TEST (2023) 32:1172–1176
https://doi.org/10.1007/s11749-023-00880-z

DISCUSSION

# Comments on: Statistical inference and large-scale multiple testing for high-dimensional regression models

Ye Tian[1] · Yang Feng[2]

Received: 20 July 2023 / Accepted: 3 August 2023 / Published online: 29 August 2023

**Mathematics Subject Classification** 62F12 · 62J05

## 1 Introduction

We extend our congratulations to the authors for their outstanding survey on statistical inference for high-dimensional regression models. Their comprehensive work serves as a valuable overview on the existing literature, encompassing the origins of high-dimensional inference as well as recent advancements in the field.

The significance of high-dimensional inference cannot be overstated, considering the soaring number of features in modern datasets over the past two decades. In response to this challenge, penalized regression models such as Ridge regression, Lasso, SCAD, group Lasso, and elastic net were proposed and widely adopted in practical applications. While consistent estimation and prediction are crucial, they alone do not provide a complete picture in scientific investigations. To establish convincing conclusions, it is imperative to quantify the uncertainties associated with these penalized regression models using tools like confidence intervals and hypothesis testing. Additionally, inference techniques offer valuable insights for effective variable selection.

In this paper, the authors have thoughtfully summarized the historical foundations of debiased methods and presented recent progress in high-dimensional inference and simultaneous hypothesis testing. Section 2 delves into the inference on coefficients of linear regression models, as well as binary-outcome models such as logistic regression and Probit regression models. Building upon this, Section 3 explores inference on

Comments on: https://doi.org/10.1007/s11749-023-00870-1.

✉ Yang Feng
yang.feng@nyu.edu

[1] Department of Statistics, Columbia University, New York, USA

[2] Department of Biostatistics, School of Global Public Health, New York University, New York, USA

linear and quadratic functionals of regression coefficients. The authors then proceed to discuss simultaneous inference and multiple testing in high-dimensional generalized linear models (GLMs) in Sect. 4. Lastly, Sect. 5 provides an overview of several related inference problems.

Of particular interest is the authors' mention of learning from multiple heterogeneous regression models in the concluding section. In the subsequent part of our discussion, we will focus on this topic, emphasizing the importance of statistical inference from heterogeneous regression models and highlighting the associated challenges. In the end, we will outline potential avenues for future research in this area.

## 2 Knowledge transfer in statistical inference

The problem of learning from multiple heterogeneous regression models has been extensively studied for over two decades, and researchers from various domains have investigated a few related learning problems that focus on distinct objectives, such as transfer learning, multi-task learning, meta-learning, and federated learning. These problems specifically address the challenge of learning from models that share similarities but are not identical. This differs from distributed learning, where the aim is to learn from the same model across different local machines.

In the context of high-dimensional problems, leveraging information from similar models can provide several advantages. First, it can enhance the effective sample size by combining observations from multiple sources. This has the potential to relax the stringent sample size requirements typically associated with high-dimensional inference. Additionally, incorporating information from similar models can lead to more accurate quantification of uncertainty in high-dimensional models, facilitating scientific discoveries.

Although numerous empirical and theoretical studies have explored knowledge transfer from similar models over the past two decades, most of them have focused on parameter estimation and out-of-sample prediction, with limited discussion on inference in this context. However, a few noteworthy studies have addressed the inference problem. For instance, Tian and Feng (2022) and Li et al. (2023) proposed inference methods for multiple generalized linear models (GLMs). Liu et al. (2021) investigated false discovery rate (FDR) and false positive rate (FPR) control when learning from multiple regression models with shared support (group sparsity). Li et al. (2022b) discussed inference for Gaussian graphical models, while Li et al. (2022a) examined inference under a transfer learning setting when the source data are unlabeled. Additionally, Zhou et al. (2022) explored inference problems within the context of semi-supervised learning and causal inference. Some of these studies justified that inference from multiple similar models can provide more accurate estimation of certain components in confidence intervals (Tian and Feng 2022) and relax the sample size condition (Li et al. 2022a, b; Li et al. 2023). However, it is important to note that the entry-wise confidence interval (CI) obtained in these papers still has a width of

$O(n_0^{-1/2})$, where $n_0$ represents the sample size of the target model in the transfer learning problem. Therefore, these CIs do not improve upon the classical CI obtained from considering only the target model, in terms of the rate. Furthermore, the optimality of these CIs remains unknown.

To shed light on the challenges involved in performing statistical inference within the context of transfer learning, let us consider a simple example. Suppose we step back and examine a low-dimensional regression problem where we observe i.i.d. data $\{\boldsymbol{x}_i^{(0)}, y_i^{(0)}\}_{i=1}^{n_0}$ and $\{\boldsymbol{x}_i^{(1)}, y_i^{(1)}\}_{i=1}^{n_1}$ from linear regression models:

$$y^{(k)} = (\boldsymbol{x}^{(k)})^T \boldsymbol{\beta}^{(k)} + \epsilon_i^{(k)},$$

where $\boldsymbol{x}^{(k)} \in \mathbb{R}^p$, $\boldsymbol{x}^{(k)} \sim N(0, \boldsymbol{\Sigma})$, $\epsilon_i^{(k)} \sim N(0, \sigma^2) \perp \boldsymbol{x}^{(k)}$, $p$ is fixed, $n_1 \gtrsim n_0$, and $k = 0, 1$. We assume that the target and source models are similar in the sense that $\|\boldsymbol{\beta}^{(1)} - \boldsymbol{\beta}^{(0)}\|_2 = h$, where $h$ is unknown and may change with $n_0$ and $n_1$. For simplicity, we assume that $\boldsymbol{\Sigma}$ and $\sigma$ are known. The models with $k = 0$ and $k = 1$ are referred to as the target and source models, respectively.

The objective is to construct confidence intervals (CIs) for each entry of the target coefficient $\boldsymbol{\beta}^{(0)}$ with the shortest width.

Before delving into the CIs, we need to establish a point estimator for $\boldsymbol{\beta}^{(0)}$. Several intuitive options can be considered:

(i) Target-only OLS: $\widehat{\boldsymbol{\beta}}^{(0)} = [(\boldsymbol{X}^{(0)})^T \boldsymbol{X}^{(0)}]^{-1} (\boldsymbol{X}^{(0)})^T \boldsymbol{Y}^{(0)}$;

(ii) Weighted average of target and source OLS: $\widehat{\boldsymbol{\beta}}_{\text{wa}}^{(0)} = \frac{n_0}{n_0+n_1}\widehat{\boldsymbol{\beta}}^{(0)} + \frac{n_1}{n_0+n_1}\widehat{\boldsymbol{\beta}}^{(1)}$, where $\widehat{\boldsymbol{\beta}}^{(1)} = [(\boldsymbol{X}^{(1)})^T \boldsymbol{X}^{(1)}]^{-1} (\boldsymbol{X}^{(1)})^T \boldsymbol{Y}^{(1)}$;

(iii) An adaptive estimator: $\widehat{\boldsymbol{\beta}}_{\text{ad}}^{(0)} = \widehat{\boldsymbol{\beta}}_{\text{wa}}^{(0)}$ if $\|\widehat{\boldsymbol{\beta}}^{(1)} - \widehat{\boldsymbol{\beta}}^{(0)}\|_2 \leq \sqrt{1/n_0}$, and $\widehat{\boldsymbol{\beta}}_{\text{ad}}^{(0)} = \widehat{\boldsymbol{\beta}}^{(0)}$ otherwise.

It can be shown that under certain regularity conditions, we have $\|\widehat{\boldsymbol{\beta}}^{(0)} - \boldsymbol{\beta}^{(0)}\|_2 = O_p(\sqrt{1/n_0})$ and $\|\widehat{\boldsymbol{\beta}}_{\text{wa}}^{(0)} - \boldsymbol{\beta}^{(0)}\|_2 = O_p(\sqrt{1/(n_0+n_1)} + h)$. The adaptive estimator $\widehat{\boldsymbol{\beta}}_{\text{ad}}^{(0)}$ combines $\widehat{\boldsymbol{\beta}}^{(0)}$ and $\widehat{\boldsymbol{\beta}}_{\text{wa}}^{(0)}$, achieving the $\ell_2$-minimax estimation error of $O_p(\sqrt{1/(n_0+n_1)} + h \wedge \sqrt{1/n_0})$. However, constructing a valid CI based on $\widehat{\boldsymbol{\beta}}_{\text{ad}}^{(0)}$ poses a non-trivial challenge. For $\widehat{\boldsymbol{\beta}}^{(0)}$ and $\widehat{\boldsymbol{\beta}}_{\text{wa}}^{(0)}$, some basic algebra reveals that:

$$\mathbb{P}\Big(\beta_j^{(0)} \in \Big[\widehat{\beta}_j^{(0)} - \frac{\sigma}{\sqrt{n_0}}\Omega_{jj}\alpha_{q/2}, \widehat{\beta}_j^{(0)} + \frac{\sigma}{\sqrt{n_0}}\Omega_{jj}\alpha_{q/2}\Big]\Big) \to 1 - q,$$
$$\mathbb{P}\Big(\bar{\beta}_j \in \Big[\widehat{\beta}_{\text{wa},j}^{(0)} - \frac{\sigma}{\sqrt{n_0+n_1}}\Omega_{jj}\alpha_{q/2}, \widehat{\beta}_{\text{wa},j}^{(0)} + \frac{\sigma}{\sqrt{n_0+n_1}}\Omega_{jj}\alpha_{q/2}\Big]\Big) \to 1 - q,$$

where $\bar{\boldsymbol{\beta}} = \frac{n_0}{n_0+n_1}\boldsymbol{\beta}^{(0)} + \frac{n_1}{n_0+n_1}\boldsymbol{\beta}^{(1)}$, $\boldsymbol{\Omega} = \boldsymbol{\Sigma}^{-1}$, and $\alpha_{q/2}$ is the $q/2$-upper quantile of $N(0, 1)$. In cases where $h \ll \sqrt{1/(n_0+n_1)}$, these results imply that $\mathbb{P}(\beta_j^{(0)} \in [\widehat{\beta}_{\text{wa},j}^{(0)} - \frac{\sigma}{\sqrt{n_0+n_1}}\Omega_{jj}\alpha_{q/2}, \widehat{\beta}_{\text{wa},j}^{(0)} + \frac{\sigma}{\sqrt{n_0+n_1}}\Omega_{jj}\alpha_{q/2}]) \to 1-q$. Based on these findings, it might be intuitively tempting to make the following conjectures regarding optimal confidence intervals (CIs) that provide correct coverage with the shortest length, up to constants:

(i) When $h \gtrsim \sqrt{1/n_0}$, the $(1-q)$-CI $[\widehat{\beta}_j^{(0)} - \frac{\sigma}{\sqrt{n_0}}\Omega_{jj}\alpha_{q/2}, \widehat{\beta}_j^{(0)} + \frac{\sigma}{\sqrt{n_0}}\Omega_{jj}\alpha_{q/2}]$ is optimal;

(ii) When $h \ll \sqrt{1/(n_0+n_1)}$, the $(1-q)$-CI $[\widehat{\beta}_{\mathrm{wa},j}^{(0)} - \frac{\sigma}{\sqrt{n_0+n_1}}\Omega_{jj}\alpha_{q/2}, \widehat{\beta}_{\mathrm{wa},j}^{(0)} + \frac{\sigma}{\sqrt{n_0+n_1}}\Omega_{jj}\alpha_{q/2}]$ is optimal.

Unfortunately, conjecture (i) is incorrect. For example, consider a scenario where $\boldsymbol{\beta}^{(0)}$ and $\boldsymbol{\beta}^{(1)}$ differ only at one entry but are identical at all the remaining entries. In such cases, the $(1-q)$-CI $[\widehat{\beta}_{\mathrm{wa},j}^{(0)} - \frac{\sigma}{\sqrt{n_0+n_1}}\Omega_{jj}\alpha_{q/2}, \widehat{\beta}_{\mathrm{wa},j}^{(0)} + \frac{\sigma}{\sqrt{n_0+n_1}}\Omega_{jj}\alpha_{q/2}]$ may be narrower than the classical CI $[\widehat{\beta}_j^{(0)} - \frac{\sigma}{\sqrt{n_0}}\Omega_{jj}\alpha_{q/2}, \widehat{\beta}_j^{(0)} + \frac{\sigma}{\sqrt{n_0}}\Omega_{jj}\alpha_{q/2}]$ for those identical entries. This discrepancy arises because the similarity metric, the $\ell_2$-norm, which is useful for prediction, may not be compatible with entry-wise inference, whereas the $\ell_\infty$-norm is more directly related. Moreover, in cases where $\sqrt{1/n_0} \gg h \gtrsim \sqrt{1/(n_0+n_1)}$, it appears to be nontrivial to formulate a valid conjecture based on the aforementioned estimators.

In high-dimensional scenarios where the number of predictors $p$ grows with $n_0$ and $n_1$, the discrepancy between the similarity metric (the $\ell_2$-norm) and the $\ell_\infty$-norm becomes more pronounced, making inference even more complex. Additionally, the introduction of penalization introduces extra bias, which compounds the bias resulting from the differences between the target and source models. The interplay of these two biases exacerbates the challenges associated with inference.

In conclusion, statistical inference from multiple similar models remains relatively underexplored in the transfer learning and multi-task learning literature, despite its practical importance. The gap between theoretical studies and practical significance highlights the need for further attention and research in this promising direction. In addition, some related possible future work includes studying the inference problem for unsupervised learning (Tian et al. 2022) and under other types of similarity measures (Gu et al. 2022; Tian et al. 2023).

## References

Gu T, Han Y, Duan R (2022) Robust angle-based transfer learning in high dimensions. arXiv:2210.12759

Li S, Cai TT, Li H (2022a) Estimation and inference with proxy data and its genetic applications. arXiv:2201.03727

Li S, Cai TT, Li H (2022b) Transfer learning in large-scale gaussian graphical models with false discovery rate control. J Am Stat Assoc 1–13

Li S, Zhang L, Cai TT, Li H (2023) Estimation and inference for high-dimensional generalized linear models with knowledge transfer. J Am Stat Assoc 1–12

Liu M, Xia Y, Cho K, Cai T (2021) Integrative high dimensional multiple testing with heterogeneity under data sharing constraints. J Mach Learn Res 22(1):5607–5632

Tian Y, Feng Y (2022) Transfer learning under high-dimensional generalized linear models. J Am Stat Assoc 1–14

Tian Y, Weng H, Feng Y (2022) Unsupervised multi-task and transfer learning on gaussian mixture models. arXiv:2209.15224

Tian Y, Gu Y, Feng Y (2023) Learning from similar linear representations: adaptivity, minimaxity, and robustness. arXiv:2303.17765

Zhou D, Liu M, Li M, Cai T (2022) Doubly robust augmented model accuracy transfer inference with high dimensional features. arXiv:2208.05134

**Publisher's Note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.