setup.
- In high-dimension, data require sophisticated variable selection methods accounting for i) **noise accumulation**, ii) **spurious correlation**, and iii) **incidental endogeneity** (**Fan et al, 2012**) which makes the aforementioned penalty-based methods inappropriate.

# Screening procedure - ii

- Among screening procedures adopting **marginal maximum likelihood**, the one proposed by **Fan and Song (2010)** is proved, under some general conditions:
  - to be consistent and efficient in GLM with logit link function
  - to enjoy the SSP for the case of NP-Dimensionality.
- Given the spatial component, Marginal MPLE (MMPLE) reads as:

$$\tilde{\beta}_h^{MMPLE} = \underset{\beta_h}{\text{argmax}} \prod_{i=1}^{n} P(Y_i \mid X_{ih}, \mathbf{Y}_{-i}), \tag{3}$$

where $\mathbf{Y}_{-i} = [Y_1, ..., Y_{i-1}, Y_{i+1}, ..., Y_n]$ and $X_{ih}$ is the $i-$th observation of the $h-$th covariate.

- SSP for the autologistic can be written as:

$$\mathbb{P}_{n\to\infty}\{\mathcal{M}_* \subset \hat{\mathcal{M}}_{\gamma_n}\} \to 1 \tag{4}$$

where $\mathcal{M}_* = \{1 \leq h \leq p_n : \beta_h \neq 0\}$ is the set of true important variables with associated coefficients $\beta_*$
$\hat{\mathcal{M}}_{\gamma_n} = \{1 \leq h \leq p_n :\mid \tilde{\beta}_h^{MMPLE} \mid \geq \gamma_n\}$ is the estimated set, given a pre-specified threshold $\gamma_n$.

# Simulation

- We employ an algorithm to evaluate SSP of MMPLE
  - by the **Median of the Minimum Model Size** (MMMS) of marginal estimated coefficients, along with its associated **Robust Standard Deviation** (RSD), as in **Fan and Song (2010)**.
  - by do not specifying parameter $\gamma_n$: we replace $\hat{\mathcal{M}}_{\gamma_n}$ with $\hat{\mathcal{M}}$ being the smallest set including ordered (descending) estimated coefficients such that the set $\mathcal{M}_*$ is a subset of it.

- Design of experiment:
  - Perfect sampling coupling from the past (CFTP) (**Propp and Wilson, 1996**) for generating the sample values for $\mathbf{Y}$, which better accounts for the dependence in $\mathbf{Y}$ compared to MCMC methods (**Hughes et al, 2011**).
  - $w_{ij}$: realization from $Bern(s)$ process, with $s$=0.1 (sparsity).
  - All the $\mathbf{X}$'s are realizations of an *i.i.d.* process $N(0, 1)$.
  - We generate $p = 1000$ covariates.
  - $m = \{3, 6\}$ non-zero coefficients (relevant covariates).
  - spatial dependence: $\eta = \{0, 0.1, 0.2, 0.3, 0.5\}$.
  - $k = 200$ iterations on a sample size of $n = \{200, 500\}$.

# Results

- According to the results, we find that:
  - for a moderate size of relevant covariates ($m = 3$), SSP is guaranteed even for large levels of SAR and moderate sample size ($n = 200$),
  - SSP performance becomes poor when the number of relevant covariates increases ($m = 6$) and the sample size is small ($n = 200$)
    - Under this setting, MMMS is way larger than $m$ and RSD is also high.
    - However, by increasing the sample size to $n = 500$, SSP is again guaranteed, when $m = 6$.

# Concluding remarks

- These results may be useful for practitioners in the context of bank failure prediction
  - because we restrict the **extent** in which the use of a **screening procedure** based on a "pseudo" marginal approach for selecting relevant covariates in autologistic is **appropriate**.

- As a further development we may think of:
  - deriving a methodological strategy to increase the performance of the proposed screening procedure when a large number of relevant covariates and a small sample size are assumed,
  - propose this method to make the variable selection in **Fu et al (2013)** feasible even when $p > n$.


Altman, E. I., *Financial ratios, discriminant analysis and the prediction of corporate bankruptcy*, The journal of finance, 23(4): 589-609, 1968.


Andreano, M.S., Benedetti, R., Mazzitelli, A. and Piersimoni, F. *Spatial autocorrelation and clusters in modelling corporate bankruptcy of manufacturing firms*, Economia e Politica Industriale, 45(4): 475-491, 2018.


Amendola, A., Giordano, F., Parrella, M.L. and Restaino, M., *Variable selection in high-dimensional regression: a nonparametric procedure for business failure prediction*, Applied Stochastic Models in Business and Industry, 33 (4): 355-368, 2017.


Berg, D., *Bankruptcy prediction by generalized additive models*. Applied Stochastic Models in Business and Industry, 23(2), 129-143, 2007.


Besag, J.,*Spatial interaction and the statistical analysis of lattice systems*. Journal of the Royal Statistical Society: Series B (Methodological), 36(2): 192-225, 1974.


Besag, J., *Statistical analysis of non-lattice data*, Journal of the Royal Statistical Society: Series D (The Statistician), 24(3): 179-195, 1975.


Caragea, P. C., and Kaiser, M. S., *Autologistic models with interpretable parameters*, Journal of agricultural, biological, and environmental statistics, 14(3): 281, 2009.


Cressie, N., *Statistics for spatial data*, John Wiley & Sons, 2015.


Fan, J., and Lv, J., *Sure independence screening for ultrahigh dimensional feature space*. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 70(5): 849-911, 2008.


Fan, J., and Song, R., *Sure independence screening in generalized linear models with NP-dimensionality*, The Annals of Statistics, 38(6): 3567-3604, 2010.


Fan, J., Feng, Y., & Tong, X., *A road to classification in high dimensional space: the regularized optimal affine discriminant*. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 74(4), 745-771, 2012.


Fu, R., Thurman, A. L., Chu, T., Steen-Adams, M. M. and Zhu, J., *On estimation and selection of autologistic regression models via penalized pseudolikelihood*, Journal of agricultural, biological, and environmental statistics, 18(3): 429-449, 2013.


Hughes, J., Haran, M. and Caragea, P. C., *Autologistic models for binary data on a lattice*, Environmetrics, 22(7): 857-871, 2011.


McCullagh, P. and Nelder, J.A., *Generalized Linear Models*, 2nd eds., Chapman and Hall/CRC, 1989.


Ogata, Y., & Tanemura, M., *Likelihood analysis of spatial point patterns*. Journal of the Royal Statistical Society: Series B (Methodological), 46(3), 496-518, 1984.


Ohlson, J. A., *Financial ratios and the probabilistic prediction of bankruptcy*, Journal of accounting research, 109-131, 1980.


Propp JG, Wilson DB., *Exact sampling with coupled Markov chains and applications to statistical mechanics*. Random Structures and Algorithms, 9:223–252, 1996.


Shin, K. S., Lee, T. S., and Kim, H. J., *An application of support vector machines in bankruptcy prediction model*, Expert systems with applications, 28(1): 127-135, 2005.


Tibshirani, R., *Regression shrinkage and selection via the lasso*, Journal of the Royal Statistical Society: Series B (Methodological), 58(1): 267-288, 1996.

# Spatial dependence

- Firm's performance is not independent from the performance of other firms located in space, due to the presence of geographical proximity.
- SAR (first type) may emerge when the response at location $i$ is dependent with the response at location $j$, for $j$ neighbour of $i$ (i.e., $Y$ is not an *i.i.d.* process).
- SAR (second type) may also emerge while omitting spatially autocorrelated variables among the explanatories. In this case $\epsilon_i$ is dependent with $\epsilon_j$, for $j$ neighbor of $i$.
- Ignoring first type of SAR leads to bias on model's parameters.
- Ignoring second type of SAR leads to bias on model parameters' standard errors.

Go back to slide

# Simulation's algorithm

**Data:** Chosen values of $n$, $\eta$, $\beta_1, ..., \beta_p$, $\mu$ and $\sigma$, $s$, $m$ and $\mathbf{X}$ determined from the design of the experiment.

1. For $k = 1$:
   1. simulate $\mathbf{Y}$ with CFTP
   2. estimate $\hat{\beta}_1^{MMPLE}, ..., \hat{\beta}_p^{MMPLE}$ as in eq. 3
   3. order (descending) $\hat{\beta}_1^{MMPLE}, ..., \hat{\beta}_p^{MMPLE}$ in terms of their absolute value
   4. find the minimum model size such that important variables $\mathbf{X}_1, ..., \mathbf{X}_m$ are all included in the estimated set $\hat{\mathcal{M}}$
   5. update $k = k + 1$
2. **if** $k < 200$ **then**
   repeat points 1(a) – 1(d)
   **else**
   compute MMMS with associated RSD

Table: Algorithm adopted to evaluate SSP with MMMS using MMPLE and autologistic models.

Go back to slide

# Results' table

| n | $\eta$ | MMMS (RSD) | n | $\eta$ | MMMS (RSD) |
|---|---|---|---|---|---|
| m $= 3$, | $\beta_* = [1,1,1]^T$ | | m $= 6$, | $\beta_* = [1,1,1,1,1,1]^T$ | |
| 200 | 0.0 | 3(1) | 200 | 0.0 | 10(9) |
| 200 | 0.1 | 3(1) | 200 | 0.1 | 44(47) |
| 200 | 0.2 | 3(0) | 200 | 0.2 | 29(36) |
| 200 | 0.3 | 3(0) | 200 | 0.3 | 30(36) |
| 200 | 0.5 | 3(1) | 200 | 0.5 | 38(42) |
| 500 | 0.0 | 3(0) | 500 | 0.0 | 6(0) |
| 500 | 0.1 | 3(0) | 500 | 0.1 | 6(0) |
| 500 | 0.2 | 3(0) | 500 | 0.2 | 6(0) |
| 500 | 0.3 | 3(0) | 500 | 0.3 | 6(0) |
| 500 | 0.5 | 3(0) | 500 | 0.5 | 6(0) |

Table: MMMS and the associated RSD (in parenthesis) of the experiment for the MMPLE autologistic, $k = 200$ and $p = 1000$.

Go back to slide