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

Naitong Chen

November 22, 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 (note that the paper also covers the heteroscedastic case where the error ϵ_i depends on x_i for all $i = 1, \ldots, n$, but we focus on the homoscedastic case in this report). $\beta^\star$ 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 $\mathrm{Var}(\epsilon) = \sigma^2 < \infty$, 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 implicitely 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

In the case where the error distribution is heavy-tailed and asymmettic, by using the absolute loss in place of the squared loss, we hope to obtain an regression estimator that is more robust to outliers at the cost of introducing some bias. The magnitude of this bias is dependent on the distribution of the error. However, note that the OLS estimator is unbiased for all error distributions with mean 0 and a finite variance. These observations motivate the idea of using a loss function that balances the unbiasedness of an OLS estimator and the robustness of an LAD estimator. One way of achieving this balance is through the use of the Huber loss with a blending parameter $\alpha \geq 0$:

$$l_\alpha(x) = \begin{cases} 2\alpha^{-1}|x| - \alpha^{-2} & \text{if } |x| > \alpha^{-1} \\ x^2 & \text{if } |x| \leq \alpha^{-1} \end{cases}. \tag{4}$$

In the Huber loss, α controls the blending between squared and absolute loss: $\alpha = 0$ corresponds to the squared loss and $\alpha = \infty$ corresponds to the absolute loss. Fan et al. (2017) proposes the penalized robust approximate (RA) quadratic loss using the Huber loss for high dimensional mean regression problems with an asymmetric and heavy-tailed error distribution. When L1 regularization is chosen to be the penalty term, for some $\alpha, \lambda \geq 0$, we arrive at the RA-lasso estimator

$$\hat{\beta}_{RA-lasso} = \arg\min_{\beta \in \mathbb{R}^p} \frac{1}{n} \sum_{i=1}^{n} l_\alpha \left(y_i - x_i^T \beta\right) + \lambda \|\beta\|_1.$$

By tuning the α parameter, we can trade off the balance between being unbiased and being robust to outliers. Note that the Huber loss Eq. (4) is convex and differentiable at $x = \alpha^{-1}$. Therefore, we can used gradient-based optimization to obtain the RA-lasso solution for some fixed α and λ.

2.1 Discussion on theoretical results

In this section, we provide a high level sketch of the theoretical guarantee on the quality of $\hat{\beta}_{RA-lasso}$ and offer connections to the motivation behind RA-lasso.

In Fan et al. (2017), the performance of $\hat{\beta}_{RA-lasso}$ is assessed through $\|\hat{\beta}_{RA-lasso} - \beta^\star\|_2$. More specifically, we can decompose this quantity as follows.

$$\|\hat{\beta}_{RA-lasso} - \beta^\star\|_2 \leq \|\beta_\alpha^\star - \beta^\star\|_2 + \|\hat{\beta}_{RA-lasso} - \beta_\alpha^\star\|_2,$$

where $\beta_\alpha^\star = \arg\min_{\beta \in \mathbb{R}^p} \mathbb{E}\left[l_\alpha(y - x^T\beta)\right]$. We call the first term in the above bound the approximation error and the second term the estimation error. Building on top of this, theorem 3 of Fan et al. (2017) states that, under the conditions of lemmas 1 and 3, for some $q \in (0, 1]$ and $k \geq 2$, with probability at least $1 - c_1 \exp(-c_2 n)$,

$$\|\hat{\beta}_{RA-lasso} - \beta^\star\|_2 \leq d_1 \alpha^{k-1} + d_2 \sqrt{R_q} \left(\frac{\log p}{n}\right)^{\frac{1}{2} - \frac{q}{4}}, \tag{5}$$

where the first term is a bound on the approximation error and the second term is a bound on the estimation error. The technical conditions and the selection of constants in full detail can be found in Section 2 of Fan et al. (2017). It is, however, worth pointing out that this above bound only applies to α and λ values satisfying the conditions in lemmas 1 and 3.

It is easy to see that, when the error distribution is symmetric (i.e. when the mean and median of the error distribution are euqal), $\beta_\alpha^\star = \beta^\star$ for all α, which implies that the approximation error is 0. When the error distribution is asymmetric, the bound on the approximation error can be thought of as the amount of bias introduced through the use of the Huber loss with blending parameter α. Note that this bound increases with α. Indeed, the larger we set α, the closer the Huber loss is to the absolute loss, which implies a larger bias induced by the RA-lasso estimator. This aligns with our intuition on the trade off between being unbiased and robust to outliers. The bound on the estimation error, on the other hand, is independent of α, and it converges to 0 as the number of observations n goes to infinity. By controlling the rate at which α increases, it is possible to have $\|\hat{\beta}_{RA-lasso} - \beta^\star\|_2$ converging to 0 at the rate of the bound on the estimation error from Eq. (5), which is the same as the optimal rate under the light tail situation Raskutti et al. (2011).

Before presenting the procedure for applying the RA-lasso estimator, we note that this estimator can be used as a mean estimator by treating it as a univariate linear regression problem where the covariate equals 1. Similarly, by noting that each element of the covariance matrix σ_{ij} for $i, j \in \{1, \cdots, p\}$ can be written as an expectation of the product between the ith and jth element of the centred covariate vector, the RA-lasso estimator can then also be used to estiamte the covariance matrix. Concentration inequalities for both of these cases are developed in Fan et al. (2017). Through this lens, Fan et al. (2017) also extended an mean estimator of heavy-tailed distributions developed in Catoni (2012) to the linear regression setting and developed a corresponding high probability bound similar to that of Eq. (5). However, these extensions are out of the scope of this report, and are thus not discussed in any more details here.

2.2 Practical implementation

While the bound from Eq. (5) aligns with our intuition on the trade off between unbiasedness and robustness against outliers, this result only holds for specific values of α and λ. In addition, the choice of α and λ values depend on universal constants that are typically not known in practice. As a result, Fan et al. (2017) suggests selecting α and λ by a two-dimensional grid search using cross-validation for a set of values that minimize the mean-squared error or an information-based criterion. More specifically, it is suggested that the search grid is formed by values that are equally spaced in log scale. Although not explicitely stated in the paper, it is recommended to form a search grid that contains candidate α and λ values ranging across multiple orders of magnitude. Algorithm 1 outlines the procedure for applying the RA-lasso estimator, where the hyperparameters α and λ are chosen from candidate values $(\alpha_i)_{i=1}^{C_\alpha}$ and $(\lambda_i)_{i=1}^{C_\lambda}$ using K-fold cross-validation with the mean-squared

loss. Note that we write

$$\forall j \in \{1, \cdots, p\}, \quad x^j = \begin{bmatrix} x_{1j} & \cdots & x_{nj} \end{bmatrix}^T \in \mathbb{R}^n, \quad \text{and} \quad y = \begin{bmatrix} y_1 & \cdots & y_n \end{bmatrix}^T \in \mathbb{R}^n.$$

Note that since the RA-lasso objective is convex, we can use composite gradient descent to obtain the RA-lasso estimate (using optimization packages such as `CVXR` in `R` or `CVX` in `MATLAB`). Fan et al. (2017) shows that this algorithm, with high probability, has the same convergence rate as Eq. (5).

Algorithm 1 RA-lasso procedure

Require: $(x_i, y_i)_{i=1}^n$, $(\alpha_i)_{i=1}^{C_\alpha}$, $(\lambda_i)_{i=1}^{C_\lambda}$, K

1. standardize the response and each of the p predictors to have mean 0 and unit variance
   $\tilde{y} = (y - \texttt{mean}(y))/\texttt{sd}(y)$;
   $\forall j = 1, \cdots, p, \quad \tilde{x}^j = (x^j - \texttt{mean}(x^j))/\texttt{sd}(x^j)$
2. randomly split the index set $\mathcal{I} = \{1, \ldots, n\}$ into K (equally sized) disjoint sets $\mathcal{I}_1, \cdots, \mathcal{I}_K \subset \mathcal{I}$ such that $\mathcal{I}_1 \cup \cdots \cup \mathcal{I}_K = \mathcal{I}$
3. perform K-fold cross-validation
   `cv_loss = matrix(0,`C_α, C_λ`)`
   for `i in 1,...,`C_α **do**
   for `j in 1,...,`C_λ **do**
   for `k in 1,...,K` **do**
   $\hat{\beta} = \arg\min_{\beta \in \mathbb{R}^p} \frac{1}{n - |\mathcal{I}_k|} \sum_{m \notin \mathcal{I}_k} l_{\alpha_i}\left(\tilde{y}_m - \tilde{x}_m^T \beta\right) + \lambda_j \|\beta\|_1$
   `cv_loss[i,j]` $+= \sum_{m \in \mathcal{I}_k} \left(\tilde{y}_m - \tilde{x}_m^T \hat{\beta}\right)^2$
   end for
   end for
   end for
   $i^\star, j^\star =$ `which(cv_loss == min(cv_loss), arr.ind = TRUE)`
4. refit the model using all data with the selected hyperparameters
   $\tilde{\beta}_{RA-lasso} = \arg\min_{\beta \in \mathbb{R}^p} \frac{1}{n} \sum_{i=1}^n l_{\alpha_{i^\star}}\left(\tilde{y}_i - \tilde{x}_i^T \beta\right) + \lambda_{j^\star} \|\beta\|_1$
5. destandardize the coefficients back to original scale (optional)
   $\hat{\beta}_{RA-lasso,0} = \texttt{mean}(y) - \sum_{j=1}^p \frac{\texttt{sd}(y)}{\texttt{sd}(x^j)} \texttt{mean}(x^j) \tilde{\beta}_{RA-lasso,j}$ ▷ intercept term
   $\forall j = 1, \ldots, p, \quad \hat{\beta}_{RA-lasso,j} = \frac{\texttt{sd}(y)}{\texttt{sd}(x^j)} \tilde{\beta}_{RA-lasso,j}$ ▷ slope terms

3 Discussion on simulation

4 Reproduced and improved simulation studies

5 Conclusion

References

Olivier Catoni. Challenging the empirical mean and empirical variance: a deviation study. In *Annales de l'IHP Probabilités et statistiques*, volume 48, pages 1148–1185, 2012.

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.

Garvesh Raskutti, Martin J Wainwright, and Bin Yu. Minimax rates of estimation for high-dimensional linear regression over ℓ_q-balls. *IEEE transactions on information theory*, 57 (10):6976–6994, 2011.