The Least-Squares Estimator for Shuffled Linear Regression is Intractable or Inconsistent (or Both)

Abubakar Abid[1], James Zou[2]

[1]Department of Electrical Engineering, Stanford University
[2]Department of Biomedical Data Sciences, Stanford University

October 31, 2018

Abstract

The purpose of this technical note is to concisely describe the statistical and computational limitations of the common least-squares (LS) estimator, when applied to the problem of shuffled linear regression. Here, we reference results that show that the LS estimator is NP-hard when the dimensionality of the input features is greater than one. When the dimensionality of the features *is* one, the LS estimator is nevertheless inconsistent. These results motivate the development of other estimators for shuffled regression, and we reference works in this direction.

1 Introduction

This note concerns *shuffled* linear regression, a variant of classical linear regression in which the mutual ordering of the input features and labels is not known. Alternatively, one can consider the labels as perturbed by an unknown permutation during the generative process. More concretely, the learning setting is defined as follows: we observe (or choose) a matrix of input features $X \in \mathbb{R}^{n \times d}$, and observe a vector of output labels $\boldsymbol{y} \in \mathbb{R}^n$ that is generated as follows:

$$\boldsymbol{y} = \Pi_0 X \boldsymbol{w}_0 + \boldsymbol{e} \tag{1}$$

where Π_0 is an unknown $n \times n$ permutation matrix, $\boldsymbol{w}_0 \in \mathbb{R}^d$ are unknown coefficients, and $\boldsymbol{e} \in \mathbb{R}^n$ is additive Gaussian noise drawn from $\mathcal{N}(0, \sigma^2)$. Here, n is the number of data points, and d is the dimensionality of the input features.

The generative model described by (1) and illustrated in Figure 1 is motivated by certain experiments and datasets that simultaneously analyze a large number of objects, such as flow cytometry [1]. Shuffled regression is also useful in other contexts where the order of measurements is unknown, such as simultaneous pose-and-correspondence estimation in computer vision [2] and in relative dating from archaeological samples [3]. An important setting where the feasibility of shuffled regression raises concern is data de-anonymization, such as of public medical records, which are sometimes shuffled to preserve privacy [4].

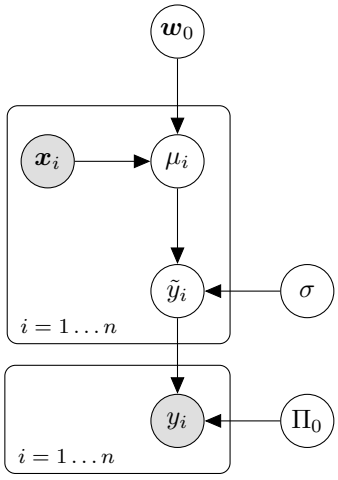

Figure 1: **Generative model for shuffled linear regression**. We show the generative model for the data as a graphical model, where arrows signify probabilistic or deterministic dependencies. Observed data are shaded and unobserved variables are clear. The parameters outside of the rounded rectangles are shared across all n data points. The relationship between variables is as follows: $\mu_i = \boldsymbol{x}_i \cdot \boldsymbol{w}_0$, $\tilde{y}_i = \mu_i + \mathcal{N}(\mu_i, \sigma^2)$, $\boldsymbol{y} = \Pi_0 \cdot \tilde{\boldsymbol{y}}$. The goal of shuffled linear regression is to estimate the latent variable w_0. This figure is reproduced from [5] with permission.

2 Least-Squares (LS) Estimator

The standard estimator for shuffled linear regression is the least-squares (LS) estimator, which has also been referred to as the "maximum-likelihood estimator" [6], and denotes the following optimization problem to estimate the linear coefficients:

$$\hat{\boldsymbol{w}} = \arg\min_{\boldsymbol{w}} \min_{\Pi \in \mathcal{P}} ||\boldsymbol{y} - \Pi X \boldsymbol{w}||^2, \tag{2}$$

where $\mathcal{P}$ is the set of all $n \times n$ permutation matrices. It is a natural extension of the ordinary least squares (OLS) estimator for shuffled linear regression. Approximations to the LS estimator based on alternating optimization have been studied in [5, 7].

3 Limitations of the LS Estimator

3.1 NP-hardness: $d > 1$

In [6], the authors studied the computational properties of the LS estimator. Their results, as described in Theorem 4 of the original paper, showed that there exist inputs X for which the shuffled linear regression problem is NP-hard, which they proved as a reduction from the NP-hard partition problem. While the authors did not prove these results for randomly-chosen input features, they conjectured that the same NP-hardness would extend to the random design case.

The authors state, however, that their hardness results do not apply when the dimensionality of the input data is 1. In fact, they propose a simple algorithm that runs in $O(n \log n)$ time, which sorts the labels $\boldsymbol{y}$ in the same order as the input features X, and then performs ordinary linear regression. However, as we note next, the LS estimator suffers a different problem when $d = 1$.

3.2 Inconsistency: $d = 1$

In [8], the authors specifically examined the statistical properties of the LS estimator for the case of $d = 1$. They proved that, in the presence of additive noise, that the LS estimator is inconsistent. For

the inconsistency results, stated in Theorem 1 of the original paper, the authors chose a random-design setting for X. Yet, their empirical results showed that the *amplification bias* of $\hat{w}$, which they proved occurs in the case of infinite samples, was also present in fixed-design input features. The authors conjectured that their results regarding inconsistency extended to higher dimensions.

4 Conclusion and Future Work

In tandem, the two results that we have referenced in this paper show that, for any dimensionality, the least-squares estimator suffers a statistical or computational limitation. However, this is not meant to suggest that there exists *no* consistent, efficient estimator for shuffled linear regression. For example, the authors of [8] suggest a simple estimator based on the method of moments for the case $d = 1$ which is both efficient and consistent under mild conditions.

These results motivate the development of other estimators for shuffled linear regression. For example, the authors of [5] and [9] develop separate polynomial-time approximations to the maximum-likelihood estimate of the weights. However, these estimators yield rough approximations, and we believe that an open area of research remains to develop regression algorithms that can be practically applied to real datasets with shuffled labels.

References

[1] Howard M Shapiro. *Practical Flow Cytometry*. John Wiley & Sons, 2005.

[2] Philip David, Daniel Dementhon, Ramani Duraiswami, and Hanan Samet. Softposit: Simultaneous pose and correspondence determination. *International Journal of Computer Vision*, 59(3):259–284, 2004.

[3] William S Robinson. A method for chronologically ordering archaeological deposits. *American Antiquity*, 16(4):293–301, 1951.

[4] Jingquan Li and Michael J Shaw. Protection of health information in data mining. *International journal of Healthcare Technology and Management*, 6(2):210–222, 2004.

[5] Abubakar Abid and James Zou. A stochastic expectation-maximization approach to shuffled linear regression. *Allerton Conference on Communication, Control, and Computing*, 2018.

[6] Ashwin Pananjady, Martin J Wainwright, and Thomas A Courtade. Linear regression with an unknown permutation: Statistical and computational limits. *Allerton Conference on Communication, Control, and Computing*, 2016.

[7] Guanyu Wang, Jiang Zhu, Rick S Blum, Paolo Braca, and Zhiwei Xu. Maximum likelihood signal amplitude estimation based on permuted blocks of differently binary quantized observations of a signal in noise. *arXiv preprint arXiv:1706.01174*, 2017.

[8] Abubakar Abid, Ada Poon, and James Zou. Linear regression with shuffled labels. *arXiv preprint arXiv:1705.01342*, 2017.

[9] Daniel J Hsu, Kevin Shi, and Xiaorui Sun. Linear regression without correspondence. In *Advances in Neural Information Processing Systems*, pages 1531–1540, 2017.