Qualifying Paper Report for Estimation of High Dimensional Mean Regression in the Absence of Symmetry and Light Tail Assumptions

Naitong Chen

November 14, 2022

1 Introduction

Linear regression is an easy-to-use and highly interpretable statistical inference method that has found itself in a wide range of applications. In the fields such as genomics and proteomics, we are often presented with high dimensional problems where the number of predictors is much greater than the number of observations. In such settings, linear regression in its simplest form may not have a unique solution, and so directly applying this method may not yield outputs that are interpretable for statistical inference. While a number of methods have been developed to address this issue, the corresponding theoretical guarantees on the inference quality for most of these methods rely on the assumption that the error distribution is symmetric and light-tailed. However, these assumptions may not be reasonable in practice, thus challenging the reliability of these methods. In this report, we discuss the regression estimator RA-lasso developed in Fan et al. (2017), which is designed to handle high dimensional data whose underlying error distribution may be neither symmetric nor light-tailed.

This report is organized as follows: for the remainder of this section, we more carefully motivate the RA-lasso estimator. In Section 2, we present the proposed method along with some intuition on its theoretical guarantees as well as some practical considerations when applying this method. We comment on their simulation studies in terms of how well they justify the claims made in the paper in Section 3 and reproduce their simulation studies in Section 4. This report is concluded with some final discussions of RA-lasso in Section 5.

In this report, we consider the linear regression model

$$y_i = x_i^T \beta^\star + \epsilon_i$$

where $\{x_i\}_{i=1}^n$ are independent and identically distributed p-dimensional covariate vectors and $\{\epsilon_i\}_{i=1}^n$ are independent and identically distributed errors. $\beta^\star$ is a p-dimensional regression coefficient vector that is assumed to be weakly sparse. In other words, many elements of $\beta^\star$ is exatly or close to 0. We assume that $p \gg n$ and that both the distributions of x and ϵ have mean 0. The goal is to find $\hat{\beta} \in \mathbb{R}^p$ such that $x^T \hat{\beta}$ is close to $\mathbb{E}_\epsilon[y|x] = x^T \beta^\star$, the conditional expectation of the response y given x.

If we consider $\{x_i\}_{i=1}^n$ to be fixed and further assume that $\text{Var}(\epsilon) = \sigma^2 < \infty$, then

$$\hat{\beta}_{OLS} \in \underset{\beta \in \mathbb{R}^p}{\arg\min} \frac{1}{n} \sum_{i=1}^n \left(y_i - x_i^T \beta\right)^2, \tag{1}$$

the ordinary least squares (OLS) estimator is the linear unbiased estimator for $\beta^\star$ with the lowest possible variance. While it is tempting to use OLS for all regression problems, as mentioned above, in the high dimensional setting, there may not be a unique solution to Eq. (1). Thus the OLS estimator in high dimensional settings likely will not be interpretable.

To handle this problem of having potentially multiple solutions, given some $\lambda > 0$, an L1

regularization term can be added to Eq. (1) to construct a different estimator:

$$\hat{\beta}_{LASSO} \in \underset{\beta \in \mathbb{R}^p}{\arg\min} \frac{1}{n} \sum_{i=1}^{n} \left(y_i - x_i^T \beta\right)^2 + \lambda \|\beta\|_1. \tag{2}$$

This is called the least absolute shrinkage and selection operator (LASSO). This method encourages sparsity in the estimated parameter $\hat{\beta}_{LASSO}$ with the λ parameter controlling the level of sparsity. This estimator has been shown to be consistent (i.e. $\hat{\beta}_{LASSO}$ converges in probability to $\beta^\star$ as n goes to infinity) under certain assumptions. However, this method is not robust to outliers where the response is unusually large or small compared to other observations with similar covariate vectors. We can see this from Eq. (2). In particular, the squared term means that this objective function penalizes large differences between the observed response y_i and fitted value $x_i^T \beta$ more so than it does smaller differences. As a result, a couple of outliers in the data set may cause the resulting estimated parameter to change drastically. This sensitivity to outliers is also reflected in one of the assumptions required for the LASSO to be consistent. Namely, one of these assumptions is that the distribution of ϵ is light-tailed, i.e., the tail of the error distribution does not decay slower than that of an exponential distribution. In other words, for the LASSO to be consistent, it is required that when the covariates of some observations are (close to) identical, there is not a substantial amount of variation among the corresponding responses.

In order to bypass this assumption of light-tailed error, which may not be reasonable in many cases in practice, a solution is to replace the squared loss with the absolute loss:

$$\hat{\beta}_{LAD} \in \underset{\beta \in \mathbb{R}^p}{\arg\min} \frac{1}{n} \sum_{i=1}^{n} \left|y_i - x_i^T \beta\right|. \tag{3}$$

This least absolute deviance (LAD) estimator is associated with estimating the conditional median of the response y given some covariate x. Median, as an alternative measure of centre, is more robust to outliers than the mean. Therefore, we can expect $\hat{\beta}_{LAD}$ to be less sensitive to the outliers described above. Note that we can combine Eq. (3) with an L1 regularization term term in order to get a robust estimate of the regression parameters that is also sparse.

However, for the LAD estimator combined with L1 regularization to be a good substitute for the LASSO, we are implicitly asking the median of the error distribution to be close to the mean. In the case that the median of the error distribution is equal to the mean, we are essentially making the implicit assumption that the error distribution is symmetric. When this implicit assumption does not hold, using the LAD estimator in place of the OLS estimator will introduce some systematic bias to our estimated regression parameter. Therefore, it is desirable to develop a high dimensional mean regresison estimator that is consistent even when the underlying error distribution is neither symmetric nor light-tailed.

2 Proposed method

2.1 Discussion on theoretical results

2.2 Practical implementation

3   Discussion on simulation

4 Reproduced and improved simulation studies

5   Conclusion

References

Jianqing Fan, Quefeng Li, and Yuyan Wang. Estimation of high dimensional mean regression in the absence of symmetry and light tail assumptions. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 79(1):247–265, 2017.