Sparse networks through regularised regressions

Mauro Bernardi and Michele Costola

Abstract We propose a Bayesian approach to the problem of variable selection and shrinkage in high dimensional sparse regression models where the regularisation method is an extension of a previous LASSO. The model allows us to include a large number of institutions which improves the identification of the relationship and maintains at the same time the flexibility of the univariate framework. Furthermore, we obtain a weighted directed network since the adjacency matrix is built "row by row" using for each institutions the posterior inclusion probabilities of the other institutions in the system.

1 Introduction

Models with High–dimensional data where the number of parameters is larger than the size dimension represent one of the most prominent research field in econometrics and statistics. The seminal paper of [3] introduced the least absolute shrinkage and selection operator (LASSO), one of the most popular method that can simultaneously perform parameters estimation and selection in regression models. Then, scholars began to develop sparse estimators in high–dimensions. Among the most important shrinkage methods proposed in the literature there are the least angle regression (LARS) of [1], the adaptive LASSO of [5] and the group LASSO of [4]. In this paper, we propose a Bayesian approach to the problem of variable selection and shrinkage in high dimensional causal sparse regression models where the regularisation method is an extension of a previous LASSO in a Bayesian framework. The

Mauro Bernardi
Department of Statistical Sciences, University of Padova and Istituto per le Applicazioni del Calcolo "Mauro Picone" - CNR, Roma, Italy, e-mail: mauro.bernardi@unipd.it

Michele Costola
Research Center SAFE, House of Finance, Goethe University Frankfurt am Main, Germany, e-mail: costola@safe.uni-frankfurt.de

model allows us to extend the pairwise Granger causality in the network estimation by including a large number of institutions which improves the identification of the relationship and maintains at the same time the flexibility of the univariate framework. Furthermore, we obtain a weighted directed network since the adjacency matrix is built "row by row" using for each institutions the posterior inclusion probabilities of the other institutions in the network.

2 The model

Let $\mathbf{y} = (y_1, y_2, \ldots, y_T)'$ be the vector of observations on the scalar response variable Y, $\mathbf{X} = (\mathbf{x}_1', \mathbf{x}_2', \ldots, \mathbf{x}_T')'$ be the $(T \times p)$ matrix of observations on the p covariates, i.e., $\mathbf{x}_{j,t} = \left(x_{j,1}, x_{j,2}, \ldots, x_{j,p}\right)$ while $\mathbf{Z} = (\mathbf{z}_1', \mathbf{z}_2', \ldots, \mathbf{z}_T')'$ be the $(T \times q)$ matrix of observations on predetermined variables which may include the lagged values of the endogenous variable Y up to the p–th lag. We consider the following regression model

$$\pi\left(\mathbf{y} \mid \mathbf{X}, \mu, \alpha, \beta, \sigma_\varepsilon^2\right) = \mathsf{N}\left(\mathbf{y} \mid \iota_T \mu + \mathbf{Z}\alpha + \mathbf{X}\beta, \sigma_\varepsilon^2\right), \tag{1}$$

where ι_T is the $T \times 1$ vector of unit elements, $\mu \in \mathbb{R}$ denotes the parameter related to the intercept of the model, $\alpha = (\alpha_1, \alpha_2, \ldots, \alpha_q)' \in \mathbb{R}^q$ and $\beta = (\beta_1, \beta_2, \ldots, \beta_p)' \in \mathbb{R}^p$ are vectors of regression parameters and $\sigma_\varepsilon^2 \in \mathbb{R}^+$ is the scale parameter. Hereafter, we distinguish between the vector of predetermined variable $\mathbf{z}$ and that of covariates $\mathbf{X}$ because of the role they play within the context of Granger causality we consider. Specifically, in what follows, we assume that the parameters corresponding to the predetermined variable cannot be excluded from the regression while those corresponding to the covariates can also be excluded.

2.1 Spike–and–Slab EM

Using standard notation, let γ be the p–vector where $\gamma_j = 1$ if the j–th covariate $\mathbf{X}_j$ is included as explanatory variable in the regression model and $\gamma_j = 0$, otherwise. Assuming that $\gamma_j \sim \mathsf{Ber}(\omega)$, the prior distribution for β_j, $j = 1, 2, \ldots, p$ can be written as the mixture

$$\pi(\beta_j \mid \tau, \sigma_\varepsilon, \omega) = (1 - \omega)\,\delta_0(\beta_j) + \omega \mathsf{DE}(\beta_j \mid \tau, \sigma_\varepsilon), \tag{2}$$

where $\delta_0(\beta_j)$ is a point mass at zero and DE denotes the doubly–exponential distribution with probability density function

$$\mathsf{DE}(x \mid \tau, \sigma_\varepsilon) = \frac{\tau}{\sigma_\varepsilon} \exp\left\{-\frac{\tau|x|}{\sigma_\varepsilon}\right\} \mathbb{1}_{(-\infty,\infty)}(x), \tag{3}$$

where $\tau \in \mathbb{R}^+$ acts as the shrinkage parameter in the Lasso framework and σ_ε is the scale parameter. The regression model defined in equation (1) with the spike and slab ℓ_1 prior defined in equation (2) becomes

$$\pi\left(\mathbf{y} \mid \mathbf{X}, \mu, \alpha, \beta, \sigma_\varepsilon^2\right) = \mathsf{N}\left(\mathbf{y} \mid \iota_T \mu + \mathbf{Z}\alpha + \mathbf{X}\beta, \sigma_\varepsilon^2\right) \tag{4}$$

$$\pi\left(\mu \mid \tau, \sigma_\varepsilon\right) = \mathsf{DE}\left(\mu \mid \tau, \sigma_\varepsilon\right) \tag{5}$$

$$\pi\left(\alpha \mid \tau, \sigma_\varepsilon\right) = \prod_{j=1}^{q} \mathsf{DE}\left(\alpha_j \mid \tau, \sigma_\varepsilon\right) \tag{6}$$

$$\pi\left(\beta \mid \tau, \sigma_\varepsilon, \omega\right) = \prod_{j=1}^{p} \Big[\left(1-\omega\right) \delta_0\left(\beta_j\right) + \omega \mathsf{DE}\left(\beta_j \mid \tau, \sigma_\varepsilon\right) \Big]. \tag{7}$$

It is worth noting that the prior distributions in equations (5)–(7) allow the corresponding parameters to be always included into the model specification. The definition of the model is completed by the specification of the prior on the remaining parameters $\left(\sigma_\varepsilon^2, \tau, \omega\right)$. The scale parameter σ_ε and the shrinkage parameter τ, as well as the prior inclusion probability ω are parameters that have to be estimated. Common choices for the prior on those parameters are: $\sigma_\varepsilon^2 \sim \mathsf{IG}\left(\sigma_\varepsilon^2 \mid \lambda_\sigma, \eta_\sigma\right)$, $\tau \sim \mathsf{G}\left(\tau \mid \lambda_\tau, \eta_\tau\right)$ and $\omega \sim \mathsf{B}e\left(\omega \mid \lambda_\omega, \eta_\omega\right)$. where $(\lambda_\sigma, \eta_\sigma, \lambda_\tau, \eta_\tau, \lambda_\omega, \eta_\omega)$ are prior hyperparameters. Hereafter, $\vartheta = \left(\mu, \alpha, \beta, \sigma_\varepsilon^2, \tau, \omega\right)$ collects all the unknown parameters that should be estimated.

The EM algorithm consists of two major steps, one for expectation (E–step) and one for maximisation (M–step), see [2]. At the $(m+1)$–th iteration the EM algorithm proceeds as follows:

(i) **E–step:** computes the conditional expectation of the complete–data log–likelihood given the observed data $\{y_t, \mathbf{z}_t, \mathbf{x}_t\}_{t=1}^T$ and the m–th iteration parameters updates $\vartheta^{(m)}$

$$\mathscr{Q}\left(\vartheta, \vartheta^{(m)}\right) = \mathbb{E}_{\vartheta^{(m)}}\left[\log \mathscr{L}_c\left(\vartheta\right) \mid \{y_t, \mathbf{z}_t, \mathbf{x}_t\}_{t=1}^T\right]; \tag{8}$$

*(ii)***M–step:** choose $\vartheta^{(m+1)}$ by maximising (8) with respect to ϑ

$$\vartheta^{(m+1)} = \arg\max_{\vartheta} \mathscr{Q}\left(\vartheta, \vartheta^{(m)}\right). \tag{9}$$

3 Application to Network analysis

We can define a network as a set of nodes $V_t = \{1, 2, \ldots, n_t\}$ and directed edges between nodes. The network can be represented through an n_t-dimensional adjacency matrix A_t, with the element $a_{ijt} = 1$ if there is an edge from i directed to j with $i, j \in V_t$ and 0 otherwise. The matrix A_t represents the weighted network estimated by using the proposed model where the linkages are estimated by the inclusion probability above a given threshold c,

$$A = \begin{bmatrix} 0 & p_{1,2} & \cdots & p_{1,j} & p_{1,n_t} \\ \vdots & \ddots & \cdots & \cdots & \vdots \\ \vdots & \cdots & \ddots & . & \vdots \\ p_{i,1} & \cdots & \cdots & \ddots & p_{i,n_t} \\ p_{n_t,1} & \cdots & \cdots & \cdots & 0 \end{bmatrix} \tag{10}$$

The aim to the analysis is to show that our methodology avoid the over - and misidentification of the linkages of the pairwise approach. As the reference measure for comparison, we consider the density of the network in each period d_t, defined as

$$d_t = \frac{1}{2n_t(n_t-1)} \sum_{i=1}^{n_t} \sum_{j=1}^{n_t} a_{ijt}. \tag{11}$$

$t = 1,\ldots,T$. When $(d_t - d_{t-1}) > 0$, there is an increase of system interconnectedness.

Acknowledgement

The author acknowledges financial support from the Marie Skłodowska-Curie Actions, European Union, Seventh Framework Program HORIZON 2020 under REA grant agreement n.707070. He also gratefully acknowledges research support from the Research Center SAFE, funded by the State of Hessen initiative for research LOEWE.

References

[1] Bradley Efron, Trevor Hastie, Iain Johnstone, and Robert Tibshirani. Least angle regression. *Ann. Statist.*, 32(2):407–499, 04 2004.

[2] Geoffrey McLachlan and Thriyambakam Krishnan. *The EM algorithm and extensions*, volume 382. John Wiley & Sons, 2007.

[3] Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society, Series B*, 58:267–288, 1996.

[4] Ming Yuan and Yi Lin. Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 68(1):49–67, 2006.

[5] Hui Zou. The adaptive lasso and its oracle properties. *Journal of the American Statistical Association*, 101(476):1418–1429, 2006.