3rd *BYMAT Conference* (2020)

A comparative study of robust regularization methods based on minimum density power and Rényi divergence losses

María Jaenada
Universidad Complutense de Madrid
mjaenada@ucm.es

Abstract: Over the last decades several regularization methods have been developed for sparse high-dimensional regression models. The influence of outliers is particularly awkward in the high dimensional context and so certain robust methods have been considered. Regularization methods simultaneously perform the model selection and the estimation of regression coefficients, merging a loss function based on the residuals and a penalty function inducing sparsity. Different penalties have been proposed, such as LASSO or Adaptive LASSO, a variant which improves the oracle model selection property, or non-concave penalties such as SCAD or MCP, which demostrably overcome the bias problem of the LASSO. We propose to examine robust losses with the various proposals for the penalties, leading to the differents estimating methods, namely the minimun density power divergence (DPD) and Rényi psedudodistance (RP) estimator penalized with LASSO, adaptative LASSO and SCAD. We develop an estimating algorithm for each method, focusing on their differences and similarities. Finally, we study the performance of the methods throught a simulation study.

Resumen: En las últimas décadas se han desarrollado varios métodos de regularización para el modelo lineal de regresión con datos de alta dimensión. La influencia de los datos atípicos en la estimación es particularmente perjudicial en el contexto de datos de alta dimensión, y por tanto se han considerado métodos robustos de estimación. Los métodos de regularización llevan a cabo simultáneamente la selección de variables y la estimación paramétrica mediante la combinación de una función de pérdida, basada en los residuos del modelo, y una función de penalización que induce la selección de variables. Han sido propuestas distintas penalizaciones como las penalizaciones LASSO y LASSO Adaptativo, una variante que mejora las propiedades oráculo del estimador, o penalizaciones no cóncavas como SCAD o MCP, que resuelven el problema de sesgo que presenta la penalización LASSO. Se propone examinar las pérdidas robustas con distintas funciones de penalización, dando lugar a distintos estimadores, a saber, el estimador de mínima potencia (DPD) y de mínima pseudodistancia de Rényi penalizado con LASSO, LASSO adaptativo y SCAD . Se desarrolla un algoritmo de estimación para cada método, señalando sus diferencias y similitudes. Por último, se estudia el comportamiento de los métodos a través de un estudio de simulación.

Keywords: high-dimensional linear regression models, adaptive LASSO estimator, density power divergence loss, Rényi pseudodistance, variable selection.

MSC2010: 62F35.

Acknowledgements: This research is supported by the Spanish Grants PGC2018-095194-B-100 and FPU19/01824.

Reference: Jaenada, María. "A comparative study of robust regularization methods based on minimum density power and Rényi divergence losses". In: *TEMat monográficos*, 2 (2021): *Proceedings of the 3rd BYMAT Conference*, pp. 195-198. ISSN: 2660-6003. URL: https://temat.es/monograficos/article/view/vol2-p195.

1. Introducción

We consider the high-dimensional linear regression model (LRM) given by

$$Y_i = \boldsymbol{X}_i^T \boldsymbol{\beta} + U_i, \quad i = 1, \dots, n, \tag{1}$$

where $\boldsymbol{X}_i = (X_{i1}, \dots, X_{ip})^T$ are the explanatory variables or covariates, $\boldsymbol{\beta} = (\beta_1, \dots, \beta_p)^T \in \mathbb{R}^p$ is the vector of unknown regression coefficients and the U_is are random noise with $\boldsymbol{U} = (U_1, \dots, U_n) \in \mathbb{R}^n$ being normally distributed with null mean vector and variance covariance matrix $\sigma^2 \boldsymbol{I}_n$.

The term high-dimensional data is used when the number of explanatory variables, p, is greater than the number of observations by nonpolynomial dimensionality. On the other hand, sparse models are those whose number of true non-zero regression parameters is very low respect to the total number of covariates. This situation is accurate to real-life problems in several areas, such as genetics and genomic, bioinformatics, neuroimaging or chemometrics. Finally, it is known that contaminated data could worsen the estimation of the regression parameters. To avoid this issue, we need to develop robust estimating procedures. In this line, we are following the ideas by Castilla et al. (2020) [1] and Ghosh et al. (2020) [2].

The main awkward of the high dimensional regression models is the variable selection. As the number of possible models grows exponentially, information criteria are not suitable to choose the best model. Hence, regularization methods are clearly more convenient in these settings. Regularization methods introduce a penalty term, which penalizes the absolute value of the regression coefficients, on the objective function to achieve simultaneously model selection and parameter estimation. Regularization methods for sparse high-dimensional data analysis are characterized by loss functions measuring data fits and penalty terms constraining model parameters. In LRM, we estimate the parameter vector $(\boldsymbol{\beta}, \sigma) \in \mathbb{R}^{p+1}$ by minimizing an objective function of the form

$$Q_{n,\lambda}(\boldsymbol{\beta}, \sigma) = L_n(\boldsymbol{\beta}, \sigma) + \sum_{j=1}^{p} p_{\lambda_n}(|\beta_j|), \tag{2}$$

which consists of a data fit functional $L_n(\boldsymbol{\beta}, \sigma)$, called loss function, and a penalty function $\sum_{j=1}^{p} p_{\lambda_n}(|\beta_j|)$, assessing the physical plausibility of $\boldsymbol{\beta}$ and controlling the complexity of the fitted model in order to avoid overfitting. A regularization parameter λ_n ($\lambda_n \geq 0$) regulates the penalty. From a practical point of view, the regularization parameter is chosen using some information criterion or by cross-validation.

The most common penalties are $p_{\lambda_n}(s) = s^2$ for Ridge estimator and $p_{\lambda_n}(s) = |s|$ for the LASSO estimator. The first one does not achieve model selection as it is unable to detect the null regression coefficients, but is more convenient when there is multicolinearity. Further, there have been several generalizations of the LASSO penalty yielding consistent estimator of the active set under much weaker conditions. In this vein, we also consider the Adaptative-LASSO and the SCAD (smoothly clipped absolute deviation) penalties.

Respect to the loss function, the most common is the least squares function obtained by the maximum likelihood criterion. The lack of robustness of this quadratic function is known, so it must be replaced by a robust loss so as to limit the impact of contamination in the data.

2. Robust losses

Let us consider the linear regression model (1) on which we assume that $Y|\boldsymbol{X} = \boldsymbol{x}$ follows a normal $\mathcal{N}(\boldsymbol{x}^T\boldsymbol{\beta}, \sigma^2)$ depending on the regression parameter, and let us consider a random sample $(Y_i, \boldsymbol{X}_i)_{1,\dots,n}$ from the model whose empirical distribution is G_n.

The minimum distance approach aims to minimize "some kind of measure of the distance or the divergence" between the proposed distribution of $Y|\boldsymbol{X} = \boldsymbol{x}$ and its empirical version. We use two of these measures of proximity between two distributions, namely the density power divergence (DPD) and the Rényi's pseudodistance (RP). These two measures take the following form for the linear regression model:

$$L_{n,\alpha}^{\mathrm{DPD}}(\boldsymbol{\beta}, \sigma) = \frac{1}{(2\pi)^{\alpha/2}\sigma^{\alpha}} \left(\frac{1}{\sqrt{\alpha+1}} - \frac{\alpha+1}{\alpha} \frac{1}{n} \sum_{i=1}^{n} \exp\left\{ -\alpha \frac{(y_i - \boldsymbol{x}_i^T\boldsymbol{\beta}))^2}{2\sigma^2} \right\} \right) + \frac{1}{\alpha}, \tag{3}$$

$$L_{n,\alpha}^{\mathrm{RP}}(\boldsymbol{\beta}, \sigma) = \frac{1}{n} \sum_{i=1}^{n} -\sigma^{\frac{-\alpha}{\alpha+1}} \exp\left(-\frac{\alpha}{2} \left(\frac{y_i - \boldsymbol{x}_i^T\boldsymbol{\beta}}{\sigma} \right)^2 \right), \tag{4}$$

where $f_{\boldsymbol{x}^T\boldsymbol{\beta},\sigma^2}$ denotes the normal density with mean $\boldsymbol{x}^T\boldsymbol{\beta}$ and variance σ^2. Note that both depend on a tunning parameter $\alpha > 0$ which controls the trade-off between efficiency and robustness. The minimun DPD estimator (MDPDE) $(\hat{\boldsymbol{\beta}}_\alpha^{\text{DPD}}, \hat{\sigma}_\alpha^{\text{DPD}})$ and the minimun RP estimator (MRPE) $(\hat{\boldsymbol{\beta}}_\alpha^{\text{RP}}, \hat{\sigma}_\alpha^{\text{RP}})$ are defined as the values $(\boldsymbol{\beta}, \sigma)$ minimizing (3) and (4), respectively. Even more, both measures can be defined at $\alpha = 0$ as the log-likelihood function taking continuous limit in α. Hence, both approaches include the maximum likelihood estimator (MLE) for the value $\alpha = 0$. From a practical point of view, the main difference between these measures lies in the estimation of σ^2.

3. Penalized MDPDE and MRPE.

The regularization methods based on DPD and RP losses are constructed by including a penalty term to the objective function so as to achieve simultaneously model selection and parameter estimation. Therefore, our objective function is $Q_{n,\alpha,\lambda}(\boldsymbol{\beta}) = \tilde{L}_{n,\alpha,\lambda}(\boldsymbol{\beta}) + \sum_{j=1}^{p} p_\lambda(|\beta_j|)$ for a robust loss $\tilde{L}_{n,\alpha,\lambda}(\boldsymbol{\beta})$ (DPD or RP loss) and a penalty function $p_\lambda(\cdot)$. We are considering three different penalties to compare their performance, namely LASSO, Adaptive LASSO and a non-concave penalty SCAD.

- LASSO panalty: $p_\lambda(\beta_j) = \lambda \sum_{j=1}^{p} |\beta_j|$.
- Adaptive LASSO penalty: $p_\lambda(\beta_j) = \lambda \sum_{j=1}^{p} \frac{1}{|\tilde{\beta}_j|} \cdot |\beta_j|$, where $\tilde{\boldsymbol{\beta}}$ is a robust estimate of $\boldsymbol{\beta}$.
- Non-concave penalty SCAD: $p_\lambda(|\beta_j|) = \begin{cases} \lambda|\beta_j| & \text{if } |\beta_j| \leq \lambda, \\ \frac{2a\lambda|\beta_j| - |\beta_j|^2 - \lambda^2}{2(a-1)} & \text{if } \lambda < |\beta_j| \leq a\lambda, \\ \frac{(a+1)\lambda^2}{2} & \text{if } a\lambda < |\beta_j|, \end{cases}$ where $a = 3.7$.

3.1. Robustness of the proposed estimators

Local robustness of an estimator can be studied through its influence function (IF). The IF measures the possible asymptotic bias in the estimation due to an infinitesimal contamination, and an estimator is said robust if its IF is bounded. We can verify that the IF of the proposed estimators is bounded for $\alpha > 0$ and non-bounded for $\alpha = 0$ corresponding to the MLE. Figure 1 shows the IF of the MDPDEs and MRPEs for univariate linear regression with $\sigma_0 = 1$, $x_t = 1$ and $\mathbb{E}[x^2] = 1$. The abscissa axis corresponds to the perturbation $u = y - x\beta$ and the ordinate axis corresponds to the IF value.

4. Estimating algorithm

The basic idea of our proposed algorithm is to iteratively minimize the objective $Q_n(\boldsymbol{\beta}, \sigma)$ in two steps: we first update the current solution of the regression parameter $\boldsymbol{\beta}$ and then we minimize the error deviance σ. For the first step, we combine MM-algorithm and coordinate descent algorithm, adapted to each situation, so as to update $\boldsymbol{\beta}$. As mentioned before, this update is similar for both proposed losses, DPD and RP. For the second step, we approximate a solution of the estimating equations of σ, obtained by equating the first derivative of the objective function to zero.

5. Simulation Study

We finally carry out a simulation study so as to evaluate the robustness and efficiency of the proposal penalized MDPDE MNPRPE under the LRM. We also estimate the regression parameters $(\boldsymbol{\beta}, \sigma)$ using other existing robust and non-robust methods of high-dimensional LRM to compare their performances with our proposed method. For each one of the estimators, we calculate the mean square error (MSE) for the true non-zero and zero coefficients separately, Absolute Prediction Bias using an unused test sample generated in the same way as train data, True Positive proportion, True Negative proportion and Model Size of the estimated regression coefficient $\hat{\boldsymbol{\beta}}$, and Estimation Error of the estimate $\hat{\sigma}$.

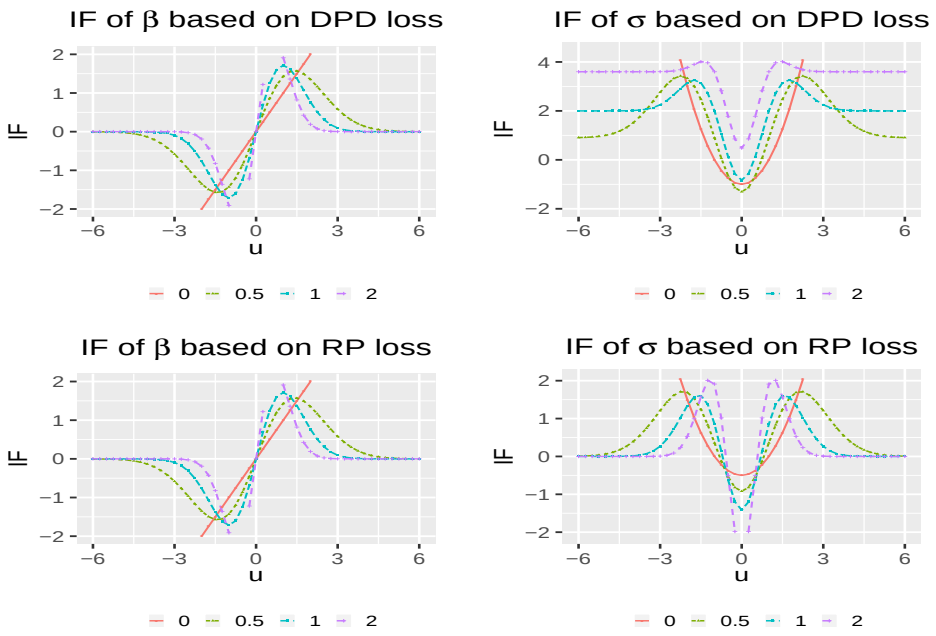

Figure 1: IF of the MDPDE for beta (upper left) and sigma (upper right), and IF of the MRPE for beta (bottom left) and sigma (bottom right).

Further, in order to examine the efficiency loss against non-robust methods in absence of any contamination, as well as compare the performance in the presence of contamination in the data, we consider different scenarios for data contamination, besides the pure data setting, including contaminating data in the responde variable Y and the explanatory variables $\boldsymbol{X}$.

The simulation results show the gain in robustness when the parameter α increases, as well as the improvement that the Adaptative LASSO and SCAD penalty entail for the variable selection. We conclude that the proposed estimators are very competitive to the classical MLE, and moreover, they perform better with contaminated data.

References

[1] Castilla, Elena; Ghosh, Abhik; Jaenada, María, and Pardo, Leandro. "On regularization methods based on Rényi's pseudodistances for sparse high-dimensional linear regression models". In: *arXiv e-prints* (2020). arXiv: 2007.15929 [math.ST].

[2] Ghosh, Abhik; Jaenada, Maria, and Pardo, Leandro. "Robust adaptive variable selection in ultra-high dimensional linear regression models". In: *arXiv e-prints* (2020). arXiv: 2004.05470 [stat.ME].