Turning the information-sharing dial: efficient inference from different data sources

Emily C. Hector and Ryan Martin
Department of Statistics, North Carolina State University

Abstract

A fundamental aspect of statistics is the integration of data from different sources. Classically, Fisher and others were focused on *how* to integrate homogeneous (or only mildly heterogeneous) sets of data. More recently, as data are becoming more accessible, the question of *if* data sets from different sources should be integrated is becoming more relevant. The current literature treats this as a question with only two answers: integrate or don't. Here we take a different approach, motivated by information-sharing principles coming from the shrinkage estimation literature. In particular, we deviate from the do/don't perspective and propose a dial parameter that controls the *extent* to which two data sources are integrated. How far this dial parameter should be turned is shown to depend, for example, on the informativeness of the different data sources as measured by Fisher information. In the context of generalized linear models, this more nuanced data integration framework leads to relatively simple parameter estimates and valid tests/confidence intervals. Moreover, we demonstrate both theoretically and empirically that setting the dial parameter according to our recommendation leads to more efficient estimation compared to other binary data integration schemes.

Keywords: Data enrichment, generalized linear models, Kullback–Leibler divergence, ridge regression, transfer learning.

1 Introduction

Statistics aims to glean insights about a population by aggregating samples of individual observations, so *data integration* is at the core of the subject. In recent years, a keen interest in combining data—or statistical inferences—from multiple studies has emerged in both statistical and domain science research (Chatterjee et al., 2016; Jordan et al., 2018; Yang and Ding, 2019; Michael et al., 2019; Lin and Lu, 2019; Chen et al., 2019; Tang et al., 2020; Cahoon and Martin, 2020). While each sample is typically believed to be a collection of observations from one population, there is no reason to believe another sample would also come from the same population. This difficulty has given rise to a large body of literature for performing data integration in the presence of heterogeneity; see Claggett et al. (2014); Wang et al. (2018); Duan et al. (2019); Cai et al. (2019); Hector and Song (2020) and references therein for examples. These methods take an all-or-nothing approach to data integration: either the two sources of data can be integrated or not. This binary view of what can, or even should, be integrated is at best impractical and at worst useless when confronted with the messy realities of real data. Indeed, two samples are often similar enough that inferences can be improved with a joint analysis despite differences in the underlying populations. It does us statistical harm to think in these binary terms when data sets come from similar but not identical populations: assuming data integration is wholly invalid leads to reduced efficiency whereas assuming it is wholly valid can unknowingly produce biased estimates (Higgins and Thompson, 2002). It is thus not practical to think of two data sets as being either entirely from the same population or not. This begs the obvious but non-trivial (multi-part) question: What and how much can be learned from a sample given additional samples from *similar* populations, and how to carry out this learning process?

Towards answering this question, we consider a simple setup with two data sets: one for which a generalized linear model has already been fit, and another for which we wish to fit the same generalized linear model. (The case with more than two data sets is discussed in Section 2.1 below.) The natural question is if inference based on the second data set can be improved in some sense by incorporating results from the analysis of the first data set. This problem has been known under many different names, of which "transfer learning" is the most recently popular (Pan and Yang, 2010; Zhuang et al., 2021). The predominant transfer learning approach uses freezing and unfreezing of layers in a neural network to improve prediction in a target data set given a single source data set (Tan et al., 2018; Weiss et al., 2016). The prevailing insight into why deep transfer learning performs well in practice is that neural networks learn general data set features in the first few layers that are not specific to any task, while deeper layers are task specific (Yosinski et al., 2014; Dube et al., 2020; Raghu et al., 2019). This insight, however, does not give any intuition into or quantification of the similarity of the source and target data sets, primarily because of a lack of model interpretability. More importantly, this approach fails to provide the uncertainty quantification required for statistical inference. Similarly, transfer learning for high-dimensional regression in generalized linear models aims to improve predictions and often does not provide valid inference (Li et al., 2020, 2022; Tian and Feng, 2022). Another viewpoint casts this problem in an online inference framework (Schifano et al., 2016; Toulis and Airoldi, 2017; Luo and Song, 2021) that assumes the true value of the parameter of interest in two sequentially considered data sets is the same.

In contrast, we aim to tailor the inference in one data set according to the amount of relevant information in the other data set. Our goal is to provide a nuanced approach to data integration, one in which the integration tunes itself according to the amount of information available. In this sense, we do not view data sets as being from the same population or not, nor data integration as being valid or invalid, and we especially do not aim to provide a final inference that binarizes data integration into a valid or invalid approach. Rather, our perspective is that there is a continuum of more to less useful integration of information from which to choose, where we use the term "useful" to mean that we are minimizing bias and variance of model parameter estimates.

This new and unusual perspective motivates our definition of *information-driven* regularization based on a certain Kullback–Leibler divergence penalty. This penalty term allows us to precisely quantify what and how much information is borrowed from the first data set when inferring parameters in the second data set. We introduce a so-called "dial" parameter that controls how much information is borrowed from the first data set to improve inference in the second data set. We prove that there exists a range of values of the dial parameter such that our proposed estimator has reduced mean squared error over the maximum likelihood estimator that only considers the second data set. This striking result indicates that there is always a benefit to integrating the data sets, but that the amount of information integrated depends entirely on the similarity between the source and target. Based on this result, we propose a choice of the dial parameter that calibrates the bias-variance trade-off to minimize the mean squared error, and show how to construct confidence intervals with our biased parameter estimator. Finally, we demonstrate empirically the superiority of our approach over alternative approaches, and show empirically that our estimator is robust to differences between the source and target data sets. Due to its disjointed nature, relevant literature is discussed throughout.

We describe the problem set-up and proposed information-driven shrinkage approach in Section 2. Section 3 gives the specific form of our estimator in the linear and logistic regression models. Theoretical properties of our estimator are established in Section 4, demonstrating its efficiency compared to the maximum likelihood estimator. We investigate the empirical performance of our estimator through simulations in Section 5 and a data analysis in Section 6. An R package implementing our methods is available at `github.com/ehector/ISEDI`.

2 A continuum of information integration

2.1 Problem setup

Consider two populations, which we will denote as Populations 1 and 2. Units are randomly sampled from the two populations and the features $\boldsymbol{X}_{ij} \in \mathbb{R}^p$ and $y_{ij} \in \mathbb{R}$ are measured on unit i from Population j, with $i = 1, \ldots, n_j$ and $j = 1, 2$. Note that, while the values of these measurements will vary across i and j, the features being measured are the same across the two populations. Write $\mathcal{D}_j = \{(\boldsymbol{X}_{ij}, y_{ij}) : i = 1, \ldots, n_j\}$ for the data set sampled from Population j, for $j = 1, 2$. Note that we have not assumed any relationship between the two populations, only that we have access to independent samples consisting of measurements

of the same features in the two populations. The reader can keep in mind the prototypical example where a well designed observational study or clinical trial has collected a set of high quality target data $\mathcal{D}_2$ and the investigator has at his/her disposal a set of source data $\mathcal{D}_1$ of unknown source and quality to use for improved inference.

While our focus is on the case of one source data set, our information-sharing method described below can be applied more generally. Indeed, if an investigator has M-many source data sets from M populations measuring the same outcome and features, as in Section 6 below, then they might consider creating a single source data set by concatenating these M data sets. The proposed information-sharing method can then be applied to the concatenated source data set. This of course has advantages and disadvantages. On the one hand, if all the data sets—source and target—are mostly homogeneous, then our proposed information-sharing leads to a substantial efficiency gain through concatenation; similarly, if the data sets are mostly heterogeneous, then there is effectively no risk since our proposed information-sharing procedure will down-weight the source data set and inference will rely mostly on the target. On the other hand, if there are groups of source data sets that are homogeneous within and heterogeneous across, then the picture is far less clear: the concatenated source data set lacks some of the nuance of the individual source data sets which, depending on the circumstances, could improve or worsen the efficiency of the inference in the target data. We investigate this potential mix of homogeneous and heterogeneous source data sets and make general recommendations in Appendix B.2, but leave it up to the individual investigator to determine if concatenation is justified in their particular application. Alternatives to concatenation include data-driven methods for detecting heterogeneous source data sets (Li et al., 2022; Tian and Feng, 2022), which are discussed in Section 4.3 in the context of "negative transfer."

Since the features measured in data sets $\mathcal{D}_1$ and $\mathcal{D}_2$ are the same, it makes sense to consider fitting identical generalized linear models to the two data sets. These models assume the conditional probability density/mass function for y_{ij}, given $\boldsymbol{X}_{ij}$, is of the form

$$f_j(y_{ij}; \boldsymbol{\beta}_j, \gamma_j) = c(y_{ij}; \gamma_j) \, \exp\big[d(\gamma_j)^{-1}\{y_{ij}\boldsymbol{X}_{ij}^\top \boldsymbol{\beta}_j - b(\boldsymbol{X}_{ij}^\top \boldsymbol{\beta}_j)\}\big], \tag{1}$$

$i = 1, \ldots, n_j$, $j = 1, 2$, where b, c, and d are known functions, $\boldsymbol{\beta}_j$ is the quantity of primary interest, and γ_j is a nuisance parameter that will not receive much attention in what follows. Since inference of γ_j is not of interest and we can appropriately estimate γ_j based on $\mathcal{D}_j$, we do not concern ourselves with how integration affects estimates of γ_j and we assume $\gamma_1 \neq \gamma_2$. The conditional distribution's dependence on $\boldsymbol{X}_{ij}$ is implicit in the "j" subscript on f_j. We will not be concerned with the marginal distribution of $\boldsymbol{X}_{ij}$ so, as is customary, we assume throughout that this marginal distribution does not depend on $(\boldsymbol{\beta}_j, \gamma_j)$, and there is no need for notation to express this marginal. We assume that $\boldsymbol{X}_j = (\boldsymbol{X}_{1j}, \ldots, \boldsymbol{X}_{n_j j}) \in \mathbb{R}^{p \times n_j}$, $j = 1, 2$, is full rank.

Of course, the two data sets can be treated separately and, for example, the standard likelihood-based inference can be carried out. In particular, the maximum likelihood estimator (MLE) can be obtained as

$$(\widehat{\boldsymbol{\beta}}_j, \widehat{\gamma}_j) = \arg\max_{\boldsymbol{\beta}_j, \gamma_j} f_j^{(n_j)}(\boldsymbol{y}_j; \boldsymbol{\beta}_j, \gamma_j), \quad j = 1, 2,$$

where $f_j^{(n_j)}(\boldsymbol{y}_j; \boldsymbol{\beta}_j, \gamma_j) = \prod_{i=1}^{n_j} f_j(y_{ij}; \boldsymbol{\beta}_j, \gamma_j)$ is the joint density/likelihood function based

on data set $\mathcal{D}_j$, $j = 1, 2$. The MLEs, together with the corresponding observed Fisher information matrices, can be used to construct approximately valid tests and confidence regions for the respective $\boldsymbol{\beta}_j$ parameters. This would be an appropriate course of action if the two data sets in question had nothing in common.

Here, however, with lots in common between the two data sets, we have a different situation in mind. We imagine that the analysis of data set $\mathcal{D}_1$ has been carried out first, resulting in the MLE $\widehat{\boldsymbol{\beta}}_1$, among other things. Then the question of interest is whether the analysis of data set $\mathcal{D}_2$ ought to depend in some way on the results of $\mathcal{D}_1$'s analysis and, if so, how. While the problem setup is similar to the domain adaptation problem frequently seen in binary classification (Cai and Wei, 2021; Reeve et al., 2021), we emphasize that our interest lies in improving efficiency of inference of $\boldsymbol{\beta}_2$. More specifically, can the inference of $\boldsymbol{\beta}_2$ based on $\mathcal{D}_2$ be improved through the incorporation of the results from $\mathcal{D}_1$'s analysis?

2.2 Information-driven regularization

Our stated goal is to improve inference of $\boldsymbol{\beta}_2$ in $\mathcal{D}_2$ given the analysis of $\mathcal{D}_1$. Further inspection of equation (1) reveals that the primary difference between f_1 and f_2 is in the difference between $\widehat{\boldsymbol{\beta}}_1$ and $\boldsymbol{\beta}_2$. Thus, at first glance, the similarity between $\mathcal{D}_1$ and $\mathcal{D}_2$ is primarily driven by the closeness of $\widehat{\boldsymbol{\beta}}_1$ and $\boldsymbol{\beta}_2$. Intuitively, if $\widehat{\boldsymbol{\beta}}_1$ and $\boldsymbol{\beta}_2$ are close, say $\|\widehat{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}_2\|_2^2$ is small, then some gain in the inference of $\boldsymbol{\beta}_2$ can be expected if the inference in $\mathcal{D}_1$ is taken into account. This rationale motivates a potential objective to maximize $f_2^{(n_2)}(\boldsymbol{y}_2; \boldsymbol{\beta}_2, \gamma_j)$ under the constraint $\|\boldsymbol{\beta}_2 - \widehat{\boldsymbol{\beta}}_1\|^2 < c$ for some constant c. The choice of c then reflects one's belief in the similarity or dissimilarity between $\mathcal{D}_1$ and $\mathcal{D}_2$, and data-driven selection of c relies on the distance $\|\boldsymbol{\beta}_2 - \widehat{\boldsymbol{\beta}}_1\|^2$.

We must ask ourselves, however, if closeness between $\boldsymbol{\beta}_2$ and $\widehat{\boldsymbol{\beta}}_1$ is sufficient for us to integrate information from $\mathcal{D}_2$ into estimation of $\boldsymbol{\beta}_2$. For example, would we choose to constrain $\boldsymbol{\beta}_2$ to be close to $\widehat{\boldsymbol{\beta}}_1$ if n_1 were small? How about if $\boldsymbol{X}_1\boldsymbol{X}_1^\top$ is small, reflecting uninformative features and resulting in a large variance of $\widehat{\boldsymbol{\beta}}_1$? These intuitive notions of what we consider to be *informative* highlight a gap in our argument so far: elements that are "close" in the parameter space may not be close in an information-theoretic sense, and vice-versa. This is best visualized by Figure 1, which plots two negative log-likelihood functions for $\boldsymbol{\beta}_1$, one being based on a more informative data set than the other. The true $\boldsymbol{\beta}_2$ is also displayed there and, as expected, is different from the MLE $\widehat{\boldsymbol{\beta}}_1$ based on $\mathcal{D}_1$; for visualization purposes, we have arranged for $\widehat{\boldsymbol{\beta}}_1$ to be the same for both the more and less informative data sets. The plot reveals that the more informative data can better distinguish the two points $\boldsymbol{\beta}_2$ and $\widehat{\boldsymbol{\beta}}_1$, in terms of quality of fit, than can the less informative data—so it is not just the distance between $\boldsymbol{\beta}_2$ and $\widehat{\boldsymbol{\beta}}_1$ that matters! Intuitively, estimation of $\boldsymbol{\beta}_2$ should account not only for its distance to $\widehat{\boldsymbol{\beta}}_1$ but also the "sharpness" of the likelihood, or the amount of information in $\mathcal{D}_1$.

This motivates our proposal of a distance-driven regularization that takes both the ambient geometry of the parameter space and aspects of the statistical model into consideration. That is, we propose a regularization based on a sort of "statistical distance." The most common such distance, closely related to the Fisher information and the associated information geometry (Amari and Nagaoka, 2000; Nielsen, 2020), is the *Kullback–Leibler* (KL)

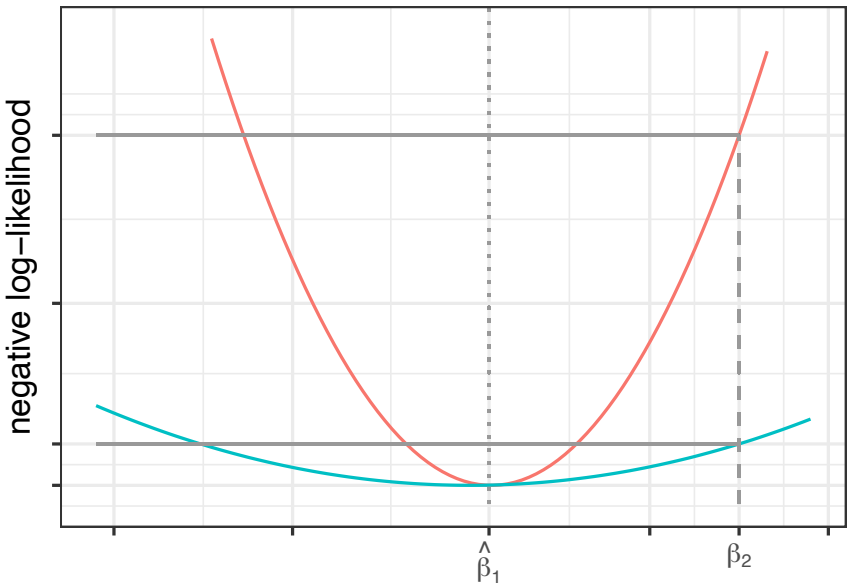

Figure 1: An illustration of the likelihoods for more (red) and less (blue) informative data.

divergence which, in our case, is given by

$$\begin{aligned}\mathrm{KL}_{n_1}(\boldsymbol{\beta}; \widehat{\boldsymbol{\beta}}_1, \gamma_1) &= \int \log \frac{f_1^{(n_1)}(\boldsymbol{y}; \widehat{\boldsymbol{\beta}}_1, \gamma_1)}{f_1^{(n_1)}(\boldsymbol{y}; \boldsymbol{\beta}, \gamma_1)} \, f_1^{(n_1)}(\boldsymbol{y}; \widehat{\boldsymbol{\beta}}_1, \gamma_1) \, d\boldsymbol{y} \\ &= \frac{1}{d(\gamma_1)} \sum_{i=1}^{n_1} \{ b'(\boldsymbol{X}_{i1}^\top \widehat{\boldsymbol{\beta}}_1)(\boldsymbol{X}_{i1}^\top \widehat{\boldsymbol{\beta}}_1 - \boldsymbol{X}_{i1}^\top \boldsymbol{\beta}) + b(\boldsymbol{X}_{i1}^\top \boldsymbol{\beta}) - b(\boldsymbol{X}_{i1}^\top \widehat{\boldsymbol{\beta}}_1) \},\end{aligned}$$

where $b'(\boldsymbol{X}_{i1}^\top \boldsymbol{\beta}) = \mathbb{E}_{\boldsymbol{\beta}, \gamma_1}(y_{i1} \mid \boldsymbol{X}_{i1})$, a standard property satisfied by all natural exponential families. Our proposal, then, is to use the aforementioned KL divergence to tie the estimates based on the two data sets together, i.e., to produce an estimator that solves the following constrained optimization problem:

$$\text{maximize } n_2^{-1} \log f_2^{(n_2)}(\boldsymbol{y}_2; \boldsymbol{\beta}, \gamma_2) \quad \text{subject to} \quad n_1^{-1} \mathrm{KL}_{n_1}(\boldsymbol{\beta}; \widehat{\boldsymbol{\beta}}_1, \gamma_1) \leq \epsilon, \tag{2}$$

where $\epsilon \geq 0$ is a constant to be specified by the user. The KL divergence measures the entropy of $f_1^{(n_1)}(\boldsymbol{y}; \boldsymbol{\beta}, \gamma_1)$ relative to $f_1^{(n_1)}(\boldsymbol{y}; \widehat{\boldsymbol{\beta}}_1, \gamma_1)$. As the Hessian of the KL divergence, the Fisher information describes the local shape of the KL divergence and quantifies its ability to discriminate between $f_1^{(n_1)}(\boldsymbol{y}; \boldsymbol{\beta}, \gamma_1)$ and $f_1^{(n_1)}(\boldsymbol{y}; \widehat{\boldsymbol{\beta}}_1, \gamma_1)$.

A key difference between what is being proposed here and other regularization strategies in the literature is that our penalty term is "data-dependent" or, perhaps more precisely, relative to $\mathcal{D}_1$. That is, it takes the *information* in $\mathcal{D}_1$ into account when estimating $\boldsymbol{\beta}_2$, hence the name information-driven regularization. Of course, the extent of the information-driven regularization, i.e., how much information is shared between $\mathcal{D}_1$ and $\mathcal{D}_2$, is determined by ϵ, so this choice is crucial. We will address this, and the desired properties of the proposed estimator, in Section 4 below.

To solve the optimization problem in equation (2), we propose the estimator

$$\widetilde{\boldsymbol{\beta}}_2(\lambda) = \arg\max_{\boldsymbol{\beta}} O(\boldsymbol{\beta}; \lambda),$$

where the objective function is

$$\begin{aligned} O(\boldsymbol{\beta}; \lambda) &= n_2^{-1} \log f_2^{(n_2)}(\boldsymbol{y}_2; \boldsymbol{\beta}, \gamma_2) - \lambda n_1^{-1} \mathrm{KL}_{n_1}(\boldsymbol{\beta}; \widehat{\boldsymbol{\beta}}_1, \gamma_1) \\ &= \frac{1}{n_2} \sum_{i=1}^{n_2} \left\{ y_{i2} \boldsymbol{X}_{i2}^\top \boldsymbol{\beta} - b(\boldsymbol{X}_{i2}^\top \boldsymbol{\beta}) \right\} \\ &\quad - \frac{\lambda}{n_1} \sum_{i=1}^{n_1} \{ b'(\boldsymbol{X}_{i1}^\top \widehat{\boldsymbol{\beta}}_1)(\boldsymbol{X}_{i1}^\top \widehat{\boldsymbol{\beta}}_1 - \boldsymbol{X}_{i1}^\top \boldsymbol{\beta}) + b(\boldsymbol{X}_{i1}^\top \boldsymbol{\beta}) - b(\boldsymbol{X}_{i1}^\top \widehat{\boldsymbol{\beta}}_1) \}, \end{aligned} \tag{3}$$

for a *dial parameter* $\lambda \geq 0$. Thanks to the general niceties of natural exponential families, this objective function is convex and, therefore, is straightforward to optimize. The addition of the KL penalty term in equation (3) essentially introduces λ-discounted units of information from $\mathcal{D}_1$ into the inference based on $\mathcal{D}_2$. Note that we specifically avoid use of the term "tuning parameter" to make a clear distinction between our approach and the classical penalized regression approach. The addition of the KL divergence penalty intentionally introduces a bias in the estimator $\widetilde{\boldsymbol{\beta}}_2(\lambda)$ for $\boldsymbol{\beta}_2$ which we hope—and later show—will be offset by a reduction in variance. This is the same basic bias–variance trade-off that motivates other shrinkage estimation strategies, such as ridge regression and lasso, but with one key difference: our motivation is information-sharing, so we propose to shrink toward a $\mathcal{D}_1$-dependent target whereas existing methods' motivation is structural constraints (e.g., sparsity), so they propose to shrink toward a fixed target. With this alternative perspective comes new questions: does adding a small amount of $\mathcal{D}_1$-dependent bias lead to efficiency gains? Below we will identify a range of the dial parameter such that the use of the external information in $\mathcal{D}_1$ reduces the variance of $\widetilde{\boldsymbol{\beta}}_2(\lambda)$ enough so that we gain efficiency even with the small amount of added bias. The parameter λ thus acts as a dial to calibrate the trade-off between bias and variance in $\widetilde{\boldsymbol{\beta}}_2(\lambda)$.

The maximum of $O(\boldsymbol{\beta}; \lambda)$ coincides with the root of the estimating function

$$\begin{aligned} \boldsymbol{\Psi}(\boldsymbol{\beta}; \lambda) &= \frac{1}{n_2} \sum_{i=1}^{n_2} \boldsymbol{X}_{i2} \{ y_{i2} - b'(\boldsymbol{X}_{i2}^\top \boldsymbol{\beta}) \} - \frac{\lambda}{n_1} \sum_{i=1}^{n_1} \boldsymbol{X}_{i1} \{ b'(\boldsymbol{X}_{i1}^\top \boldsymbol{\beta}) - b'(\boldsymbol{X}_{i1}^\top \widehat{\boldsymbol{\beta}}_1) \} \\ &= n_2^{-1} \boldsymbol{X}_2 \{ \boldsymbol{y}_2 - \boldsymbol{\mu}_2(\boldsymbol{\beta}) \} - \lambda n_1^{-1} \boldsymbol{X}_1 \{ \boldsymbol{\mu}_1(\boldsymbol{\beta}) - \boldsymbol{\mu}_1(\widehat{\boldsymbol{\beta}}_1) \}, \end{aligned}$$

where $\boldsymbol{\mu}_j(\boldsymbol{\beta})$ is the n_j-vector of (conditional) mean responses, given $\boldsymbol{X}_j$, for $j = 1, 2$. The root is the value of $\boldsymbol{\beta}$ that satisfies

$$n_2^{-1} \boldsymbol{X}_2 \boldsymbol{y}_2 + \lambda n_1^{-1} \boldsymbol{X}_1 \boldsymbol{\mu}_1(\widehat{\boldsymbol{\beta}}_1) = n_2^{-1} \boldsymbol{X}_2 \boldsymbol{\mu}_2(\boldsymbol{\beta}) + \lambda n_1^{-1} \boldsymbol{X}_1 \boldsymbol{\mu}_1(\boldsymbol{\beta}).$$

From this equation, we see that a solution $\boldsymbol{\beta}$ must be such that a certain linear combination of $\boldsymbol{\mu}_1(\boldsymbol{\beta})$ and $\boldsymbol{\mu}_2(\boldsymbol{\beta})$ agrees with the same linear combination of $\boldsymbol{\mu}_1(\widehat{\boldsymbol{\beta}}_1)$ and $\boldsymbol{y}_2$, two natural estimators of the mean responses in $\mathcal{D}_1$ and $\mathcal{D}_2$, respectively. In Section 3, we give the specific form of the estimating function for two common generalized linear models: linear and logistic regression.

2.3 Remarks

2.3.1 Comparison to Wasserstein distance

The Wasserstein distance has enjoyed recent popularity in the transfer learning literature; see for example Shen et al. (2018); Yoon et al. (2020); Cheng et al. (2020); Zhu et al. (2021). An example illustrates why we prefer the KL divergence to the Wasserstein distance. Suppose $f_1(y_{i1}; \boldsymbol{\beta}, \gamma_1)$ is the Gaussian density with mean $\boldsymbol{X}_{i1}^\top \boldsymbol{\beta}$ and variance $\gamma_1 = \sigma_1^2$. The KL divergence defined above is

$$\mathrm{KL}_{n_1}(\boldsymbol{\beta}; \widehat{\boldsymbol{\beta}}_1, \gamma_1) = (2\sigma_1^2)^{-1}(\widehat{\boldsymbol{\beta}}_1 - \boldsymbol{\beta})^\top \boldsymbol{X}_1 \boldsymbol{X}_1^\top (\widehat{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}). \tag{4}$$

Clearly, this takes into consideration not only the distance between $\boldsymbol{\beta}$ and $\widehat{\boldsymbol{\beta}}_1$, but also other relevant information-related aspects of the data set $\mathcal{D}_1$. By contrast, following Olkin and Pukelsheim (1982), the 2-Wasserstein distance between the two Gaussian joint densities $f_1^{(n_1)}(\cdot; \widehat{\boldsymbol{\beta}}_1, \sigma_1^2)$ and $f_1^{(n_1)}(\cdot; \boldsymbol{\beta}, \sigma_1^2)$ is

$$\begin{aligned}
&\mathcal{W}\{f_1^{(n_1)}(\cdot; \widehat{\boldsymbol{\beta}}_1, \sigma_1^2), f_1^{(n_1)}(\cdot; \boldsymbol{\beta}, \sigma_1^2)\} \\
&\qquad = \|\widehat{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}\|_2^2 + \mathrm{trace}\{\sigma_1^2 \boldsymbol{I}_{n_1} + \sigma_1^2 \boldsymbol{I}_{n_1} - 2\sigma_1 \sigma_1 \boldsymbol{I}_{n_1}\} \\
&\qquad = \|\widehat{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}\|_2^2.
\end{aligned}$$

In this example, the Wasserstein distance is only a function of the distance between the means, and fails to take into account any other aspects of $\mathcal{D}_1$, in particular, it does not depend on the observed Fisher information in $\mathcal{D}_1$. Replacing the KL divergence in our proposal above with the 2-Wasserstein distance would, therefore, reduce to a simple ridge regression formulated with $\mathcal{D}_1$-dependent shrinkage target $\widehat{\boldsymbol{\beta}}_1$. As discussed above, such an approach would not satisfactorily achieve our objectives.

2.3.2 Connection to data-dependent penalties

The dependence of the constraint in equation (2) on the data $\mathcal{D}_1$ is unusual in contrast to more common constraints on parameters. Our formulation leads to a penalty term in $O(\boldsymbol{\beta}; \lambda)$ that depends on data. This approach allows an appealing connection to the empirical Bayes estimation framework, which allows us to treat information gained from analyzing $\mathcal{D}_1$ as prior knowledge when analyzing $\mathcal{D}_2$. Through this lens, $\boldsymbol{\beta}_2 \mapsto \exp\{-\lambda \mathrm{KL}_{n_1}(\boldsymbol{\beta}_2; \widehat{\boldsymbol{\beta}}_1, \gamma_1)\}$ acts like a prior density and maximizing $O(\boldsymbol{\beta}_2; \lambda)$ is equivalent to maximizing the corresponding posterior density for $\boldsymbol{\beta}_2$, i.e., $\widehat{\boldsymbol{\beta}}_2(\lambda)$ is the maximum *a posteriori* estimator of $\boldsymbol{\beta}_2$. The dial parameter λ adjusts the impact of the prior and reflects the experimenter's belief in the similarity of $\mathcal{D}_1$ and $\mathcal{D}_2$.

The prior term $\mathrm{KL}_{n_1}(\boldsymbol{\beta}; \widehat{\boldsymbol{\beta}}_1, \gamma_1)$ is itself a measure of the excess surprise (Baldi and Itti, 2010) between the prior information $f_1^{(n_1)}(\boldsymbol{y}_1; \widehat{\boldsymbol{\beta}}_1, \gamma_1)$ and the likelihood $f_1^{(n_1)}(\boldsymbol{y}_1; \boldsymbol{\beta}, \gamma_1)$ in $\mathcal{D}_1$. This observation leads to an intuitive understanding of our KL prior: if the prior $f_1^{(n_1)}(\boldsymbol{y}_1; \widehat{\boldsymbol{\beta}}_1, \gamma_1)$ and the likelihood $f_1^{(n_1)}(\boldsymbol{y}_1; \boldsymbol{\beta}, \gamma_1)$ are close for all values of $\boldsymbol{\beta}$, then the prior probabilities are small for all values of $\boldsymbol{\beta}$ and the prior is weakly informative.

2.3.3 Distinction from data fusion

Apart from the literature that considers (a subset of) effects to be *a priori* heterogeneous (e.g. Liu et al., 2015; Tang and Song, 2020), homogeneous (e.g. Xie et al., 2011) or somewhere in between (e.g. Shen et al., 2020), some methods consider fusion of individual feature effects. Broadly, fusion approaches jointly estimate $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2$ by shrinking them towards each other. At first blush, these methods may appear similar to our approach, so we take the time to highlight two key differences.

Primarily, these methods differ in that they do not consider estimation in $\mathcal{D}_1$ to be fixed to $\widehat{\boldsymbol{\beta}}_1$, and can jointly re-estimate both parameters for improved efficiency. This is quite different from our goal, which is to quantify the utility of a first, already analyzed data set $\mathcal{D}_1$ in improving inference in a second data set $\mathcal{D}_2$. Our approach has the advantage that we do not require the model to be correctly specified in $\mathcal{D}_1$, which endows our method with substantial robustness properties which are lacking for fusion approaches.

Moreover, data fusion's underlying premise is to exploit feature clustering structure across data sources, thereby determining which *features* should be combined (Bondell and Reich, 2009; Shen and Huang, 2010; Tibshirani and Taylor, 2011; Ke et al., 2015; Tang and Song, 2016). In particular, this leads to the integration of some feature effects but not others. Different sets of feature effects are estimated from different sets of data, which does not provide the desired quantification of the similarity between data sets.

3 Examples

3.1 Gaussian linear regression

For the Gaussian linear model, the KL divergence is given in (4). In this case, the objective function in (3) is given by

$$\begin{aligned} O(\boldsymbol{\beta};\lambda) = n_2^{-1}\boldsymbol{\beta}^\top \boldsymbol{X}_2\boldsymbol{y}_2 - (2n_2)^{-1}\boldsymbol{\beta}^\top \boldsymbol{X}_2\boldsymbol{X}_2^\top\boldsymbol{\beta} \\ - \lambda(2n_1)^{-1}(\widehat{\boldsymbol{\beta}}_1 - \boldsymbol{\beta})^\top \boldsymbol{X}_1\boldsymbol{X}_1^\top(\widehat{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}). \end{aligned}$$

Optimizing O corresponds to finding the root of the estimating function

$$\boldsymbol{\Psi}(\boldsymbol{\beta};\lambda) = n_2^{-1}(\boldsymbol{X}_2\boldsymbol{y}_2 - \boldsymbol{X}_2\boldsymbol{X}_2^\top\boldsymbol{\beta}) - \lambda n_1^{-1}(\boldsymbol{X}_1\boldsymbol{X}_1^\top\boldsymbol{\beta} - \boldsymbol{X}_1\boldsymbol{X}_1^\top\widehat{\boldsymbol{\beta}}_1).$$

Let $\boldsymbol{G}_j = n_j^{-1}\boldsymbol{X}_j\boldsymbol{X}_j^\top$, $j = 1, 2$, denote the scaled Gram matrices. Then the solution to the estimating equation $\boldsymbol{\Psi}(\boldsymbol{\beta};\lambda) = 0$ is

$$\begin{aligned} \widetilde{\boldsymbol{\beta}}_2(\lambda) &= \big(\boldsymbol{G}_2 + \lambda\boldsymbol{G}_1\big)^{-1}\big(n_2^{-1}\boldsymbol{X}_2\boldsymbol{y}_2 + \lambda\boldsymbol{G}_1\widehat{\boldsymbol{\beta}}_1\big) \\ &= \big(\boldsymbol{G}_2 + \lambda\boldsymbol{G}_1\big)^{-1}\big(\boldsymbol{G}_2\widehat{\boldsymbol{\beta}}_2 + \lambda\boldsymbol{G}_1\widehat{\boldsymbol{\beta}}_1\big), \end{aligned} \tag{5}$$

where $\widehat{\boldsymbol{\beta}}_2 = (\boldsymbol{X}_2\boldsymbol{X}_2)^{-1}\boldsymbol{X}_2\boldsymbol{y}_2$ is the MLE based on $\mathcal{D}_2$ only. The estimator $\widetilde{\boldsymbol{\beta}}_2(\lambda)$ in equation (5) progressively grows closer to $\widehat{\boldsymbol{\beta}}_1$ as more weight (through λ) is given to $\mathcal{D}_1$. This is desirable behavior: λ acts as a "dial" allowing us to tune our preference towards $\mathcal{D}_1$ or $\mathcal{D}_2$. When $\boldsymbol{\beta}_1 = \boldsymbol{\beta}_2$, we recognize $\widetilde{\boldsymbol{\beta}}_2(\lambda)$ as a generalization of the best linear unbiased estimator.

The estimator $\widetilde{\boldsymbol{\beta}}_2(\lambda)$ does not rely on individual level data from the first data set, $\mathcal{D}_1$; all that is needed is the sample size, the estimate, and the observed Fisher information. Thus, our proposed procedure is privacy preserving and can be implemented in a meta-analytic setting where only summaries of $\mathcal{D}_1$ and $\mathcal{D}_2$ are available.

This estimator also takes the familiar form of a generalized ridge estimator (Hoerl and Kennard, 1970; Hemmerle, 1975) and benefits from many well-known properties; see the excellent review of van Wieringen (2021). The estimator $\widetilde{\boldsymbol{\beta}}_2(\lambda)$ is a convex combination of the MLEs from $\mathcal{D}_1$ and $\mathcal{D}_2$, with weights determined by the corresponding observed Fisher information matrices. This formulation allows us to identify the crucial role that λ plays in balancing not only the bias but the variance of $\widehat{\boldsymbol{\beta}}_2$, a role reminiscent of that played by the tuning parameter for ridge regression (Theobald, 1974; Farebrother, 1976).

Our estimator $\widetilde{\boldsymbol{\beta}}_2(\lambda)$ can be rewritten as $\widetilde{\boldsymbol{\beta}}_2(\lambda) = \widetilde{\boldsymbol{W}}_\lambda \widehat{\boldsymbol{\beta}}_2 + (\boldsymbol{I}_p - \widetilde{\boldsymbol{W}}_\lambda)\widehat{\boldsymbol{\beta}}_1$ with

$$\widetilde{\boldsymbol{W}}_\lambda = n_2^{-1}\big(n_2^{-1}\boldsymbol{X}_2\boldsymbol{X}_2^\top + \lambda n_1^{-1}\boldsymbol{X}_1\boldsymbol{X}_1^\top\big)^{-1}\boldsymbol{X}_2\boldsymbol{X}_2^\top$$

In contrast, Chen et al. (2015) consider pooling $\mathcal{D}_1$ and $\mathcal{D}_2$ to jointly estimate $\boldsymbol{\beta}_2$ and $\boldsymbol{\beta}_1 = \boldsymbol{\beta}_2 + \boldsymbol{\delta}_2$ with some penalty $\|\boldsymbol{X}_2\boldsymbol{\delta}_2\|_2^2$, where $\boldsymbol{\delta}_2 = \boldsymbol{\beta}_1 - \boldsymbol{\beta}_2$. Their estimator is given by $\widehat{\boldsymbol{\beta}}_2(\boldsymbol{W}_\lambda) = \boldsymbol{W}_\lambda\widehat{\boldsymbol{\beta}}_2 + (\boldsymbol{I}_p - \boldsymbol{W}_\lambda)\widehat{\boldsymbol{\beta}}_1$, where

$$\boldsymbol{W}_\lambda = \big(\boldsymbol{X}_1\boldsymbol{X}_1^\top + \lambda\boldsymbol{X}_2\boldsymbol{X}_2^\top + \lambda\boldsymbol{X}_1\boldsymbol{X}_1^\top\big)^{-1}\big(\boldsymbol{X}_1\boldsymbol{X}_1^\top + \lambda\boldsymbol{X}_2\boldsymbol{X}_2^\top\big).$$

When $n_1 = n_2$, $\widetilde{\boldsymbol{W}}_\lambda \preceq \boldsymbol{W}_\lambda$ if and only if $(1-\lambda)\boldsymbol{I}_p \preceq (\boldsymbol{X}_2\boldsymbol{X}_2^\top)^{-1}\boldsymbol{X}_1\boldsymbol{X}_1^\top$. This is clearly always true when $\lambda \geq 1$. When $\lambda < 1$, $\widetilde{\boldsymbol{W}}_\lambda \preceq \boldsymbol{W}_\lambda$ if $\mathcal{D}_1$ is informative relative to $\mathcal{D}_2$, and $\widetilde{\boldsymbol{W}}_\lambda \succeq \boldsymbol{W}_\lambda$ if $\mathcal{D}_1$ is uninformative relative to $\mathcal{D}_2$. Therefore, our estimator assigns more weight to $\mathcal{D}_1$ than Chen et al. (2015)'s estimator when $\mathcal{D}_1$ is more informative, and less when it is uninformative, as desired.

3.2 Bernoulli logistic regression

For the standard logistic regression model, the KL divergence is given by

$$\mathrm{KL}_{n_1}(\boldsymbol{\beta}; \widehat{\boldsymbol{\beta}}_1) = \sum_{i=1}^{n_1}\left[\frac{\exp(\boldsymbol{X}_{i1}^\top\widehat{\boldsymbol{\beta}}_1)}{1+\exp(\boldsymbol{X}_{i1}^\top\widehat{\boldsymbol{\beta}}_1)}(\boldsymbol{X}_{i1}^\top\widehat{\boldsymbol{\beta}}_1 - \boldsymbol{X}_{i1}^\top\boldsymbol{\beta}) + \log\Big\{\frac{1+\exp(\boldsymbol{X}_{i1}^\top\boldsymbol{\beta})}{1+\exp(\boldsymbol{X}_{i1}^\top\widehat{\boldsymbol{\beta}}_1)}\Big\}\right].$$

Then the objective function in (3) can be written as

$$\begin{aligned} O(\boldsymbol{\beta};\lambda) = {} & \frac{1}{n_2}\sum_{i=1}^{n_2}\big[y_{i2}\boldsymbol{X}_{i2}^\top\boldsymbol{\beta} - \log\{1+\exp(\boldsymbol{X}_{i2}^\top\boldsymbol{\beta})\}\big] \\ & - \frac{\lambda}{n_1}\sum_{i=1}^{n_1}\left[\frac{\exp(\boldsymbol{X}_{i1}^\top\widehat{\boldsymbol{\beta}}_1)}{1+\exp(\boldsymbol{X}_{i1}^\top\widehat{\boldsymbol{\beta}}_1)}(\boldsymbol{X}_{i1}^\top\widehat{\boldsymbol{\beta}}_1 - \boldsymbol{X}_{i1}^\top\boldsymbol{\beta}) + \log\Big\{\frac{1+\exp(\boldsymbol{X}_{i1}^\top\boldsymbol{\beta})}{1+\exp(\boldsymbol{X}_{i1}^\top\widehat{\boldsymbol{\beta}}_1)}\Big\}\right]. \end{aligned}$$

To optimize O, we find the root of the estimating function given by

$$\boldsymbol{\Psi}(\boldsymbol{\beta};\lambda) = n_2^{-1}\boldsymbol{X}_2\big\{\boldsymbol{y}_2 - \mathrm{expit}(\boldsymbol{X}_2^\top\boldsymbol{\beta})\big\} - \lambda n_1^{-1}\boldsymbol{X}_1\big\{\mathrm{expit}(\boldsymbol{X}_1^\top\boldsymbol{\beta}) - \mathrm{expit}(\boldsymbol{X}_1^\top\widehat{\boldsymbol{\beta}}_1)\big\},$$

where $\mathrm{expit}(\boldsymbol{z})$ is the vector obtained by applying $z \mapsto e^z/(1+e^z)$ to each component of the vector $\boldsymbol{z}$. The estimator $\widetilde{\boldsymbol{\beta}}_2(\lambda)$ is the solution to $\boldsymbol{\Psi}(\boldsymbol{\beta};\lambda) = \boldsymbol{0}$. Of course, there is no closed-form solution, but the solution can be obtained numerically.

4 Theoretical support

4.1 Objectives

A distinguishing feature of our perspective here is that we treat $\mathcal{D}_1$ as fixed. In particular, we treat $\widehat{\boldsymbol{\beta}}_1$ as a known constant. A relevant feature in the analysis below is the difference $\boldsymbol{\delta} = \boldsymbol{\beta}_2 - \widehat{\boldsymbol{\beta}}_1$, which is just one measure of the similarity/dissimilarity between $\mathcal{D}_1$ and $\mathcal{D}_2$. Of course, $\boldsymbol{\delta}$ is unknown because $\boldsymbol{\beta}_2$ is unknown, but $\boldsymbol{\delta}$ can be estimated if necessary. Note also that we do not assume $\|\boldsymbol{\delta}\|_2$ to be small.

Thanks to well-known properties of KL divergence, $\boldsymbol{\beta} \mapsto \mathbb{E}\{O(\boldsymbol{\beta}; \lambda)\}$ is concave, so it has a unique maximum, denoted by $\boldsymbol{\beta}_2^\star(\lambda)$. If the true value of $\boldsymbol{\beta}$ in $\mathcal{D}_2$ is $\boldsymbol{\beta}_2$, then clearly $\boldsymbol{\beta}_2^\star(\lambda) \neq \boldsymbol{\beta}_2$ when $\lambda \neq 0$ and $\boldsymbol{\delta} \neq \mathbf{0}$. On the other hand, it is easy to verify that, as $\lambda \to 0^+$, the objective function satisfies $\mathbb{E}\{O(\boldsymbol{\beta}; \lambda)\} \to \mathbb{E}\{O(\boldsymbol{\beta}; 0)\}$ uniformly on compact subsets of the parameter space of $\boldsymbol{\beta}$, so, as expected, $\lim_{\lambda \to 0^+} \boldsymbol{\beta}_2^\star(\lambda) = \boldsymbol{\beta}_2$. As stated in Section 2.2, our objective is to find a range of the dial parameter λ — depending on $\boldsymbol{\delta}$, etc. — such that the use of the external information in $\mathcal{D}_1$ sufficiently reduces the variance of $\widetilde{\boldsymbol{\beta}}_2(\lambda)$ so as to overcome the addition of a small amount of bias to $\widetilde{\boldsymbol{\beta}}_2(\lambda)$.

We first consider the Gaussian linear model for its simplicity, which allows for exact (non-asymptotic) results. The ideas extend to generalized linear models, but there we will need asymptotic approximations as $n_2 \to \infty$. Throughout, we use the notational convention $\boldsymbol{M}^2 = \boldsymbol{M}^\top \boldsymbol{M}$ for a matrix $\boldsymbol{M}$.

Proofs of all the results are given in Appendix A.

4.2 Exact results in the Gaussian linear model

In the Gaussian case, the expected estimating function evaluated at $\boldsymbol{\beta}_2^\star(\lambda)$ is

$$\begin{aligned}
\mathbf{0} &= \mathbb{E}_{\boldsymbol{\beta}_2}\big\{\boldsymbol{\Psi}(\boldsymbol{\beta}_2^\star(\lambda); \lambda)\big\} \\
&= n_2^{-1} \boldsymbol{X}_2 \big\{\mathbb{E}_{\boldsymbol{\beta}_2}(\boldsymbol{y}_2) - \boldsymbol{X}_2^\top \boldsymbol{\beta}_2^\star(\lambda)\big\} - \lambda n_1^{-1} \boldsymbol{X}_1 \big\{\boldsymbol{X}_1 \boldsymbol{\beta}_2^\star(\lambda) - \boldsymbol{X}_1 \widehat{\boldsymbol{\beta}}_1\big\} \\
&= \boldsymbol{G}_2\{\boldsymbol{\beta}_2 - \boldsymbol{\beta}_2^\star(\lambda)\} - \lambda \boldsymbol{G}_1 \{\boldsymbol{\beta}_2^\star(\lambda) - \widehat{\boldsymbol{\beta}}_1\}.
\end{aligned}$$

Denote $\boldsymbol{S}(\lambda) = \boldsymbol{G}_2 + \lambda \boldsymbol{G}_1$. Rearranging, we obtain

$$\boldsymbol{\beta}_2^\star(\lambda) = \boldsymbol{S}^{-1}(\lambda)\big(\boldsymbol{G}_2 \boldsymbol{\beta}_2 + \lambda \boldsymbol{G}_1 \widehat{\boldsymbol{\beta}}_1\big) = \boldsymbol{\beta}_2 - \lambda \boldsymbol{S}^{-1}(\lambda) \boldsymbol{G}_1 \boldsymbol{\delta}.$$

Thus, $\boldsymbol{\beta}_2^\star(\lambda) = \boldsymbol{\beta}_2$ when $\lambda = 0$ or $\boldsymbol{\delta} = \mathbf{0}$. In general, however, our proposed estimator $\widetilde{\boldsymbol{\beta}}_2(\lambda)$ is estimating $\boldsymbol{\beta}_2^\star(\lambda) \neq \boldsymbol{\beta}_2$, so a practically important question is, if we already have an unbiased estimator, $\widehat{\boldsymbol{\beta}}_2$, of $\boldsymbol{\beta}_2$, then why would we introduce a biased estimator? Theorem 1 below establishes that there exists a range of $\lambda > 0$ for which the mean squared error (MSE) of $\widetilde{\boldsymbol{\beta}}_2(\lambda)$ is strictly less than that of $\widehat{\boldsymbol{\beta}}_2$. Details on the λ range over which the efficiency gain is achieved are given in the third paragraph following the theorem statement.

Theorem 1. *There exists a range of $\lambda > 0$ on which the mean squared error of $\widetilde{\boldsymbol{\beta}}_2(\lambda)$ is strictly less than the mean squared error of $\widehat{\boldsymbol{\beta}}_2$.*

This result is striking when considered in the context of data integration: we have shown that it is always "better" to integrate two sources of data than to use only one, even when the two are substantially different. We also claim that our estimator's gain in efficiency is robust to heterogeneity between the two data sets; we will return to this point in Section 4.3 to make general remarks in the context of transfer learning.

Of course, the reader familiar with Stein shrinkage (Stein, 1956; James and Stein, 1960) may not be surprised by our Theorem 1. Our result has a similar flavor to Stein's paradox (Efron and Morris, 1977), i.e., that some amount of shrinkage always leads to a more efficient estimator compared to a (weighted) sample average, here $\widehat{\boldsymbol{\beta}}_2$. In their presciently titled paper, "Combining possibly related estimation problems," Efron and Morris (1973) anticipated the ubiquity of results like our Theorem 1 that show estimation efficiency gains when combining different but related data sets.

The proof of Theorem 1 shows that $\text{MSE}\{\widetilde{\boldsymbol{\beta}}_2(\lambda)\} < \text{MSE}(\widehat{\boldsymbol{\beta}}_2)$ for all $\lambda > 0$ when $\boldsymbol{\delta} = \mathbf{0}$; when $\boldsymbol{\delta} \neq \mathbf{0}$, it proves that $\text{MSE}\{\widetilde{\boldsymbol{\beta}}_2(\lambda)\}$ is monotonically decreasing over the range $[0, \lambda^\star]$ and monotonically increasing over the range $[\lambda^\star, \infty)$ for some $\lambda^\star > 0$ that does not have a closed-form. The proof does, however, provide a bound for $\lambda^\star$:

$$\lambda^\star > \frac{\sigma_2^2}{n_2} \frac{\min_{r=1,\ldots,p} \kappa_r g_{r2}^{-1}}{\max_{r=1,\ldots,p} \delta_r^2}, \tag{6}$$

where $g_{r2} > 0$, $r = 1, \ldots, p$, the eigenvalues of $\boldsymbol{G}_2$ in increasing order and κ_r, $r = 1, \ldots, p$, the eigenvalues of $\boldsymbol{G}_2^{1/2}\boldsymbol{G}_1^{-1}\boldsymbol{G}_2^{1/2}$ in decreasing order. That is, the MSE of $\widetilde{\boldsymbol{\beta}}_2(\lambda)$ will be less than that of $\widehat{\boldsymbol{\beta}}_2$ if λ is no more than the right-hand side of (6). From the above expression, we see that if elements of $\boldsymbol{\delta}$ are large in absolute value, then the improvement in MSE only occurs for a small range of λ. Moreover, if n_2^{-1}, κ_r or g_{r2}^{-1} are small then the range of λ such that $\text{MSE}\{\widetilde{\boldsymbol{\beta}}_2(\lambda)\} < \text{MSE}(\widehat{\boldsymbol{\beta}}_2)$ is small: in other words, if $\mathcal{D}_2$ is highly informative then very little weight should be given to $\mathcal{D}_1$, regardless of how informative $\mathcal{D}_1$ is. This is intuitively appealing and practically useful because it provides a loose guideline based on the informativeness of $\mathcal{D}_2$ for how much improvement can be obtained using $\mathcal{D}_1$.

In practice, we propose to find a data-driven version of the minimizer, $\lambda^\star$, of the MSE. For this, we minimize an empirical version of the MSE based on plug-in estimators, i.e.,

$$\tilde{\lambda} = \arg\min_{\lambda > 0} \big[n_2^{-1} \widehat{\sigma}_2^2 \,\text{trace}\{\boldsymbol{S}^{-2}(\lambda)\boldsymbol{G}_2\} + \lambda^2 \text{trace}\{\boldsymbol{G}_1 \boldsymbol{S}^{-2}(\lambda)\boldsymbol{G}_1 \widehat{\boldsymbol{\delta}^2}\} \big],$$

the minimizer of the estimated mean squared error of $\widetilde{\boldsymbol{\beta}}_2(\lambda)$, where

$$\widehat{\sigma}_2^2 = \frac{1}{n_2 - p} \sum_{i=1}^{n_2} (y_{i2} - \boldsymbol{X}_{i2}^\top \widehat{\boldsymbol{\beta}}_2)^2,$$

is the usual estimate of the error variance σ_2^2, and $\widehat{\boldsymbol{\delta}^2}$ is a bias-adjusted estimate of $\boldsymbol{\delta}\boldsymbol{\delta}^\top$ computed as follows. If we define $\widehat{\boldsymbol{\delta}}_p = \widehat{\boldsymbol{\beta}}_2 - \widehat{\boldsymbol{\beta}}_1$, then the equality $\mathbb{E}_{\boldsymbol{\beta}_2}(\widehat{\boldsymbol{\delta}}_p \widehat{\boldsymbol{\delta}}_p^\top) = \boldsymbol{\delta}\boldsymbol{\delta}^\top + \sigma_2^2 n_2^{-1} \boldsymbol{G}_2^{-1}$ implies that $\widehat{\boldsymbol{\delta}}_p \widehat{\boldsymbol{\delta}}_p^\top$ is a positively biased estimator of $\boldsymbol{\delta}\boldsymbol{\delta}^\top$. Similar to Vinod (1987) and Chen et al. (2015), we estimate $\boldsymbol{\delta}\boldsymbol{\delta}^\top$ with

$$\widehat{\boldsymbol{\delta}^2} = \big(\widehat{\boldsymbol{\delta}}_p \widehat{\boldsymbol{\delta}}_p^\top - \widehat{\sigma}_2^2 n_2^{-1} \boldsymbol{G}_2^{-1} \big)_+,$$

where $(\boldsymbol{A})_+ = \boldsymbol{P}\mathbf{diag}\{\max(\lambda_i, 0)\}_{i=1}^p \boldsymbol{P}^\top$ for a symmetric matrix $\boldsymbol{A} \in \mathbb{R}^{p\times p}$ with eigendecomposition $\boldsymbol{P}\,\mathbf{diag}(\lambda_i)_{i=1}^p\,\boldsymbol{P}^\top$. If all eigenvalues are negative, we use $\widehat{\boldsymbol{\delta}^2} = \widehat{\boldsymbol{\delta}}_p\widehat{\boldsymbol{\delta}}_p^\top$. This bias-adjusted approach to estimating MSE is related to Stein's unbiased risk estimator (SURE, Stein, 1981); see also Vinod (1987) for a discussion in the ridge regression setting. The minimizer $\tilde{\lambda}$ (almost) always exists since $\widehat{\boldsymbol{\delta}^2}$ is non-zero with probability 1. That the minimizer is strictly positive, even if $\mathcal{D}_1$ is minimally- or non-informative, might be counter-intuitive; but this is a consequence of Theorem 1, which shows that we need only consider the set of strictly positive λ values to improve inference in $\mathcal{D}_2$. Finally, we show that constructing exact confidence intervals for $\boldsymbol{\beta}_2$ using a debiased version of $\widetilde{\boldsymbol{\beta}}_2(\lambda)$ reduces to the traditional inference using the MLE. It follows from the proof of Theorem 1 that,

$$\widetilde{\boldsymbol{\beta}}_2(\lambda) + \lambda\boldsymbol{S}^{-1}(\lambda)\boldsymbol{G}_1(\boldsymbol{\beta}_2 - \widehat{\boldsymbol{\beta}}_1) \sim \mathcal{N}\{\boldsymbol{\beta}_2, n_2^{-1}\sigma_2^2\boldsymbol{S}^{-1}(\lambda)\boldsymbol{G}_2\boldsymbol{S}^{-1}(\lambda)\}, \tag{7}$$

$\lambda \geq 0$, and therefore, using the definition of $\widetilde{\boldsymbol{\beta}}_2(\lambda)$ in equation (5),

$$\{\boldsymbol{I}_p - \lambda\boldsymbol{S}^{-1}(\lambda)\boldsymbol{G}_1\}^{-1}\boldsymbol{S}^{-1}(\lambda)\boldsymbol{G}_2\widehat{\boldsymbol{\beta}}_2 \sim \mathcal{N}\{\boldsymbol{\beta}_2, (\sigma_2^2/n_2)\boldsymbol{D}(\lambda)\},$$

with

$$\boldsymbol{D}(\lambda) = \{\boldsymbol{I}_p - \lambda\boldsymbol{S}^{-1}(\lambda)\boldsymbol{G}_1\}^{-1}\boldsymbol{S}^{-1}(\lambda)\boldsymbol{G}_2\boldsymbol{S}^{-1}(\lambda)\{\boldsymbol{I}_p - \lambda\boldsymbol{G}_1\boldsymbol{S}^{-1}(\lambda)\}^{-1}.$$

Algebra reveals that $\{\boldsymbol{I}_p - \lambda\boldsymbol{S}^{-1}(\lambda)\boldsymbol{G}_1\}^{-1}\boldsymbol{S}^{-1}(\lambda)\boldsymbol{G}_2 = \boldsymbol{I}_p$, which finally yields the familiar result: $\widehat{\boldsymbol{\beta}}_2 \sim \mathcal{N}\{0, n_2^{-1}\sigma_2^2\boldsymbol{G}_2^{-1}\}$. Therefore, confidence intervals based on equation (7) reduce to confidence intervals obtained from the MLE $\widehat{\boldsymbol{\beta}}_2$ and familiar Gaussian sampling distribution. That is, the debiasing that ensures the confidence interval is centered (on average) at $\boldsymbol{\beta}_2$ effectively negates the gain in efficiency of our estimator. As remarked by Obenchain (1977) in ridge regression, our shrinkage does not produce "shifted" confidence regions, and the MLE is the most suitable choice if one desires valid confidence intervals. Nonetheless, we will consider in Sections 4.3 and 5 if the biased estimator $\widetilde{\boldsymbol{\beta}}_2(\lambda)$ can be used to derive confidence intervals with *asymptotic* nominal coverage as $n_2 \to \infty$.

4.3 Asymptotic results in generalized linear models

Next we investigate the asymptotic properties of $\widetilde{\boldsymbol{\beta}}_2(\lambda)$ in generalized linear models as $n_2 \to \infty$. For $j = 1, 2$, let

$$\boldsymbol{A}(\boldsymbol{X}_j^\top\boldsymbol{\beta}) = \mathbf{diag}\big\{h'(\boldsymbol{X}_{ij}^\top\boldsymbol{\beta})\big\}_{i=1}^{n_j} = \mathbf{diag}\big\{b''(\boldsymbol{X}_{ij}^\top\boldsymbol{\beta})\big\}_{i=1}^{n_j} \in \mathbb{R}^{n_j\times n_j},$$

where $h(\boldsymbol{X}_{i1}^\top\boldsymbol{\beta}) = b'(\boldsymbol{X}_{ij}^\top\boldsymbol{\beta})$ is the inverse of the link function. Define $\boldsymbol{\Delta}(\boldsymbol{\beta}) = \{h(\boldsymbol{X}_{i1}^\top\boldsymbol{\beta}) - h(\boldsymbol{X}_{i1}^\top\widehat{\boldsymbol{\beta}}_1)\}_{i=1}^{n_1}$ and $\boldsymbol{S}(\boldsymbol{\beta};\lambda) = -\partial\boldsymbol{\Psi}(\boldsymbol{\beta};\lambda)/(\partial\boldsymbol{\beta})$. Then

$$\begin{aligned}\boldsymbol{S}(\boldsymbol{\beta};\lambda) &= \frac{1}{n_2}\sum_{i=1}^{n_2}\boldsymbol{X}_{i2}\boldsymbol{X}_{i2}^\top h'(\boldsymbol{X}_{i2}^\top\boldsymbol{\beta}) + \frac{\lambda}{n_1}\sum_{i=1}^{n_1}\boldsymbol{X}_{i1}\boldsymbol{X}_{i1}^\top h'(\boldsymbol{X}_{i1}^\top\boldsymbol{\beta}) \\ &= n_2^{-1}\boldsymbol{X}_2\boldsymbol{A}(\boldsymbol{X}_2^\top\boldsymbol{\beta})\boldsymbol{X}_2^\top + \lambda n_1^{-1}\boldsymbol{X}_1\boldsymbol{A}(\boldsymbol{X}_1^\top\boldsymbol{\beta})\boldsymbol{X}_1^\top.\end{aligned}$$

We assume the following conditions.

(C1) $\lim_{n_2\to\infty} n_2^{-1}\sum_{i=1}^{n_2}\{h(\boldsymbol{X}_{i2}^\top\boldsymbol{\beta}_2)\boldsymbol{X}_{i2}^\top\boldsymbol{\beta} - b(\boldsymbol{X}_{i2}^\top\boldsymbol{\beta})\}$ exists and is finite.

(C2) For any $\boldsymbol{\beta}$ between $\boldsymbol{\beta}_2$ and $\boldsymbol{\beta}_2^\star(\lambda)$ inclusive, the two matrices defined below exist and are positive definite:

$$\boldsymbol{v}_1(\boldsymbol{\beta}) = n_1^{-1}\boldsymbol{X}_1\boldsymbol{A}(\boldsymbol{X}_1^\top\boldsymbol{\beta})\boldsymbol{X}_1^\top \quad \text{and} \quad \boldsymbol{v}_2(\boldsymbol{\beta}) = \lim_{n_2\to\infty} n_2^{-1}\boldsymbol{X}_2\boldsymbol{A}(\boldsymbol{X}_2^\top\boldsymbol{\beta})\boldsymbol{X}_2^\top.$$

Denote by $\boldsymbol{s}(\boldsymbol{\beta};\lambda) = \lim_{n_2\to\infty}\mathbb{E}_{\boldsymbol{\beta}_2}\{\boldsymbol{S}(\boldsymbol{\beta};\lambda)\} = \lim_{n_2\to\infty}\boldsymbol{S}(\boldsymbol{\beta};\lambda) = \boldsymbol{v}_2(\boldsymbol{\beta}) + \lambda\boldsymbol{v}_1(\boldsymbol{\beta})$. Recall that the asymptotic variance of the MLE $\widehat{\boldsymbol{\beta}}_2$ is $d(\gamma_2)\boldsymbol{v}_2^{-1}(\boldsymbol{\beta}_2)$.

Lemma 1 is a standard result stating that the minimizer of our objective function converges to the minimizer of its expectation.

Lemma 1. *If Condition (C1) holds, then* $\widetilde{\boldsymbol{\beta}}_2(\lambda) - \boldsymbol{\beta}_2^\star(\lambda) = O_p(n_2^{-1/2})$, *for each* $\lambda \geq 0$.

Since we already have that $\lim_{\lambda\to 0^+}\boldsymbol{\beta}_2^\star(\lambda) = \boldsymbol{\beta}_2$ as $n_2\to\infty$, an immediate consequence of Lemma 1 is that $\widetilde{\boldsymbol{\beta}}_2(\lambda) \xrightarrow{p} \boldsymbol{\beta}_2$ as $n_2\to\infty$ and $\lambda\to 0^+$. More can be said, however, about the local behavior of $\widetilde{\boldsymbol{\beta}}_2(\lambda)$ depending on how quickly λ vanishes with n_2.

Lemma 2. *If Conditions (C1)–(C2) hold, and* $\lambda = O(n_2^{-1/2})$, *then as* $n_2\to\infty$,

$$n_2^{1/2}\boldsymbol{J}^{1/2}(\boldsymbol{\beta}_2;\lambda)\big\{\widetilde{\boldsymbol{\beta}}_2(\lambda) - \boldsymbol{\beta}_2 + \lambda n_1^{-1}\boldsymbol{S}^{-1}(\boldsymbol{\beta}_2;\lambda)\boldsymbol{X}_1\boldsymbol{\Delta}(\boldsymbol{\beta}_2)\big\} \xrightarrow{d} \mathcal{N}(\boldsymbol{0},\boldsymbol{I}_p),$$

where $\boldsymbol{J}(\boldsymbol{\beta};\lambda) = d(\gamma_2)\boldsymbol{S}^\top(\boldsymbol{\beta};\lambda)\boldsymbol{v}_2^{-1}(\boldsymbol{\beta})\boldsymbol{S}(\boldsymbol{\beta};\lambda)$ *is the observed Godambe information matrix (Godambe, 1991). Moreover, if* $\lambda = o(n_2^{-1/2})$, *then*

$$n_2^{1/2}\boldsymbol{J}^{1/2}(\boldsymbol{\beta}_2;\lambda)\{\widetilde{\boldsymbol{\beta}}_2(\lambda) - \boldsymbol{\beta}_2\} \xrightarrow{d} \mathcal{N}(\boldsymbol{0},\boldsymbol{I}_p), \quad n_2\to\infty.$$

To summarize, if λ vanishes not too rapidly, then there is a bias effect in the first-order asymptotic distribution approximation of $\widetilde{\boldsymbol{\beta}}_2$; if λ vanishes rapidly, then there is no bias effect in the first-order approximation. In either case, however, there is a reduction in the variance due to the combining of information in $\mathcal{D}_1$ and $\mathcal{D}_2$. This gain in efficiency can be seen, at least intuitively, by looking at the Godambe information matrix $\boldsymbol{J}(\boldsymbol{\beta}_2;\lambda)$. Indeed, since $\lim_{n_2\to\infty}\boldsymbol{S}(\boldsymbol{\beta};\lambda) = \boldsymbol{v}_2(\boldsymbol{\beta}) + \lambda\boldsymbol{v}_1(\boldsymbol{\beta})$ has eigenvalues strictly larger than those of $\boldsymbol{v}_2(\boldsymbol{\beta})$, it follows that $\boldsymbol{J}(\boldsymbol{\beta}_2;\lambda)$ has eigenvalues strictly smaller than those of $d(\gamma_2)\boldsymbol{v}_2^{-1}(\boldsymbol{\beta}_2)$. The following theorem, our main result, confirms the above intuition.

Theorem 2. *If Conditions (C1)–(C2) hold, then there exists a range of* $\lambda > 0$ *values, with upper limit* $O(n_2^{-1/2})$, *on which the asymptotic mean squared error (aMSE) of* $\widetilde{\boldsymbol{\beta}}_2(\lambda)$ *is strictly less than that of* $\widehat{\boldsymbol{\beta}}_2$.

The proof of Theorem 2 shows that $\text{aMSE}\{\widetilde{\boldsymbol{\beta}}_2(\lambda)\} < \text{aMSE}(\widehat{\boldsymbol{\beta}}_2)$ for all $\lambda > 0$ when $\boldsymbol{\delta} = \boldsymbol{0}$; when $\boldsymbol{\delta}\neq\boldsymbol{0}$, it proves that the aMSE of $\widetilde{\boldsymbol{\beta}}_2(\lambda)$ is monotonically decreasing over the range $[0,\lambda^\star]$ and monotonically increasing over the range $[\lambda^\star,\infty)$, for some $\lambda^\star > 0$ that does not have a closed-form. But the proof of Theorem 2 does provide a bound:

$$\lambda^\star > \frac{d(\gamma_2)}{n_2}\frac{\min_{r=1,\ldots,p}\{\kappa_r(\boldsymbol{\beta}_2)g_{r2}^{-1}(\boldsymbol{\beta}_2)\}}{\max\mathbf{diag}\{\boldsymbol{v}_1^{-1}(\boldsymbol{\beta}_2)\boldsymbol{X}_1\boldsymbol{\Delta}(\boldsymbol{\beta}_2)\boldsymbol{\Delta}^\top(\boldsymbol{\beta}_2)\boldsymbol{X}_1^\top\boldsymbol{v}_1^{-1}(\boldsymbol{\beta}_2)\}/n_1^2}, \tag{8}$$

where $g_{r2}(\boldsymbol{\beta}) > 0$, $r = 1, \ldots, p$, are the eigenvalues in increasing order of $\boldsymbol{v}_2(\boldsymbol{\beta})$ and $\kappa_r(\boldsymbol{\beta})$, $r = 1, \ldots, p$, are the eigenvalues in decreasing order of $\boldsymbol{v}_2^{1/2}(\boldsymbol{\beta})\boldsymbol{v}_1^{-1}(\boldsymbol{\beta})\boldsymbol{v}_2^{1/2}(\boldsymbol{\beta})$. This implies that the aMSE of $\widetilde{\boldsymbol{\beta}}_2(\lambda)$ is less than that of $\widehat{\boldsymbol{\beta}}_2$ for λ no more than the expression on the right-hand side of (8). The denominator of (8) is the nonlinear analogue of $\max_{r=1,\ldots,p} \delta_r^2$. To see this, by the mean value theorem we can write $\boldsymbol{X}_1\boldsymbol{\Delta}(\boldsymbol{\beta}_2) = n_1\boldsymbol{v}_1(\boldsymbol{c}_2)\boldsymbol{\delta}$ for some vector $\boldsymbol{c}_2$ between $\widehat{\boldsymbol{\beta}}_1$ and $\boldsymbol{\beta}_2$. Then the denominator can be rewritten as the maximum entry of

$$\mathbf{diag}\{\boldsymbol{v}_1^{-1}(\boldsymbol{\beta}_2)\boldsymbol{v}_1(\boldsymbol{c}_2)\boldsymbol{\delta}\boldsymbol{\delta}^\top\boldsymbol{v}_1(\boldsymbol{c}_2)\boldsymbol{v}_1^{-1}(\boldsymbol{\beta}_2)\}.$$

Recall that $\boldsymbol{v}_1(\boldsymbol{\beta})$ is the derivative of $n_1^{-1}\boldsymbol{\mu}_1(\boldsymbol{\beta})$. Thus, the denominator divides the distance $\boldsymbol{\delta}$ by the rate of change of the link function. When $\boldsymbol{\delta}$ is small, $\boldsymbol{v}_1^{-1}(\boldsymbol{\beta}_2)\boldsymbol{v}_1(\boldsymbol{c}_2) \approx \boldsymbol{I}_p$ and we recover the denominator in (6). From this, we draw the same conclusion as in the Gaussian linear model: if $\mathcal{D}_2$ is highly informative, then little weight should be given to $\mathcal{D}_1$, regardless of how informative $\mathcal{D}_1$ is.

For a data-driven choice of $\lambda^\star$, we propose

$$\begin{aligned}\tilde{\lambda} = \arg\min_{\lambda>0}\big[&d(\widehat{\gamma}_2)\mathrm{trace}\{\boldsymbol{S}^{-2}(\widehat{\boldsymbol{\beta}}_2;\lambda)\boldsymbol{V}_2(\widehat{\boldsymbol{\beta}}_2)\} \\ &+ n_2 n_1^{-2}\lambda^2\boldsymbol{\Delta}^\top(\widehat{\boldsymbol{\beta}}_2)\boldsymbol{X}_1^\top\boldsymbol{S}^{-2}(\widehat{\boldsymbol{\beta}}_2;\lambda)\boldsymbol{X}_1\boldsymbol{\Delta}(\widehat{\boldsymbol{\beta}}_2)\big],\end{aligned} \tag{9}$$

the minimizer of the estimated aMSE of $\widetilde{\boldsymbol{\beta}}_2(\lambda)$, where $\widehat{\gamma}_j$ the maximum likelihood estimator of γ_j in $\mathcal{D}_j$, $j = 1, 2$, and $\boldsymbol{V}_2(\boldsymbol{\beta}) = n_2^{-1}\boldsymbol{X}_2\boldsymbol{A}(\boldsymbol{X}_2^\top\boldsymbol{\beta})\boldsymbol{X}_2^\top$. Since we are assuming n_2 is large, we do not adjust for finite sample bias in $\boldsymbol{\Delta}(\widehat{\boldsymbol{\beta}}_2)\boldsymbol{\Delta}^\top(\widehat{\boldsymbol{\beta}}_2)$ as we did in the Gaussian linear model setting. Again, the minimizer always exists since $\boldsymbol{\Delta}(\widehat{\boldsymbol{\beta}}_2) \neq \mathbf{0}$ with probability 1.

As suggested in Section 4.2, Theorem 2 endows our approach with robustness to model misspecification or lack of information in $\mathcal{D}_1$ by guaranteeing (under conditions) that our approach always improves efficiency. This is far superior to traditional transfer learning which can suffer from degraded performance in $\mathcal{D}_2$, or "negative transfer" (Torrey and Shavlik, 2009). As a remedy, Li et al. (2022) and Tian and Feng (2022) develop methods that detect informative data sets $\mathcal{D}_1$ to avoid the pitfalls of negative transfer. Their approaches can handle settings with multiple candidate data sets $\mathcal{D}_1$: they find those most "similar" to $\mathcal{D}_2$ and use these for transferring. Unfortunately, their approaches can still result in degraded performance, as illustrated in Section 5.

Interestingly, the bound in equation (8) is $O(n_2^{-1})$, implying that the choice of λ that maximizes the asymptotic efficiency of $\widetilde{\boldsymbol{\beta}}_2(\lambda)$ is a $\lambda^\star$ strictly between $O(n_2^{-1/2})$ and $O(n_2^{-1})$; numerical results that help to confirm this are given in Section 5 below. This suggests the bias introduced by $\lambda^\star$ should be ignorable, while still affording us a small gain in efficiency. In conjunction with Lemma 2, this suggests a path to inference using the biased estimator $\widetilde{\boldsymbol{\beta}}_2(\tilde{\lambda})$. We propose to construct large-sample $100(1-\alpha)\%$ confidence intervals for $\boldsymbol{\beta}_2$ using

$$\widetilde{\boldsymbol{\beta}}_2(\tilde{\lambda}) \pm z_{\alpha/2} n_2^{-1/2}\boldsymbol{j}^{-1/2}\{\widehat{\boldsymbol{\beta}}_2;\tilde{\lambda}\}, \tag{10}$$

with $z_{\alpha/2}$ the $1-\frac{\alpha}{2}$ quantile of the standard normal distribution and $\boldsymbol{j}^{-1/2}(\boldsymbol{\beta};\lambda)$ the square root of the diagonal elements in $\{d(\widehat{\gamma}_2)\boldsymbol{S}^\top(\boldsymbol{\beta};\lambda)\boldsymbol{V}_2^{-1}(\boldsymbol{\beta})\boldsymbol{S}(\boldsymbol{\beta};\lambda)\}^{-1}$. Since $\boldsymbol{S}(\boldsymbol{\beta};\lambda) \succeq \boldsymbol{V}_2(\boldsymbol{\beta})$ in Loewner order, it follows that the confidence interval based on $\widetilde{\boldsymbol{\beta}}_2(\tilde{\lambda})$ in equation (10) is statistically more efficient than the classical confidence interval based on the sampling distribution of the MLE $\widehat{\boldsymbol{\beta}}_2$.

5 Empirical investigation

5.1 Objectives, setup, and take-aways

We examine the empirical performance of our proposed information-based shrinkage estimator (ISE) $\widetilde{\boldsymbol{\beta}}_2(\lambda)$ through simulations in linear and logistic regression models. For all settings, we compute $\widetilde{\boldsymbol{\beta}}_2(\lambda)$ for a range of λ values, and compute $\tilde{\lambda}$ and $\widetilde{\boldsymbol{\beta}}_2(\tilde{\lambda})$. We construct large sample 95% confidence intervals using equation (10). Finally, we investigate the robustness of our proposed approach to model misspecification in the assumed model for analyzing $\mathcal{D}_1$. Throughout, true values $\boldsymbol{\beta}_1$ and $\boldsymbol{\delta}$ are randomly sampled from suitable uniform distributions.

In addition to comparison with the MLE $\widehat{\boldsymbol{\beta}}_2$, we show that the behavior of our estimator $\widetilde{\boldsymbol{\beta}}_2(\lambda)$ more closely aligns with the intuition developed in Section 2 than competitor approaches, and that it is more efficient. In both the linear and logistic models, we compare our approach to the Trans-GLM estimator $\widehat{\boldsymbol{\beta}}_2^{TG}$ in Algorithm 2 of Tian and Feng (2022) with L_1 regularization (their default) using the R package and function `glmtrans`, and the pooled MLE $\widehat{\boldsymbol{\beta}}_2^{p}$ that estimates $\boldsymbol{\beta}_2$ with $\mathcal{D}_1$ and $\mathcal{D}_2$ by assuming $\boldsymbol{\beta}_2 = \boldsymbol{\beta}_1$. In the linear model, we compare our approach to the estimator $\widehat{\boldsymbol{\beta}}_2(\boldsymbol{W}_\lambda)$ of Chen et al. (2015) described in Section 3.1 over the same range of λ values as our estimator, and to $\widehat{\boldsymbol{\beta}}_2(\boldsymbol{W}_{\widehat{\lambda}})$ with $\widehat{\lambda}$ selected by minimizing the predictive mean squared error of $\widehat{\boldsymbol{\beta}}_2(\boldsymbol{W}_\lambda)$ using their bias-adjusted plug-in estimate (their version of our $\tilde{\lambda}$). We also compare our approach to the Trans-Lasso estimator $\widehat{\boldsymbol{\beta}}_2^{TL}$ of Li et al. (2022) (using their software defaults): their approach assumes all covariates are Gaussian with mean zero, and so we center our outcome and estimate $\boldsymbol{\beta}_2$ without its intercept. In the logistic model, we compare our approach to that of Zheng et al. (2019), who propose the estimator $\widehat{\boldsymbol{\beta}}_2(\boldsymbol{W}) = \boldsymbol{W}\widehat{\boldsymbol{\beta}}_2 + (\boldsymbol{I}_p - \boldsymbol{W})\widehat{\boldsymbol{\beta}}_1$ with the matrix $\boldsymbol{W}$ given by

$$\big[(\widehat{\boldsymbol{\beta}}_1 - \widehat{\boldsymbol{\beta}}_2)(\widehat{\boldsymbol{\beta}}_1 - \widehat{\boldsymbol{\beta}}_2)^\top + \big\{\boldsymbol{X}_1\boldsymbol{A}(\boldsymbol{X}_1^\top\widehat{\boldsymbol{\beta}}_1)\boldsymbol{X}_1^\top\big\}^{-1} + \big\{\boldsymbol{X}_2\boldsymbol{A}(\boldsymbol{X}_2^\top\widehat{\boldsymbol{\beta}}_2)\boldsymbol{X}_2^\top\big\}^{-1}\big]^{-1} \\ \big[(\widehat{\boldsymbol{\beta}}_1 - \widehat{\boldsymbol{\beta}}_2)(\widehat{\boldsymbol{\beta}}_1 - \widehat{\boldsymbol{\beta}}_2)^\top + \big\{\boldsymbol{X}_1\boldsymbol{A}(\boldsymbol{X}_1^\top\widehat{\boldsymbol{\beta}}_1)\boldsymbol{X}_1^\top\big\}^{-1}\big].$$

Across all simulations, we report 95% confidence interval empirical coverage (CP) of $\widetilde{\boldsymbol{\beta}}_2(\tilde{\lambda})$, mean $\tilde{\lambda}$ and its standard error, and empirical MSE (eMSE). Li et al. (2022); Chen et al. (2015); Zheng et al. (2019) do not provide a method for constructing confidence intervals, prohibiting comparison of inferential properties of these methods to ours. While Tian and Feng (2022) offer an alternative approach to construct confidence intervals for each element of $\boldsymbol{\beta}_2$, their proposal is not designed to yield a point estimator.

We show that the mean squared error of $\widetilde{\boldsymbol{\beta}}_2(\tilde{\lambda})$ is smaller than that of $\widehat{\boldsymbol{\beta}}_2$ and various competitors when n_2 is not too small. As expected, $\tilde{\lambda}$ is larger when n_2 or $\boldsymbol{\delta}$ are smaller. We show that, when n_2 is large, $\mathcal{D}_1$ is informative or $\boldsymbol{\delta}$ is small, the bias of our estimator $\widetilde{\boldsymbol{\beta}}_2(\tilde{\lambda})$ is negligible and empirical coverage reaches nominal levels. This phenomenon is unique to our approach: our choice of $\tilde{\lambda}$, in contrast to competitors, guarantees that the bias is negligible across a range of practical settings.

Finally, an additional simulation in Appendix B.2 investigates the trade-offs of concatenation versus single-source transfer learning when more than one source data set is available. We summarize the key take-aways from this investigation in the discussion of Section 7.

5.2 Setting I: Varying sample size

We study the performance of $\widetilde{\boldsymbol{\beta}}_2(\lambda)$ in the linear model with varying degrees of likelihood information in $\mathcal{D}_1$ relative to $\mathcal{D}_2$. We vary $n_1 \in \{50, 100, 500\}$, $n_2 \in \{50, 100, 500\}$: for each (n_1, n_2) pair, we simulate one data set $\mathcal{D}_1 = \{y_{i1}, \boldsymbol{X}_{i1}\}_{i=1}^{n_1}$. For each simulated $\mathcal{D}_1$, we generate 1000 data sets $\mathcal{D}_2 = \{y_{i2}, \boldsymbol{X}_{i2}\}_{i=1}^{n_2}$. Features $\boldsymbol{X}_{ij} \in \mathbb{R}^{11}$ consist of an intercept and ten continuous features independently generated from a standard Gaussian distribution. Outcomes are simulated from the Gaussian distribution with $\mathbb{E}(Y_{i2}) = \boldsymbol{X}_{ij}^\top \boldsymbol{\beta}_j$ and $\mathbb{V}(Y_{ij}) = 1$. True parameter values are set to $\boldsymbol{\beta}_1 = (1, -1.8, 2.6, 1.4, -3.6, 3.5, 2.4, -3.3, 1.8, -3.4, 2.8)$, $\boldsymbol{\delta} = (0.2, 0.1, 0.2, -0.1, 0.1, -0.1, 0.2, 0.2, 0.2, -0.1, 0.1)$ and $\boldsymbol{\beta}_2 = \widehat{\boldsymbol{\beta}}_1 + \boldsymbol{\delta}$. MLEs $\widehat{\boldsymbol{\beta}}_1$ and corresponding values $\boldsymbol{\beta}_2$ are reported in Appendix B.1. For each $\mathcal{D}_2$, we compute $\widetilde{\boldsymbol{\beta}}_2(\lambda)$ and $\widehat{\boldsymbol{\beta}}_2(\boldsymbol{W}_\lambda)$ for λ a sequence of 500 evenly spaced values in $[0, 5]$ for $n_2 = 50, 100$ and $[0, 0.2]$ for $n_2 = 500$.

Empirical mean squared errors (eMSE) of $\widetilde{\boldsymbol{\beta}}_2(\lambda)$, $\widehat{\boldsymbol{\beta}}_2(\boldsymbol{W}_\lambda)$ and $\widehat{\boldsymbol{\beta}}_2$ are depicted in Figure 2. Both $\widetilde{\boldsymbol{\beta}}_2(\lambda)$ and $\widehat{\boldsymbol{\beta}}_2(\boldsymbol{W}_\lambda)$ have smaller eMSE than $\widehat{\boldsymbol{\beta}}_2$ across all (n_1, n_2) pairs. For fixed n_2, the eMSEs of $\widetilde{\boldsymbol{\beta}}_2(\lambda)$ and $\widehat{\boldsymbol{\beta}}_2(\boldsymbol{W}_\lambda)$ decrease modestly as n_1 increases. For fixed n_1, the eMSEs of $\widetilde{\boldsymbol{\beta}}_2(\lambda)$ and $\widehat{\boldsymbol{\beta}}_2(\boldsymbol{W}_\lambda)$ decrease much more drastically as n_2 increases. For fixed n_1, the minimum of eMSE$\{\widetilde{\boldsymbol{\beta}}_2(\lambda)\}$ is achieved at smaller λ for larger n_2, with eMSE$\{\widetilde{\boldsymbol{\beta}}_2(\lambda)\} <$ eMSE$(\widehat{\boldsymbol{\beta}}_2)$ for shrinking ranges of λ values as n_2 increases. This is consistent with the intuition developed in Sections 2 and 4: little weight should be placed on $\mathcal{D}_1$, regardless of how informative it is, when $\mathcal{D}_2$ is highly informative. In contrast, $\widehat{\boldsymbol{\beta}}_2(\boldsymbol{W}_\lambda)$ continues to place a large weight on $\mathcal{D}_1$ even when n_2 is large. Theoretically, this results in substantial bias of the most efficient $\widehat{\boldsymbol{\beta}}_2(\boldsymbol{W}_\lambda)$, while the bias of the most efficient $\widetilde{\boldsymbol{\beta}}_2(\lambda)$ will become negligible, allowing us to construct confidence intervals that reach nominal levels (discussed below). Practically, this suggests $\widehat{\boldsymbol{\beta}}_2(\boldsymbol{W}_\lambda)$ struggles with predicting how much efficiency gain can be expected from incorporating inference in $\mathcal{D}_1$ when n_2 is large. Our approach is advantageous when proposing practical guidelines for improving inference in data integration.

A relevant question that the theory in Section 4 is unable to answer is how large $\lambda^\star$ is as a function of n_2. That is, we have an upper bound that is $O(n_2^{-1/2})$ and a lower bound that is $O(n_2^{-1})$, but what about $\lambda^\star$ itself? For an empirical check, this simulation provides realizations of $\tilde{\lambda}$ for a range of n_2 values, so if we regress $\log \tilde{\lambda}$ against $\log n_2$, then the estimated slope coefficient would provide some information about how fast $\lambda^\star$ vanishes with n_2. Based on the results summarized in Table 1, the estimated slope is -0.77, with a 95% confidence interval $(-0.78, -0.76)$, which is well within the range $[-1.0, -0.5]$ that we were looking in. This provides empirical support of the claim that $\lambda^\star$ is strictly between our lower and upper bounds, which are $O(n_2^{-1})$ and $O(n_2^{-1/2})$, respectively.

We show in Table 1 that our proposed $\tilde{\lambda}$ results in a reduced eMSE by reporting eMSE of $\widetilde{\boldsymbol{\beta}}_2(\tilde{\lambda})$, $\widehat{\boldsymbol{\beta}}_2(\boldsymbol{W}_{\tilde{\lambda}})$, $\widehat{\boldsymbol{\beta}}_2^{TG}$, $\widehat{\boldsymbol{\beta}}_2^{TL}$, $\widehat{\boldsymbol{\beta}}_2^{p}$ and $\widehat{\boldsymbol{\beta}}_2$ (Monte Carlo standard errors, MCse, of eMSE are reported in Table 11 of the Appendix). Aside from the pooled MLE, our estimator achieves the smallest eMSE when $n_2 \leq n_1$. The Trans-GLM approach of Tian and Feng (2022) apparently gives too much weight to $\mathcal{D}_1$ and suffers from negative transfer in most settings. We report the mean value of $\tilde{\lambda}$ and the median (over the 11 features) CP of $\widetilde{\boldsymbol{\beta}}_2(\tilde{\lambda})$, $\widehat{\boldsymbol{\beta}}_2^{p}$ and $\widehat{\boldsymbol{\beta}}_2$ over the 1000 simulated $\mathcal{D}_2$ in Table 2. Coverage of $\widetilde{\boldsymbol{\beta}}_2(\tilde{\lambda})$ reaches the nominal 95% coverage when $n_2 = 500$. As discussed above, this is a consequence of the negligible bias when n_2 is

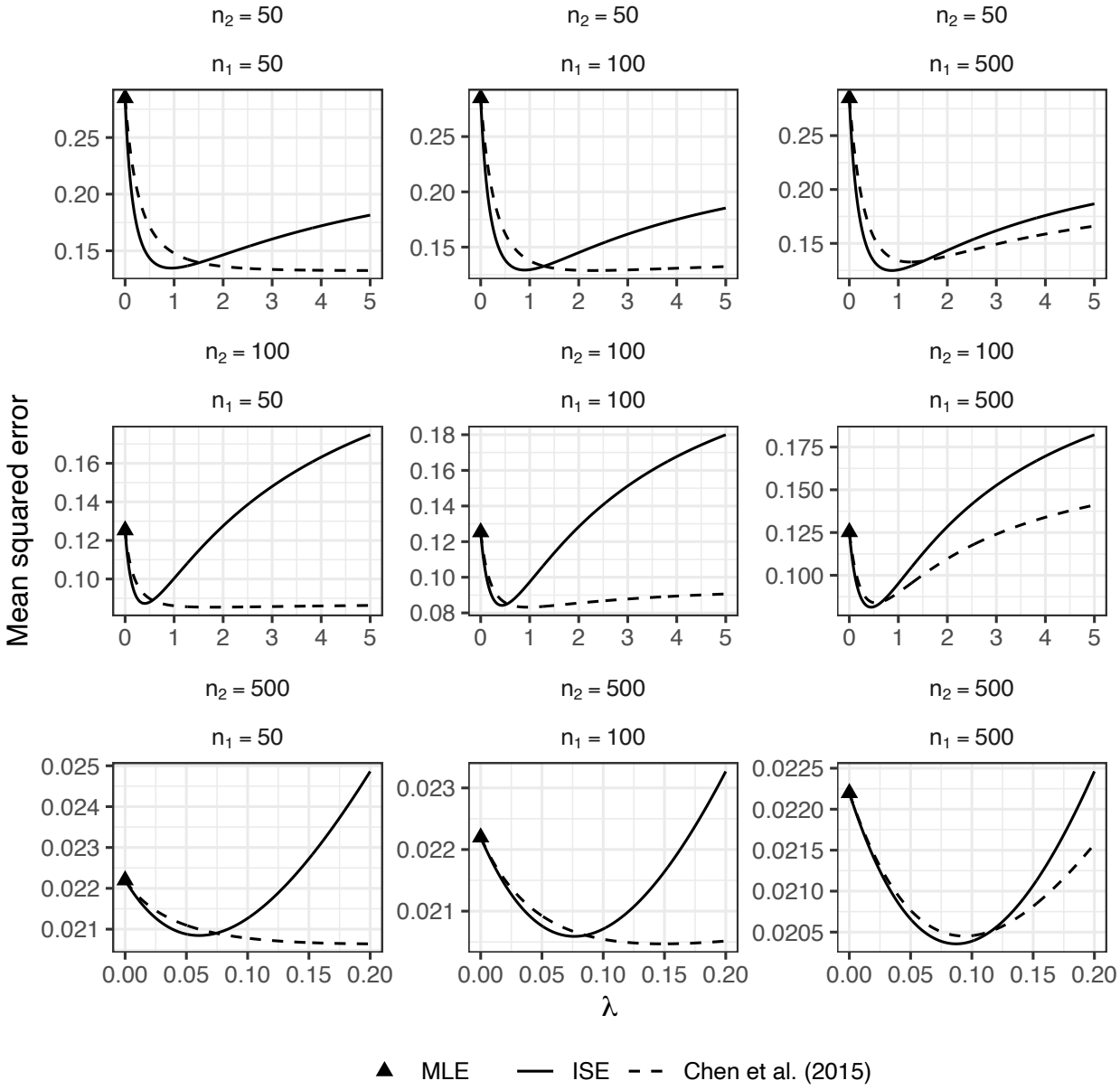

Figure 2: eMSE of $\widetilde{\boldsymbol{\beta}}_2(\lambda)$, $\widehat{\boldsymbol{\beta}}_2(\boldsymbol{W}_\lambda)$ and $\widehat{\boldsymbol{\beta}}_2$ over 1000 simulated $\mathcal{D}_2$ in Setting I.

large or one of $\tilde{\lambda}$, $\boldsymbol{\delta}$ is sufficiently small. The bias incurred by assuming $\boldsymbol{\beta}_1 = \boldsymbol{\beta}_2$ leads to inflated eMSE of the pooled MLE and severe under-coverage when $n_2 > n_1 > 50$. Of course, the value of $\boldsymbol{\delta}$ is not known in practice and the use of the pooled analysis runs the substantial risk of introducing a large bias that is not offset by a reduction in variance.

5.3 Setting II: Varying feature information

We study the performance of $\widetilde{\boldsymbol{\beta}}_2(\lambda)$ in the logistic model with varying degrees of likelihood information in $\mathcal{D}_1$ relative to $\mathcal{D}_2$ when $\boldsymbol{\delta} = \mathbf{0}$ and when $\boldsymbol{\delta} \neq \mathbf{0}$. We fix $(n_1, n_2) = (500, 500)$

n_1	n_2	eMSE$\cdot 10^2$					
		$\widetilde{\boldsymbol{\beta}}_2(\tilde{\lambda})$	$\widehat{\boldsymbol{\beta}}_2(\boldsymbol{W}_{\widehat{\lambda}})$	$\widehat{\boldsymbol{\beta}}_2^{TG}$	$\widehat{\boldsymbol{\beta}}_2^{TL}$	$\widehat{\boldsymbol{\beta}}_2^{p}$	$\widehat{\boldsymbol{\beta}}_2$
50	50	17.1	17.2	32.4	27.7	13.5	28.5
	100	9.69	9.56	20.8	12.6	8.80	12.5
	500	2.16	2.11	2.23	2.21	2.13	2.22
100	50	16.5	16.8	27.3	24.2	14.5	28.5
	100	9.34	9.33	15.8	11.0	9.74	12.5
	500	2.12	2.09	2.26	2.10	2.33	2.22
500	50	16.0	16.6	23.6	24.2	21.6	28.5
	100	9.06	9.24	12.7	11.0	18.2	12.5
	500	2.10	2.10	2.38	2.12	6.93	2.22

Table 1: Setting I: eMSE of $\widetilde{\boldsymbol{\beta}}_2(\tilde{\lambda})$, $\widehat{\boldsymbol{\beta}}_2(\boldsymbol{W}_{\widehat{\lambda}})$, $\widehat{\boldsymbol{\beta}}_2^{TG}$, $\widehat{\boldsymbol{\beta}}_2^{TL}$, $\widehat{\boldsymbol{\beta}}_2^{p}$, $\widehat{\boldsymbol{\beta}}_2$.

n_1	n_2	$\tilde{\lambda}$ (s.e.)	CP		
			$\widetilde{\boldsymbol{\beta}}_2(\tilde{\lambda})$	$\widehat{\boldsymbol{\beta}}_2^{p}$	$\widehat{\boldsymbol{\beta}}_2$
50	50	$4.87 \cdot 10^{-1}$ $(3.81 \cdot 10^{-1})$	88.4	94.9	94.5
	100	$3.05 \cdot 10^{-1}$ $(1.93 \cdot 10^{-1})$	90.8	93.5	94.4
	500	$6.16 \cdot 10^{-2}$ $(1.90 \cdot 10^{-2})$	94.5	95.4	94.9
100	50	$4.97 \cdot 10^{-1}$ $(3.45 \cdot 10^{-1})$	88.1	93.3	94.5
	100	$3.31 \cdot 10^{-1}$ $(1.74 \cdot 10^{-1})$	88.7	90.3	94.4
	500	$7.62 \cdot 10^{-2}$ $(1.92 \cdot 10^{-2})$	93.5	93.2	94.9
500	50	$4.89 \cdot 10^{-1}$ $(3.07 \cdot 10^{-1})$	89.7	24.7	94.5
	100	$3.37 \cdot 10^{-1}$ $(1.62 \cdot 10^{-1})$	91.7	39.6	94.4
	500	$8.50 \cdot 10^{-2}$ $(1.87 \cdot 10^{-2})$	93.9	69.1	94.9

Table 2: Setting I: mean $\tilde{\lambda}$ (s.e.), %CP of $\widetilde{\boldsymbol{\beta}}_2(\tilde{\lambda})$, $\widehat{\boldsymbol{\beta}}_2^{TL}$, $\widehat{\boldsymbol{\beta}}_2^{p}$, $\widehat{\boldsymbol{\beta}}_2$.

and vary $\boldsymbol{X}_1\boldsymbol{X}_1^\top$ by simulating two data sets $\mathcal{D}_1$: one with large $\boldsymbol{X}_1\boldsymbol{X}_1^\top$ and one with small $\boldsymbol{X}_1\boldsymbol{X}_1^\top$. For each simulated $\mathcal{D}_1$, we generate 1000 data sets $\mathcal{D}_2 = \{y_{i2}, \boldsymbol{X}_{i2}\}_{i=1}^{n_2}$. Features $\boldsymbol{X}_{ij} \in \mathbb{R}^5$ consist of an intercept and four continuous features independently generated from $\mathcal{N}(0, 0.75^2)$ and $\mathcal{N}(0, 3^2)$ distributions for $\boldsymbol{X}_1\boldsymbol{X}_1^\top$ small and large respectively. Outcomes are simulated with mean $\mathbb{E}(Y_{i2}) = \text{expit}(\boldsymbol{X}_{ij}^\top\boldsymbol{\beta}_j)$ from the Bernoulli distribution. True parameter values are set to $\boldsymbol{\beta}_1 = (1, -1.8, -1.2, 1.6, 0.2)$ and $\boldsymbol{\beta}_2 = \widehat{\boldsymbol{\beta}}_1 + \boldsymbol{\delta}$. We let $\boldsymbol{\delta}$ take two values: $\boldsymbol{\delta} = (0, 0.25, 0, -0.25, 0.25)$ ($\boldsymbol{\delta} \neq \mathbf{0}$) and $\boldsymbol{\delta} = \mathbf{0}$. MLEs $\widehat{\boldsymbol{\beta}}_1$ and corresponding values $\boldsymbol{\beta}_2$ are reported in Appendix B.1. For each $\mathcal{D}_2$, we compute $\widetilde{\boldsymbol{\beta}}_2(\lambda)$ for λ a sequence of 100 evenly spaced values in $[0, 7]$.

Empirical mean squared error (eMSE) of $\widetilde{\boldsymbol{\beta}}_2(\lambda)$ is depicted in Figure 3. When $\boldsymbol{\delta} = \mathbf{0}$, the smallest eMSE$\{\widetilde{\boldsymbol{\beta}}_2(\lambda)\}$ is achieved for λ only marginally smaller when $\mathcal{D}_1$ is less informative ($\boldsymbol{X}_1\boldsymbol{X}_1^\top$ is small): our estimator uses more or less the same amount of information from $\mathcal{D}_1$

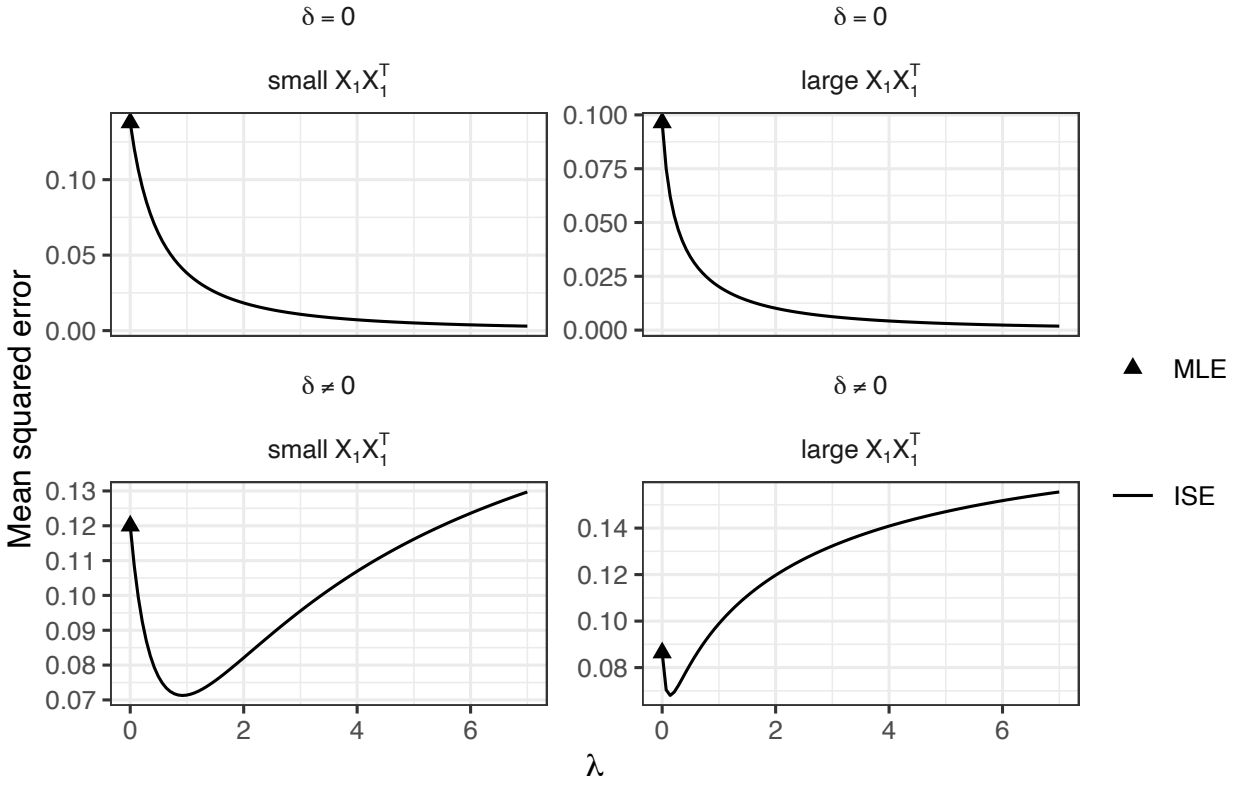

Figure 3: eMSE of $\widetilde{\boldsymbol{\beta}}_2(\lambda)$ and $\widehat{\boldsymbol{\beta}}_2$ over 1000 simulated $\mathcal{D}_2$ in Setting II.

when $\boldsymbol{\delta} = \mathbf{0}$. When $\boldsymbol{\delta} \neq \mathbf{0}$, however, the smallest eMSE$\{\widetilde{\boldsymbol{\beta}}_2(\lambda)\}$ is achieved for smaller λ when $\mathcal{D}_1$ is informative: when the two data sets are different and $\mathcal{D}_1$ provides sufficient information to discriminate between the two, our approach does not use as much information from $\mathcal{D}_1$.

We show in Table 3 that $\tilde{\lambda}$ results in a reduced eMSE by reporting eMSE of $\widetilde{\boldsymbol{\beta}}_2(\tilde{\lambda})$, $\widehat{\boldsymbol{\beta}}_2(\boldsymbol{W})$, $\widehat{\boldsymbol{\beta}}_2^{TG}$, $\widehat{\boldsymbol{\beta}}_2^{p}$ and $\widehat{\boldsymbol{\beta}}_2$ (MCse of eMSE are reported in Table 12 of the Appendix). Aside from the pooled MLE, our estimator achieves the smallest eMSE across all settings. As a sample mean, the eMSE is approximately normally distributed given the large number of simulations (1000 Monte Carlo replicates). An approximate 95% confidence interval for the MSE can be constructed following the formula eMSE $\pm$ 1.96 MCse and two-sample z-tests show that, for example, the MSEs for $\widehat{\boldsymbol{\beta}}_2^{TG}$ and $\widehat{\boldsymbol{\beta}}_2$ are not statistically significantly different at level 0.05 when $\boldsymbol{\delta} = \mathbf{0}$. When $\boldsymbol{X}_1\boldsymbol{X}_1^\top$ is large, $\mathcal{D}_1$ provides a substantial amount of information on $\boldsymbol{\beta}_2$. When $\boldsymbol{\delta} = \mathbf{0}$, this information is used by turning the dial parameter λ to a large value, whereas when $\boldsymbol{\delta} \neq \mathbf{0}$, the dial should not be turned very far. In other words, the likelihood in $\mathcal{D}_1$ is "sharp", and the distance between $f_1^{(n_1)}(\boldsymbol{y}_1; \widehat{\boldsymbol{\beta}}_1, \gamma_1)$ and $f_1^{(n_1)}(\boldsymbol{y}_1; \boldsymbol{\beta}_2, \gamma_1)$ is either small (when $\boldsymbol{\delta} = \mathbf{0}$) or large (when $\boldsymbol{\delta} \neq \mathbf{0}$). On the other hand, when $\boldsymbol{X}_1\boldsymbol{X}_1^\top$ is small, there is insufficient information in $\mathcal{D}_1$ to distinguish $\boldsymbol{\beta}_2$ and $\widehat{\boldsymbol{\beta}}_1$, and more weight can be placed on $\mathcal{D}_1$ with little loss of efficiency when $\boldsymbol{\delta} \neq \mathbf{0}$. Our estimator uses the abundance of information in $\mathcal{D}_1$ differently based on the value of $\boldsymbol{\delta}$. In Table 4, we also report the mean value of $\tilde{\lambda}$ and the median (over the 5 features) CP of $\widetilde{\boldsymbol{\beta}}_2(\lambda)$, $\widehat{\boldsymbol{\beta}}_2^{p}$ and $\widehat{\boldsymbol{\beta}}_2$ over the 1000

$\boldsymbol{\delta}$	$\boldsymbol{X}_1\boldsymbol{X}_1^\top$	eMSE·10				
		$\widetilde{\boldsymbol{\beta}}_2(\tilde{\lambda})$	$\widehat{\boldsymbol{\beta}}_2(\boldsymbol{W})$	$\widehat{\boldsymbol{\beta}}_2^{TG}$	$\widehat{\boldsymbol{\beta}}_2^{p}$	$\widehat{\boldsymbol{\beta}}_2$
$= \mathbf{0}$	large	0.310	0.751	0.971	0.202	0.964
	small	0.464	1.11	1.32	0.381	1.38
$\neq \mathbf{0}$	large	0.773	0.805	0.930	0.989	0.864
	small	0.849	1.06	1.20	0.714	1.20

Table 3: Setting II: eMSE of $\widetilde{\boldsymbol{\beta}}_2(\tilde{\lambda})$, $\widehat{\boldsymbol{\beta}}_2(\boldsymbol{W})$, $\widehat{\boldsymbol{\beta}}_2^{TG}$, $\widehat{\boldsymbol{\beta}}_2^{p}$, $\widehat{\boldsymbol{\beta}}_2$.

$\boldsymbol{\delta}$	$\boldsymbol{X}_1\boldsymbol{X}_1^\top$	$\tilde{\lambda}$ (s.e.)	CP		
			$\widetilde{\boldsymbol{\beta}}_2(\tilde{\lambda})$	$\widehat{\boldsymbol{\beta}}_2^{p}$	$\widehat{\boldsymbol{\beta}}_2$
$= \mathbf{0}$	large	2.83 (6.42)	95.1	99.7	95.3
	small	2.27 (3.20)	95.3	99.0	95.5
$\neq \mathbf{0}$	large	0.120 (0.266)	93.1	77.1	94.7
	small	0.720 (0.476)	80.5	93.8	94.9

Table 4: Setting II: mean $\tilde{\lambda}$ (s.e.), %CP of $\widetilde{\boldsymbol{\beta}}_2(\tilde{\lambda})$, $\widehat{\boldsymbol{\beta}}_2^{p}$, $\widehat{\boldsymbol{\beta}}_2$.

simulated $\mathcal{D}_2$. Empirical coverage of $\widetilde{\boldsymbol{\beta}}_2(\tilde{\lambda})$ reaches the nominal level in all but the most difficult setting, when $\mathcal{D}_1$ and $\mathcal{D}_2$ are dissimilar and $\mathcal{D}_1$ is less informative. The CP and eMSE of the pooled MLE reveals over- and under-coverage when $\boldsymbol{\delta} = \mathbf{0}$ ($\boldsymbol{\beta}_2 = \widehat{\boldsymbol{\beta}}_1$) and $\boldsymbol{\delta} \neq \mathbf{0}$ ($\boldsymbol{\beta}_2 \neq \widehat{\boldsymbol{\beta}}_1$), respectively.

5.4 Setting III: Varying parameter distance

In Setting III, we study the performance of $\widetilde{\boldsymbol{\beta}}_2(\lambda)$ in the logistic model as a function of $\boldsymbol{\delta}$ with both independent and correlated features. With independent features, we fix $(n_1, n_2) = (500, 500)$ and generate one data set $\mathcal{D}_1 = \{y_{i1}, \boldsymbol{X}_{i1}\}_{i=1}^{n_1}$ with $\boldsymbol{\beta}_1 = (1, -0.5, 0.5)$, yielding $\widehat{\boldsymbol{\beta}}_1 = (0.966, -0.563, 0.551)$. Source features $\boldsymbol{X}_{i1} \in \mathbb{R}^3$ consist of an intercept and two continuous features independently generated from a normal distribution with mean 0 and variance 4. Target features $\boldsymbol{X}_{i2} \in \mathbb{R}^3$ consist of an intercept and two continuous features independently generated from a standard normal distribution. The correlation between features in source and target data sets is $\rho = 0$. With correlated features, we fix $(n_1, n_2) = (500, 100)$ and generate one data set $\mathcal{D}_1 = \{y_{i1}, \boldsymbol{X}_{i1}\}_{i=1}^{n_1}$ with $\boldsymbol{\beta}_1 = (1, -0.5, 0.5)$, yielding $\widehat{\boldsymbol{\beta}}_1 = (0.881, -0.516, 0.571)$. Source and target features $\boldsymbol{X}_{ij} \in \mathbb{R}^3$ consist of an intercept and two continuous features generated from a multivariate Gaussian distribution with mean 0, correlation $\rho = 0.4$ and variance 1. With independent and correlated features, we set $\boldsymbol{\beta}_2 = \widehat{\boldsymbol{\beta}}_1 + \boldsymbol{\delta}$ and let $\boldsymbol{\delta}$ take values $\boldsymbol{\delta}_1 = (0, 0, 0)$, $\boldsymbol{\delta}_2 = (0, 1, 0)$ and $\boldsymbol{\delta}_3 = (0, 2, 0)$. For each value $\boldsymbol{\delta}_j$, we generate 1000 data sets $\mathcal{D}_2 = \{y_{i2}, \boldsymbol{X}_{i2}\}_{i=1}^{n_2}$. Outcomes are simulated with $\mathbb{E}(Y_{ij}) = \text{expit}(\boldsymbol{X}_{ij}^\top \boldsymbol{\beta}_j)$ from the Bernoulli distribution, with features generated as in the source data set for independent and correlated features, respectively. For each $\mathcal{D}_2$, we

compute $\widetilde{\boldsymbol{\beta}}_2(\lambda)$ for λ a sequence of 100 evenly spaced values in $[0, 10]$, $[0, 0.02]$ and $[0, 0.01]$ for $\boldsymbol{\delta}_1, \boldsymbol{\delta}_2, \boldsymbol{\delta}_3$ respectively with independent features, and λ a sequence of 100 evenly spaced values in $[0, 10]$, $[0, 0.6]$ and $[0, 0.2]$ for $\boldsymbol{\delta}_1, \boldsymbol{\delta}_2, \boldsymbol{\delta}_3$ respectively with correlated features.

Empirical mean squared error (eMSE) of $\widetilde{\boldsymbol{\beta}}_2(\lambda)$ is depicted in Figure 4. The smallest

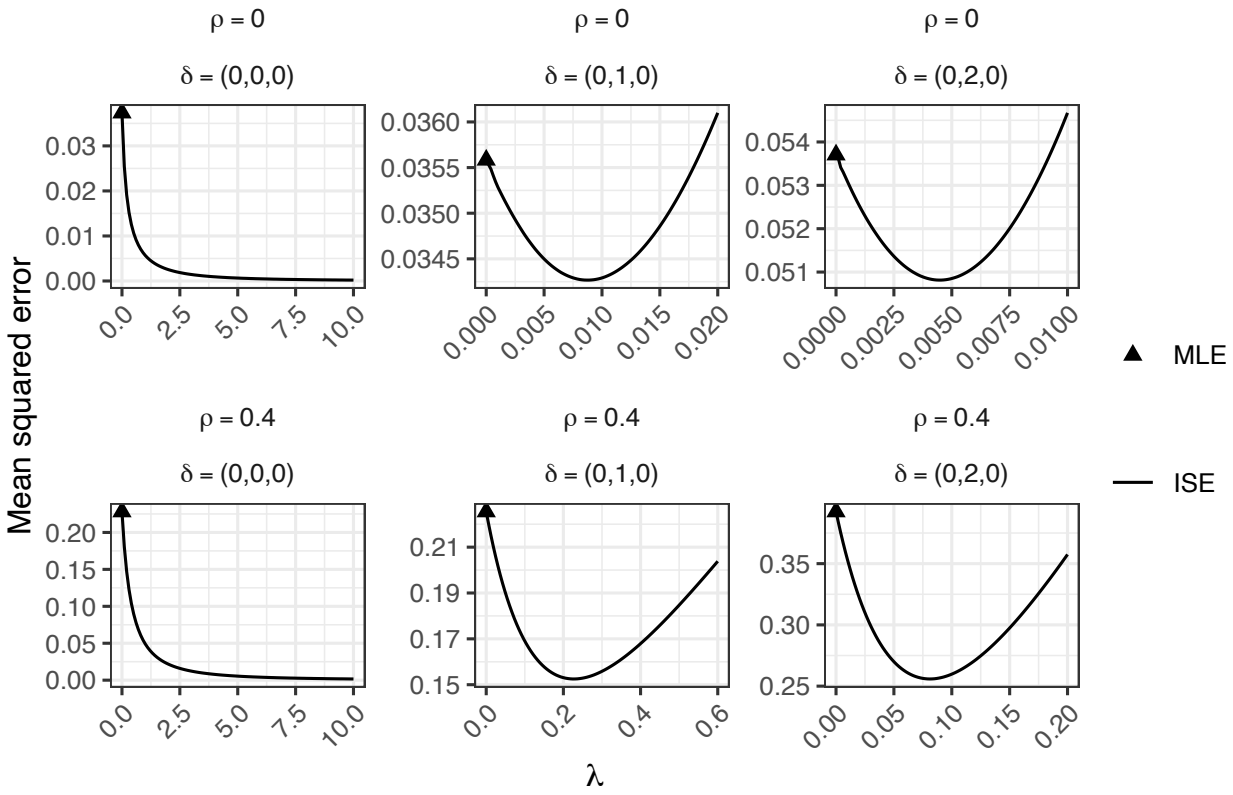

Figure 4: eMSE of $\widetilde{\boldsymbol{\beta}}_2(\lambda)$ and $\widehat{\boldsymbol{\beta}}_2$ over 1000 simulated $\mathcal{D}_2$ in Setting III.

eMSE$\{\widetilde{\boldsymbol{\beta}}_2(\lambda)$ is achieved for smaller values of λ when $\boldsymbol{\delta}$ is larger. We show in Table 5 that $\tilde{\lambda}$ results in a reduced eMSE by reporting eMSE of $\widetilde{\boldsymbol{\beta}}_2(\tilde{\lambda})$, $\widehat{\boldsymbol{\beta}}_2(\boldsymbol{W})$, $\widehat{\boldsymbol{\beta}}_2^{TG}$, $\widehat{\boldsymbol{\beta}}_2^{p}$ and $\widehat{\boldsymbol{\beta}}_2$ (Monte Carlo standard errors of eMSE are reported in Table 13 of the Appendix). Our estimator's efficiency gain remains robust to substantial differences between $\mathcal{D}_1$ and $\mathcal{D}_2$, and our proposed $\tilde{\lambda}$ consistently minimizes eMSE over competitors. In Table 6, we report mean value of $\tilde{\lambda}$ and the median (over the 2 features) CP of $\widetilde{\boldsymbol{\beta}}_2(\tilde{\lambda})$, $\widehat{\boldsymbol{\beta}}_2^{p}$ and $\widehat{\boldsymbol{\beta}}_2$ over the 1000 simulated $\mathcal{D}_2$. Empirical coverage of $\widetilde{\boldsymbol{\beta}}_2(\tilde{\lambda})$ reaches the nominal level in all settings. The pooled MLE shows inflated eMSE when $\boldsymbol{\delta} \neq \mathbf{0}$. With independent features, $\boldsymbol{\beta}_2$ is over- and under-covered when $\boldsymbol{\delta} = \mathbf{0}$ and $\boldsymbol{\delta} \neq \mathbf{0}$, respectively, whereas it is over-covered for all values of $\boldsymbol{\delta}$ with correlated features.

5.5 Setting IV: Robustness to model misspecification

In Setting IV, we study the robustness of the gain in MSE of $\widetilde{\boldsymbol{\beta}}_2(\lambda)$ in the linear model under model misspecification. We fix $(n_1, n_2) = (500, 500)$ and generate three data sets

$\boldsymbol{\delta}$	eMSE$\cdot 10^2$				
	$\widetilde{\boldsymbol{\beta}}_2(\tilde{\lambda})$	$\widehat{\boldsymbol{\beta}}_2(\boldsymbol{W})$	$\widehat{\boldsymbol{\beta}}_2^{TG}$	$\widehat{\boldsymbol{\beta}}_2^{p}$	$\widehat{\boldsymbol{\beta}}_2$
(i) independent features, $\rho = 0$					
$(0,0,0)$	1.45	2.55	4.21	0.572	3.73
$(0,1,0)$	3.53	3.84	3.64	58.3	3.56
$(0,2,0)$	5.21	30.2	5.22	259	5.37
(ii) correlated features, $\rho = 0.4$					
$(0,0,0)$	8.34	16.9	28.1	0.546	22.8
$(0,1,0)$	18.9	29.5	21.0	67.1	22.5
$(0,2,0)$	31.6	146	31.0	287	39.2

Table 5: Setting III: eMSE of $\widetilde{\boldsymbol{\beta}}_2(\tilde{\lambda})$, $\widehat{\boldsymbol{\beta}}_2(\boldsymbol{W})$, $\widehat{\boldsymbol{\beta}}_2^{TG}$, $\widehat{\boldsymbol{\beta}}_2^{p}$, $\widehat{\boldsymbol{\beta}}_2$.

$\boldsymbol{\delta}$	$\tilde{\lambda}$ (s.e.)	CP		
		$\widetilde{\boldsymbol{\beta}}_2(\tilde{\lambda})$	$\widehat{\boldsymbol{\beta}}_2^{p}$	$\widehat{\boldsymbol{\beta}}_2$
(i) independent features, $\rho = 0$				
$(0,0,0)$	3.43 (26.2)	94.3	100	94.8
$(0,1,0)$	$5.56 \cdot 10^{-3}$ $(1.18 \cdot 10^{-3})$	94.7	89.4	95.5
$(0,2,0)$	$2.05 \cdot 10^{-3}$ $(1.93 \cdot 10^{-4})$	94.9	7.20	95.0
(ii) correlated features, $\rho = 0.4$				
$(0,0,0)$	3.22 (9.96)	94.9	100	95.1
$(0,1,0)$	0.197 (0.162)	95.2	99.9	95.6
$(0,2,0)$	0.0483 (0.0183)	95.2	99.3	95.5

Table 6: Setting III: mean $\tilde{\lambda}$ (s.e.), %CP of $\widetilde{\boldsymbol{\beta}}_2(\tilde{\lambda})$, $\widehat{\boldsymbol{\beta}}_2^{p}$, $\widehat{\boldsymbol{\beta}}_2$.

$\mathcal{D}_1 = \{y_{i1}, \boldsymbol{X}_{i1}\}_{i=1}^{n_1}$ constructed as follows:

(i) (Cauchy) outcomes are generated as $y_{i1} = \boldsymbol{X}_{i1}\boldsymbol{\beta}_1 + \epsilon_{ij}$ with ϵ_{ij} independent standard Cauchy random variables;

(ii) (dropped Z) outcomes y_{i1} are generated from a Gaussian distribution with mean $\boldsymbol{X}_{i1}\boldsymbol{\beta}_1 + Z_{i1}$ and variance $\mathbb{V}(Y_{i1}) = 1$, where Z_{i1} is generated from a standard Gaussian distribution and acts as an unmeasured confounder;

(iii) (X^2) outcomes are generated as $y_{i1} = \mathbf{diag}(\boldsymbol{X}_{i1}\boldsymbol{X}_{i1}^\top)\boldsymbol{\beta}_1 + \epsilon_{ij}$, i.e., the relationship between outcome and squared features is linear, with ϵ_{ij} independent standard Gaussian random variables.

True parameter values are set to $\boldsymbol{\beta}_1 = (1, -1.8, 2.6, 1.4, -3.6, 3.5, 2.4, -3.3, 1.8, -3.4, 2.8, 1)$ and $\boldsymbol{\beta}_2 = \boldsymbol{\beta}_1 + (0, 0.1, 0, -0.1, 0.1, -0.1, 0, 0, 0, -0.1, 0.1)$, which differs from previous settings; we prefer to use $\boldsymbol{\beta}_1$ to define $\boldsymbol{\beta}_2$ since $\widehat{\boldsymbol{\beta}}_1$ will be substantially biased in these misspecified settings. MLEs $\widehat{\boldsymbol{\beta}}_1$ and corresponding values $\boldsymbol{\delta}$ are reported in Appendix B.1. For each simulated $\mathcal{D}_1$, we generate 1000 data sets $\mathcal{D}_2 = \{y_{i2}, \boldsymbol{X}_{i2}\}_{i=1}^{n_2}$ as in Setting I. For each $\mathcal{D}_2$, we compute $\widetilde{\boldsymbol{\beta}}_2(\lambda)$ and $\widehat{\boldsymbol{\beta}}_2(\boldsymbol{W}_\lambda)$ for λ a sequence of 100 evenly spaced values in $[0, 0.02]$, $[0, 0.4]$ and $[0, 5 \times 10^{-4}]$ for misspecifications (i), (ii) and (iii) respectively.

Empirical mean squared errors (eMSE) of $\widetilde{\boldsymbol{\beta}}_2(\lambda)$, $\widehat{\boldsymbol{\beta}}_2(\boldsymbol{W}_\lambda)$ and $\widehat{\boldsymbol{\beta}}_2$ are depicted in Figure 5.

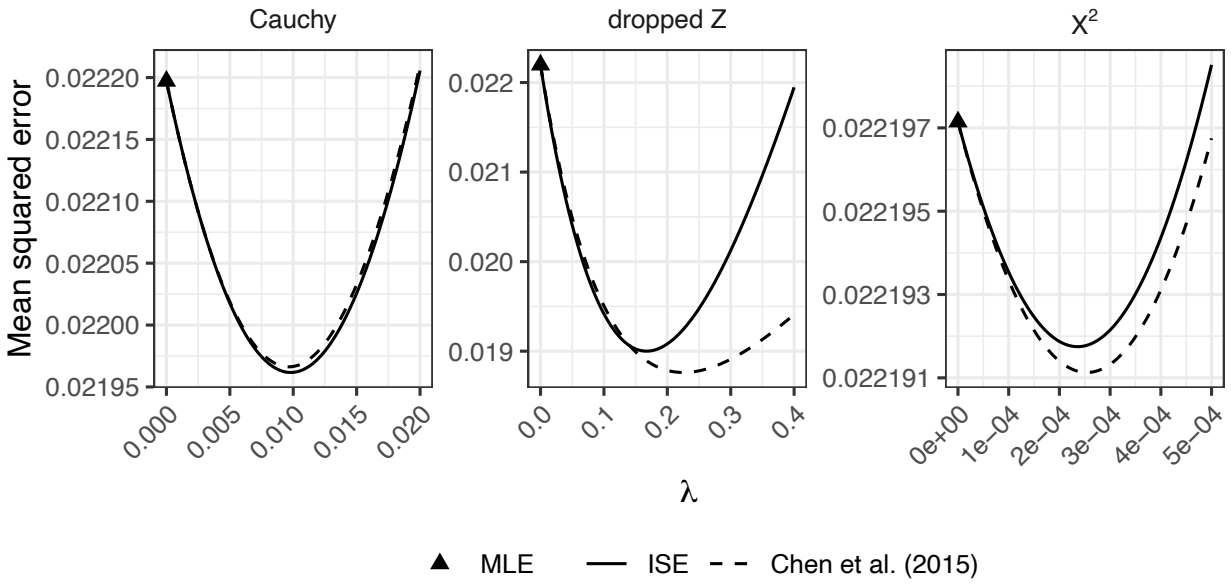

Figure 5: eMSE of $\widetilde{\boldsymbol{\beta}}_2(\lambda)$, $\widehat{\boldsymbol{\beta}}_2(\boldsymbol{W}_\lambda)$ and $\widehat{\boldsymbol{\beta}}_2$ over 1000 simulated $\mathcal{D}_2$ in Setting IV.

The estimator proposed by Chen et al. (2015) appears slightly more efficient than our approach when the mean model is misspecified (dropped Z and X^2). We show in Table 7 that $\tilde{\lambda}$ results in a reduced eMSE by reporting eMSE of $\widetilde{\boldsymbol{\beta}}_2(\tilde{\lambda})$, $\widehat{\boldsymbol{\beta}}_2(\boldsymbol{W}_{\hat{\lambda}})$, $\widehat{\boldsymbol{\beta}}_2^{TG}$, $\widehat{\boldsymbol{\beta}}_2^{TL}$, $\widehat{\boldsymbol{\beta}}_2^{p}$ and $\widehat{\boldsymbol{\beta}}_2$ (Monte Carlo standard errors of eMSE are reported in Table 11 of the Appendix). Our estimator's gain in efficiency is robust when $f_1^{(n_1)}$ is misspecified (Cauchy misspecification): as with Stein shrinkage, our use of the KL divergence in the shrinkage is robust to the misspecification of the density $f_1^{(n_1)}$. Moreover, our approach is robust to unmeasured confounders in $\mathcal{D}_1$ (dropped Z misspecification), encouraging the use of $\widetilde{\boldsymbol{\beta}}_2(\tilde{\lambda})$ when $\mathcal{D}_2$ only measures a subset of the features measured in $\mathcal{D}_1$. Finally, when mean models in $\mathcal{D}_1$ and $\mathcal{D}_2$ are sizeably different (X^2 misspecification), our approach reverts to the MLE in $\mathcal{D}_2$ with a very small amount of shrinkage. In Table 8, we also report mean value of $\tilde{\lambda}$ and the median (over the 11 features) CP of $\widetilde{\boldsymbol{\beta}}_2(\tilde{\lambda})$, $\widehat{\boldsymbol{\beta}}_2^{p}$ and $\widehat{\boldsymbol{\beta}}_2$ over the 1000 simulated $\mathcal{D}_2$. Empirical coverage of $\widetilde{\boldsymbol{\beta}}_2(\tilde{\lambda})$ reaches the nominal level in all settings. The Trans-Lasso estimator occasionally has smaller eMSE than our estimator; it does not, however, suggest a path to inference, so that the reduction in the MSE may give false confidence in the strength of a result that may not in fact replicate. The CP and eMSE of the pooled MLE show inflated eMSE in all settings, with over-coverage, nominal coverage and under-coverage in the Cauchy, dropped Z, and X^2 cases, respectively.

6 Real data analysis

We present a real-data illustration of our information-based shrinkage estimator $\widetilde{\boldsymbol{\beta}}_2(\lambda)$ in an analysis of data from the multi-center eICU Collaborative Research Database (Pollard

Case	eMSE·10^2					
	$\widetilde{\boldsymbol{\beta}}_2(\tilde{\lambda})$	$\widehat{\boldsymbol{\beta}}_2(\boldsymbol{W}_{\hat{\lambda}})$	$\widehat{\boldsymbol{\beta}}_2^{TG}$	$\widehat{\boldsymbol{\beta}}_2^{TL}$	$\widehat{\boldsymbol{\beta}}_2^{p}$	$\widehat{\boldsymbol{\beta}}_2$
Cauchy	2.20	2.20	2.22	2.15	60.0	2.22
dropped Z	1.99	1.96	2.37	1.93	3.68	2.22
X^2	2.22	2.22	2.22	2.29	$230 \cdot 10$	2.22

Table 7: Setting IV: eMSE of $\widetilde{\boldsymbol{\beta}}_2(\tilde{\lambda})$, $\widehat{\boldsymbol{\beta}}_2(\boldsymbol{W}_{\hat{\lambda}})$, $\widehat{\boldsymbol{\beta}}_2^{TG}$, $\widehat{\boldsymbol{\beta}}_2^{TL}$, $\widehat{\boldsymbol{\beta}}_2^{p}$, $\widehat{\boldsymbol{\beta}}_2$.

Case	$\tilde{\lambda}$ (s.e.)	CP		
		$\widetilde{\boldsymbol{\beta}}_2(\tilde{\lambda})$	$\widehat{\boldsymbol{\beta}}_2^{p}$	$\widehat{\boldsymbol{\beta}}_2$
Cauchy	$9.49 \cdot 10^{-3}$ $(1.44 \cdot 10^{-3})$	94.7	100	94.7
dropped Z	$1.58 \cdot 10^{-1}$ $(4.66 \cdot 10^{-2})$	93.6	95.3	93.6
X^2	$2.44 \cdot 10^{-4}$ $(3.45 \cdot 10^{-5})$	95.1	0	95.1

Table 8: Setting IV: mean $\tilde{\lambda}$ (s.e.), %CP of $\widetilde{\boldsymbol{\beta}}_2(\tilde{\lambda})$, $\widehat{\boldsymbol{\beta}}_2^{p}$, $\widehat{\boldsymbol{\beta}}_2$.

et al., 2018) maintained by the Philips eICU Research Institute. The database consists of data from patients admitted to one of several intensive care units (ICUs) throughout the continental United States in 2014 and 2015. Information on data access and pre-processing is provided in Appendix C.

Our analysis focuses on estimating the association between death and baseline features for patients suffering from cardiac arrest upon admission to the ICU. Inclusion criteria are described in Appendix C. Our first population consists of ICU admissions at hospitals located in the western United States, and the data set $\mathcal{D}_1$ consists of cardiac arrest ICU admissions at these 34 hospitals, $n_1 = 575$. (The source data set $\mathcal{D}_1$ is a concatenation of source data from 34 hospitals, as discussed in Section 2.1; our justification for this concatenation is that these records are from hospitals in the same geographic region, so substantial heterogeneity between hospitals would not be expected.) The data set $\mathcal{D}_2$ consists of cardiac arrest ICU admissions at one hospital in the southern United States, $n_2 = 145$. The probability of death μ_{ij} for participant i in data set $\mathcal{D}_j$ is modeled by

$$\log \frac{\mu_{ij}}{1-\mu_{ij}} = \beta_{0j} + \beta_{1j}\,\mathrm{sex}_{ij} + \beta_{2j}\,\mathrm{age}_{ij} + \beta_{3j}\,\mathrm{ethnicity}_{ij} + \beta_{4j}\,\mathrm{BMI}_{ij},$$

with sex_{ij}, age_{ij}, $\mathrm{ethnicity}_{ij}$ and BMI_{ij} the sex (1 for male, 0 for female), age (in years), ethnicity (1 for African American, 0 for Caucasian) and body mass index (in $\mathrm{kg/m^2}$) of participant i in data set $\mathcal{D}_j$, respectively.

Maximum likelihood estimation based on $\mathcal{D}_1$ reveals that sex and age are significantly associated with death following cardiac arrest at hospitals in the west at level 0.05, with estimated effects $\widehat{\beta}_{1j} = -0.39$ (standard error, se = 0.18) and $\widehat{\beta}_{2j} = 0.022$ (se = 0.0073). The MLEs and approximate standard errors for $\mathcal{D}_1$ and $\mathcal{D}_2$ are displayed in Table 9. While associations between the outcome and age and sex are in the same direction in $\mathcal{D}_1$ and $\mathcal{D}_2$, the estimates in the latter are not significant.

Covariate	West hospitals $\mathcal{D}_1$	South hospital $\mathcal{D}_2$
intercept	−1.3 (0.65)	−1.1 (1.3)
sex	**−0.39** (0.18)	−0.17 (0.36)
age	**0.022** (0.0073)	0.017 (0.015)
ethnicity	−0.29 (0.39)	0.27 (0.36)
BMI	−0.00069 (0.011)	0.019 (0.024)

Table 9: Maximum likelihood estimates (standard errors) of feature effects in the data sets $\mathcal{D}_1$ and $\mathcal{D}_2$ of cardiac arrest ICU admissions at the hospitals in the west and the hospital in the south, respectively. Bolded estimates are significant at level 0.05.

Our proposed information-shrinkage approach can be used to borrow information from $\mathcal{D}_1$ to improve the efficiency of estimates in $\mathcal{D}_2$. Trace plots of the ISE $\widetilde{\boldsymbol{\beta}}_2(\lambda)$ with shaded 95% (pointwise) confidence bands are plotted in Figure 6 for a sequence of $\lambda \in [0, 10]$. The minimizer $\tilde{\lambda}$ of the estimated aMSE is also depicted there and shows that our proposed estimator $\widetilde{\boldsymbol{\beta}}_2(\tilde{\lambda})$ leads to statistically significant estimates of the sex and age effects in $\mathcal{D}_2$, at level 0.05. The 95% confidence intervals based on equation (10) for the sex and age effects are $(-0.44, -0.28)$ and $(0.017, 0.024)$ respectively.

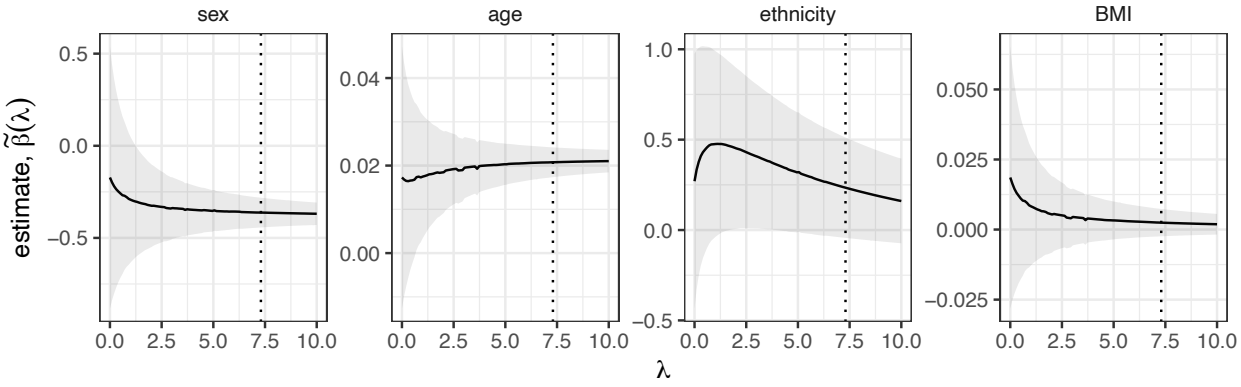

Figure 6: ISE $\widetilde{\boldsymbol{\beta}}_2(\lambda)$ of the sex, age, ethnicity and BMI effects in $\mathcal{D}_2$, the data set of cardiac arrest ICU admissions at the southern U.S. hospital. The dotted vertical line gives the value of the minimizer of the estimated aMSE, $\tilde{\lambda}$.

We also report estimates of the intercept, sex, age, ethnicity and BMI effects for the estimators $\widetilde{\boldsymbol{\beta}}_2(\tilde{\lambda})$, $\widehat{\boldsymbol{\beta}}_2(\boldsymbol{W})$, $\widehat{\boldsymbol{\beta}}_2^{TG}$, $\widehat{\boldsymbol{\beta}}_2^{p}$ and $\widehat{\boldsymbol{\beta}}_2$ in Table 10. The estimator $\widehat{\boldsymbol{\beta}}_2^{TG}$ uses the source detection algorithm in Tian and Feng (2022) and finds that data from all 34 hospitals in the west are transferable. Our proposed ISE appears close to the pooled and target-only estimates. In contrast, $\widehat{\boldsymbol{\beta}}_2(\boldsymbol{W})$ exhibits substantial bias in the age and BMI effects, whereas the debiasing step in $\widehat{\boldsymbol{\beta}}_2^{TG}$ appears to fail and returns effect estimates that are zero for the ethnicity and BMI covariates. In fairness, $\widehat{\boldsymbol{\beta}}_2^{TG}$ is designed for a different setting than ours, as discussed in Section 4.3. Specifically, the estimator is designed to improve target data

	$\widetilde{\boldsymbol{\beta}}_2(\tilde{\lambda})$	$\widehat{\boldsymbol{\beta}}_2(\boldsymbol{W})$	$\widehat{\boldsymbol{\beta}}_2^{TG}$	$\widehat{\boldsymbol{\beta}}_2^{p}$	$\widehat{\boldsymbol{\beta}}_2$
intercept	−1.3	−0.94	0.15	−1.2	−1.1
sex	−0.36	−0.17	−0.085	−0.35	−0.17
age	0.021	19	0.0070	0.020	0.017
ethnicity	0.23	0.063	0	0.36	0.27
BMI	0.0025	9.1	0	0.0038	0.019

Table 10: Estimates of feature effects in the target data set $\mathcal{D}_2$ of cardiac arrest ICU admissions at the hospital in the south.

set predictions in high-dimensional transfer learning problems with multiple heterogeneous sources, and so the algorithm may be over-designed for our setting. We can nonetheless compare predictive metrics of the estimators. Area under the receiver operating characteristic curve (AUC) (standard error) are 57.6% (4.79%), 58.6% (4.74%), 56.7% (4.81%), 57.4% (4.81%) and 59.6% (4.77%) using estimates $\widetilde{\boldsymbol{\beta}}_2(\tilde{\lambda})$, $\widehat{\boldsymbol{\beta}}_2(\boldsymbol{W})$, $\widehat{\boldsymbol{\beta}}_2^{TG}$, $\widehat{\boldsymbol{\beta}}_2^{p}$ and $\widehat{\boldsymbol{\beta}}_2$, respectively. While not designed for prediction, our approach does not give worse predictive performance—as measured by AUC—in comparison to $\widehat{\boldsymbol{\beta}}_2(\boldsymbol{W})$, $\widehat{\boldsymbol{\beta}}_2^{p}$, and $\widehat{\boldsymbol{\beta}}_2^{TG}$, despite the latter being designed specifically to improve prediction. Further, our information shrinkage estimate $\widetilde{\boldsymbol{\beta}}_2(\tilde{\lambda})$ appears less biased than $\widehat{\boldsymbol{\beta}}_2(\boldsymbol{W})$ and $\widehat{\boldsymbol{\beta}}_2^{TG}$, facilitating estimation and inference.

7 Conclusion

This paper proposes a new approach to data integration, focusing not on *if* two data sets should be integrated, but on the *extent* to which inference based on the second data set could be made more efficient by leveraging information in the first. This *extent* is controlled by a dial parameter λ. We proposed a λ-dependent estimator in generalized linear models, offered a data-driven choice of λ, and established theoretical support for our claims that this new estimator is more efficient than that which ignores the first data set, even when the underlying populations differ. Our new, more nuanced data integration framework not only matches statistical intuition on notions of informativeness but empirically out-performs the state-of-the-art techniques. Our proposed approach yields relatively simple parameter estimates, yet performs powerfully in practice and is intuitively related to a broad scope of inferential frameworks.

A unique feature of our approach is the ability to essentially ignore $\mathcal{D}_1$ when n_2 is large. This is important. We have already discussed that our approach is protected from negative transfer. Almost equally important, our approach says something useful about when to expect efficiency gains by incorporating inference from another data set. Intuition tells us that the contribution from another data set should vanish as the sample size n_2 grows when $\boldsymbol{\delta} \neq \mathbf{0}$, which is typical in practice; this intuition underpins the foundations of efficiency results for maximum likelihood estimation. This is borne out by our information-shrinkage approach because we are accounting for the relative *information* in $\mathcal{D}_1$ with respect to $\mathcal{D}_2$.

It is somewhat surprising, however, that this is not the case for all shrinkage estimators, as evidenced in Section 5. While the estimator in Chen et al. (2015) has a smaller asymptotic variance than our proposed estimator when n_2 is large, theirs is asymptotically biased and does not offer a path to inference. In contrast, we guarantee that, asymptotically, the bias is vanishing and approximately valid confidence intervals are available and can be computed.

Another key point is that the bias is controlled by choosing $\lambda^\star$ *at the right rate*, namely between $O(n_2^{-1})$ and $O(n_2^{-1/2})$. This rate guarantees a gain in efficiency, even when the bias is negligible. The need for data integration is driven by the high cost of data collection, and a gain in efficiency, however small, can make the difference between a null finding and a new discovery.

In contrast to data fusion discussed in Section 2.3.3, our approach limits its consideration to data sets $\mathcal{D}_1$ and $\mathcal{D}_2$ as the units of integration, rather than features. While elements of $\widetilde{\boldsymbol{\beta}}_2(\lambda)$ will shrink unevenly depending on feature information, we do not allow the user to differentially shrink estimates using feature-specific dial parameters. It is unclear how to incorporate feature-specific dial parameters because the penalty in (3) is data-dependent, i.e. a summation over independent units $i = 1, \ldots, n_1$ rather than parameters.

Our focus on generalized linear models is motivated by a desire to balance practical utility and the insights we can obtain into the efficiency gains expected from integrating data sets. Our approach can surely be applied in problems that fall outside the generalized linear model framework, but the theory may be substantially complicated by, e.g., the lack of a closed form for the KL divergence. The substantial and practically useful intuition developed in the present paper, for example on the rate of $\lambda^\star$, should be helpful in guiding extensions to more general models in future work.

We envision that our proposed method can be extended to the setting with high-dimensional predictors ($p > n_2$). A special case of interest considers the setting where the correctly specified model in the source $\mathcal{D}_1$ depends on a set of $q > n_1$ features, but the correctly specified model in the target $\mathcal{D}_2$ depends only on a subset of $p < n_2$ of these features. Then, our information-based shrinkage estimation can be carried out using only these p features without modification. This is due to the fact that the model in $\mathcal{D}_1$ need not be correctly specified for us to borrow information from $\mathcal{D}_1$. Theorems 1 and 2 and the simulation results in Setting IV of Section 5 support this solution, although further investigation is required to determine how much efficiency can be gained when more than one feature differs between $\mathcal{D}_2$ and $\mathcal{D}_1$. More generally, extensions to our work may consider the setting where the correctly specified models in both data sets $\mathcal{D}_1, \mathcal{D}_2$ depend on p features, $p \geq n_2 \vee n_1$. We envision that our approach can be extended by regularizing the estimator $\widetilde{\boldsymbol{\beta}}_2(\lambda)$ of $\boldsymbol{\beta}_2$. It is doubtful that the bias introduced by this regularization decreases quickly enough to yield valid inference, and so we expect to lose many of the appealing properties we derived for our estimator $\widetilde{\boldsymbol{\beta}}_2(\tilde{\lambda})$. Additional theoretical study is required to investigate the implementation of debiasing strategies in our data integrative setting.

As seen in Section 6, the multiple source setting is very realistic. When the number of source data sets is small, an investigator may compare our proposed source concatenation approach to, for example, one that uses each source individually, one at a time. Appendix B.2 shows that our suggested concatenation strategy is safe in the sense that it is theoretically guaranteed to protect against negative transfer, and efficient in that it generally outperforms the integration of a single source data set. We also argue that concatenation is a reasonable

de facto approach in all but the most extreme cases, e.g., where one source data set obviously matches the target data much better than the other source data sets. A further question centers around the optimal source configuration to minimize the MSE of the estimator. In Appendix B.2, we consider the idea of choosing a source data configuration by minimizing our estimated MSE. That is, in the linear model case, define the function

$$\mathcal{S} \mapsto n_{\mathcal{T}}^{-1}\widehat{\sigma}_{\mathcal{T}}^2 \,\mathrm{trace}\{\boldsymbol{S}^{-2}(\tilde{\lambda})\boldsymbol{G}_{\mathcal{T}}\} + \tilde{\lambda}^2\mathrm{trace}\{\boldsymbol{G}_{\mathcal{S}}\boldsymbol{S}^{-2}(\tilde{\lambda})\boldsymbol{G}_{\mathcal{S}}\widehat{\boldsymbol{\delta}^2}\},$$

where the input $\mathcal{S}$ could be the full concatenation of source data sets, any one of the individual source data sets, or some intermediate configuration consisting of a concatenation of only some of the source data sets, and the right-hand side is the minimized estimated MSE for this source configuration. The same idea can be applied in the generalized linear model case, with an expression slightly more complicated than that above. Then the idea is to choose $\widehat{\mathcal{S}}$ as the minimizer of the above objective function. We show in Appendix B.2 that this strategy reliably selects the most efficient source configuration. The minimized estimated MSE can form the basis for this comparison by quantifying the effectiveness of the data integration: a smaller value signals that a large amount of information can be safely borrowed and should therefore be preferred. As with all model selection approaches, the downstream inference does not account for the model selection and so an investigator should be wary of over-interpreting confidence intervals/standard errors. A natural extension of our concatenation proposal would incorporate features of data-driven heterogeneous source detection methods (Li et al., 2022; Tian and Feng, 2022) to perform inference in the multi-source setting.

Acknowledgments

The authors thank the reviewer and associate editor for their valuable feedback, which led to an improved manuscript. The first author's work is partially supported by the U. S. National Science Foundation, grant DMS–2337943. The second author's work is partially supported by the U. S. National Science Foundation, grant SES–2051225.

A Proofs

A.1 Proof of Theorem 1

We start with the mean squared error of $\widetilde{\boldsymbol{\beta}}_2(\lambda)$. First,

$$\{\widetilde{\boldsymbol{\beta}}_2(\lambda) - \boldsymbol{\beta}_2\}^\top\{\widetilde{\boldsymbol{\beta}}_2(\lambda) - \boldsymbol{\beta}_2\} = \{\boldsymbol{G}_2(\widehat{\boldsymbol{\beta}}_2 - \boldsymbol{\beta}_2) + \lambda\boldsymbol{G}_1(\widehat{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}_2)\}^\top(\boldsymbol{G}_2 + \lambda\boldsymbol{G}_1)^{-2} \\ \{\boldsymbol{G}_2(\widehat{\boldsymbol{\beta}}_2 - \boldsymbol{\beta}_2) + \lambda\boldsymbol{G}_1(\widehat{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}_2)\}.$$

Define $\boldsymbol{M} = \boldsymbol{G}_2^{-1/2}\boldsymbol{G}_1\boldsymbol{G}_2^{-1/2}$ and $\boldsymbol{R} = \boldsymbol{G}_1^{-1}\boldsymbol{G}_2$. Then

$$\begin{aligned}\mathrm{MSE}\{\widetilde{\boldsymbol{\beta}}_2(\lambda)\} &= \mathbb{E}_{\boldsymbol{\beta}_2}[\{\widetilde{\boldsymbol{\beta}}_2(\lambda) - \boldsymbol{\beta}_2\}^\top\{\widetilde{\boldsymbol{\beta}}_2(\lambda) - \boldsymbol{\beta}_2\}] \\ &= n_2^{-1}\sigma_2^2\mathrm{trace}\{\boldsymbol{S}^{-2}(\lambda)\boldsymbol{G}_2\} + \lambda^2\boldsymbol{\delta}^\top\boldsymbol{G}_1\boldsymbol{S}^{-2}(\lambda)\boldsymbol{G}_1\boldsymbol{\delta} \\ &= n_2^{-1}\sigma_2^2\mathrm{trace}\{(\boldsymbol{I}_p + \lambda\boldsymbol{M})^{-2}\boldsymbol{G}_2^{-1}\} + \lambda^2\mathrm{trace}\{(\boldsymbol{R} + \lambda\boldsymbol{I}_p)^{-2}\boldsymbol{\delta}\boldsymbol{\delta}^\top\} \\ &= v(\lambda) + b(\lambda).\end{aligned} \tag{11}$$

The term $v(\lambda)$ is the sum of the variances of the (weighted) least squares estimates from $\mathcal{D}_1$ and $\mathcal{D}_2$. On the other hand, $b(\lambda)$ is the squared distance from $\widehat{\boldsymbol{\beta}}_1$ to $\boldsymbol{\beta}_2$ and will be $\mathbf{0}$ when $\lambda = 0$ or $\widehat{\boldsymbol{\beta}}_1 = \boldsymbol{\beta}_2$. Clearly, $v(\lambda)$ is monotonically decreasing and $b(\lambda)$ is monotonically increasing. Note that, as expected,

$$v(0) + b(0) = n_2^{-1}\sigma_2^2\mathrm{trace}(\boldsymbol{G}_2^{-1}) = \mathrm{MSE}(\widehat{\boldsymbol{\beta}}_2).$$

When $\boldsymbol{\delta} = \mathbf{0}$, we therefore have $\mathrm{MSE}\{\widetilde{\boldsymbol{\beta}}_2(\lambda)\} < \mathrm{MSE}(\widehat{\boldsymbol{\beta}}_2)$ for all $\lambda > 0$. Throughout the rest of the proof, we consider the case when $\boldsymbol{\delta} \neq \mathbf{0}$.

Since $\boldsymbol{G}_2$ is symmetric positive definite, it has an invertible and symmetric square root denoted by $\boldsymbol{G}_2^{1/2}$. From

$$\boldsymbol{G}_2^{1/2}\boldsymbol{R}\boldsymbol{G}_2^{-1/2} = \boldsymbol{G}_2^{1/2}\boldsymbol{G}_1^{-1}\boldsymbol{G}_2^{1/2} = \boldsymbol{M}^{-1},$$

we have that $\boldsymbol{R}$ and $\boldsymbol{M}^{-1}$ are similar matrices. By symmetry and, therefore, orthogonal diagonalizability of the latter, the former is diagonalizable, and $\boldsymbol{R}$ and $\boldsymbol{M}^{-1}$ share eigenvalues. Let $\kappa_r > 0$, $r = 1, \ldots, p$, the eigenvalues of $\boldsymbol{M}^{-1}$ and $\boldsymbol{R}$ in decreasing order. Denote by $\boldsymbol{K} = \mathbf{diag}\{\kappa_r\}_{r=1}^p$ and $\boldsymbol{P}$ the matrices of eigenvalues and eigenvectors of $\boldsymbol{M}^{-1}$, with $\boldsymbol{P}^\top = \boldsymbol{P}^{-1}$ such that $\boldsymbol{M}^{-1} = \boldsymbol{P}^\top\boldsymbol{K}\boldsymbol{P}$. Then the mean squared error of $\widetilde{\boldsymbol{\beta}}_2(\lambda)$ can be rewritten as

$$\mathrm{MSE}\{\widetilde{\boldsymbol{\beta}}_2(\lambda)\} = n_2^{-1}\sigma_2^2\mathrm{trace}\{(\boldsymbol{I}_p + \lambda\boldsymbol{K}^{-1})^{-2}\boldsymbol{P}\boldsymbol{G}_2^{-1}\boldsymbol{P}^\top\} + \mathrm{trace}\{(\lambda^{-1}\boldsymbol{R} + \boldsymbol{I}_p)^{-2}\boldsymbol{\delta}\boldsymbol{\delta}^\top\}.$$

We see that $\mathrm{MSE}\{\widetilde{\boldsymbol{\beta}}_2(\lambda)\}$ is monotonically decreasing on an interval $[0, \lambda^\star]$ and monotonically increasing on an interval $[\lambda^\star, \infty)$ for some $\lambda^\star > 0$. Our aim is to find a bound on $\lambda^\star$, so that $\mathrm{MSE}\{\widetilde{\boldsymbol{\beta}}_2(\lambda)\} < \mathrm{MSE}(\widehat{\boldsymbol{\beta}}_2)$ for all λ less than that bound.

Denote by $(\delta_r^2)_{r=1}^p$ the values of $\mathbf{diag}(\boldsymbol{\delta}\boldsymbol{\delta}^\top)$ sorted in increasing order, i.e. $0 \leq \delta_1^2 \leq \ldots \leq \delta_p^2$. Let $g_{r2} > 0$, $r = 1, \ldots, p$, the eigenvalues of $\boldsymbol{G}_2$ in increasing order. Then the bias and variance functions satisfy

$$\begin{aligned}b(\lambda) &\leq \lambda^2\sum_{r=1}^p \frac{\delta_r^2}{(\kappa_r + \lambda)^2} \\ v(\lambda) &\leq \frac{\sigma_2^2}{n_2}\sum_{r=1}^p \frac{g_{r2}^{-1}}{(1 + \lambda\kappa_r^{-1})^2} = \frac{\sigma_2^2}{n_2}\sum_{r=1}^p \frac{g_{r2}^{-1}\kappa_r^2}{(\kappa_r + \lambda)^2},\end{aligned}$$

using Von Neumann's trace inequality and the fact that the eigenvalues of $\boldsymbol{A}^\top\boldsymbol{A}$ are the squared eigenvalues of $\boldsymbol{A}$ for a positive definite matrix $\boldsymbol{A}$. Therefore

$$\mathrm{MSE}\{\widetilde{\boldsymbol{\beta}}_2(\lambda)\} = v(\lambda) + b(\lambda) \leq \sum_{r=1}^p \frac{\sigma_2^2 g_{r2}^{-1}\kappa_r^2 + \lambda^2 n_2\delta_r^2}{n_2(\kappa_r + \lambda)^2}.$$

Denote the upper bound on the right-hand side by $U(\lambda)$. We proceed to show that there exists a $\lambda > 0$ such that $\mathrm{MSE}\{\widetilde{\boldsymbol{\beta}}_2(\lambda)\} \leq U(\lambda) < \mathrm{MSE}(\widehat{\boldsymbol{\beta}}_2)$.

Towards this, we examine the monotonicity of $U(\lambda)$. First,

$$\frac{\partial}{\partial \lambda} U(\lambda) = 2 \sum_{r=1}^{p} \frac{\kappa_r (\lambda n_2 \delta_r^2 - \sigma_2^2 g_{r2}^{-1} \kappa_r)}{n_2 (\kappa_r + \lambda)^3}.$$

From

$$\lim_{\lambda \to 0^+} \frac{\partial}{\partial \lambda} U(\lambda) = -\frac{2\sigma_2^2}{n_2} \sum_{r=1}^{p} \frac{g_{r2}^{-1}}{\kappa_r} < 0,$$

the upper bound $U(\lambda)$ immediately decreases as λ moves away from 0. To find a range of λ such that $\mathrm{MSE}\{\widetilde{\boldsymbol{\beta}}_2(\lambda)\} \leq U(\lambda) < \mathrm{MSE}(\widehat{\boldsymbol{\beta}}_2)$, it is therefore sufficient to find the range of λ over which $U(\lambda)$ is decreasing, due to the fact that, at $\lambda = 0$,

$$\mathrm{MSE}\{\widetilde{\boldsymbol{\beta}}_2(0)\} = U(0) = \mathrm{MSE}(\widehat{\boldsymbol{\beta}}_2).$$

The range of λ such that $U(\lambda)$ is decreasing corresponds to the range of λ values for which the derivative of $U(\lambda)$ is negative. This, in turn, corresponds to the range of λ satisfying

$$\lambda \sum_{r=1}^{p} \frac{\kappa_r \delta_r^2}{(\kappa_r + \lambda)^3} < \frac{\sigma_2^2}{n_2} \sum_{r=1}^{p} \frac{g_{r2}^{-1} \kappa_r^2}{(\kappa_r + \lambda)^3}.$$

This is satisfied by

$$\lambda < \frac{\sigma_2^2}{n_2} \frac{\min_{r=1,\ldots,p}(\kappa_r g_{r2}^{-1})}{\max_{r=1,\ldots,p}(\delta_r^2)}. \tag{12}$$

Equation (12) gives a range of λ values such that $\mathrm{MSE}\{\widetilde{\boldsymbol{\beta}}_2(\lambda)\} < \mathrm{MSE}(\widehat{\boldsymbol{\beta}}_2)$.

A.2 Proof of Lemma 1

First, observe that $-O(\boldsymbol{\beta}; \lambda)$ is a convex function of $\boldsymbol{\beta} \in \Theta$, where $\Theta \in \mathbb{R}^p$ is an open convex set. Define

$$\begin{aligned} o_i(\boldsymbol{\beta}; \lambda) = {} & y_{i2} \boldsymbol{X}_{i2}^\top \boldsymbol{\beta} - b(\boldsymbol{X}_{i2}^\top \boldsymbol{\beta}) \\ & - \frac{\lambda}{n_1} \sum_{i=1}^{n_1} \left\{ b'(\boldsymbol{X}_{i1}^\top \widehat{\boldsymbol{\beta}}_1)(\boldsymbol{X}_{i1}^\top \widehat{\boldsymbol{\beta}}_1 - \boldsymbol{X}_{i1}^\top \boldsymbol{\beta}) + b(\boldsymbol{X}_{i1}^\top \boldsymbol{\beta}) - b(\boldsymbol{X}_{i1}^\top \widehat{\boldsymbol{\beta}}_1) \right\}, \end{aligned}$$

so that $-O(\boldsymbol{\beta}; \lambda) = n_2^{-1} \sum_{i=1}^{n_2} -o_i(\boldsymbol{\beta}; \lambda)$. By the law of large numbers, for each fixed $\boldsymbol{\beta} \in \Theta$, $O(\boldsymbol{\beta}; \lambda)$ converges in probability to

$$\lim_{n_2 \to \infty} \frac{1}{n_2} \sum_{i=1}^{n_2} \mathbb{E}_{\boldsymbol{\beta}_2} o_i(\boldsymbol{\beta}; \lambda) = \lim_{n_2 \to \infty} \frac{1}{n_2} \sum_{i=1}^{n_2} \left\{ h(\boldsymbol{X}_{i2}^\top \boldsymbol{\beta}_2) \boldsymbol{X}_{i2}^\top \boldsymbol{\beta} - b(\boldsymbol{X}_{i2}^\top \boldsymbol{\beta}) \right\}$$

$$-\frac{\lambda}{n_1}\sum_{i=1}^{n_1}\left\{b'(\boldsymbol{X}_{i1}^\top\widehat{\boldsymbol{\beta}}_1)(\boldsymbol{X}_{i1}^\top\widehat{\boldsymbol{\beta}}_1-\boldsymbol{X}_{i1}^\top\boldsymbol{\beta})+b(\boldsymbol{X}_{i1}^\top\boldsymbol{\beta})-b(\boldsymbol{X}_{i1}^\top\widehat{\boldsymbol{\beta}}_1)\right\}.$$

By convexity, it follows from Lemma 1 in Hjort and Pollard (1993) that the above convergence is uniform in $\boldsymbol{\beta}$ over compact subsets $K \subset \Theta$. Proofs are given in Andersen and Gill (1982) or Pollard (1991). Consistency of $\widetilde{\boldsymbol{\beta}}_2(\lambda)$ and the rate $\widetilde{\boldsymbol{\beta}}_2(\lambda)-\boldsymbol{\beta}_2^\star(\lambda)=O_p(n_2^{-1/2})$ follow directly from the conditions and Theorem 2.2 in Hjort and Pollard (1993).

A.3 Proof of Lemma 2

By a Taylor's expansion of $\mathbf{0}=\boldsymbol{\Psi}\{\widetilde{\boldsymbol{\beta}}_2(\lambda);\lambda\}$ around $\boldsymbol{\beta}_2$, we obtain

$$\begin{aligned}\mathbf{0}&=\boldsymbol{\Psi}(\boldsymbol{\beta}_2;\lambda)-\boldsymbol{S}(\boldsymbol{\beta}_2;\lambda)\{\widetilde{\boldsymbol{\beta}}_2(\lambda)-\boldsymbol{\beta}_2\}+O_p\{\|\widetilde{\boldsymbol{\beta}}_2(\lambda)-\boldsymbol{\beta}_2\|^2\}\\&=\boldsymbol{\Psi}(\boldsymbol{\beta}_2;\lambda)-\boldsymbol{S}(\boldsymbol{\beta}_2;\lambda)\{\widetilde{\boldsymbol{\beta}}_2(\lambda)-\boldsymbol{\beta}_2\}+O_p\{\|\widetilde{\boldsymbol{\beta}}_2(\lambda)-\boldsymbol{\beta}_2^\star(\lambda)\|^2+\|\boldsymbol{\beta}_2^\star(\lambda)-\boldsymbol{\beta}_2\|^2\\&\qquad+2\|\widetilde{\boldsymbol{\beta}}_2(\lambda)-\boldsymbol{\beta}_2^\star(\lambda)\|\|\boldsymbol{\beta}_2^\star(\lambda)-\boldsymbol{\beta}_2\|\}.\end{aligned}\tag{13}$$

We bound the higher-order terms in the above expansion. By continuity of $\boldsymbol{\Psi}(\boldsymbol{\beta};\lambda)$, there exists a vector $\boldsymbol{c}_\lambda\in\mathbb{R}^p$ between $\boldsymbol{\beta}_2$ and $\boldsymbol{\beta}_2^\star(\lambda)$ such that

$$\boldsymbol{\beta}_2-\boldsymbol{\beta}_2^\star(\lambda)=-\boldsymbol{S}^{-1}(\boldsymbol{c}_\lambda;\lambda)\{\boldsymbol{\Psi}(\boldsymbol{\beta}_2;\lambda)-\boldsymbol{\Psi}(\boldsymbol{\beta}_2^\star(\lambda);\lambda)\}.$$

Recall that $\boldsymbol{\beta}_2^\star(\lambda)$ is the unique solution to $\mathbb{E}_{\boldsymbol{\beta}_2}\{\boldsymbol{\Psi}(\boldsymbol{\beta}_2^\star(\lambda);\lambda)\}=\mathbf{0}$. Taking expectations,

$$\boldsymbol{\beta}_2-\boldsymbol{\beta}_2^\star(\lambda)=-\boldsymbol{S}^{-1}(\boldsymbol{c}_\lambda;\lambda)\mathbb{E}_{\boldsymbol{\beta}_2}\{\boldsymbol{\Psi}(\boldsymbol{\beta}_2;\lambda)\}.\tag{14}$$

We decompose $\boldsymbol{\Psi}(\boldsymbol{\beta};\lambda)=\boldsymbol{U}_2(\boldsymbol{\beta})-\lambda\boldsymbol{P}_1(\boldsymbol{\beta})$ into the difference of the score function $\boldsymbol{U}_2(\boldsymbol{\beta})$ in $\mathcal{D}_2$ and a (non-random) penalty term $\boldsymbol{P}_1(\boldsymbol{\beta})$:

$$\boldsymbol{U}_2(\boldsymbol{\beta})=n_2^{-1}\boldsymbol{X}_2\{\boldsymbol{y}_2-\boldsymbol{\mu}_2(\boldsymbol{\beta})\},\quad \boldsymbol{P}_1(\boldsymbol{\beta})=n_1^{-1}\boldsymbol{X}_1\{\boldsymbol{\mu}_1(\boldsymbol{\beta})-\boldsymbol{\mu}_1(\widehat{\boldsymbol{\beta}}_1)\}.$$

Since $\boldsymbol{\beta}_2$ is the true value of $\boldsymbol{\beta}$ in $\mathcal{D}_2$, i.e., $\mathbb{E}_{\boldsymbol{\beta}_2}(\boldsymbol{Y}_2)=\boldsymbol{X}_2^\top\boldsymbol{\beta}_2=\boldsymbol{\mu}_2(\boldsymbol{\beta}_2)$, clearly $\mathbb{E}_{\boldsymbol{\beta}_2}\{\boldsymbol{U}_2(\boldsymbol{\beta}_2)\}=\mathbf{0}$. This implies

$$\begin{aligned}\mathbb{E}_{\boldsymbol{\beta}_2}\{\boldsymbol{\Psi}(\boldsymbol{\beta}_2;\lambda)\}&=-\lambda n_1^{-1}\boldsymbol{X}_1\{h(\boldsymbol{X}_{i1}^\top\boldsymbol{\beta}_2)-h(\boldsymbol{X}_{i1}^\top\widehat{\boldsymbol{\beta}}_1)\}_{i=1}^{n_1}\\&=-\lambda n_1^{-1}\boldsymbol{X}_1\boldsymbol{\Delta}(\boldsymbol{\beta}_2)\\&=-\lambda n_1^{-1}\boldsymbol{X}_1\boldsymbol{A}(\boldsymbol{X}_1^\top\boldsymbol{c}_2)\boldsymbol{X}_1^\top\boldsymbol{\delta}\\&=-\lambda\boldsymbol{v}_1(\boldsymbol{c}_2)\boldsymbol{\delta},\end{aligned}\tag{15}$$

where the third line is by the mean value theorem with $\boldsymbol{c}_2\in\mathbb{R}^p$ a vector between $\boldsymbol{\beta}_2$ and $\widehat{\boldsymbol{\beta}}_1$. Plugging (15) into (14) gives $\boldsymbol{\beta}_2-\boldsymbol{\beta}_2^\star(\lambda)=\lambda\boldsymbol{S}^{-1}(\boldsymbol{c}_\lambda;\lambda)\boldsymbol{v}_1(\boldsymbol{c}_2)\boldsymbol{\delta}$. Taking the limit as $n_2\to\infty$ on both sides yields

$$\boldsymbol{\beta}_2-\boldsymbol{\beta}_2^\star(\lambda)=\lambda\left\{\boldsymbol{v}_2(\boldsymbol{c}_\lambda)+\lambda\boldsymbol{v}_1(\boldsymbol{c}_\lambda)\right\}^{-1}\boldsymbol{v}_1(\boldsymbol{c}_2)\boldsymbol{\delta}.$$

By condition (C2), the eigenvalues of $\boldsymbol{v}_j(\boldsymbol{c}_\lambda)$ are bounded away from 0, $j = 1, 2$. If $\lambda = O(n_2^{-1/2})$, then we obtain $\boldsymbol{\beta}_2^\star(\lambda) - \boldsymbol{\beta}_2 = O(n_2^{-1/2})$. Plugging this rate into equation (13) and using Lemma 1,

$$\begin{aligned}\mathbf{0} &= \boldsymbol{\Psi}(\boldsymbol{\beta}_2; \lambda) - \boldsymbol{S}(\boldsymbol{\beta}_2; \lambda)\{\widetilde{\boldsymbol{\beta}}_2(\lambda) - \boldsymbol{\beta}_2\} + O_p(n_2^{-1}) + O(n_2^{-1}) + O_p(n_2^{-1/2})O(n_2^{-1/2}) \\ &= \boldsymbol{\Psi}(\boldsymbol{\beta}_2; \lambda) - \boldsymbol{S}(\boldsymbol{\beta}_2; \lambda)\{\widetilde{\boldsymbol{\beta}}_2(\lambda) - \boldsymbol{\beta}_2\} + O_p(n_2^{-1}) + O(n_2^{-1}).\end{aligned}$$

Rearranging gives

$$n_2^{1/2}\{\widetilde{\boldsymbol{\beta}}_2(\lambda) - \boldsymbol{\beta}_2\} = \boldsymbol{S}^{-1}(\boldsymbol{\beta}_2; \lambda) n_2^{1/2} \boldsymbol{\Psi}(\boldsymbol{\beta}_2; \lambda) + O_p(n_2^{-1/2}) + O(n_2^{-1/2}). \tag{16}$$

We examine the asymptotic behavior of $n_2^{1/2}\boldsymbol{\Psi}(\boldsymbol{\beta}_2; \lambda)$. For $i = 1, \ldots, n_2$, define

$$\boldsymbol{\psi}_i(\boldsymbol{\beta}; \lambda) = \boldsymbol{X}_{i2}\big\{y_{i2} - h(\boldsymbol{X}_{i2}^\top \boldsymbol{\beta})\big\} - \frac{\lambda}{n_1}\sum_{j=1}^{n_1} \boldsymbol{X}_{j1}\big\{h(\boldsymbol{X}_{j1}^\top \boldsymbol{\beta}) - h(\boldsymbol{X}_{j1}^\top \widehat{\boldsymbol{\beta}}_1)\big\},$$

such that $\boldsymbol{\Psi}(\boldsymbol{\beta}; \lambda) = n_2^{-1}\sum_{i=1}^{n_2} \boldsymbol{\psi}_i(\boldsymbol{\beta}; \lambda)$. Since

$$\begin{aligned}\lim_{n_2 \to \infty} \frac{1}{n_2}\sum_{i=1}^{n_2} \mathbb{V}_{\boldsymbol{\beta}_2}\{\boldsymbol{\psi}_i(\boldsymbol{\beta}; \lambda)\} &= \lim_{n_2 \to \infty} \frac{1}{n_2}\sum_{i=1}^{n_2} \boldsymbol{X}_{i2}\mathbb{V}_{\boldsymbol{\beta}_2}\left\{Y_{i2} - h\left(\boldsymbol{X}_{i2}^\top \boldsymbol{\beta}\right)\right\}\boldsymbol{X}_{i2}^\top \\ &= \lim_{n_2 \to \infty} \frac{1}{n_2} d(\gamma_2)\boldsymbol{X}_2 \boldsymbol{A}(\boldsymbol{X}_2^\top \boldsymbol{\beta}_2)\boldsymbol{X}_2^\top \\ &= d(\gamma_2)\boldsymbol{v}_2(\boldsymbol{\beta}_2),\end{aligned}$$

by equation (15) and conditions (C1)–(C2),

$$n_2^{1/2}\{\boldsymbol{\Psi}(\boldsymbol{\beta}_2; \lambda) + \lambda n_1^{-1}\boldsymbol{X}_1 \boldsymbol{\Delta}(\boldsymbol{\beta}_2)\} \xrightarrow{d} \mathcal{N}\{\mathbf{0}, d(\gamma_2)\boldsymbol{v}_2(\boldsymbol{\beta}_2)\}.$$

In other words,

$$n_2^{1/2}\boldsymbol{\Psi}(\boldsymbol{\beta}_2; \lambda) = -\lambda n_1^{-1} n_2^{1/2}\boldsymbol{X}_1 \boldsymbol{\Delta}(\boldsymbol{\beta}_2) + d(\gamma_2)^{1/2}\boldsymbol{v}_2^{1/2}(\boldsymbol{\beta}_2)\boldsymbol{Z} + o_p(1),$$

where $\boldsymbol{Z} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I}_p)$. By equation (16),

$$\begin{aligned}n_2^{1/2}\{\widetilde{\boldsymbol{\beta}}_2(\lambda) - \boldsymbol{\beta}_2\} &= \boldsymbol{S}^{-1}(\boldsymbol{\beta}_2; \lambda) d(\gamma_2)^{1/2}\boldsymbol{v}_2^{1/2}(\boldsymbol{\beta}_2)\boldsymbol{Z} \\ &\quad - \lambda n_1^{-1} n_2^{1/2}\boldsymbol{S}^{-1}(\boldsymbol{\beta}_2; \lambda)\boldsymbol{X}_1 \boldsymbol{\Delta}(\boldsymbol{\beta}_2) + o_p(1) + O_p(n_2^{-1/2}) + O(n_2^{-1/2}).\end{aligned} \tag{17}$$

As $n_2 \to \infty$, using symmetry of $\boldsymbol{S}(\boldsymbol{\beta}; \lambda)$, it follows that

$$n_2^{1/2} d(\gamma_2)^{-1/2}\boldsymbol{S}^\top(\boldsymbol{\beta}_2; \lambda)\boldsymbol{v}_2^{-1/2}(\boldsymbol{\beta}_2)\big\{\widetilde{\boldsymbol{\beta}}_2(\lambda) - \boldsymbol{\beta}_2 + \lambda n_1^{-1}\boldsymbol{S}^{-1}(\boldsymbol{\beta}_2; \lambda)\boldsymbol{X}_1 \boldsymbol{\Delta}(\boldsymbol{\beta}_2)\big\}$$

converges in distribution to $\mathcal{N}(\mathbf{0}, \boldsymbol{I}_p)$, which proves the first claim. When $\lambda = o(n_2^{-1/2})$, the above display can be re-expressed as

$$n_2^{1/2} d(\gamma_2)^{-1/2}\boldsymbol{S}^\top(\boldsymbol{\beta}_2; \lambda)\boldsymbol{v}_2^{-1/2}(\boldsymbol{\beta}_2)\big\{\widetilde{\boldsymbol{\beta}}_2(\lambda) - \boldsymbol{\beta}_2\big\} + o(1).$$

Then the second claim follows from the first together with Slutsky's theorem.

A.4 Proof of Theorem 2

We start with the asymptotic mean squared error of $\widetilde{\boldsymbol{\beta}}_2(\lambda)$. By (17) in the proof of Lemma 2,

$$\text{aMSE}\{\widetilde{\boldsymbol{\beta}}_2(\lambda)\} = d(\gamma_2)\text{trace}\{\boldsymbol{s}^{-2}(\boldsymbol{\beta}_2;\lambda)\boldsymbol{v}_2(\boldsymbol{\beta}_2)\} + n_2\lambda^2\boldsymbol{\delta}^\top\boldsymbol{v}_1(\boldsymbol{c}_2)\boldsymbol{s}^{-2}(\boldsymbol{\beta}_2;\lambda)\boldsymbol{v}_1(\boldsymbol{c}_2)\boldsymbol{\delta},$$

where $\boldsymbol{c}_2 \in \mathbb{R}^p$ is a vector between $\boldsymbol{\beta}_2$ and $\widehat{\boldsymbol{\beta}}_1$ such that $\boldsymbol{v}_1(\boldsymbol{c}_2)\boldsymbol{\delta} = n_1^{-1}\boldsymbol{X}_1\boldsymbol{\Delta}(\boldsymbol{\beta}_2)$. Notice that $\boldsymbol{s}(\boldsymbol{\beta}_2;\lambda) = \boldsymbol{v}_2(\boldsymbol{\beta}_2) + \lambda\boldsymbol{v}_1(\boldsymbol{\beta}_2)$. Analogous to the Gaussian linear model case, define $\boldsymbol{M}(\boldsymbol{\beta}) = \boldsymbol{v}_2^{-1/2}(\boldsymbol{\beta})\boldsymbol{v}_1(\boldsymbol{\beta})\boldsymbol{v}_2^{-1/2}(\boldsymbol{\beta})$, $\boldsymbol{R}(\boldsymbol{\beta}) = \boldsymbol{v}_1^{-1}(\boldsymbol{\beta})\boldsymbol{v}_2(\boldsymbol{\beta})$, and $\boldsymbol{C} = \boldsymbol{v}_1^{-1}(\boldsymbol{\beta}_2)\boldsymbol{v}_1(\boldsymbol{c}_2)$. Then the aMSE can be rewritten as

$$\text{aMSE}\{\widetilde{\boldsymbol{\beta}}_2(\lambda)\} = v(\lambda) + b(\lambda),$$

where

$$\begin{aligned} v(\lambda) &= d(\gamma_2)\text{trace}[\{\boldsymbol{I}_p + \lambda\boldsymbol{M}(\boldsymbol{\beta}_2)\}^{-2}\boldsymbol{v}_2^{-1}(\boldsymbol{\beta}_2)] \\ b(\lambda) &= n_2\lambda^2\boldsymbol{\delta}^\top\boldsymbol{C}^\top\{\boldsymbol{R}(\boldsymbol{\beta}_2) + \lambda\boldsymbol{I}_p\}^{-2}\boldsymbol{C}\boldsymbol{\delta}. \end{aligned}$$

The $v(\lambda)$ term is the sum of the variances of the (weighted) least squares estimates from $\mathcal{D}_1$ and $\mathcal{D}_2$, whereas $b(\lambda)$ is the squared distance from $\widehat{\boldsymbol{\beta}}_1$ to $\boldsymbol{\beta}_2$ and will be 0 when $\lambda = 0$ or $\boldsymbol{\delta} = \boldsymbol{0}$. Clearly, $v(\lambda)$ is monotonically decreasing, $b(\lambda)$ is monotonically increasing, and

$$v(0) + b(0) = d(\gamma_2)\text{trace}\{\boldsymbol{v}_2^{-1}(\boldsymbol{\beta}_2)\} = \text{aMSE}(\widehat{\boldsymbol{\beta}}_2).$$

When $\boldsymbol{\delta} = \boldsymbol{0}$, we therefore have $\text{aMSE}\{\widetilde{\boldsymbol{\beta}}_2(\lambda)\} < \text{aMSE}(\widehat{\boldsymbol{\beta}}_2)$ for all $\lambda > 0$. Throughout the rest of the proof, we consider the case when $\boldsymbol{\delta} \neq \boldsymbol{0}$.

Since $\boldsymbol{v}_2(\boldsymbol{\beta})$ is symmetric positive definite, it has an invertible and symmetric square root denoted by $\boldsymbol{v}_2^{1/2}(\boldsymbol{\beta})$. From

$$\boldsymbol{v}_2^{1/2}(\boldsymbol{\beta})\boldsymbol{R}(\boldsymbol{\beta})\boldsymbol{v}_2^{-1/2}(\boldsymbol{\beta}) = \boldsymbol{v}_2^{1/2}(\boldsymbol{\beta})\boldsymbol{v}_1^{-1}(\boldsymbol{\beta})\boldsymbol{v}_2^{1/2}(\boldsymbol{\beta}) = \boldsymbol{M}^{-1}(\boldsymbol{\beta}),$$

we have that $\boldsymbol{R}(\boldsymbol{\beta})$ and $\boldsymbol{M}^{-1}(\boldsymbol{\beta})$ are similar matrices. By symmetry, and therefore orthogonal diagonalizability, of the latter, the former is diagonalizable, and $\boldsymbol{R}(\boldsymbol{\beta})$ and $\boldsymbol{M}^{-1}(\boldsymbol{\beta})$ share eigenvalues. Let $\kappa_r(\boldsymbol{\beta}) > 0$, $r = 1, \ldots, p$ the (common) eigenvalues of $\boldsymbol{M}^{-1}(\boldsymbol{\beta})$ and $\boldsymbol{R}(\boldsymbol{\beta})$ in decreasing order. Denote by $\boldsymbol{K}(\boldsymbol{\beta}) = \textbf{diag}\{\kappa_r(\boldsymbol{\beta})\}_{r=1}^p$ and $\boldsymbol{P}(\boldsymbol{\beta})$ the matrices of eigenvalues and eigenvectors of $\boldsymbol{M}^{-1}(\boldsymbol{\beta})$, with $\boldsymbol{P}^\top(\boldsymbol{\beta}) = \boldsymbol{P}^{-1}(\boldsymbol{\beta})$ such that $\boldsymbol{M}^{-1}(\boldsymbol{\beta}) = \boldsymbol{P}^\top(\boldsymbol{\beta})\boldsymbol{K}(\boldsymbol{\beta})\boldsymbol{P}(\boldsymbol{\beta})$. Then

$$\begin{aligned} v(\lambda) &= d(\gamma_2)\text{trace}[\{\boldsymbol{I}_p + \lambda\boldsymbol{K}^{-1}(\boldsymbol{\beta}_2)\}^{-2}\boldsymbol{P}(\boldsymbol{\beta}_2)\boldsymbol{v}_2^{-1}(\boldsymbol{\beta}_2)\boldsymbol{P}^\top(\boldsymbol{\beta}_2)] \\ b(\lambda) &= n_2\boldsymbol{\delta}^\top\boldsymbol{C}^\top\{\lambda^{-1}\boldsymbol{R}(\boldsymbol{\beta}_2) + \boldsymbol{I}_p\}^{-2}\boldsymbol{C}\boldsymbol{\delta}. \end{aligned}$$

Recall that $\boldsymbol{v}_1(\boldsymbol{c}_2)\boldsymbol{\delta} = n_1^{-1}\boldsymbol{X}_1\boldsymbol{\Delta}(\boldsymbol{\beta}_2)$. Denote by $\{\delta_r^2(\boldsymbol{\beta}_2)\}_{r=1}^p$ the values of

$$\begin{aligned} \textbf{diag}(\boldsymbol{C}\boldsymbol{\delta}\boldsymbol{\delta}^\top\boldsymbol{C}^\top) &= \textbf{diag}\{\boldsymbol{v}_1^{-1}(\boldsymbol{\beta}_2)\boldsymbol{v}_1(\boldsymbol{c}_2)\boldsymbol{\delta}\boldsymbol{\delta}^\top\boldsymbol{v}_1(\boldsymbol{c}_2)\boldsymbol{v}_1^{-1}(\boldsymbol{\beta}_2)\} \\ &= n_1^{-2}\textbf{diag}\{\boldsymbol{v}_1^{-1}(\boldsymbol{\beta}_2)\boldsymbol{X}_1\boldsymbol{\Delta}(\boldsymbol{\beta}_2)\boldsymbol{\Delta}^\top(\boldsymbol{\beta}_2)\boldsymbol{X}_1^\top\boldsymbol{v}_1^{-1}(\boldsymbol{\beta}_2)\}, \end{aligned}$$

sorted in increasing order, i.e. $0 \leq \delta_1^2(\boldsymbol{\beta}) \leq \ldots \leq \delta_p^2(\boldsymbol{\beta})$. Let $g_{r2}(\boldsymbol{\beta})$, $r = 1, \ldots, p$, the eigenvalues of $\boldsymbol{v}_2(\boldsymbol{\beta})$ in increasing order. Then the bias and variance functions satisfy

$$n_2\lambda^2 \sum_{r=1}^{p} \frac{\delta_{p-r+1}^2(\boldsymbol{\beta}_2)}{\{\kappa_r(\boldsymbol{\beta}_2) + \lambda\}^2} \leq b(\lambda) \leq n_2\lambda^2 \sum_{r=1}^{p} \frac{\delta_r^2(\boldsymbol{\beta}_2)}{\{\kappa_r(\boldsymbol{\beta}_2) + \lambda\}^2}$$
$$d(\gamma_2) \sum_{r=1}^{p} \frac{g_{p-r+1\ 2}^{-1}(\boldsymbol{\beta}_2)\kappa_{p-r+1}^2(\boldsymbol{\beta}_2)}{\{\kappa_r(\boldsymbol{\beta}_2) + \lambda\}^2} \leq v(\lambda) \leq d(\gamma_2) \sum_{r=1}^{p} \frac{g_{r2}^{-1}(\boldsymbol{\beta}_2)\kappa_r^2(\boldsymbol{\beta}_2)}{\{\kappa_r(\boldsymbol{\beta}_2) + \lambda\}^2},$$

using Von Neumann's trace inequality and the fact that the eigenvalues of $\boldsymbol{A}^2$ are the squared eigenvalues of $\boldsymbol{A}$ for a positive definite matrix $\boldsymbol{A}$. Then we have a bound on the aMSE,

$$L(\lambda) \leq \text{aMSE}\{\widetilde{\boldsymbol{\beta}}_2(\lambda)\} \leq U(\lambda),$$

where

$$L(\lambda) = \sum_{r=1}^{p} \frac{d(\gamma_2)g_{p-r+1\ 2}^{-1}(\boldsymbol{\beta}_2)\kappa_{p-r+1}^2(\boldsymbol{\beta}_2) + n_2\lambda^2\delta_{p-r+1}^2(\boldsymbol{\beta}_2)}{\{\kappa_r(\boldsymbol{\beta}_2) + \lambda\}^2}$$
$$U(\lambda) = \sum_{r=1}^{p} \frac{d(\gamma_2)g_{r2}^{-1}(\boldsymbol{\beta}_2)\kappa_r^2(\boldsymbol{\beta}_2) + n_2\lambda^2\delta_r^2(\boldsymbol{\beta}_2)}{\{\kappa_r(\boldsymbol{\beta}_2) + \lambda\}^2},$$

are the sums of the lower and upper bounds on $v(\lambda)$ and $b(\lambda)$ above, respectively. We calculate the derivatives:

$$\frac{\partial}{\partial\lambda}L(\lambda) = 2\sum_{r=1}^{p} \frac{2\delta_{p-r+1}(\boldsymbol{\beta}_2)\{\delta_{p-r+1}(\boldsymbol{\beta}_2)\kappa_r(\boldsymbol{\beta}_2)n_2\lambda - g_{p-r+1\ 2}^{-1}\kappa_{p-r+1}^2(\boldsymbol{\beta}_2)\}}{\{\kappa_r(\boldsymbol{\beta}_2) + \lambda\}^3}$$
$$\frac{\partial}{\partial\lambda}U(\lambda) = 2\sum_{r=1}^{p} \frac{\kappa_r(\boldsymbol{\beta}_2)\{\lambda n_2\delta_r^2(\boldsymbol{\beta}_2) - d(\gamma_2)g_{r2}^{-1}(\boldsymbol{\beta}_2)\kappa_r(\boldsymbol{\beta}_2)\}}{\{\kappa_r(\boldsymbol{\beta}_2) + \lambda\}^3}.$$

Since the derivative of the lower bound $L(\lambda)$ is positive for large enough λ, $L(\lambda)$ is increasing for large enough λ. From

$$\lim_{\lambda\to 0^+} \frac{\partial}{\partial\lambda}U(\lambda) = -2d(\gamma_2)\sum_{r=1}^{p} \frac{g_{r2}^{-1}(\boldsymbol{\beta}_2)}{\kappa_r(\boldsymbol{\beta}_2)} < 0,$$

we find that $U(\lambda)$ is decreasing in a neighborhood of the origin. All together, we find that $\text{aMSE}\{\widetilde{\boldsymbol{\beta}}_2(\lambda)\}$ is monotonically decreasing on an interval $[0, \lambda^\star]$ and monotonically increasing on an interval $[\lambda^\star, \infty)$ for some $\lambda^\star > 0$.

Next we show that the upper limit on the range of λ on which there is an efficiency gain is $O(n_2^{-1/2})$. Let $\lambda > 0$ be such that $\text{aMSE}\{\widetilde{\boldsymbol{\beta}}_2(\lambda)\} = \text{aMSE}(\widehat{\boldsymbol{\beta}}_2)$. Then

$$\begin{aligned}
0 &= v(\lambda) + b(\lambda) - \{v(0) + b(0)\} \\
&= d(\gamma_2)\text{trace}[\{\boldsymbol{I}_p + \lambda\boldsymbol{M}(\boldsymbol{\beta}_2)\}^{-2}\boldsymbol{v}_2^{-1}(\boldsymbol{\beta}_2) - \boldsymbol{v}_2^{-1}(\boldsymbol{\beta}_2)] \\
&\quad + n_2\boldsymbol{\delta}^\top\boldsymbol{C}^\top\{\lambda^{-1}\boldsymbol{R}(\boldsymbol{\beta}_2) + \boldsymbol{I}_p\}^{-2}\boldsymbol{C}\boldsymbol{\delta}
\end{aligned}$$

Recall that $\boldsymbol{C\delta}$ is a constant with respect to λ. Rearranging, λ satisfies

$$0 = d(\gamma_2)\text{trace}\big([\{\boldsymbol{I}_p + \lambda \boldsymbol{M}(\boldsymbol{\beta}_2)\}^{-2} - \boldsymbol{I}_p]\boldsymbol{v}_2^{-1}(\boldsymbol{\beta}_2)\big) + n_2\boldsymbol{\delta}^\top \boldsymbol{C}^\top \{\lambda^{-1}\boldsymbol{R}(\boldsymbol{\beta}_2) + \boldsymbol{I}_p\}^{-2}\boldsymbol{C\delta}$$

The first term on the right-hand side is a decreasing function of λ with limit 0 as $\lambda \to 0^+$ and limit $-\text{aMSE}(\widehat{\boldsymbol{\beta}}_2)$ as $\lambda \to \infty$; the first term is, therefore, effectively (a negative) constant in λ. The second term is a strictly positive increasing function of λ for all $\lambda > 0$. For this, too, to be a constant, so that the above equality is satisfied, we need $n_2\lambda^2$ to be effectively a constant. Therefore, the λ at which equality of aMSEs is achieved is $O(n_2^{-1/2})$.

Finally, we can get a lower bound on $\lambda^\star$, so that $\text{aMSE}\{\widetilde{\boldsymbol{\beta}}_2(\lambda)\} < \text{aMSE}(\widehat{\boldsymbol{\beta}}_2)$ for all λ less than that bound. Since $U(\lambda)$ is decreasing in a neighborhood of the origin, it is sufficient to find the λ values for which this derivative is negative, i.e., it is sufficient to find λ such that

$$\lambda n_2 \sum_{r=1}^{p} \frac{\kappa_r(\boldsymbol{\beta}_2)\delta_r^2(\boldsymbol{\beta}_2)}{\{\kappa_r(\boldsymbol{\beta}_2) + \lambda\}^3} < d(\gamma_2) \sum_{r=1}^{p} \frac{g_{r2}^{-1}(\boldsymbol{\beta}_2)\kappa_r^2(\boldsymbol{\beta}_2)}{\{\kappa_r(\boldsymbol{\beta}_2) + \lambda\}^3}.$$

This is satisfied by

$$\lambda < \frac{d(\gamma_2)}{n_2} \frac{\min_{r=1,\ldots,p}\{\kappa_r(\boldsymbol{\beta}_2)g_{r2}^{-1}(\boldsymbol{\beta}_2)\}}{\max_{r=1,\ldots,p}\{\delta_r^2(\boldsymbol{\beta}_2)\}}. \tag{18}$$

Equation (18) gives a range of λ values such that $\text{aMSE}\{\widetilde{\boldsymbol{\beta}}_2(\lambda)\} < \text{aMSE}(\widehat{\boldsymbol{\beta}}_2)$, so the right-hand side is a lower bound on $\lambda^\star$.

B Additional numerical results

B.1 Further results from Sections 5.2-5.5

Figures 7–8 plot $\widehat{\boldsymbol{\beta}}_1$ and $\boldsymbol{\beta}_2$ in Settings I and II, respectively. Figure 9 plots $\widehat{\boldsymbol{\beta}}_1$ and the corresponding value of $\boldsymbol{\delta}$ in Setting IV.

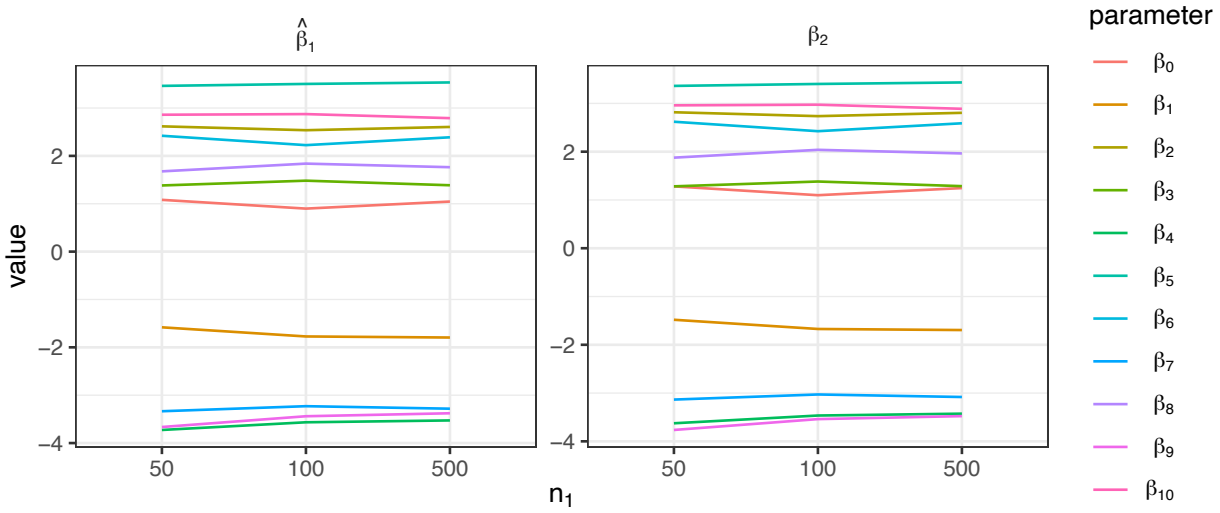

Figure 7: MLE $\widehat{\boldsymbol{\beta}}_1$ (left) with corresponding value of $\boldsymbol{\beta}_2$ (right) for each $\mathcal{D}_1$ in Setting I.

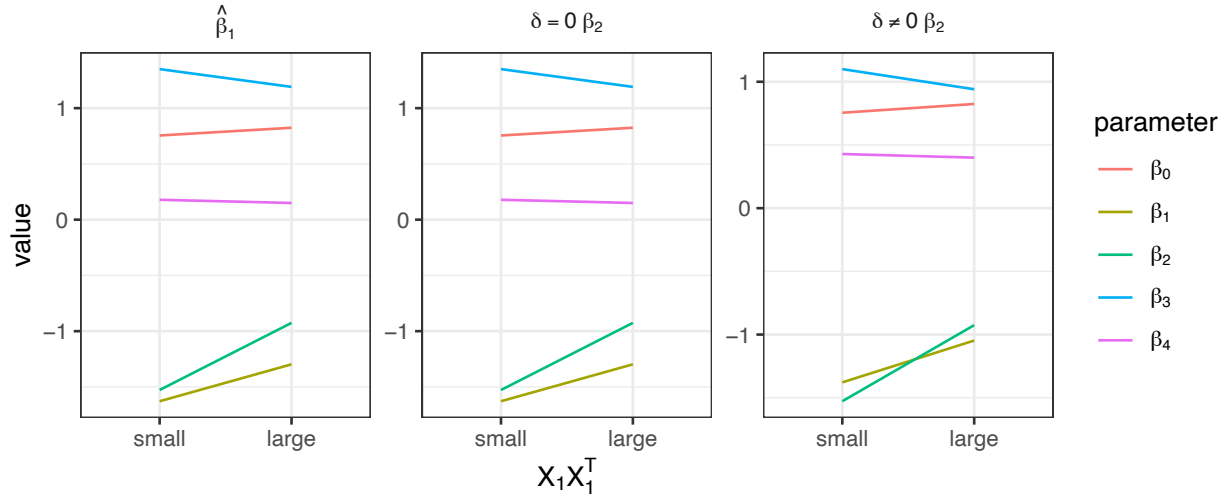

Figure 8: MLE $\widehat{\boldsymbol{\beta}}_1$ (left) with corresponding value of $\boldsymbol{\beta}_2$ (right) for each $\mathcal{D}_1$ in Setting II.

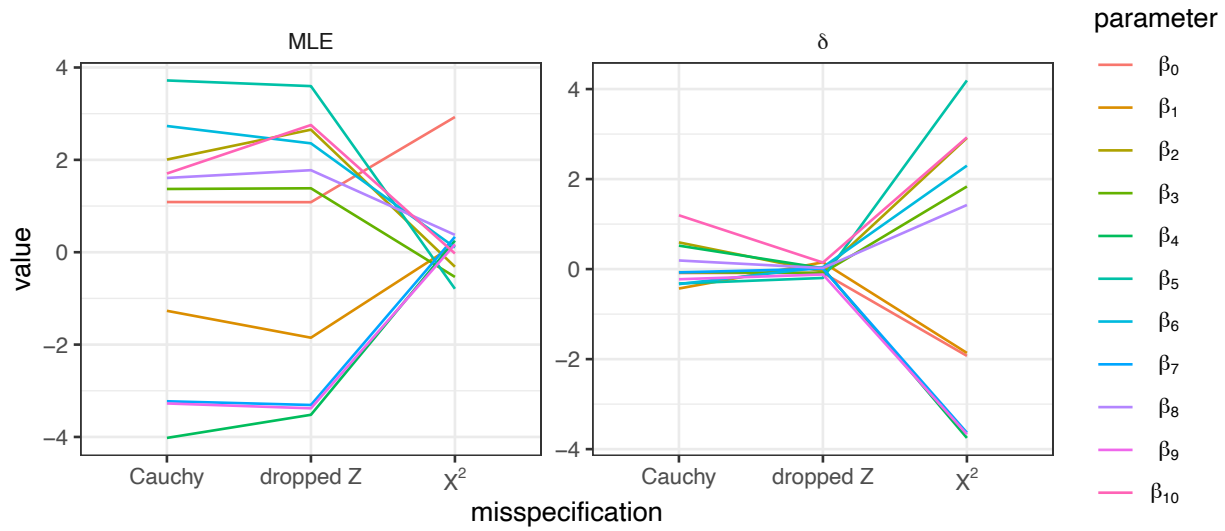

Figure 9: MLE $\widehat{\boldsymbol{\beta}}_1$ (left) with corresponding value of $\boldsymbol{\delta}$ (right) for each $\mathcal{D}_1$ in Setting IV.

Tables 11–14 give the Monte Carlo standard error of the empirical mean squared error for $\widetilde{\boldsymbol{\beta}}_2(\tilde{\lambda})$ and comparison estimators in Settings I–IV respectively.

n_1	n_2	MCse $\cdot 10^3$ of eMSE					
		$\widetilde{\boldsymbol{\beta}}_2(\tilde{\lambda})$	$\widehat{\boldsymbol{\beta}}_2(\boldsymbol{W}_{\widehat{\lambda}})$	$\widehat{\boldsymbol{\beta}}_2^{TG}$	$\widehat{\boldsymbol{\beta}}_2^{TL}$	$\widehat{\boldsymbol{\beta}}_2^{p}$	$\widehat{\boldsymbol{\beta}}_2$
	50	2.73	2.73	3.80	4.48	1.49	4.38
50	100	1.31	1.35	2.94	1.87	1.06	1.81
	500	0.295	0.289	0.316	0.321	0.284	0.31
	50	2.64	2.63	3.48	3.98	1.15	4.38
100	100	1.27	1.31	2.68	1.60	0.945	1.81
	500	0.289	0.286	0.327	0.304	0.291	0.310
	50	2.60	2.57	3.11	3.98	0.438	4.38
500	100	1.25	1.28	2.15	1.58	0.526	1.81
	500	0.284	0.285	0.313	0.304	0.384	0.310

Table 11: Setting I: Monte Carlo standard error (MCse) of eMSE of $\widetilde{\boldsymbol{\beta}}_2(\tilde{\lambda})$, $\widehat{\boldsymbol{\beta}}_2(\boldsymbol{W}_{\widehat{\lambda}})$, $\widehat{\boldsymbol{\beta}}_2^{TG}$, $\widehat{\boldsymbol{\beta}}_2^{TL}$, $\widehat{\boldsymbol{\beta}}_2^{p}$, $\widehat{\boldsymbol{\beta}}_2$.

$\boldsymbol{\delta}$	$\boldsymbol{X}_1\boldsymbol{X}_1^\top$	MCse $\cdot 10^3$ of eMSE				
		$\widetilde{\boldsymbol{\beta}}_2(\tilde{\lambda})$	$\widehat{\boldsymbol{\beta}}_2(\boldsymbol{W})$	$\widehat{\boldsymbol{\beta}}_2^{TG}$	$\widehat{\boldsymbol{\beta}}_2^{p}$	$\widehat{\boldsymbol{\beta}}_2$
$= \mathbf{0}$	large	1.37	2.24	2.54	0.586	2.35
	small	2.29	3.66	3.82	0.865	3.78
$\neq \mathbf{0}$	large	1.68	1.62	2.37	1.03	1.98
	small	1.89	2.30	3.28	1.44	3.05

Table 12: Setting II: Monte Carlo standard error (MCse) of eMSE of $\widetilde{\boldsymbol{\beta}}_2(\tilde{\lambda})$, $\widehat{\boldsymbol{\beta}}_2(\boldsymbol{W})$, $\widehat{\boldsymbol{\beta}}_2^{TG}$, $\widehat{\boldsymbol{\beta}}_2^{p}$, $\widehat{\boldsymbol{\beta}}_2$.

$\boldsymbol{\delta}$	MCse $\cdot 10^3$ of eMSE				
	$\widetilde{\boldsymbol{\beta}}_2(\tilde{\lambda})$	$\widehat{\boldsymbol{\beta}}_2(\boldsymbol{W})$	$\widehat{\boldsymbol{\beta}}_2^{TG}$	$\widehat{\boldsymbol{\beta}}_2^{p}$	$\widehat{\boldsymbol{\beta}}_2$
(i) independent features, $\rho = 0$					
$(0,0,0)$	0.679	0.894	1.23	0.190	1.03
$(0,1,0)$	0.950	0.980	1.03	1.13	0.978
$(0,2,0)$	1.50	4.30	1.48	1.87	1.58
(ii) correlated features, $\rho = 0.4$					
$(0,0,0)$	5.16	8.94	6.60	0.166	8.59
$(0,1,0)$	5.77	8.64	6.07	2.48	7.76
$(0,2,0)$	11.6	58.1	9.82	5.07	18.1

Table 13: Setting III: Monte Carlo standard error (MCse) of eMSE of $\widetilde{\boldsymbol{\beta}}_2(\tilde{\lambda})$, $\widehat{\boldsymbol{\beta}}_2(\boldsymbol{W})$, $\widehat{\boldsymbol{\beta}}_2^{TG}$, $\widehat{\boldsymbol{\beta}}_2^{p}$, $\widehat{\boldsymbol{\beta}}_2$.

Case	MCse $\cdot 10^3$ of eMSE					
	$\widetilde{\boldsymbol{\beta}}_2(\tilde{\lambda})$	$\widehat{\boldsymbol{\beta}}_2(\boldsymbol{W}_{\widehat{\lambda}})$	$\widehat{\boldsymbol{\beta}}_2^{TG}$	$\widehat{\boldsymbol{\beta}}_2^{TL}$	$\widehat{\boldsymbol{\beta}}_2^{p}$	$\widehat{\boldsymbol{\beta}}_2$
Cauchy	0.308	0.308	0.310	0.315	1.70	0.310
dropped Z	0.272	0.269	0.362	0.285	0.264	0.310
X^2	0.310	0.310	0.310	0.334	47.0	0.310

Table 14: Setting IV: Monte Carlo standard error (MCse) of eMSE of $\widetilde{\boldsymbol{\beta}}_2(\tilde{\lambda})$, $\widehat{\boldsymbol{\beta}}_2(\boldsymbol{W}_{\widehat{\lambda}})$, $\widehat{\boldsymbol{\beta}}_2^{TG}$, $\widehat{\boldsymbol{\beta}}_2^{TL}$, $\widehat{\boldsymbol{\beta}}_2^{p}$, $\widehat{\boldsymbol{\beta}}_2$.

B.2 Multi-source simulation

For the purposes of this section, we denote by $\mathcal{D}_1^{\mathcal{S}}, \ldots, \mathcal{D}_M^{\mathcal{S}}$ M source data sets and $\mathcal{D}^{\mathcal{T}}$ the target data set. Each data set $\mathcal{D}_j^{\mathcal{S}}$ is composed of features $\boldsymbol{X}_{ij}^{\mathcal{S}}$ and outcomes $y_{ij}^{\mathcal{S}}$, $i = 1, \ldots, n_j^{\mathcal{S}}$, $j = 1, \ldots, M$, and the target data set $\mathcal{D}^{\mathcal{T}}$ is composed of features $\boldsymbol{X}_i^{\mathcal{T}}$ and outcomes $y_i^{\mathcal{T}}$, $i = 1, \ldots, n^{\mathcal{T}}$.

In the notation of the main manuscript, the concatenation approach considers $\boldsymbol{X}_1 = (\boldsymbol{X}_1^{\mathcal{S}}, \ldots, \boldsymbol{X}_M^{\mathcal{S}})$, $\boldsymbol{y}_1 = ({\boldsymbol{y}_1^{\mathcal{S}}}^{\top}, \ldots, {\boldsymbol{y}_M^{\mathcal{S}}}^{\top})\top$, $\mathcal{D}_1 = \{\boldsymbol{X}_1, \boldsymbol{y}_1\}$ and $\mathcal{D}_2 = \mathcal{D}^{\mathcal{T}}$, with $n_1 = n_1^{\mathcal{S}} + \ldots + n_M^{\mathcal{S}}$ and $n_2 = n^{\mathcal{T}}$, $\boldsymbol{\beta}_2 = \boldsymbol{\beta}_{\mathcal{T}}$. We have suggested that this concatenation approach will work well when the source data sets are mostly homogeneous or mostly heterogeneous. It is possible, as described in Section 2.1, that there are groups of source data sets that are homogeneous within and heterogeneous across. When the number of source data sets is small, as is the case below, it may be possible to consider integrating the target data set with only one of the source data sets, and to compare performance across these separate integrative estimates in order to select the best approach. In this section, we investigate this question by examining the trade-offs of concatenation versus single-source transfer learning when $M = 3$ source data sets are available.

In both source and target data sets, features consist of an intercept and three continuous variables independently generated from a standard Gaussian distribution. Source outcomes $y_{ij}^{\mathcal{S}}$ are simulated from the Gaussian distribution with $\mathbb{E}(Y_{ij}^{\mathcal{S}}) = (\boldsymbol{X}_{ij}^{\mathcal{S}})^{\top}\boldsymbol{\beta}_j^{\mathcal{S}}$ and $\mathbb{V}(Y_{ij}^{\mathcal{S}}) = 1$, $j = 1, \ldots, M$, and target outcomes $y_i^{\mathcal{T}}$ are simulated from the Gaussian distribution with $\mathbb{E}(Y_i^{\mathcal{T}}) = (\boldsymbol{X}_i^{\mathcal{T}})^{\top}\boldsymbol{\beta}_{\mathcal{T}}$ and $\mathbb{V}(Y_i^{\mathcal{T}}) = 1$. We fix $n_1^{\mathcal{S}} = n_2^{\mathcal{S}} = n_3^{\mathcal{S}} = 100$, $n_{\mathcal{T}} = 50$. We simulate three source data sets $\mathcal{D}_j^{\mathcal{S}} = \{y_{ij}^{\mathcal{S}}, \boldsymbol{X}_{ij}^{\mathcal{S}}\}_{i=1}^{n_j^{\mathcal{S}}}$. We consider three settings of values for $\boldsymbol{\beta}_j^{\mathcal{S}}$ and $\boldsymbol{\beta}^{\mathcal{T}}$ that characterize the heterogeneity/homogeneity of the sources relative to the target. Specifically, defining $\boldsymbol{c} = (1, -1.8, 2.6, 1.4)$:

- In Setting I, $\boldsymbol{\beta}_1^{\mathcal{S}} = \boldsymbol{c} + (1/4, 0, 0, 0)$, $\boldsymbol{\beta}_2^{\mathcal{S}} = \boldsymbol{c} + (0, 1/4, 0, 0)$, $\boldsymbol{\beta}_3^{\mathcal{S}} = \boldsymbol{c} + (0, 0, 1, 0)$ and $\boldsymbol{\beta}^{\mathcal{T}} = \boldsymbol{c} + (0, 0, 5/4, 0)$. Here, the first and second source data sets are closer to each other than to the third source data set, which is itself closer to the target data set.

- In Setting II, $\boldsymbol{\beta}_1^{\mathcal{S}}$ and $\boldsymbol{\beta}_2^{\mathcal{S}}$ are as in Setting I, $\boldsymbol{\beta}_3^{\mathcal{S}} = \boldsymbol{c} + (0, 0, 1/4, 0)$ and $\boldsymbol{\beta}^{\mathcal{T}} = \boldsymbol{c} + (0, 0, 0, 1/4)$. All source data sets are close to each other and to the target data set.

- In Setting III, $\boldsymbol{\beta}_1^{\mathcal{S}} = \boldsymbol{c} + (1, 0, 0, 0)$, $\boldsymbol{\beta}_2^{\mathcal{S}} = \boldsymbol{c} + (0, 1, 0, 0)$, $\boldsymbol{\beta}_3^{\mathcal{S}} = \boldsymbol{c} + (0, 0, 1, 0)$ and $\boldsymbol{\beta}^{\mathcal{T}} = \boldsymbol{c} + (0, 0, 0, 1)$. All source data sets are far from each other and the target data

source data set	Setting I	Setting II	Setting III
$\mathcal{D}_1^{\mathcal{S}}$	8.68 (1.96)	6.17 (1.38)	8.53 (1.94)
$\mathcal{D}_2^{\mathcal{S}}$	8.69 (1.96)	7.24 (1.50)	8.68 (1.99)
$\mathcal{D}_3^{\mathcal{S}}$	**6.73** (1.36)	7.17 (1.55)	8.60 (1.94)
$\{\mathcal{D}_1^{\mathcal{S}}, \mathcal{D}_2^{\mathcal{S}}, \mathcal{D}_3^{\mathcal{S}}\}$	8.63 (1.93)	**5.90** (1.28)	**8.48** (1.94)
$\varnothing$	8.65 (1.98)	8.65 (1.98)	8.65 (1.98)

Table 15: Multi-source simulation: eMSE $\times 10^2$ (Monte Carlo standard error $\times 10^3$) of $\widetilde{\boldsymbol{\beta}}_{\mathcal{T}}(\tilde{\lambda})$. Minimum eMSE in each Setting is bolded.

set.

The MLEs $\widehat{\boldsymbol{\beta}}_j^{\mathcal{S}}$ are reported in Figure 10. For each setting, we generate 1000 data sets $\mathcal{D}_{\mathcal{T}} = \{y_i^{\mathcal{T}}, \boldsymbol{X}_i^{\mathcal{T}}\}_{i=1}^{n^{\mathcal{T}}}$.

For each $\mathcal{D}_{\mathcal{T}}$, we compute $\widetilde{\boldsymbol{\beta}}_2(\tilde{\lambda})$ using each source separately and using the concatenation of the sources. We also compute the target-only MLE $\widehat{\boldsymbol{\beta}}_2$, which corresponds to setting the source data set to be $\varnothing$. Empirical mean squared error (eMSE) of $\widetilde{\boldsymbol{\beta}}_2(\tilde{\lambda})$ averaged across the 1000 target data sets in each setting are reported in Table 15. In all cases, our information-shrinkage estimator is more efficient than the target-only MLE. In Setting I, the estimator with the smallest eMSE uses $\mathcal{D}_3^{\mathcal{S}}$ only, which is to be expected given that $\mathcal{D}_3^{\mathcal{S}}$ is objectively closer to the target data than the others. In Setting II, the estimator with the smallest eMSE uses the concatenation $\{\mathcal{D}_1^{\mathcal{S}}, \mathcal{D}_2^{\mathcal{S}}, \mathcal{D}_3^{\mathcal{S}}\}$. Again, this makes intuitive sense given that the three source data sets are similar and similar to the target—no reason not to concatenate. In Setting III, the estimator with the smallest eMSE uses the concatenation $\{\mathcal{D}_1^{\mathcal{S}}, \mathcal{D}_2^{\mathcal{S}}, \mathcal{D}_3^{\mathcal{S}}\}$. This, too, is not unexpected, since the concatenation of three source data sets that differ from each other and the target can be no worse than a single source data set that differs from the target.

The concatenation approach appears to be the optimal choice in most settings. Setting I is a somewhat pathological example, in that it is rare that a single source out of many is much more similar to the target than any other source. Usually, such as setting is easy to identify *a priori* based on external information such as study characteristics. Nonetheless, it is reassuring that, even in this extreme setting, the concatenation approach remains a reasonable choice. This is in fact very much by design: the selection of $\tilde{\lambda}$ theoretically guarantees that the concatenation outperforms the target-only estimation (in terms of MSE). In other words, regardless of the heterogeneity/homogeneity structure of the sources, the concatenation guarantees protection against negative transfer and is a safe approach.

Beyond guaranteeing robustness to heterogeneity/homogeneity of multiple sources, an investigator may wish to select the approach that is "best" in the sense of minimizing MSE. Specifically, treating $\widehat{\mathrm{MSE}}(\tilde{\lambda})$ as a function of the source data set configuration

$$\mathcal{S} \mapsto \underbrace{n_{\mathcal{T}}^{-1}\widehat{\sigma}_{\mathcal{T}}^2 \,\mathrm{trace}\{\boldsymbol{S}^{-2}(\tilde{\lambda})\boldsymbol{G}_{\mathcal{T}}\} + \tilde{\lambda}^2 \mathrm{trace}\{\boldsymbol{G}_{\mathcal{S}}\boldsymbol{S}^{-2}(\tilde{\lambda})\boldsymbol{G}_{\mathcal{S}}\widehat{\boldsymbol{\delta}}^2\}}_{\widehat{\mathrm{MSE}}(\tilde{\lambda})},$$

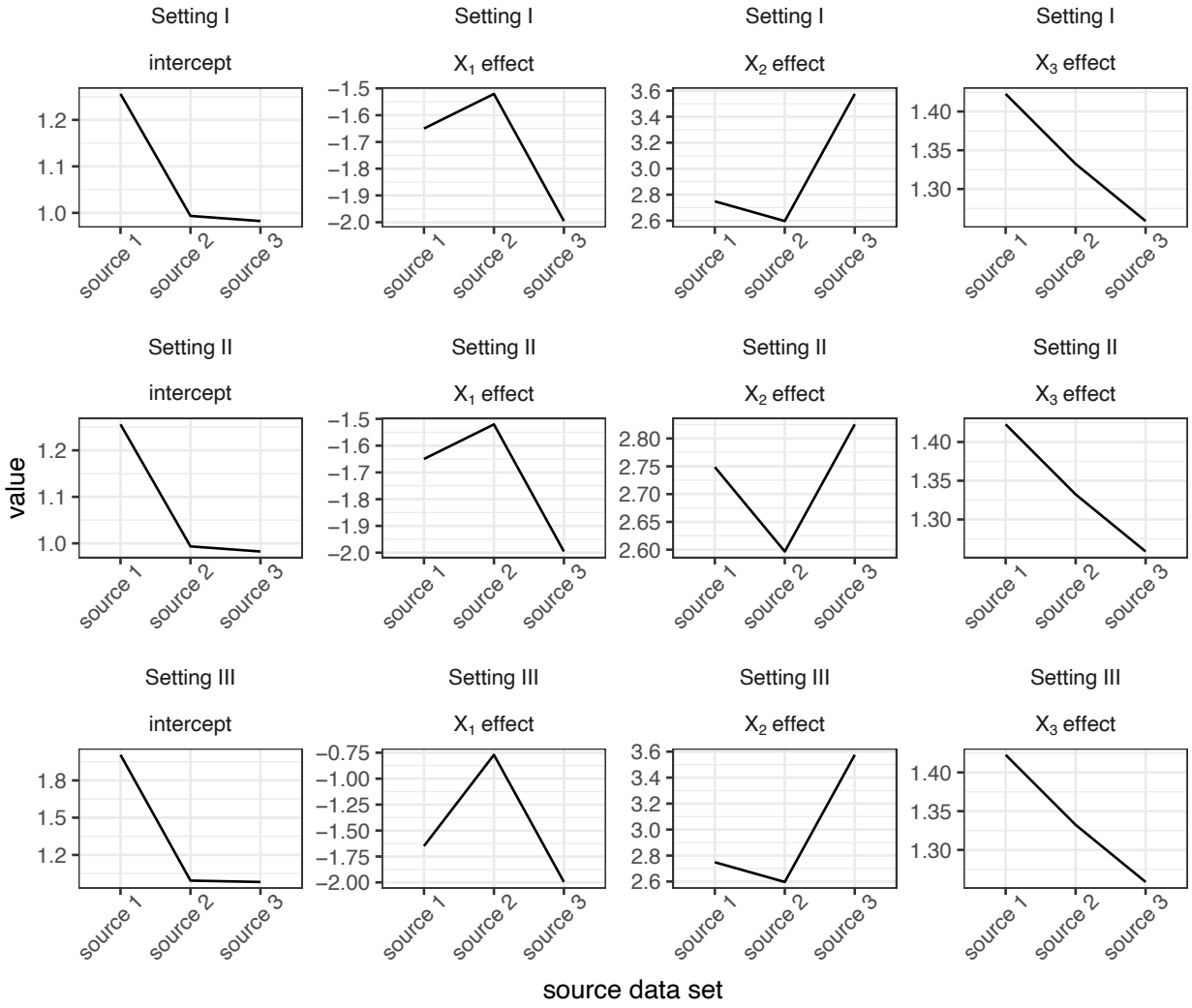

Figure 10: MLEs $\widehat{\boldsymbol{\beta}}_1^{\mathcal{S}}$, $\widehat{\boldsymbol{\beta}}_2^{\mathcal{S}}$, $\widehat{\boldsymbol{\beta}}_3^{\mathcal{S}}$ and $\widehat{\boldsymbol{\beta}}^{\mathcal{T}}$ in Settings I, II and III.

one may select the source data set configuration $\widehat{\mathcal{S}}$ to minimize the right-hand side. As $\widehat{\mathrm{MSE}}(\tilde{\lambda})$ balances bias and variance to evaluate the most efficient estimator's inferential performance, an investigator may choose the single source or concatenated source that gives the most efficient estimator. Across the 1000 simulations, we compare the value of $\widehat{\mathrm{MSE}}(\tilde{\lambda})$ across the five source configurations in each setting. In Setting I, the estimator that uses $\mathcal{D}_3^{\mathcal{S}}$ has smaller $\widehat{\mathrm{MSE}}(\tilde{\lambda})$ in 100%, 100% and 99.7% of simulations compared to the estimator that uses $\mathcal{D}_1^{\mathcal{S}}$, $\mathcal{D}_2^{\mathcal{S}}$ and $\{\mathcal{D}_1^{\mathcal{S}}, \mathcal{D}_2^{\mathcal{S}}, \mathcal{D}_3^{\mathcal{S}}\}$, respectively. In Setting II, the estimator that uses $\{\mathcal{D}_1^{\mathcal{S}}, \mathcal{D}_2^{\mathcal{S}}, \mathcal{D}_3^{\mathcal{S}}\}$ has smaller $\widehat{\mathrm{MSE}}(\tilde{\lambda})$ in 61.4%, 83.0% and 83.5% of simulations compared to the estimator that uses $\mathcal{D}_1^{\mathcal{S}}$, $\mathcal{D}_2^{\mathcal{S}}$ and $\mathcal{D}_3^{\mathcal{S}}$, respectively. In Setting III, the estimator that uses $\{\mathcal{D}_1^{\mathcal{S}}, \mathcal{D}_2^{\mathcal{S}}, \mathcal{D}_3^{\mathcal{S}}\}$ has smaller $\widehat{\mathrm{MSE}}(\tilde{\lambda})$ in 67.8%, 98.4% and 89.4% of simulations compared to the estimators that use $\mathcal{D}_1^{\mathcal{S}}$, $\mathcal{D}_2^{\mathcal{S}}$ and $\mathcal{D}_3^{\mathcal{S}}$, respectively. In other words, across all settings, the criterion $\widehat{\mathrm{MSE}}(\tilde{\lambda})$ does reasonably well in helping an investigator select the most useful source configuration in terms of minimizing MSE, showing less certainty in settings for which

the source configurations perform similarly. We therefore suggest that an investigator rely on $\widehat{\text{MSE}}(\tilde{\lambda})$ to decide among several multi-source integrative estimators and a concatenation approach.

C eICU Database information

The eICU Collaborative Research Database (Pollard et al., 2018) data are publicly available upon completion of training and authentication. Users may begin their data access request at `https://eicu-crd.mit.edu/gettingstarted/access/`.

Our analysis focuses on African American and Caucasian patients between the ages of 40 and 89 with body mass index (BMI) between 14 and 60 that were admitted through any means other than another intensive care unit (ICU).

D Bibliography

Amari, S. and Nagaoka, H. (2000). *Methods of Information Geometry*, volume 191. American Mathematical Society.

Andersen, P. and Gill, R. (1982). Cox's regression model for counting processes: a large sample study. *The Annals of Statistics*, 10(4):1100–1120.

Baldi, P. and Itti, L. (2010). Of bits and wows: A Bayesian theory of surprise with applications to attention. *Neural Networks*, 23(5):649–666.

Bondell, H. D. and Reich, B. J. (2009). Simultaneous factor selection and collapsing levels in ANOVA. *Biometrics*, 65(1):169–177.

Cahoon, J. and Martin, R. (2020). A generalized inferential model for meta-analyses based on few studies. *Statistics and Applications*, 18(2):299–316.

Cai, T., Liu, M., and Xia, Y. (2019). Individual data protected integrative regression analysis of high-dimensional heterogeneous data. *arXiv preprint arXiv:1902.06115*.

Cai, T. T. and Wei, J. (2021). Transfer learning for nonparametric classification: Minimax rate and adaptive classifier. *The Annals of Statistics*, 49(1):100–128.

Chatterjee, N., Chen, Y.-H., Maas, P., and Carroll, R. J. (2016). Constrained maximum likelihood estimation for model calibration using summary-level information from external big data sources. *Journal of the American Statistical Association*, 111(513):107–117.

Chen, A., Owen, A. B., and Shi, M. (2015). Data enriched linear regression. *Electronic Journal of Statistics*, 9(1):1078–1112.

Chen, X., Liu, W., Mao, X., and Yang, Z. (2019). Distributed high-dimensional regression under a quantile loss function. *arXiv preprint arXiv:1906.05741*.

Cheng, C., Zhou, B., Ma, G., Wu, D., and Yuan, Y. (2020). Wasserstein distance based deep adversarial transfer learning for intelligent fault diagnosis. *Neurocomputing*, 409:35–45.

Claggett, B., Xie, M., and Tian, L. (2014). Meta-analysis with fixed, unknown, study-specific parameters. *Journal of the American Statistical Association*, 109(508):1660–1671.

Duan, R., Ning, Y., and Chen, Y. (2019). Heterogeneity-aware and communication-efficient distributed statistical inference. *arXiv preprint arXiv:1912.09623*.

Dube, P., Bhattacharjee, B., Petit-Bois, E., and Hill, M. (2020). Improving transferability of deep neural networks. *Domain Adaptation for Visual Understanding*, pages 51–64.

Efron, B. and Morris, C. (1973). Combining possibly related estimation problems. *Journal of the Royal Statistical Society Series B*, 35(3):379–421.

Efron, B. and Morris, C. (1977). Stein's paradox in statistics. *Scientific American*, 236(5):119–127.

Farebrother, R. W. (1976). Further results on the mean square error of ridge regression. *Journal of the Royal Statistical Society Series B*, 38(3):248–250.

Godambe, V. P. (1991). *Estimating functions*. Oxford University Press.

Hector, E. C. and Song, P. X.-K. (2020). Doubly distributed supervised learning and inference with high-dimensional correlated outcomes. *Journal of Machine Learning Research*, 21:1–35.

Hemmerle, W. J. (1975). An explicit solution for generalized ridge regression. *Technometrics*, 17(3):309–314.

Higgins, J. P. T. and Thompson, S. G. (2002). Quantifying heterogeneity in a meta-analysis. *Statistics in Medicine*, 21(11):1539–1558.

Hjort, N. L. and Pollard, D. (1993). Asymptotics for minimisers of convex processes. *arXiv*, arXiv:1107.3806.

Hoerl, A. E. and Kennard, R. W. (1970). Ridge regression: biased estimation for nonorthogonal problems. *Technometrics*, 12(1):55–67.

James, W. and Stein, C. (1960). Estimation with quadratic loss. *Fourth Berkeley Symposium on Mathematical Statistics and Probability*, 4.1:361–379.

Jordan, M. I., Lee, J. D., and Yang, Y. (2018). Communication-efficient distributed statistical inference. *Journal of the American Statistical Association*.

Ke, Z. T., Fan, J., and Wu, Y. (2015). Homogeneity pursuit. *Journal of the American Statistical Association*, 110(509):175–194.

Li, S., Cai, T. T., and Li, H. (2020). Transfer learning in large-scale gaussian graphical models with false discovery rate control. *arXiv*, arXiv:2010.11037.

Li, S., Cai, T. T., and Li, H. (2022). Transfer learning for high-dimensional linear regression: Prediction, estimation and minimax optimality. *Journal of the Royal Statistical Society Series B*, 84(1):149–173.

Lin, L. and Lu, J. (2019). A race-dc in big data. *arXiv preprint arXiv:1911.11993*.

Liu, D., Liu, R. Y., and Xie, M. (2015). Multivariate meta-analysis of heterogeneous studies using only summary statistics: efficiency and robustness. *Journal of the American Statistical Association*, 110(509):326–340.

Luo, L. and Song, P. X.-K. (2021). Multivariate online regression analysis with heterogeneous streaming data. *Canadian Journal of Statistics*, 0(0):1–23.

Michael, H., Thornton, S., Xie, M., and Tian, L. (2019). Exact inference on the random-effects model for meta-analyses with few studies. *Biometrics*, 75(2):485–493.

Nielsen, F. (2020). An elementary introduction to information geometry. *Entropy*, 22(10):1100.

Obenchain, R. L. (1977). Classical F-tests and confidence regions for ridge regression. *Technometrics*, 19(4):429–439.

Olkin, I. and Pukelsheim, F. (1982). The distance between two random vectors with given dispersion matrices. *Linear Algebra and its Applications*, 48:257–263.

Pan, S. J. and Yang, Q. (2010). A survey on transfer learning. *IEEE Transactions on knowledge and data engineering*, 22(10):1345–1359.

Pollard, D. (1991). Asymptotics for least absolute deviation regression estimators. *Econometric Theory*, 7(2):186–199.

Pollard, T. J., Johnson, A. E. W., Raffa, J. D., Celi, L. A., Mark, R. G., and Badawi, O. (2018). The eICU Collaborative Research Database, a freely available multi-center database for critical care research. *Scientific Data*, 5(1):1–13.

Raghu, M., Zhang, C., Kleinberg, J., and Bengio, S. (2019). Transfusion: understanding transfer learning for medical imaging. *Advances in NEural Information Processing Systems 33*.

Reeve, H. W., Cannings, T. I., and Samworth, R. J. (2021). Adaptive transfer learning. *The Annals of Statistics*, 49(6):3618–3649.

Schifano, E. D., Wu, J., Wang, C., Yan, J., and Chen, M.-H. (2016). Online updating of statistical inference in the big data setting. *Technometrics*, 58(3):393–403.

Shen, J., Liu, R. Y., and Xie, M.-g. (2020). i fusion: Individualized fusion learning. *Journal of the American Statistical Association*, 115(531):1251–1267.

Shen, J., Qu, Y., Zhang, W., and Yu, Y. (2018). Wasserstein distance guided representation learning for domain adaptation. In *Thirty-Second AAAI Conference on Artificial Intelligence*, pages 3–9.

Shen, X. and Huang, H.-C. (2010). Grouping pursuit through a regularization solution surface. *Journal of the American Statistical Association*, 105(490):727–739.

Stein, C. M. (1956). Inadmissibility of the usual estimator for the mean of a multivariate normal distribution. *Proceedings of the Third Berkeley Symposium*, 1:197–206.

Stein, C. M. (1981). Estimation of the mean of a multivariate normal distribution. *The Annals of Statistics*, 9(6):1135–1151.

Tan, C., Sun, F., Kong, T., Zhang, W., Yang, C., and Liu, C. (2018). A survey on deep transfer learning. In Kůrková, V., Manolopoulos, Y., Hammer, B., Iliadis, L., and Maglogiannis, I., editors, *Artificial Neural Networks and Machine Learning – ICANN 2018*, pages 270–279, Cham. Springer International Publishing.

Tang, L. and Song, P. X.-K. (2016). Fused lasso approach in regression coefficients clustering – learning parameter heterogeneity in data integration. *Journal of Machine Learning Research*, 17(113):1–23.

Tang, L. and Song, P. X.-K. (2020). Poststratification fusion learning in longitudinal data analysis. *Biometrics*.

Tang, L., Zhou, L., and Song, P. X.-K. (2020). Distributed simultaneous inference in generalized linear models via confidence distribution. *Journal of Multivariate Analysis*, 176:104567.

Theobald, C. M. (1974). Generalizations of mean square error applied to ridge regression. *Journal of the Royal Statistical Society Series B*, 36(1):103–106.

Tian, Y. and Feng, Y. (2022). Transfer learning under high-dimensional generalized linear models. *Journal of the American Statistical Association*, doi: 10.1080/01621459.2022.2071278:1–14.

Tibshirani, R. J. and Taylor, J. (2011). The solution path of the generalized lasso. *The Annals of Statistics*, 39(3):1335–1371.

Torrey, L. and Shavlik, J. (2009). *Handbook of research on machine learning applications and trends: algorithms, methods, and techniques*, chapter Transfer learning, pages 242–264. IGI Global.

Toulis, P. and Airoldi, E. M. (2017). Asymptotic and finite-sample properties of estimators based on stochastic gradients. *The Annals of Statistics*, 45(4):1694–1727.

van Wieringen, W. N. (2021). Lecture notes on ridge regression. *arXiv*, arXiv:1509.09169v7.

Vinod, H. D. (1987). Confidence intervals for ridge regression parameters. *Time Series and Econometric Modelling*, 36:279–300.

Wang, Z., Kim, J. K., and Yang, S. (2018). Approximate bayesian inference under informative sampling. *Biometrika*, 105(1):91–102.

Weiss, K., Khoshgoftaar, T. M., and Wang, D. (2016). A survey of transfer learning. *Journal of Big Data*, 3(9):1–40.

Xie, M., Singh, K., and Strawderman, W. E. (2011). Confidence distributions and a unifying framework for meta-analysis. *Journal of the American Statistical Association*, 106(493):320–333.

Yang, S. and Ding, P. (2019). Combining multiple observational data sources to estimate causal effects. *Journal of the American Statistical Association*, 115(531):1540–1554.

Yoon, T., Lee, J., and Lee, W. (2020). Joint transfer of model knowledge and fairness over domains using wasserstein distance. *IEEE Access*, 8:123783–123798.

Yosinski, J., Clune, J., Bengio, Y., and Lipson, H. (2014). How transferable are features in deep neural networks? *Advances in Neural Information Processing Systems 27*, pages 3320–3328.

Zheng, C., Dasgupta, S., Xie, Y., Haris, A., and Chen, Y. Q. (2019). On data enriched logistic regression. *arXiv*, arXiv:1911.06380.

Zhu, Z., Wang, L., Peng, G., and Li, S. (2021). WDA: an improved Wasserstein distance-based transfer learning fault diagnosis method. *Sensors*, 21(13):4394.

Zhuang, F., Qi, Z., Duan, K., Xi, D., Zhu, Y. Z. H., Xiong, H., and He, Q. (2021). A comprehensive survey on transfer learning. *Proceedings of the IEEE*, 109(1):43–76.