

Adaptive Lasso, Transfer Lasso, and Beyond: An Asymptotic Perspective

Masaaki Takada^{*†}
 masaaki1.takada@toshiba.co.jp Hironori Fujisawa[†]
 fujisawa@ism.ac.jp

Abstract

This paper presents a comprehensive exploration of the theoretical properties inherent in the Adaptive Lasso and the Transfer Lasso. The Adaptive Lasso, a well-established method, employs regularization divided by initial estimators and is characterized by asymptotic normality and variable selection consistency. In contrast, the recently proposed Transfer Lasso employs regularization subtracted by initial estimators with the demonstrated capacity to curtail non-asymptotic estimation errors. A pivotal question thus emerges: Given the distinct ways the Adaptive Lasso and the Transfer Lasso employ initial estimators, what benefits or drawbacks does this disparity confer upon each method? This paper conducts a theoretical examination of the asymptotic properties of the Transfer Lasso, thereby elucidating its differentiation from the Adaptive Lasso. Informed by the findings of this analysis, we introduce a novel method, one that amalgamates the strengths and compensates for the weaknesses of both methods. The paper concludes with validations of our theory and comparisons of the methods via simulation experiments.

1 Introduction

We consider an ordinary high-dimensional regression problem. Let $X = (\mathbf{x}_1, \dots, \mathbf{x}_p) = (x_1^\top, \dots, x_n^\top)^\top \in \mathbb{R}^{n \times p}$ and $y \in \mathbb{R}^n$ be a feature matrix and response vector, respectively. We suppose a true model is linear with independent and identically distributed (i.i.d.) Gaussian noise, that is,

$$(1) \quad y = X\beta^* + \varepsilon, \quad \varepsilon_i \stackrel{i.i.d.}{\sim} \mathcal{N}(0, \sigma^2),$$

^{*}Toshiba Corporation

[†]The Institute of Statistical Mathematics

where $\beta^* \in \mathbb{R}^p$ is a true regression parameter and $\varepsilon \in \mathbb{R}^n$ is a Gaussian noise. We presume that β^* is sparse, and designate the active and inactive parameters as S and S^c , namely $S := \{j : \beta_j^* \neq 0\}$ and $S^c := \{j : \beta_j^* = 0\}$, respectively.

The *Lasso* [18] is a classical regression method for high-dimensional data, defined by

$$(2) \quad \hat{\beta}_n^L = \underset{\beta}{\operatorname{argmin}} \left\{ \frac{1}{n} \|y - X\beta\|_2^2 + \frac{\lambda_n}{n} \sum_j |\beta_j| \right\}.$$

Owing to ℓ_1 regularization, the solution exhibits sparsity. We denote $\hat{S}_n^L := \{j : \hat{\beta}_j^L \neq 0\}$.

Numerous theoretical studies have elucidated the strengths and limitations of the Lasso. According to asymptotic theory, the Lasso estimator is consistent if $\lambda_n = o(n)$ and is \sqrt{n} -consistent if $\lambda_n = O(\sqrt{n})$ [5]. However, [21] demonstrates that the Lasso has *inconsistent* variable selection if $\lambda_n = O(\sqrt{n})$, while it does not have \sqrt{n} -consistency if $\lambda_n = o(n)$ and $\lambda_n/\sqrt{n} \rightarrow \infty$. Hence, the Lasso cannot achieve both \sqrt{n} -consistency and consistent variable selection simultaneously (see Figure 1 left).

To improve the asymptotic properties of the Lasso, one of the most well-known methods is the *Adaptive Lasso* [21, 10], which is given by

$$(3) \quad \hat{\beta}_n^A = \underset{\beta}{\operatorname{argmin}} \left\{ \frac{1}{n} \|y - X\beta\|_2^2 + \frac{\lambda_n}{n} \sum_j w_j |\beta_j| \right\}, \quad w_j := \frac{1}{|\tilde{\beta}_j|^\gamma},$$

where $\tilde{\beta}$ is an initial estimator of the true parameter β^* and $\gamma > 0$ is a hyperparameter. We denote $\hat{S}_n^A := \{j : \hat{\beta}_j^A \neq 0\}$. If $\tilde{\beta}$ is a \sqrt{n} -consistent estimator, $\lambda_n = o(\sqrt{n})$, and $\lambda_n n^{(\gamma-1)/2} \rightarrow \infty$, then the Adaptive Lasso satisfies both \sqrt{n} -consistency and consistent variable selection, as well as asymptotic normality (Figure 1 right). This is known as the *oracle property* because it behaves as if the true active variables were given in advance. The Adaptive Lasso assumes the existence of a \sqrt{n} -consistent initial estimator and uses it as the weight of the ℓ_1 regularization.

Recently, a different use of an initial estimator has been proposed [16, 1], which is given by

$$(4) \quad \hat{\beta}_n^T = \underset{\beta}{\operatorname{argmin}} \left\{ \frac{1}{n} \|y - X\beta\|_2^2 + \frac{\lambda_n}{n} \sum_j |\beta_j| + \frac{\eta_n}{n} \sum_j |\beta_j - \tilde{\beta}_j| \right\},$$

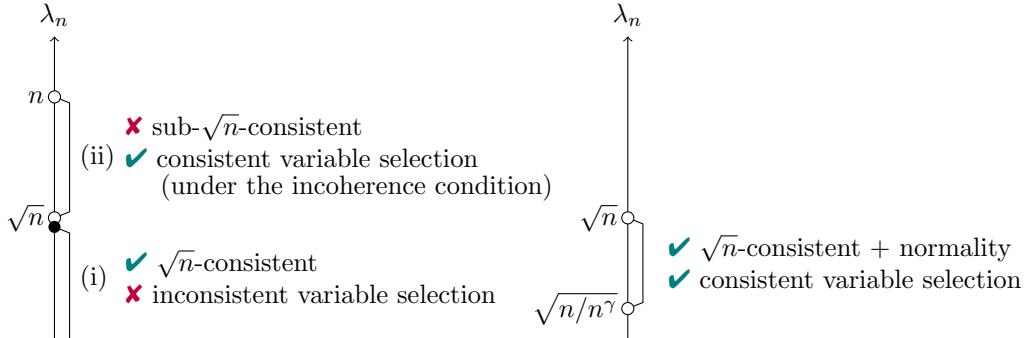


Figure 1: Phase diagrams with the order of λ_n for the Lasso (left) and the Adaptive Lasso (right). The Lasso does not achieve \sqrt{n} -consistent and consistent variable selection simultaneously, while the Adaptive Lasso satisfies both.

where $\tilde{\beta}$ is an initial estimator (“source parameter” in the field of transfer learning). We denote $\hat{S}_n^T := \{j : \hat{\beta}_j^T \neq 0\}$. This method is called *Transfer Lasso*. The first regularization term in (4) shrinks the estimator to zero and induces sparsity. The second regularization term in (4), on the other hand, shrinks the estimator to the initial estimator and induces the sparsity of changes from the initial estimator. The ℓ_1 regularization of the difference between the initial estimator and the target estimator plays a key role in sparse updating, in which only a small number of parameters are changed from the initial estimator. Non-asymptotic analysis reveals that a small $\Delta := \tilde{\beta} - \beta^*$ brings advantageous on its estimation error bounds for the Transfer Lasso over the Lasso [16].

The Adaptive Lasso and the Transfer Lasso have similarities and differences. They are similar in that they both use an initial estimator in ℓ_1 regularization. However, the way the initial estimator is used is different: the Adaptive Lasso uses the parameter “divided” by the initial estimator in the regularization, whereas Transfer Lasso uses the parameter “subtracted” by the initial estimator in the regularization. In addition, the original motivations are different: The Adaptive Lasso aims to reduce estimation bias as well as satisfy consistency in variable selection, whereas the Transfer Lasso aims to sparsify both the estimator itself and the change from the initial estimator, leveraging the knowledge of the initial estimator.

These raise major questions: How do these similarities and differences between Adaptive Lasso and Transfer Lasso affect the theoretical properties and empirical results of each method? In this paper, we highlight the asymp-

totic properties of each method and seek to answer the following research questions.

1. Does the Transfer Lasso have the same properties as the Adaptive Lasso? Specifically, does the Transfer Lasso have the oracle property that the Adaptive Lasso has?
2. Does the Transfer Lasso have different properties from the Adaptive Lasso? If so, under what conditions of initial estimators, does the Transfer Lasso have an advantage over the Adaptive Lasso, or vice versa?
3. If these two methods have their specific advantages and disadvantages, are there any ways to compensate for the disadvantages of both and to reconcile their advantages?
4. How does the asymptotic property of the estimator change as the order of the hyperparameters changes for each method?

Our theoretical analysis led us to the following findings.

1. The Transfer Lasso does not have the oracle property in general. This is an unfavorable property compared to the Adaptive Lasso.
2. The Transfer Lasso has an advantage in convergence rate if the initial estimator is estimated from sufficiently large data. The Adaptive Lasso, in contrast, does not benefit from such an initial estimator.
3. We found that a non-trivial integration of the Adaptive Lasso and the Transfer Lasso provides a combination of the benefits of both. The superiority of this integration was shown by asymptotic analysis and empirical simulations.
4. We comprehensively analyzed the relation between hyperparameters and asymptotic properties and drew phase diagrams representing them. Figure 2 illustrates the phase diagram of the Adaptive Lasso and the Transfer Lasso, and Figure 3 illustrates the phase diagram of the proposed method. These theoretical results were reproduced empirically by numerical simulations in Figure 5.

This paper discusses the above research questions in the following organization. First, we review the asymptotic properties of the Lasso and Adaptive Lasso (Section 2). Then, we define a setup for our analysis and theoretically

analyze the asymptotic properties of the Adaptive Lasso and the Transfer Lasso (Section 3). This elucidates the advantages and disadvantages of each method. Furthermore, to compensate for their disadvantages and to reconcile their advantages, we propose a novel method, which effectively integrates both of them (Section 4). We demonstrate its superiority through theoretical analysis. We then compare the Adaptive Lasso, the Transfer Lasso, and their integrated method through numerical experiments (Section 5). Finally, we provide additional discussion and conclusions (Sections 6 and 7).

Notations

Consider a vector $v \in \mathbb{R}^p$. We denote the element-wise absolute vector by $|v|$, with the j -th element given by $|v_j|$. The sign vector is represented as $\text{sgn}(v)$, with its elements being 1 for $v_j > 0$, -1 for $v_j < 0$, and 0 for $v_j = 0$. The support set of v is denoted as $\text{supp}(v)$ and defined as $\text{supp}(v) := \{j \in \{1, \dots, p\} | v_j \neq 0\}$. The ℓ_q -norm of v is expressed as $\|v\|_q = (\sum_{j=1}^p |v_j|^q)^{1/q}$.

For a matrix $M \in \mathbb{R}^{p \times p}$, we use $M \succeq O$ for a positive semi-definite matrix and $M \succ O$ for a positive definite matrix, implying $v^\top M v \geq 0$ for all $v \in \mathbb{R}^p$ and $v^\top M v > 0$ for all non-zero $v \in \mathbb{R}^p$, respectively.

Given a subset S of $\{1, \dots, p\}$, we denote its cardinality as $|S|$, and the complement set as $S^c = \{1, \dots, p\} \setminus S$. The vector v_S represents v restricted to the index set S . The matrix $M_{S_1 S_2}$ denotes the submatrix with row indices in S_1 and column indices in S_2 .

For sequences a_n and b_n , we use $a_n = O(b_n)$ to indicate that $|a_n/b_n|$ converges to a finite value, and $a_n = o(b_n)$ to signify $|a_n/b_n|$ converging to zero as $n \rightarrow \infty$.

2 Literature Review

We review some asymptotic properties for the Lasso and the Adaptive Lasso based on [5] and [21], and then present other related studies. All of the proofs in this section are essentially the same as those in [5] and [21], but for the sake of readability, we provide them in Appendix B.1.

We make the following assumption throughout this paper as in [5, 21].

Assumption 2.1.

$$(5) \quad C_n := \frac{1}{n} X^\top X \rightarrow C \succ O \quad (n \rightarrow \infty),$$

$$(6) \quad \frac{1}{n} \max_i \|x_i\|_2^2 \rightarrow 0 \ (n \rightarrow \infty).$$

Let W be a random variable of a Gaussian distribution with mean 0 and covariance $\sigma^2 C$, that is, $W \sim \mathcal{N}(0, \sigma^2 C)$.

2.1 Asymptotic Properties for the Lasso

The Lasso is given by (2). According to [5, 21], several asymptotic properties have been obtained for the Lasso: consistency (Lemma 2.2 and Corollary 2.3), convergence rate (Lemma 2.4, Corollary 2.5, Lemma 2.7, and Corollary 2.8), and variable selection consistency (Lemma 2.6).

Lemma 2.2 (Theorem 1 in [5] and Lemma 1 in [21]). *If $\lambda_n/n \rightarrow \lambda_0 \geq 0$, then*

$$(7) \quad \hat{\beta}_n^L \xrightarrow{p} \operatorname{argmin}_{\beta} \left\{ (\beta - \beta^*)^\top C(\beta - \beta^*) + \lambda_0 \sum_j |\beta_j| \right\}.$$

Corollary 2.3 (Consistency for Lasso). *If $\lambda_n = o(n)$, then $\hat{\beta}_n^L$ is consistent.*

Lemma 2.4 (Theorem 2 in [5] and Lemma 2 in [21]). *If $\lambda_n/\sqrt{n} \rightarrow \lambda_0 \geq 0$, then*

(8)

$$\begin{aligned} & \sqrt{n}(\hat{\beta}_n^L - \beta^*) \\ & \xrightarrow{d} \operatorname{argmin}_u \left\{ u^\top C u - 2u^\top W + \lambda_0 \sum_j (u_j \operatorname{sgn}(\beta_j^*) I(\beta_j^* \neq 0) + |u_j| I(\beta_j^* = 0)) \right\}. \end{aligned}$$

Corollary 2.5 (\sqrt{n} -consistency for Lasso). *If $\lambda_n = O(\sqrt{n})$, then $\hat{\beta}_n^L$ is \sqrt{n} -consistent.*

Lemma 2.6 (Inconsistent Variable Selection; Proposition 1 in [21]). *Let $\hat{S}_n^L := \{j : \hat{\beta}_j^L \neq 0\}$. If $\lambda_n/\sqrt{n} \rightarrow \lambda_0 \geq 0$, then*

$$(9) \quad \limsup_{n \rightarrow \infty} P(\hat{S}_n^L = S) \leq c < 1$$

where c is a constant.

Lemma 2.7 (Lemma 3 in [21]). *If $\lambda_n/n \rightarrow 0$ and $\lambda_n/\sqrt{n} \rightarrow \infty$, then*

(10)

$$\frac{n}{\lambda_n}(\hat{\beta}_n^{\mathcal{L}} - \beta^*) \xrightarrow{d} \operatorname{argmin}_u \left\{ u^\top C u + \sum_{j=1}^p (u_j \operatorname{sgn}(\beta_j^*) I(\beta_j^* \neq 0) + |u_j| I(\beta_j^* = 0)) \right\}.$$

Corollary 2.8 (Slower Rate Consistency for Lasso). *If $\lambda_n/n \rightarrow 0$ and $\lambda_n/\sqrt{n} \rightarrow \infty$, then the convergence rate of $\hat{\beta}_n^{\mathcal{L}}$ is slower than \sqrt{n} .*

We first obtain a convergence result for $\lambda_n = O(n)$ (Lemma 2.2). If $\lambda_n = o(n)$, then we have consistency for the Lasso (Corollary 2.3). Although $\lambda_n = o(n)$ is sufficient for consistency, it is not always \sqrt{n} -consistent. We obtain an asymptotic distribution for $\lambda_n = O(\sqrt{n})$ (Lemma 2.4). This implies \sqrt{n} -consistency for the Lasso (Corollary 2.5). Unfortunately, $\lambda_n = O(\sqrt{n})$ leads to inconsistent variable selection (Lemma 2.6). This implies that $\lambda_n = O(\sqrt{n})$ achieves \sqrt{n} -consistency but inconsistent variable selection for the Lasso. In contrast, if λ_n is greater than $O(\sqrt{n})$ and $\lambda_n = o(n)$, we obtain an asymptotic distribution (Lemma 2.7). This implies that the convergence rate is slower than \sqrt{n} (Corollary 2.8), although it can be a consistent variable selection under the incoherence conditions [21, 20].

Figure 1 (left) summarizes the asymptotic properties for the Lasso. It cannot simultaneously achieve both \sqrt{n} -consistent estimation and consistent variable selection. This is a major limitation of the Lasso and is the motivation to develop the Adaptive Lasso.

2.2 Asymptotic Properties for Adaptive Lasso

Adaptive Lasso is given by (3). It is known that the Adaptive Lasso has the so-called “oracle property” [21].

Lemma 2.9 (Oracle Property for Adaptive Lasso; Theorem 2 in [21]). *Suppose that $\hat{\beta}_n$ is a \sqrt{n} -consistent estimator. If $\lambda_n/\sqrt{n} \rightarrow 0$ and $\lambda_n n^{(\gamma-1)/2} \rightarrow \infty$, then the Adaptive Lasso estimator (2) satisfies the oracle property, that is, consistent variable selection and \sqrt{n} -consistency with asymptotic normality:*

$$(11) \quad \lim_{n \rightarrow \infty} P(\hat{S}_n^{\mathcal{A}} = S) = 1,$$

$$(12) \quad \sqrt{n}(\hat{\beta}_S^{\mathcal{A}} - \beta_S^*) \xrightarrow{d} \mathcal{N}(0, \sigma^2 C_{SS}^{-1}).$$

Proof. The proof is given in B.1.5. □

The oracle property demonstrates a clear advantage of the Adaptive Lasso over the Lasso. With a \sqrt{n} -consistent initial estimator, the Adaptive Lasso can simultaneously achieve both \sqrt{n} -consistent estimation and consistent variable selection (Figure 1 right). Thus, the Adaptive Lasso performs as well as if the true active variables were given in advance.

2.3 Other Related Work

Besides the Adaptive Lasso and the Transfer Lasso, several related methods have been studied. In this subsection, we review related methods in three categories: (I) methods with the oracle property similar to the Adaptive Lasso, (II) methods with two-stage estimation to eliminate bias, similar to the Adaptive Lasso, and (III) methods using the ℓ_1 norm to transfer knowledge about the source data, similar to the Transfer Lasso.

(I) The oracle property is known to hold not only for the Adaptive Lasso but also for the SCAD [4] and MCP [19]. These methods use nonconvex regularization, instead of using an initial estimator. Because of the nonconvexity, the algorithm converges to a local minimum and the oracle property holds only for some local minima or under restricted conditions. The Adaptive Lasso, on the other hand, uses convex regularization and always converges to a global minimum, although it requires an appropriate initial estimator.

(II) The Lasso penalizes the ℓ_1 norm of the parameters and thus introduces a bias, leading to the failure of the oracle property. Several two-step estimation methods have been proposed to eliminate the bias [13, 9, 3]. In [13], after the Lasso estimation in the first stage, the second stage is another Lasso estimation using only the selected variables. In [9], after the Lasso estimation in the first stage, the second stage is estimated by a linear combination of the first stage estimator and the OLS estimator of the selected variables. These methods are called Relaxed Lasso. [3] generalized these refitting methods as “methods that minimize the loss function with regularization and then decrease the loss function without regularization”. Based on this idea, they developed several refitting methods.

(III) Regularization of ℓ_1 -norm between target and initial estimators was proposed by [1, 11, 17] as well as the Transfer Lasso [16]. [1] corresponds to the case where $\lambda_n = 0$ in Transfer Lasso [16]. In the TransLasso [11] and its GLM extension [17], two-stage estimation methods were proposed for the case of multiple source data, where the initial estimator is estimated using

both the source and target data. The Transfer Lasso [16], in contrast, is performed on target data using the initial estimator without the need for source data.

3 Asymptotic Properties for Adaptive Lasso and Transfer Lasso

We will perform asymptotic analysis based on the following general settings throughout this paper.

Assumption 3.1. Let $m \geq 0$ be an integer satisfying $n/m \rightarrow r_0 \geq 0$. The initial estimator $\tilde{\beta}$ is a \sqrt{m} -consistent estimator and $z := \sqrt{m}(\tilde{\beta} - \beta^*)$ converges to some distribution.

Assumption 3.1 implies that the initial estimator is estimated on source data of size m , and then the final estimator is estimated on target data of size n using the initial estimator. The case $m = n$ ($r_0 = 1$) corresponds to the existing results for the Adaptive Lasso, whereas $m \gg n$ ($r_0 = 0$) corresponds to the typical transfer learning setup. The source and target data are assumed to be independent of each other. We also make assumption 2.1 in our analysis.

We note that the initial estimator $\tilde{\beta}$ is *not* a fixed (deterministic) source parameter, but an estimator (random variable). This is the same as the previous studies. The case where $\tilde{\beta}$ is fixed is discussed in Appendix A.

3.1 Asymptotic Properties for Adaptive Lasso

We provide the property of the Adaptive Lasso for an initial estimator with source data of size m . It is straightforward to extend the oracle property for \sqrt{n} -consistent initial estimators (Lemma 2.9) to \sqrt{m} -consistent initial estimators (Lemma 3.2).

Lemma 3.2 (Oracle Property for Adaptive Lasso with Different Sample Size). *Suppose that $\tilde{\beta}$ is a \sqrt{m} -consistent estimator. If $\lambda_n/\sqrt{n} \rightarrow 0$ and $\lambda_n\sqrt{m^\gamma/n} \rightarrow \infty$, then the Adaptive Lasso estimator (3) satisfies the oracle property, that is, consistent variable selection and \sqrt{n} -consistency with asymptotic normality:*

$$(13) \quad \lim_{n \rightarrow \infty} P(\hat{S}_n^A = S) = 1,$$

$$(14) \quad \sqrt{n}(\hat{\beta}_S^A - \beta_S^*) \xrightarrow{d} \mathcal{N}(0, \sigma^2 C_{SS}^{-1}).$$

Proof. The proof is given in B.2.1. \square

Furthermore, we extensively analyze the convergence rate depending on the hyperparameter λ_n . We obtain Theorem 3.3 and Corollary 3.4.

Theorem 3.3 (Asymptotic Distribution for Adaptive Lasso). *We have the following asymptotic distributions for the Adaptive Lasso estimator (3).*

(i) *If $\sqrt{m^\gamma/n} \lambda_n \rightarrow \lambda_1 \geq 0$, then*

$$(15) \quad \sqrt{n}(\hat{\beta}_n^A - \beta^*) \xrightarrow{d} \operatorname{argmin}_u \left\{ u^\top Cu - 2u^\top W + \sum_{j \in S^c} \frac{\lambda_1}{|z_j|^\gamma} |u_j| \right\}.$$

(ii) *If $\sqrt{m^\gamma/n} \lambda_n \rightarrow \infty$ and $\lambda_n/\sqrt{n} \rightarrow \lambda_0 \geq 0$, then*

$$(16) \quad \sqrt{n}(\hat{\beta}_n^A - \beta^*) \xrightarrow{d} \operatorname{argmin}_{u \in \mathcal{U}} \left\{ u^\top Cu - 2u^\top W + \sum_{j \in S} \lambda_0 \frac{\operatorname{sgn}(\beta_j^*)}{|\beta_j^*|^\gamma} u_j \right\}, \quad \mathcal{U} := \{u \mid u_{S^c} = 0\}.$$

(iii) *If $\lambda_n/\sqrt{n} \rightarrow \infty$ and $\lambda_n/n \rightarrow 0$, then*

$$(17) \quad \frac{n}{\lambda_n}(\hat{\beta}_n^A - \beta^*) \xrightarrow{d} \operatorname{argmin}_{u \in \mathcal{U}} \left\{ u^\top Cu + \sum_{j \in S} \frac{\operatorname{sgn}(\beta_j^*)}{|\beta_j^*|^\gamma} u_j \right\}, \quad \mathcal{U} := \{u \mid u_{S^c} = 0\}.$$

Proof. This is a special case of Theorem 4.1 and the proof is the same as B.3.1. \square

Corollary 3.4 (Convergence Rate for Adaptive Lasso). *We have the following convergence rates for the Adaptive Lasso estimator (3).*

- (i) *If $\sqrt{m^\gamma/n} \lambda_n \rightarrow \lambda_1 \geq 0$, then the convergence rate is \sqrt{n} .*
- (ii) *If $\sqrt{m^\gamma/n} \lambda_n \rightarrow \infty$ and $\lambda_n/\sqrt{n} \rightarrow \lambda_0 \geq 0$, then convergence rate is \sqrt{n} .*
- (iii) *If $\lambda_n/\sqrt{n} \rightarrow \infty$ and $\lambda_n/n \rightarrow 0$, then the convergence rate is n/λ_n , which is slower than \sqrt{n} .*

Lemma 3.2 shows that the oracle property still holds for \sqrt{m} -consistent estimators, $\lambda_n/\sqrt{n} \rightarrow 0$, and $\lambda_n\sqrt{m^\gamma/n} \rightarrow \infty$. In addition, Theorem 3.3 and Corollary 3.4 show that the convergence rate of the Adaptive Lasso estimator

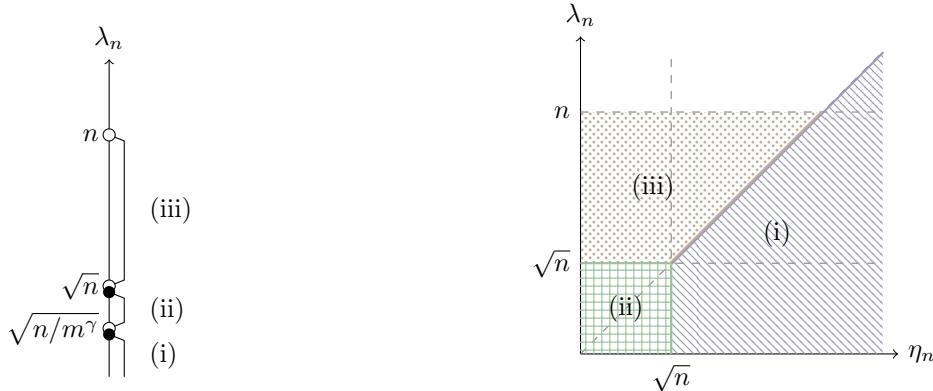


Figure 2: Phase diagrams with λ_n for the Adaptive Lasso in Lemma 3.2–Theorem 3.4 (left) and λ_n and η_n for the Transfer Lasso in Theorem 3.5–Theorem 3.11 (right). The Adaptive Lasso has \sqrt{n} -consistency in (i) and (ii) and active variable selection consistency in (ii), but the convergence rate in (iii) is slower than \sqrt{n} . The Transfer Lasso has convergence rates of \sqrt{m} , \sqrt{n} , and $n/\lambda_n (< \sqrt{n})$ for (i), (ii), and (iii) respectively. It has invariant variable selection consistency in (i) but does not have active variable selection consistency in (i) and (ii).

is equal to \sqrt{n} in the case (i) and (ii) and is less than \sqrt{n} in the case (iii). The condition of (ii) in Theorem 3.3 includes the condition of the oracle property in Lemma 3.2. Figure 2 (left) illustrates each hyperparameter region in Theorem 3.3 and Corollary 3.4.

These results imply both an advantage and a disadvantage. The advantage is that the initial estimator does not necessarily require \sqrt{n} -consistent. The Adaptive Lasso has the oracle property even when the source data is small compared to the target data ($m \lesssim n$) and the initial estimator is less than \sqrt{n} -consistent. The disadvantage of the Adaptive Lasso, however, is that it does not take full advantage even when the sample size of the source data is very large ($m \gg n$). This is because the convergence rate is equal to \sqrt{n} ($\ll \sqrt{m}$).

3.2 Asymptotic Properties for Transfer Lasso

Now we consider the asymptotic properties of the Transfer Lasso. The Transfer Lasso has two hyperparameters, λ_n and η_n , and various asymptotic properties appear depending on their values. We first obtain several asymptotic distributions in Theorem 3.5 and convergence rate in Corollary 3.6.

The illustration of the division of cases is shown in Figure 2.

Theorem 3.5 (Asymptotic distribution for Transfer Lasso). *We have the following asymptotic distributions for the Transfer Lasso estimator (4).*

(i) *If $\eta_n/\sqrt{n} \rightarrow \infty$ and $\lambda_n/\eta_n \rightarrow \rho_0$ with $0 \leq \rho_0 < 1$, then*

$$(18) \quad \sqrt{m}(\hat{\beta}_n^T - \beta^*) \xrightarrow{d} z.$$

(ii) *If $\lambda_n/\sqrt{n} \rightarrow \lambda_0 \geq 0$ and $\eta_n/\sqrt{n} \rightarrow \eta_0 \geq 0$, then*

(19)

$$\begin{aligned} & \sqrt{n}(\hat{\beta}_n^T - \beta^*) \\ & \xrightarrow{d} \underset{u}{\operatorname{argmin}} \left\{ u^\top Cu - 2u^\top W + \lambda_0 \left(\sum_{j \in S} u_j \operatorname{sgn}(\beta_j^*) + \sum_{j \in S^c} |u_j| \right) + \eta_0 \sum_{j=1}^p |u_j - \sqrt{r_0} z_j| \right\}. \end{aligned}$$

(iii) *If $\lambda_n/\sqrt{n} \rightarrow \infty$, $\lambda_n/n \rightarrow 0$, and $\eta_n/\lambda_n \rightarrow \rho'_0 \geq 0$, then*

(20)

$$\frac{n}{\lambda_n}(\hat{\beta}_n^T - \beta^*) \xrightarrow{d} \underset{u}{\operatorname{argmin}} \left\{ u^\top Cu + \sum_{j \in S} (u_j \operatorname{sgn}(\beta_j^*) + \rho'_0 |u_j|) + \sum_{j \in S^c} (1 + \rho'_0) |u_j| \right\}.$$

Proof. The proof is given in B.2.2 □

Corollary 3.6 (Convergence Rate for Transfer Lasso). *We have the following convergence rates for the Transfer Lasso estimator (4).*

- (i) *If $\eta_n/\sqrt{n} \rightarrow \infty$ and $\lambda_n/\eta_n \rightarrow \rho_0$ with $0 \leq \rho_0 < 1$, then the convergence rate is \sqrt{m} .*
- (ii) *If $\lambda_n/\sqrt{n} \rightarrow \lambda_0 \geq 0$ and $\eta_n/\sqrt{n} \rightarrow \eta_0 \geq 0$, then the convergence rate is \sqrt{n} .*
- (iii) *If $\lambda_n/\sqrt{n} \rightarrow \infty$, $\lambda_n/n \rightarrow 0$, and $\eta_n/\lambda_n \rightarrow \rho'_0$ with $0 \leq \rho'_0 < 1$, then the convergence rate is n/λ_n , which is slower than \sqrt{n} . On the other hand, if $\rho'_0 \geq 1$, then the convergence rate is faster than n/λ_n .*

Theorem 3.5 and Corollary 3.6 show that the Transfer Lasso estimators achieve a convergence rate of \sqrt{m} in the case (i). This is beneficial when source data is large ($m \gg n$) and is an advantage for the Transfer Lasso over the Adaptive Lasso.

Next, we provide the results of variable selection consistency. We first define two types of variable selection consistency.

Definition 3.7 (Active Variable Selection Consistency). We say that an estimator exhibits *consistent active variable selection* when it estimates the true active variable to be nonzero and the true inactive variable to be zero, that is,

$$(21) \quad P(\hat{S}_n = S) \rightarrow 1.$$

Conversely, we say that an estimator is an *inconsistent active variable selection* when this is not the case, that is,

$$(22) \quad \limsup_{n \rightarrow \infty} P(\hat{S}_n = S) \leq c < 1.$$

Definition 3.8 (Invariant Variable Selection Consistency). We say that an estimator exhibits *consistent invariant variable selection* when the true active variable remains invariant from the initial estimator, that is,

$$(23) \quad P(\hat{\beta}_S^T = \tilde{\beta}_S) \rightarrow 1.$$

The concepts of “active” and “invariant” variable selection consistency are distinct yet interconnected. “Active” variable selection consistency aligns with conventional variable selection consistency, propelled by the estimator’s sparsity. This property ensures the correct identification of non-zero variables. In contrast, “invariant” variable selection consistency is a unique feature of estimators like the Transfer Lasso, driven by the sparsity of the change from the initial estimator. This property ensures that the estimation of the true active variables remains unchanged and inherits the accuracy of the initial estimator. This could be an advantage when the initial estimator is sufficiently accurate. When both active and invariant variable selection consistencies are satisfied, the estimator effectively zeroes out the true inactive elements while the true active elements align with the initial estimator’s values. Consequently, sparsity is attained both in the estimator and in its change.

We present results on active/invariant variable selection consistency for the Transfer Lasso in Theorems 3.9, 3.10, and 3.11. We assume that the initial estimator $\tilde{\beta}$ may not exhibit consistent active variable selection in our variable selection analysis.

Theorem 3.9 (Inconsistent Active Variable Selection for Transfer Lasso). *Suppose that $\tilde{\beta}$ is inconsistent with active variable selection. For the cases (i) and (ii) in Theorem 3.5, the Transfer Lasso estimator (4) yields inconsistent active variable selection, that is,*

$$(24) \quad \limsup_{n \rightarrow \infty} P(\hat{S}_n^T = S) \leq c < 1,$$

where c is a constant.

Proof. The proof is given in B.2.3. \square

Theorem 3.10 (Consistent Invariant Variable Selection for Transfer Lasso). *Suppose that $\tilde{\beta}$ is inconsistent with active variable selection. For the case (i) in Theorem 3.5, the Transfer Lasso estimator (4) yields consistent invariant variable selection, that is,*

$$(25) \quad P(\hat{\beta}_S^T = \tilde{\beta}_S) \rightarrow 1.$$

Proof. The proof is given in B.2.4. \square

Theorem 3.11 (Inconsistent Invariant Variable Selection for Transfer Lasso). *Suppose that $\tilde{\beta}$ is inconsistent with active variable selection. For the case (ii) in Theorem 3.5, the Transfer Lasso estimators (4) yield inconsistent invariant variable selection, that is,*

$$(26) \quad \limsup_{n \rightarrow \infty} P(\hat{\beta}_S^T = \tilde{\beta}_S) \leq c < 1.$$

where c is a constant.

Proof. The proof is given in B.2.5. \square

Theorems 3.9, 3.10, and 3.11 unveil the benefits and drawbacks of the Transfer Lasso. Theorem 3.9 implies that the \sqrt{m} -consistent region (i) does not hold active variable selection consistency. The \sqrt{n} -consistent region (ii) does not hold as well. This is a disadvantage for the Transfer Lasso. On the other hand, Theorem 3.10 indicates that the Transfer Lasso in the case (i) has a property of consistent invariant variable selection, which the Adaptive Lasso does not have. Theorem 3.11 implies that the estimators are inconsistent in terms of invariant variable selection in the case (ii).

As shown in Figure 2, the Transfer Lasso cannot simultaneously achieve \sqrt{m} -consistency and consistent active/invariant variable selection in the regions (i), (ii), and (iii). This is why we explore a new methodology in the next section. We note that in regions other than (i), (ii), and (iii) (e.g., boundary regions), the asymptotic property is unclear. Appendix A.3 contains additional results for boundary regions. At the very least, the above results imply that \sqrt{m} -consistency and consistent active/invariant variable selection are incompatible in most regions for the Transfer Lasso.

4 Beyond Adaptive Lasso and Transfer Lasso

The Adaptive Lasso and the Transfer Lasso have their advantages and disadvantages, as seen in the previous section. The Adaptive Lasso achieves both \sqrt{n} -consistency and consistent variable selection for $m \leq n$, but its convergence rate is $\sqrt{n}(\ll \sqrt{m})$ for $m \gg n$. The Transfer Lasso, on the other hand, achieves a convergence rate of \sqrt{m} for $m \gg n$, but it results in inconsistent variable selection. Are there any ways to combine their benefits and compensate for their drawbacks?

4.1 Adaptive Transfer Lasso: A Non-Trivial Integration

To exploit their benefits and compensate for their drawbacks, we integrate the ideas of the Adaptive Lasso and the Transfer Lasso. We propose a novel method using the initial estimator $\tilde{\beta}$ as

$$(27) \quad \hat{\beta}_n^\# = \underset{\beta}{\operatorname{argmin}} \left\{ \frac{1}{n} \|y - X\beta\|_2^2 + \frac{\lambda_n}{n} \sum_j v_j |\beta_j| + \frac{\eta_n}{n} \sum_j w_j |\beta_j - \tilde{\beta}_j| \right\},$$

$$(28) \quad v_j := \frac{1}{|\tilde{\beta}_j|^{\gamma_1}}, \quad w_j := |\tilde{\beta}_j|^{\gamma_2},$$

where $\gamma_1 \geq 0$ and $\gamma_2 \geq 0$ are new hyperparameters. We denote $\hat{S}_n^\# := \{j : \hat{\beta}_j^\# \neq 0\}$. The weight $v_j = 1/|\tilde{\beta}_j|^{\gamma_1}$ is the same as that of the Adaptive Lasso, whereas the term $w_j = |\tilde{\beta}_j|^{\gamma_2}$ is a new non-trivial part. Because $w_j \rightarrow 0$ as $\tilde{\beta}_j \rightarrow 0$, the effect of transfer learning from the initial estimator disappears for inactive parameters. We call this method *Adaptive Transfer Lasso* because it is a generalization of the Adaptive Lasso and the Transfer Lasso. Indeed, if $\eta_n = 0$, then it reduces to the Adaptive Lasso, and if $\gamma_1 = \gamma_2 = 0$, then it reduces to the Transfer Lasso.

4.2 Asymptotic Properties for Adaptive Transfer Lasso

We present the asymptotic properties of the Adaptive Transfer Lasso. The assumptions are the same as for the Adaptive Lasso and the Transfer Lasso. To derive the asymptotic distribution and convergence rate, we need a more detailed case analysis than before. The illustration of the division of cases is shown in Figure 3.

Theorem 4.1 (Asymptotic Distribution for Adaptive Transfer Lasso). *We have the following asymptotic distributions for the Adaptive Transfer Lasso*

estimator (27).

(i) If $\eta_n/\sqrt{nm^{\gamma_2}} \rightarrow \infty$ and $\eta_n/\sqrt{m^{\gamma_1+\gamma_2}}\lambda_n \rightarrow \infty$, then

$$(29) \quad \sqrt{m}(\hat{\beta}_n^\# - \beta^*) \xrightarrow{d} z.$$

(ii) If $\sqrt{m^{\gamma_1}/n}\lambda_n \rightarrow \infty$, $\eta_n/\sqrt{n} \rightarrow \infty$, $\eta_n/\lambda_n \rightarrow \infty$, $\eta_n/\sqrt{m^{\gamma_1+\gamma_2}}\lambda_n \rightarrow 0$, and $\sqrt{m^{\gamma_1}}\lambda_n/\eta_n \rightarrow \rho_0 \geq 0$, then

$$(30) \quad \sqrt{m}(\hat{\beta}_n^\# - \beta^*) \xrightarrow{d} \begin{cases} 0 & \text{for } j \in S^c, \\ z_j & \text{for } j \in S. \end{cases}$$

(iii) If $\sqrt{m^{\gamma_1}/n}\lambda_n \rightarrow \lambda_1 \geq 0$ and $\eta_n/\sqrt{n} \rightarrow \eta_0 \geq 0$, then

$$(31)$$

$$(32) \quad \begin{aligned} & \sqrt{n}(\hat{\beta}_n^\# - \beta^*) \\ & \xrightarrow{d} \underset{u}{\operatorname{argmin}} \left\{ u^\top Cu - 2u^\top W + \sum_{j \in S^c} \frac{\lambda_1}{|z_j|^{\gamma_1}} |u_j| + \sum_{j \in S} \eta_0 |\beta_j^*|^{\gamma_2} |u_j - \sqrt{r_0}z_j| \right\}. \end{aligned}$$

(iv) If $\sqrt{m^{\gamma_1}/n}\lambda_n \rightarrow \lambda_1 \geq 0$, $\eta_n/\sqrt{n} \rightarrow \infty$, and $\eta_n/\sqrt{nm^{\gamma_2}} \rightarrow \eta_1 \geq 0$, then

$$(33)$$

$$(34) \quad \begin{aligned} & \sqrt{n}(\hat{\beta}_n^\# - \beta^*) \\ & \xrightarrow{d} \underset{u \in \mathcal{U}}{\operatorname{argmin}} \left\{ u^\top Cu - 2u^\top W + \sum_{j \in S^c} \left(\frac{\lambda_1}{|z_j|^{\gamma_1}} |u_j| + \eta_1 |z_j|^{\gamma_2} |u_j - \sqrt{r_0}z_j| \right) \right\}, \end{aligned}$$

$$(35)$$

$$\mathcal{U} := \{u \mid u_S = r_0 z_S\}.$$

(v) If $\sqrt{m^{\gamma_1}/n} \lambda_n \rightarrow \infty$, $\lambda_n/\sqrt{n} \rightarrow \lambda_0 \geq 0$, and $\eta_n/\sqrt{n} \rightarrow \eta_0 \geq 0$, then

(36)

$$\begin{aligned} & \sqrt{n}(\hat{\beta}_n^\# - \beta^*) \\ (37) \quad & \xrightarrow{d} \operatorname{argmin}_{u \in \mathcal{U}} \left\{ u^\top Cu - 2u^\top W + \sum_{j \in S} \left(\lambda_0 \frac{\operatorname{sgn}(\beta_j^*)}{|\beta_j^*|^{\gamma_1}} u_j + \eta_0 |\beta_j^*|^{\gamma_2} |u_j - \sqrt{r_0} z_j| \right) \right\}, \\ (38) \quad & \mathcal{U} := \{u \mid u_{S^c} = 0\}. \end{aligned}$$

(vi) If $\lambda_n/\sqrt{n} \rightarrow \infty$, $\lambda_n/n \rightarrow 0$, and $\lambda_n/\eta_n \rightarrow \infty$, then

(39)

$$\frac{n}{\lambda_n}(\hat{\beta}_n^\# - \beta^*) \xrightarrow{d} \operatorname{argmin}_{u \in \mathcal{U}} \left\{ u^\top Cu + \sum_{j \in S} \frac{\operatorname{sgn}(\beta_j^*)}{|\beta_j^*|^{\gamma_1}} u_j \right\}, \quad \mathcal{U} := \{u \mid u_{S^c} = 0\}.$$

Proof. The proof is given in B.3.1 \square

Corollary 4.2 (Convergence Rate for Adaptive Transfer Lasso). *We have the following convergence rates for the Adaptive Transfer Lasso estimator (27).*

- (i) $\eta_n/\sqrt{nm^{\gamma_2}} \rightarrow \infty$ and $\eta_n/\sqrt{m^{\gamma_1+\gamma_2}}\lambda_n \rightarrow \infty$, then the convergence rate is \sqrt{m} .
- (ii) $\sqrt{m^{\gamma_1}/n} \lambda_n \rightarrow \infty$, $\eta_n/\sqrt{n} \rightarrow \infty$, $\eta_n/\lambda_n \rightarrow \infty$, and $\eta_n/\sqrt{m^{\gamma_1+\gamma_2}}\lambda_n \rightarrow 0$, then the convergence rate is \sqrt{m} .
- (iii) If $\sqrt{m^{\gamma_1}/n} \lambda_n \rightarrow \lambda_1 \geq 0$ and $\eta_n/\sqrt{n} \rightarrow \eta_0 \geq 0$, then the convergence rate is \sqrt{n} .
- (iv) If $\sqrt{m^{\gamma_1}/n} \lambda_n \rightarrow \lambda_1 \geq 0$, $\eta_n/\sqrt{n} \rightarrow \infty$, and $\eta_n/\sqrt{nm^{\gamma_2}} \rightarrow \eta_1 \geq 0$, then the convergence rate is \sqrt{n} .
- (v) If $\sqrt{m^{\gamma_1}/n} \lambda_n \rightarrow \infty$, $\lambda_n/\sqrt{n} \rightarrow \lambda_0 \geq 0$, and $\eta_n/\sqrt{n} \rightarrow \eta_0 \geq 0$, then the convergence rate is \sqrt{n} .
- (vi) If $\lambda_n/\sqrt{n} \rightarrow \infty$, $\lambda_n/n \rightarrow 0$, and $\lambda_n/\eta_n \rightarrow \infty$, then the convergence rate is n/λ_n , which is slower than \sqrt{n} .

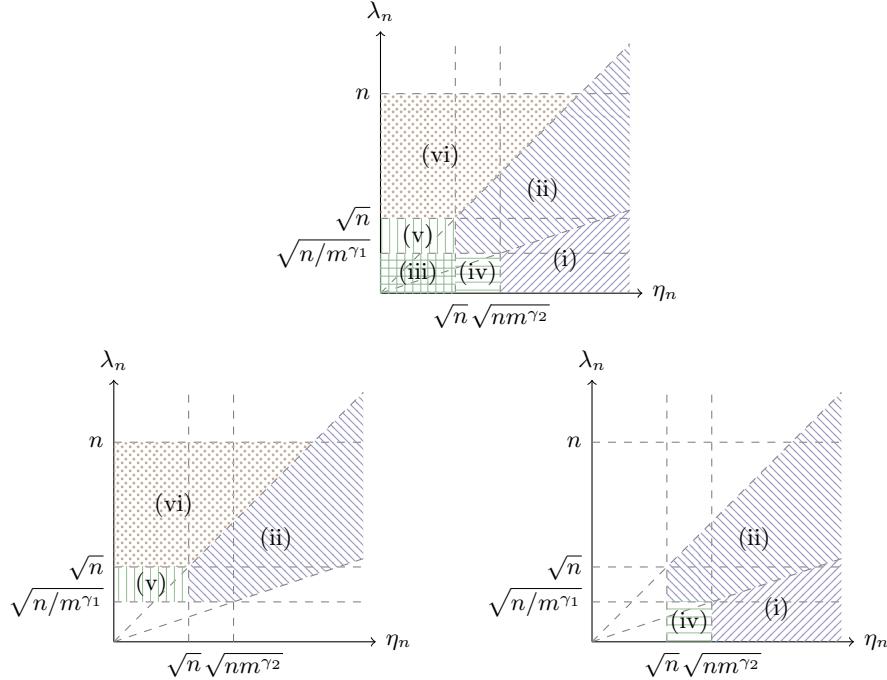


Figure 3: Phase diagrams of convergence rate (top) and active/invariant variable selection (bottom left/right) with λ_n and η_n for the Adaptive Transfer Lasso in Theorems 4.1, 4.3, 4.4, and Corollary 4.2. They are \sqrt{m} -consistent in (i - ii), \sqrt{n} -consistent in (iii - v), and sub- \sqrt{n} -consistent in (vi). They yield consistent active variable selection in (ii), (v), and (vi) (left), while consistent invariant variable selection in (i), (ii), and (iv) (right). Estimators in (ii) satisfy \sqrt{m} -consistency and active/invariant variable selection consistency.

Theorem 4.1 and Corollary 4.2 show that the Adaptive Transfer Lasso achieves a convergence rate of \sqrt{m} in the case (i) and (ii). This property is inherited from the Transfer Lasso. The asymptotic distribution for (i) is the same as the initial estimator. On the other hand, the asymptotic distribution for (ii) is remarkable. The distribution is the same as the initial estimator for the active variables, whereas is zero for the inactive variables. This implies that inactive parameters shrink to zero quickly.

We also provide the results of active/invariant variable selection consistency for the Adaptive Transfer Lasso.

Theorem 4.3 (Consistent Active Variable Selection for Adaptive Trans-

fer Lasso). For the cases (ii), (v), and (vi) in Theorem 4.1, the Adaptive Transfer Lasso yields consistent active variable selection, that is,

$$(40) \quad P(\hat{S}_n^\# = S) \rightarrow 1.$$

Proof. The proof is given in B.3.2. \square

Theorem 4.4 (Consistent Invariant Variable Selection for Adaptive Transfer Lasso). For the cases (i), (ii), and (iv) in Theorem 4.1, the Adaptive Transfer Lasso yields consistent invariant variable selection, that is,

$$(41) \quad P(\hat{\beta}_S = \tilde{\beta}_S) \rightarrow 1.$$

Proof. The proof is given in B.3.3. \square

Theorems 4.3 and 4.4 imply that both active/invariant variable selection consistency hold in the case (ii). Hence, we have the following corollary.

Corollary 4.5 (Oracle Region for Adaptive Transfer Lasso). For the case (ii) in Theorem 4.1, the Adaptive Transfer Lasso estimator satisfies

- \sqrt{m} -consistent: $\sqrt{m}(\hat{\beta}_n^\# - \beta^*)$ converges to some distributions,
- consistent active variable selection: $\hat{S}_n^\# = S$ with probability tending to 1,
- consistent invariant variable selection: $\hat{\beta}_S = \tilde{\beta}_S$ with probability tending to 1.

Corollary 4.5 shows that the Adaptive Transfer Lasso incorporates the advantages of both the Adaptive Lasso and the Transfer Lasso. The hyperparameters γ_1 and γ_2 play a crucial role in this property. If $\gamma_1 = \gamma_2 = 0$, then the region (ii) disappears and it reduces to the Transfer Lasso. If either γ_1 or γ_2 is positive, then the region (ii) appears and it holds \sqrt{m} -consistency and active/invariant variable selection consistency. Both γ_1 and γ_2 contribute to expanding the region (ii). One possible advantage of using $\gamma_2 > 0$ compared to $\gamma_1 > 0$ is that it is stable since there is no division by zero even when the initial estimator is sparse and the values are exactly zero.

Figure 3 are the phase diagrams that demonstrate the relation between hyperparameters (λ_n, η_n) and their asymptotic properties for the Adaptive Transfer Lasso. We see that the region (ii) is the intersection of the part with \sqrt{m} -consistency and the part with active/invariant variable selection consistency. Such a region exists neither in the Adaptive Lasso nor in the Transfer Lasso.

5 Empirical Results

We first empirically validate the theoretical properties. We then compare the performance of various methods through extensive simulations. Appendix C provides additional experimental results. The codes are available at <https://github.com/tkdmah/trlasso>.

5.1 Empirical Validation of Theory

In this subsection, we empirically validate the theoretical results for the Transfer Lasso and the Adaptive Transfer Lasso.

We first evaluated the ℓ_2 norm of the estimation error with respect to sample size. Theoretically, the convergence rate is \sqrt{m} , \sqrt{n} , and so on, depending on the hyperparameters. Assuming the convergence rate is $l(n)$, we have $E[\log \|\hat{\beta} - \beta^*\|_2] = \text{const.} - \log l(n)$, since $l(n)\|\hat{\beta} - \beta^*\|_2$ converges to some distribution. Therefore, by drawing a graph with $E[\log \|\hat{\beta} - \beta^*\|_2]$ on the vertical axis and $\log n$ on the horizontal axis, the convergence rate can be empirically calculated from its slope. Assuming $m = n^2$, the slope is $-1/2$ when \sqrt{n} -consistent, and -1 when \sqrt{m} -consistent.

We generated data by $y_i = x_i^\top \beta^* + \varepsilon_i$ ($i = 1, \dots, n$) where $x_i (\in \mathbb{R}^{10}) \stackrel{i.i.d.}{\sim} \mathcal{N}(0, \Sigma)$, $\Sigma_{jk} = 0.5^{|j-k|}$, $\varepsilon_i \stackrel{i.i.d.}{\sim} \mathcal{N}(0, \sigma^2)$, $\sigma = 1$, and $\beta^* = [3, 1.5, 0, 0, 2, 0, 0, \dots, 0]^\top (\in \mathbb{R}^{10})$ (as in [21]). We generated source data of size m and target data of size n with $m = n^2$ and $n = 20, 50, 100, 200, 500, 1000, 2000, 5000$. The initial estimators were obtained by the ordinary least squares using source data. The hyperparameters for each method were determined as follows according to Figures 1, 2, and 3.

- Lasso: $\lambda_n = n^{1/4}$ (i) and $\lambda_n = n^{3/4}$ (ii).
- Adaptive Lasso: $\gamma = 1$. $\lambda_n = n^{-1}$ (i), $n^{1/4}$ (ii), and $n^{3/4}$ (iii).
- Transfer Lasso: $(\lambda_n, \eta_n) = (n^{1/2}, n^{3/4})$ (i), $(n^{1/4}, n^{1/4})$ (ii), and $(n^{3/4}, n^{1/2})$ (iii).
- Adaptive Transfer Lasso: $\gamma_1 = \gamma_2 = 1$. $(\lambda_n, \eta_n) = (n^{-1/2}, n^2)$ (i), $(n^{1/2}, n^{3/2})$ (ii), $(n^{-1}, n^{1/4})$ (iii), (n^{-1}, n) (iv), $(1, n^{1/4})$ (v), and $(n^{3/4}, n^{1/2})$ (vi).

We performed each experiment ten times and evaluated their averages and standard errors.

Figure 4 shows the ℓ_2 estimation errors for the Lasso, the Adaptive Lasso, the Transfer Lasso, and the Adaptive Transfer Lasso with respect

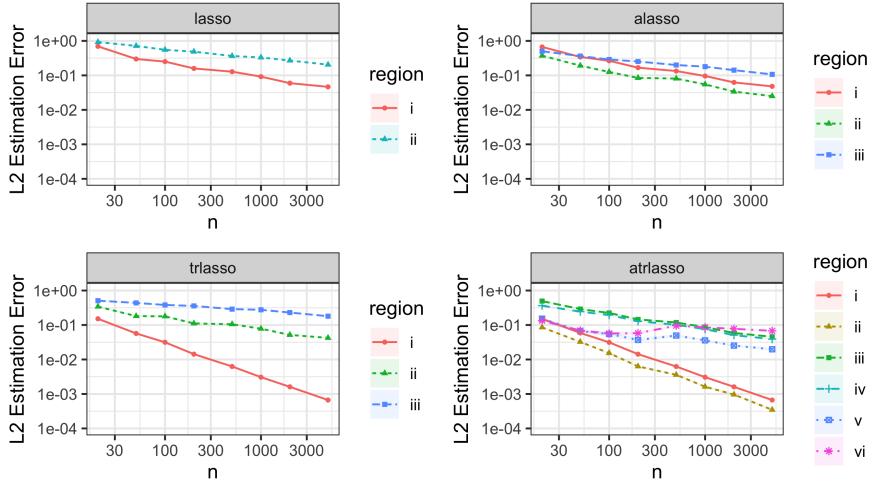


Figure 4: ℓ_2 estimation errors for the Lasso (top left), the Adaptive Lasso (top right), the Transfer Lasso (bottom left), and the Adaptive Transfer Lasso (bottom right) with respect to sample size. The convergence rates of the Transfer Lasso in the region (i) and the Adaptive Transfer Lasso in the region (i) and (ii) are \sqrt{m} (the slopes are -1), whereas the others are \sqrt{n} or less (the slope are $-1/2$ or greater).

to sample size. The slopes of the Transfer Lasso in the region (i) and the Adaptive Transfer Lasso in the region (i) and (ii) are -1 , indicating that the convergence rate was $n = \sqrt{m}$. For the other methods or regions, the slopes are -0.5 or greater, which confirms that the convergence rate is \sqrt{n} or less. These results were fully consistent with Theorems 3.5, 4.1, and 4.3.

We can observe two potential advantages of Adaptive Transfer Lasso. First, although the convergence rate (for $n \geq 500$) is \sqrt{n} in regions (v) and (vi), the estimation error is on the line of the convergence rate \sqrt{m} for $n < 500$. In other words, even in regions where the convergence rate is \sqrt{n} , the estimation error can be reduced when the sample size is small. Second, the estimation error is consistently smaller in the region (ii) than in the region (i), although the convergence rates are comparable between the two regions. This might be because the estimator in (i) is more likely to be perfectly matched to the initial estimator, whereas the estimator in (ii) is more likely to be matched to the initial estimator for active variables, but not for the inactive variables, and is more likely to be zero.

Having found that the convergence rate can be empirically evaluated

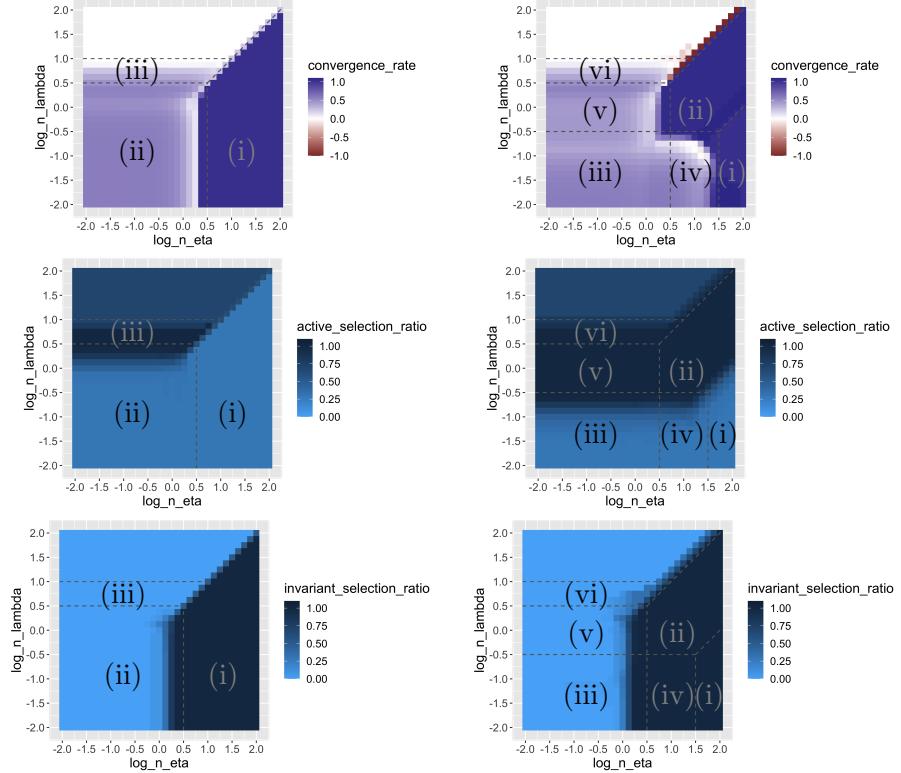


Figure 5: log–log phase diagrams of convergence rate (top), active variable selection ratio (middle), and invariant variable selection ratio (bottom) for the Transfer Lasso (left) and the Adaptive Transfer Lasso (right). These empirical results confirm the theoretical results of Figure 2 and 3.

accurately, we next empirically drew phase diagrams for the Transfer Lasso and the Adaptive Transfer Lasso as in Figure 3. The experimental setup was the same as in the previous subsection and $m = n^2$. The hyperparameters λ_n and η_n were set to n^δ with $\delta = -2, -1.75, -1.5, \dots, 1.75, 2$, respectively. The convergence rates were calculated from the slopes of the ℓ_2 errors for $n = 1000$ and $n = 5000$. We plotted the exponential parts of n in the convergence rates, taking the value 1 if \sqrt{m} -consistent and 0.5 if \sqrt{n} -consistent. Active variable selection consistency was evaluated as the ratio of correctly estimated zeros/non-zeros among all variables for $n = 5000$. Invariant variable selection consistency was evaluated by the ratio of variables that did not change from the initial estimator among the active variables for $n = 5000$.

Figure 5 illustrates the empirical phase diagrams of log–log scale for the Transfer Lasso and the Adaptive Transfer Lasso. As Theorems 3.5–3.11 suggest, the Transfer Lasso achieves both \sqrt{m} -consistency and invariant variable selection consistency in the lower right region (i), but does not have active variable selection consistency. Other regions also do not satisfy these properties simultaneously. On the other hand, in the Adaptive Transfer Lasso, the upper right region (ii) satisfies the properties of \sqrt{m} -consistency and active/invariant variable selection consistency. The empirical convergence rates and active/invariant variable selection ratios well reproduce Theorems 4.1, 4.3, and 4.4 in other regions as well. These empirical results confirm the theoretical results (Theorems 3.5–4.4, Figures 1–3).

5.2 Empirical Comparison of Methods

In this subsection, we compare the methods in various experimental settings based on hyperparameter determination by cross-validation. The experimental settings include various source/target data sample sizes, number of dimensions, signal-to-noise ratios, and initial estimators. We mainly considered two cases: one with a large amount of source data and the other with the same amount of source data as the target data.

First, we suppose that we have a large amount of source data and its sample size is $m = 10000$. The simulation setting follows the previous subsections. We used $\sigma = 1, 3, 6, 10$; $p = 10, 20, 50, 100$; and $n = 10, 20, 50, 100, 200, \dots, 5000, 10000$.

Initial estimators were obtained by the Lasso because the number of dimensions p can be greater than sample size n in this experiment. We compared other initial estimators including Ridge, Ridgeless [2, 8], and Lassoless [14, 12] in Appendix C.2. The search spaces were $\gamma = 0.5, 1, 2$ for the Adaptive Lasso; $\alpha := \lambda_n / (\lambda_n + \eta_n) = 0.75, 0.5, 0.25$ for the Transfer Lasso; and $(\gamma_1, \gamma_2) = (0.5, 0.5), (1, 1), (2, 2)$, and $\alpha := \lambda_n / (\lambda_n + \eta_n) = 0.75, 0.5, 0.25$ for the Adaptive Transfer Lasso. The hyperparameter λ_n was determined by 10-fold cross validation with $\lambda_{\min} / \lambda_{\max} = 10^{-6}$ where λ_{\max} is automatically determined by Theorem 4 in [16]. If $|\tilde{\beta}_j| \leq 10^{-3}$, then we set $|\tilde{\beta}_j| = 10^{-3}$ to avoid division by zero.

We evaluated the performance by two metrics: ℓ_2 norm for estimation error and F1 score for variable selection. The F1 score is a harmonic average of precision and recall, where precision = (the number of correct selected variables) / (the number of selected variables) and recall = (the number of correct selected variables) / (the number of true active variables). We used the F1 score because it allows us to evaluate the performance of variable

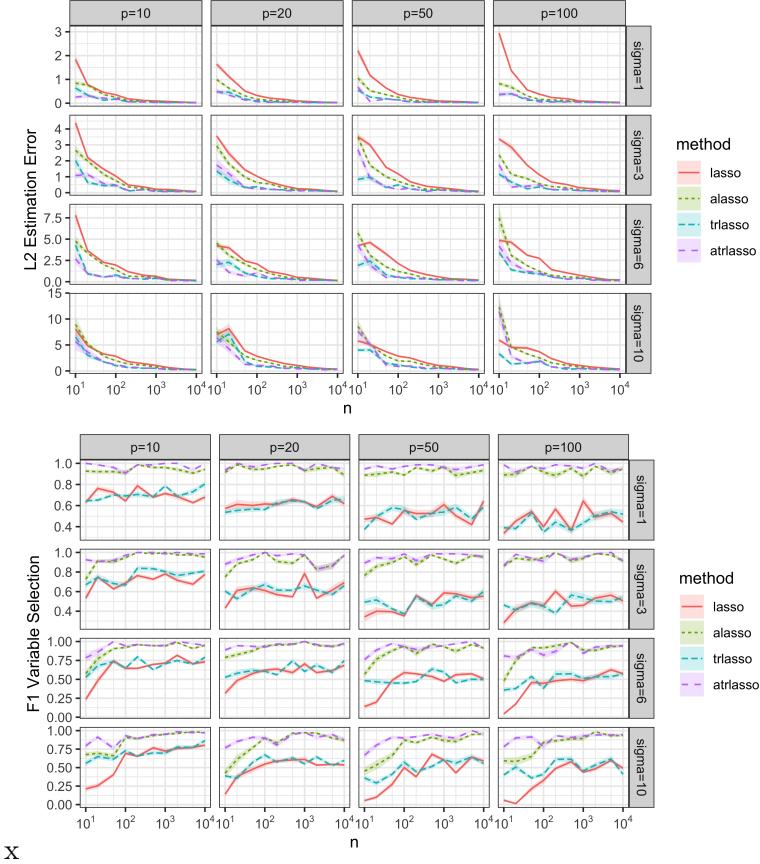


Figure 6: ℓ_2 estimation errors (top) and variable selection F1-score (bottom) for a large amount of source data.

selection even when there is an imbalance between the number of active and inactive variables. We also evaluated other metrics in Appendix C.3. They included RMSE for prediction evaluation and sensitivity, specificity, positive predictive value, and the number of active variables for feature selection evaluation.

The results are shown in Figure 6. In terms of estimation errors, the Transfer Lasso and the Adaptive Transfer Lasso outperformed the other methods. The Adaptive Lasso was superior to the Lasso, but it was inferior to the Transfer Lasso and the Adaptive Transfer Lasso. In terms of variable selection, the Adaptive Lasso and the Adaptive Transfer Lasso outperformed the others, and the Adaptive Transfer Lasso was slightly superior

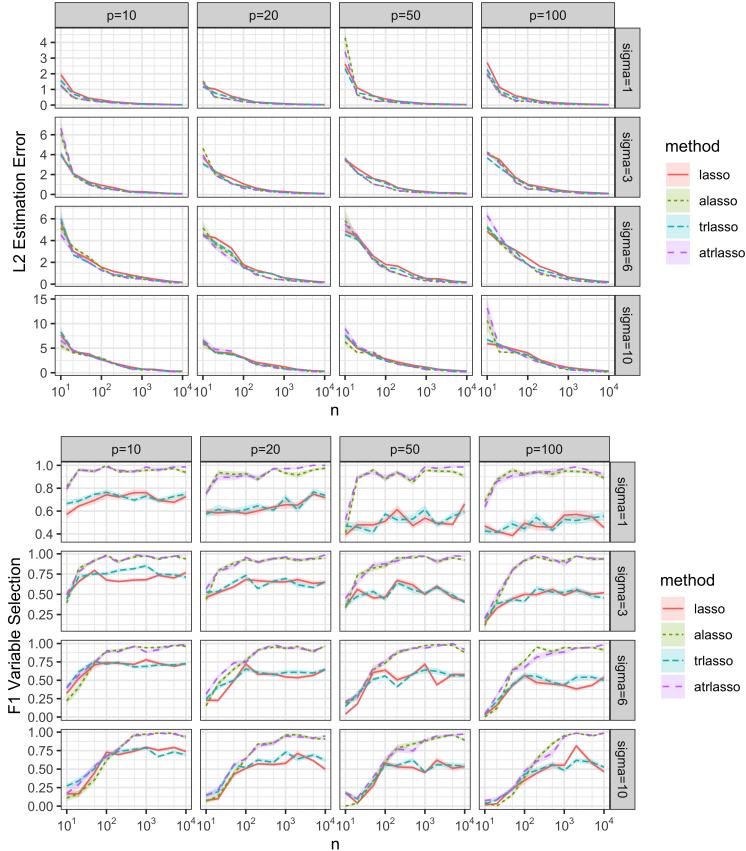


Figure 7: ℓ_2 estimation errors (top) and variable selection F1-scores (bottom) for a medium amount of source data.

to the Adaptive Lasso. These results imply the superiority of the Adaptive Transfer Lasso with initial estimators using large amounts of source data. The Adaptive Lasso, however, does not fully utilize the initial estimators in this setting.

Next, we supposed that we have a medium amount of source data and its sample size is the same as the target data. We used the same data generation process, comparison methods, and performance measurements as above.

The results are shown in Figure 7. All methods were comparable in terms of estimation errors, but in terms of variable selection, the Adaptive Lasso and the Adaptive Transfer Lasso were superior to those of the others. The Adaptive Lasso and the Adaptive Transfer Lasso had similar performances

for both estimation error and variable selection. This is consistent with our theoretical analyses.

6 Discussion

We discuss additional comparisons among methods from two perspectives; regularization contours and prior distributions. We also discuss future work.

6.1 Regularization Contours

The regularization contours help to intuitively capture the strength and pattern of regularization. Figure 8 shows their contours with an initial estimator with a small value $\tilde{\beta}_1 = 0.5$ and a large value $\tilde{\beta}_2 = 2$.

The contours of the Adaptive Lasso are pointed at the coordinate axes (where some elements are zero) and are especially sharp where the initial estimator is small. The contours of the Transfer Lasso are pointed at the points where some elements are zero or equal to the initial estimator, but they are not so sharp. The contours for the Adaptive Transfer Lasso are pointed where some elements are zero or equal to the initial estimator, and the sharpness varies depending on the hyperparameters. These observations indicate that the Adaptive Transfer Lasso flexibly changes the strength of regularization depending on the initial estimator.

6.2 Prior Distribution

In Bayesian perspectives, the Lasso regularization (2) can be seen as a negative log-likelihood of Laplace prior,

$$(42) \quad \lambda|\beta_j| = -\log P(\beta_j; \lambda) + \text{const.}, \quad P(z; \lambda) := \frac{\lambda}{2} \exp(-\lambda|z|).$$

A similar view is possible for the Adaptive Lasso, the Transfer Lasso, and the Adaptive Transfer Lasso. Most generally, the prior distribution of the Adaptive Transfer Lasso is given by

$$(43) \quad \lambda v_j |\beta_j| + \eta w_j |\beta_j - \tilde{\beta}_j| = -\log P(\beta_j; \lambda, \eta, v_j, w_j, \tilde{\beta}_j) + \text{const.},$$

$$(44) \quad P(\beta_j; \lambda, \eta, v_j, w_j, \tilde{\beta}_j) := \frac{1}{Z} \exp \left(-\lambda v_j |\beta_j| - \eta w_j |\beta_j - \tilde{\beta}_j| \right),$$

$$(45) \quad Z := \frac{2\lambda v_j}{\lambda^2 v_j^2 - \eta^2 w_j^2} \exp \left(-\eta w_j |\tilde{\beta}_j| \right) - \frac{2\eta w_j}{\lambda^2 v_j^2 - \eta^2 w_j^2} \exp \left(-\lambda v_j |\tilde{\beta}_j| \right).$$

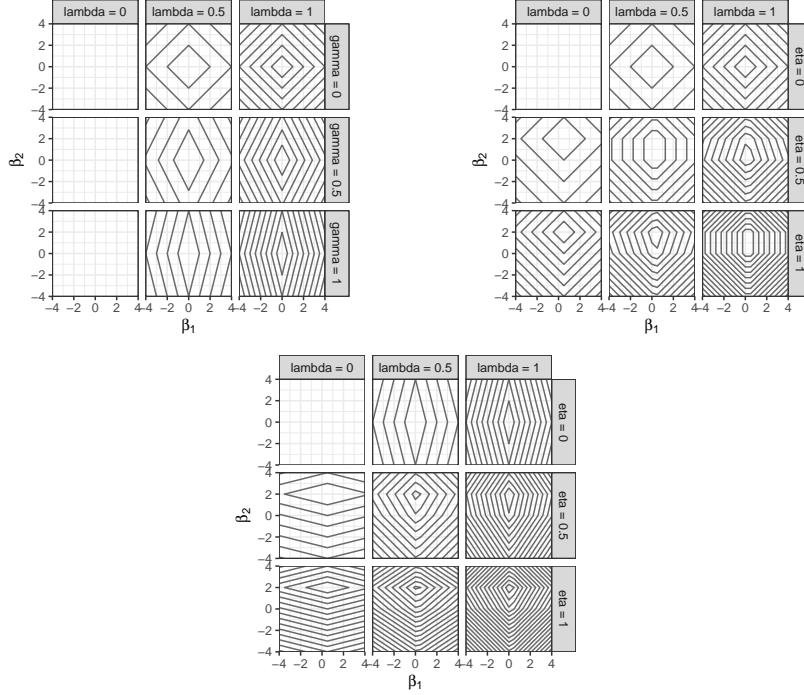


Figure 8: Regularization contours for the Adaptive Lasso (top left), the Transfer Lasso (top right), and the Adaptive Transfer Lasso (bottom) with initial estimators $\tilde{\beta} = [0.5, 2]^\top$. Hyperparameters are $\lambda_n = 0, 0.5, 1$ (from left to right); $\eta_n = 0, 0.5, 1$ (from top to bottom); and $\gamma_1 = \gamma_2 = 1$ for the Adaptive Transfer Lasso.

The prior distributions for the Adaptive Lasso, the Transfer Lasso, and the Adaptive Transfer Lasso are shown in Figures 9. The prior distributions for the Adaptive Lasso are all sharp at zero, and the distributions become steeper as the initial estimator decreases. This means that the Adaptive Lasso controls the variance of the prior distribution based on how close to zero the initial estimator is. The prior distribution for the Transfer Lasso is sharp at two points: zero and the initial estimator. When the initial estimator is small, it is nearly the same as that for the Lasso, but when the initial estimator is large, the sharpness changes depending on the magnitudes of λ and η . The prior distribution for the Adaptive Transfer Lasso is somewhat different from that of the Transfer Lasso. The prior distribution tends to peak at zero when the initial estimator is close to zero, whereas the prior distribution tends to peak at that initial estimator when the initial estimator

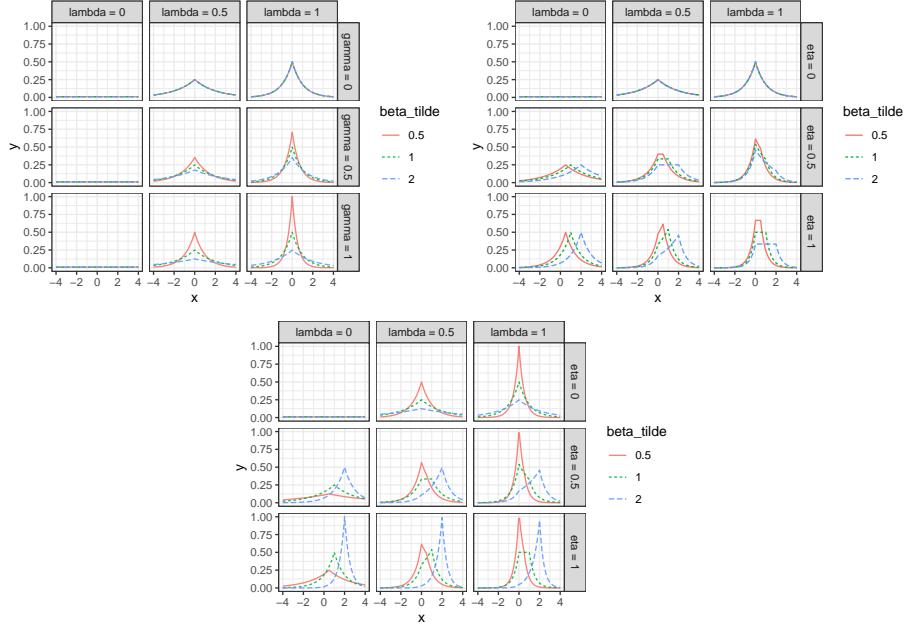


Figure 9: Prior distributions for the Adaptive Lasso (top left), the Transfer Lasso (top right), and the Adaptive Transfer Lasso (bottom) with various initial estimators. Hyperparameters are $\gamma_1 = \gamma_2 = 1$ for the Adaptive Transfer Lasso.

is far from zero. This suggests that the Adaptive Transfer Lasso can make full use of the information from the initial estimator and achieves accurate active and varying variable selection.

6.3 Future Work

In our asymptotic analysis, we considered the case where p is fixed and n is infinitely divergent. The oracle property of Adaptive Lasso [21] can be extended to the case for $p \gg n$ by high-dimensional asymptotic theory [10], under different kinds of assumptions. As future research, it would be interesting to see whether this can be extended to the Transfer Lasso and the Adaptive Transfer Lasso.

In addition, we assumed that the initial estimator is consistent in our asymptotic analysis. When the initial estimator is incorrectly specified, performance deteriorates significantly for the Adaptive Lasso, but not so much for the Transfer Lasso. It would be interesting to theoretically verify

this property.

7 Conclusion

The Adaptive Lasso and the Transfer Lasso are similar but have their advantages and disadvantages from the viewpoint of an asymptotic perspective. We proposed the Adaptive Transfer Lasso, which has advantages over the Adaptive Lasso and the Transfer Lasso. We confirmed it in numerical simulations.

Acknowledgment

The research is the collaborative work of Toshiba Corporation and The Institute of Statistical Mathematics, based on funding from Toshiba Corporation.

References

- [1] Hamsa Bastani. Predicting with proxies: Transfer learning in high dimension. *Management Science*, 67(5):2964–2984, 2021.
- [2] Mikhail Belkin, Daniel Hsu, and Ji Xu. Two models of double descent for weak features. *SIAM Journal on Mathematics of Data Science*, 2(4):1167–1180, 2020.
- [3] Evgenii Chzhen, Mohamed Hebiri, and Joseph Salmon. On lasso refitting strategies. *Bernoulli*, 25(4A):3175–3200, 2019.
- [4] Jianqing Fan and Runze Li. Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American statistical Association*, 96(456):1348–1360, 2001.
- [5] Wenjiang Fu and Keith Knight. Asymptotics for lasso-type estimators. *The Annals of statistics*, 28(5):1356–1378, 2000.
- [6] Charles J Geyer. On the asymptotics of constrained m-estimation. *The Annals of statistics*, pages 1993–2010, 1994.
- [7] Charles J Geyer. On the asymptotics of convex stochastic optimization. *Unpublished manuscript*, 37, 1996.

- [8] Trevor Hastie, Andrea Montanari, Saharon Rosset, and Ryan J Tibshirani. Surprises in high-dimensional ridgeless least squares interpolation. *The Annals of Statistics*, 50(2):949–986, 2022.
- [9] Trevor Hastie, Robert Tibshirani, and Ryan Tibshirani. Best subset, forward stepwise or lasso? analysis and recommendations based on extensive comparisons. *Statistical Science*, 35(4):579–592, 2020.
- [10] Jian Huang, Shuangge Ma, and Cun-Hui Zhang. Adaptive lasso for sparse high-dimensional regression models. *Statistica Sinica*, pages 1603–1618, 2008.
- [11] Sai Li, T Tony Cai, Hongzhe Li, et al. Transfer learning for high-dimensional linear regression: Prediction, estimation and minimax optimality. *Journal of the Royal Statistical Society Series B*, 84(1):149–173, 2022.
- [12] Yue Li and Yuting Wei. Minimum ℓ_1 -norm interpolators: Precise asymptotics and multiple descent. *arXiv preprint arXiv:2110.09502*, 2021.
- [13] Nicolai Meinshausen. Relaxed lasso. *Computational Statistics & Data Analysis*, 52(1):374–393, 2007.
- [14] Partha P Mitra. Understanding overfitting peaks in generalization error: Analytical risk curves for ℓ_2 and ℓ_1 penalized interpolation. *arXiv preprint arXiv:1906.03667*, 2019.
- [15] David Pollard. Asymptotics for least absolute deviation regression estimators. *Econometric Theory*, 7(2):186–199, 1991.
- [16] Masaaki Takada and Hironori Fujisawa. Transfer learning via ℓ_1 regularization. *Advances in Neural Information Processing Systems*, 33:14266–14277, 2020.
- [17] Ye Tian and Yang Feng. Transfer learning under high-dimensional generalized linear models. *Journal of the American Statistical Association*, pages 1–14, 2022.
- [18] Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1):267–288, 1996.

- [19] C-H ZHANG. Nearly unbiased variable selection under minimax concave penalty. *The Annals of Statistics*, 38:894–942, 2010.
- [20] Peng Zhao and Bin Yu. On model selection consistency of lasso. *The Journal of Machine Learning Research*, 7:2541–2563, 2006.
- [21] Hui Zou. The adaptive lasso and its oracle properties. *Journal of the American statistical association*, 101(476):1418–1429, 2006.

Appendices

A Additional Asymptotic Properties

In this section, we describe additional theoretical results that are not described in the main text. In Sections A.1 and A.2, we describe the asymptotic properties of Adaptive Lasso and Transfer Lasso, respectively, when the source parameters are deterministic (fixed). In Section A.3, we discuss additional results for Transfer Lasso when the hyperparameters are at boundary values.

A.1 Adaptive Lasso with deterministic Source Parameter

To avoid zero division, we define Adaptive Lasso with a deterministic source parameter $\tilde{\beta}$ as

$$(46) \quad \hat{\beta}_n^A = \underset{\beta: \beta_j=0 \text{ for } \tilde{\beta}_j=0}{\operatorname{argmin}} \left\{ Z_n^A(\beta; \tilde{\beta}, \lambda_n, \gamma) := \frac{1}{n} \|y - X\beta\|_2^2 + \frac{\lambda_n}{n} \sum_{j: \tilde{\beta}_j \neq 0} w_j |\beta_j| \right\},$$

and $w_j := 1/|\tilde{\beta}_j|^\gamma$ for $\tilde{\beta}_j \neq 0$. By the definition (46) and Lemma 2.9, we can easily obtain Lemma A.1 and Corollary A.2.

Lemma A.1 (\sqrt{n} -consistency for the Adaptive Lasso with Deterministic Source Parameter). *Suppose that $\tilde{S} = S$. If $\lambda_n/\sqrt{n} \rightarrow \lambda_0 \geq 0$, then*

$$(47)$$

$$\begin{aligned} & \sqrt{n}(\hat{\beta}_{Sn}^A - \beta_S^*) \\ & \xrightarrow{d} \underset{u}{\operatorname{argmin}} \left\{ u_S^\top C_{SS} u_S - 2u_S^\top W_S + \lambda_0 \sum_{j: \tilde{\beta}_j \neq 0} w_j (u_j \operatorname{sgn}(\beta_j^*) I(\beta_j^* \neq 0) + |u_j| I(\beta_j^* = 0)) \right\}, \end{aligned}$$

where $W \sim \mathcal{N}(0, \sigma^2 C)$, and $\hat{\beta}_{S^c n}^A = 0$.

Corollary A.2 (Oracle Property for Adaptive Lasso with Deterministic Source Parameter). *Suppose that $\tilde{S} = S$. If $\lambda_n = o(\sqrt{n})$, then the Adaptive Lasso estimator satisfies the oracle property.*

Based on Lemma A.1 and Corollary A.2, the Adaptive Lasso can satisfy the oracle property even for deterministic source parameters. However, the source parameter must satisfy exact support recovery, which is very restrictive.

A.2 Transfer Lasso with deterministic Source Parameter

Now we consider the asymptotic properties of Transfer Lasso with deterministic source parameter. Let $T := \{j : \tilde{\beta}_j \neq \beta_j^*\}$. We can identify the range of its estimator (Theorem A.3) for a general condition. The inequality in the probability is not necessarily tight.

Theorem A.3 (Estimation Range for Transfer Lasso). *If $(\lambda_n + \eta_n)/\sqrt{n} \rightarrow \infty$ and $(\lambda_n - \eta_n)/\sqrt{n} \rightarrow 0$, then the transfer lasso estimates satisfy*

$$(48) \quad \lim_{n \rightarrow \infty} P \left(\min \left\{ 0, \tilde{\beta}_j \right\} \leq \hat{\beta}_j \leq \max \left\{ 0, \tilde{\beta}_j \right\} \right) = 1.$$

Proof. The proof is given in B.4.1. \square

To obtain an asymptotic distribution and consistent variable selection, we need to impose the condition that $\text{sgn}(\tilde{\beta}_j - \beta_j^*) = \text{sgn}(\beta_j^*)$ for $\forall j \in S \cap T$. This requires “over-estimation” for the initial estimator. We obtain Theorem A.4 under this condition.

Theorem A.4 (Asymptotic distribution and active/varying variable consistency with deterministic source parameters). *Suppose that $\text{sgn}(\tilde{\beta}_j - \beta_j^*) = \text{sgn}(\beta_j^*)$ for $\forall j \in S \cap T$. If $(\lambda_n + \eta_n)/\sqrt{n} \rightarrow \infty$ and $(\lambda_n - \eta_n)/\sqrt{n} \rightarrow 0$, then the Transfer Lasso estimator satisfies*

$$(49) \quad \sqrt{n}(\hat{\beta} - \beta^*) \xrightarrow{d} \underset{u \in \mathcal{U}}{\text{argmin}} \left\{ -2u^\top W + u^\top Cu \right\},$$

$$(50) \quad W \sim \mathcal{N}(0, \sigma^2 C), \quad \mathcal{U} := \left\{ u \in \mathbb{R}^p \mid \begin{array}{l} u_j = 0 \text{ for } \forall j \in S^c \cap T^c \\ u_j \tilde{\beta}_j \geq 0 \text{ for } \forall j \in S^c \cap T \\ u_j \beta_j^* \leq 0 \text{ for } \forall j \in S \cap T^c \end{array} \right\},$$

and

$$(51) \quad \begin{cases} \lim_{n \rightarrow \infty} P \left(0 < \hat{\beta}_j < \tilde{\beta}_j \text{ or } \tilde{\beta}_j < \hat{\beta}_j < 0 \right) = 1 & \text{for } \forall j \in (S \cap T) \\ \lim_{n \rightarrow \infty} P \left(0 < \hat{\beta}_j \leq \tilde{\beta}_j \text{ or } \tilde{\beta}_j \leq \hat{\beta}_j < 0 \right) = 1 & \text{for } \forall j \in (S \cap T^c) \\ \lim_{n \rightarrow \infty} P \left(0 \leq \hat{\beta}_j < \tilde{\beta}_j \text{ or } \tilde{\beta}_j < \hat{\beta}_j \leq 0 \right) = 1 & \text{for } \forall j \in (S^c \cap T) \\ \lim_{n \rightarrow \infty} P \left(\hat{\beta}_j = 0 \right) = 1 & \text{for } \forall j \in (S^c \cap T^c). \end{cases}$$

Proof. The proof is given in B.4.2. \square

Theorem A.4 shows that the asymptotic distribution in (49) and (50) is a truncated Gaussian-mixture distribution. Each distribution in mixtures must satisfy 1) Gaussian distribution for $j \in S \cap T$; 2) truncated Gaussian distribution truncated at zero, or delta distribution at zero for $j \in S^c \cap T$ and $j \in S \cap T^c$; and 3) delta distribution at zero for $j \in S^c \cap T^c$.

In addition, Theorem A.4 indicates that the true active variables ($j \in S$) and the true varying variables ($j \in T$) can be recovered asymptotically if the source parameters of both active and varying variables have the same sign as the true variables and have larger absolute values than the true variables.

Asymptotic normality (instead of a truncated normal mixture distribution) can be obtained under a more restrictive condition (Theorem A.5).

Theorem A.5 (Oracle property (asymptotic normality and active/varying variable selection consistency)). *Suppose that $S = T$, and $\text{sgn}(\tilde{\beta}_j - \beta_j^*) = \text{sgn}(\beta_j^*)$ for $\forall j \in S (= T)$. If $(\lambda_n + \eta_n)/\sqrt{n} \rightarrow \infty$ and $(\lambda_n - \eta_n)/\sqrt{n} \rightarrow 0$, then the Transfer Lasso estimates satisfy the oracle property, that is,*

$$(52) \quad \lim_{n \rightarrow \infty} P(\hat{S}_n = S) = \lim_{n \rightarrow \infty} P(\hat{T}_n = T) = 1,$$

$$(53) \quad \sqrt{n}(\hat{\beta}_S - \beta_S^*) \xrightarrow{d} \mathcal{N}(0, \sigma^2 C_{SS}^{-1}).$$

Proof. The proof is given in B.4.3. \square

Theorem A.5 is the oracle property for the Transfer Lasso. The Adaptive Lasso requires \sqrt{n} -consistency for the initial estimator. In contrast, the Transfer Lasso requires variable consistency, sign consistency, and “over-estimation” for the source parameters. This yields \sqrt{n} -consistency as well as variable selection consistency for both active and varying variables.

A.3 Transfer Lasso with Initial Estimator in Boundary Region

We have an additional result on the asymptotic property for Transfer Lasso with an initial estimator when the hyperparameters are at boundary values.

Theorem A.6. *Suppose that $\tilde{\beta}$ is a \sqrt{m} -consistent estimator and define $z := \sqrt{m}(\tilde{\beta}_m - \beta^*)$. Suppose that $n/m \rightarrow 0$. If $\lambda_n/\sqrt{n} \rightarrow \infty$, $\eta_n/\lambda_n \rightarrow 1$, and $(\lambda_n - \eta_n)/\sqrt{n} \rightarrow \delta_0$, then*

$$(54) \quad \sqrt{n} \left(\hat{\beta}_n^T - \beta^* \right) \xrightarrow{d} \underset{u \in \mathcal{U}}{\operatorname{argmin}} \left\{ u_S^\top C_{SS} u_S - 2u_S^\top W_S - \delta_0 \sum_{j \in S} |u_j| \right\},$$

$$(55) \quad \mathcal{U} := \left\{ u \in \mathbb{R}^p \mid \beta_j^* u_j \leq 0 \text{ for } \forall j \in S \text{ and } u_j = 0 \text{ for } \forall j \in S^c \right\}.$$

In addition, $\hat{\beta}_n^T$ results in inconsistent active variable selection.

Proof. The proof is given in B.5.1. \square

Theorem A.6 implies \sqrt{n} -consistency but inconsistent variable selection for the Transfer Lasso.

B Proofs

B.1 Proofs of Lasso and Adaptive Lasso

In these proofs, the superscripts of each method may be omitted when it is obvious. For example, $\hat{\beta}_n^L$, $\hat{\beta}_n^A$, $\hat{\beta}_n^T$, and $\hat{\beta}_n^\#$ may be simply written as $\hat{\beta}_n$.

B.1.1 Proof of Lemma 2.2

We have

$$(56) \quad Z_n^L(\beta; \lambda_n) := \frac{1}{n} \|y - X\beta\|_2^2 + \frac{\lambda_n}{n} \sum_j |\beta_j|$$

$$(57) \quad = \frac{1}{n} \|X\beta^* + \varepsilon - X\beta\|_2^2 + \frac{\lambda_n}{n} \sum_j |\beta_j|$$

$$(58) \quad = (\beta - \beta^*)^\top \left(\frac{1}{n} X^\top X \right) (\beta - \beta^*) - \frac{2}{n} (\beta - \beta^*)^\top X^\top \varepsilon + \frac{1}{n} \|\varepsilon\|_2^2 + \frac{\lambda_n}{n} \sum_j |\beta_j|.$$

Let

$$(59) \quad V(\beta, \lambda_0) := (\beta - \beta^*)^\top C(\beta - \beta^*) + \lambda_0 \sum_j |\beta_j|.$$

Then, we have

$$(60) \quad Z_n^{\mathcal{L}}(\beta; \lambda_n) - V(\beta; \lambda_0) - \sigma^2$$

$$(61) \quad = (\beta - \beta^*)^\top \left(\frac{1}{n} X^\top X - C \right) (\beta - \beta^*) - \frac{2}{n} (\beta - \beta^*)^\top X^\top \varepsilon$$

$$(62) \quad + \left(\frac{1}{n} \|\varepsilon\|_2^2 - \sigma^2 \right) + \left(\frac{\lambda_n}{n} - \lambda_0 \right) \sum_j |\beta_j|$$

$$(63) \quad \rightarrow 0 \quad (n \rightarrow \infty; \text{ pointwise convergence})$$

Since $Z_n^{\mathcal{L}}$ is convex, based on [15], we have

$$(64) \quad \sup_{\beta \in K} |Z_n^{\mathcal{L}}(\beta; \lambda_n) - V(\beta; \lambda_0) - \sigma^2| \xrightarrow{P} 0 \text{ for any compact set } K.$$

Let $\beta_n^{(0)} := \operatorname{argmin}_{\beta} Z_n^{\mathcal{L}}(\beta; 0)$. Because $\beta_n^{(0)} = O_P(1)$ and $\|\hat{\beta}_n^{\mathcal{L}}\|_1 \leq \|\hat{\beta}_n^{(0)}\|_1$, we have $\hat{\beta}_n^{\mathcal{L}} = O_P(1)$ and thus

$$(65) \quad \hat{\beta}_n^{\mathcal{L}} = \operatorname{argmin}_{\beta} Z_n^{\mathcal{L}}(\beta; \lambda_n) \xrightarrow{P} \operatorname{argmin}_{\beta} V(\beta; \lambda_0).$$

B.1.2 Proof of Lemma 2.4

Let $u := \sqrt{n}(\beta - \beta^*)$. We have

$$(66) \quad Z_n^{\mathcal{L}}(\beta; \lambda_n) := \frac{1}{n} \|y - X\beta\|_2^2 + \frac{\lambda_n}{n} \sum_j |\beta_j|$$

$$(67) \quad = \frac{1}{n} \left\| \varepsilon - \frac{1}{\sqrt{n}} Xu \right\|_2^2 + \frac{\lambda_n}{n} \sum_j \left| \beta_j^* + \frac{u_j}{\sqrt{n}} \right|,$$

and

$$(68)$$

$$(69) \quad \begin{aligned} & Z_n^{\mathcal{L}}(\beta; \lambda_n) - Z_n^{\mathcal{L}}(\beta^*; \lambda_n) \\ &= \frac{1}{n} \left\| \varepsilon - \frac{1}{\sqrt{n}} Xu \right\|_2^2 - \frac{1}{n} \|\varepsilon\|_2^2 + \frac{\lambda_n}{n} \sum_j \left(\left| \beta_j^* + \frac{u_j}{\sqrt{n}} \right| - |\beta_j^*| \right) \end{aligned}$$

$$(70)$$

$$\begin{aligned} &= \underbrace{\frac{1}{n} \left\{ u^\top \left(\frac{1}{n} X^\top X \right) u - 2u^\top \left(\frac{1}{\sqrt{n}} X^\top \varepsilon \right) + \frac{\lambda_n}{\sqrt{n}} \sum_j (|u_j + \sqrt{n}\beta_j^*| - |\sqrt{n}\beta_j^*|) \right\}}_{=: V_n(u)}. \end{aligned}$$

Assuming $\lambda_n/\sqrt{n} \rightarrow \lambda_0$, we obtain

$$(71) \quad V_n(u) \xrightarrow{d} u^\top Cu - 2u^\top W + \lambda_0 \sum_j (u_j \operatorname{sgn}(\beta_j^*) I(\beta_j^* \neq 0) + |u_j| I(\beta_j^* = 0)) =: V(u).$$

Since $V_n(u)$ is convex and $V(u)$ has a unique minimum, based on [7], we have

$$(72) \quad \operatorname{argmin}_u V_n(u) \xrightarrow{d} \operatorname{argmin}_u V(u).$$

On the other hand, we have

$$(73) \quad \hat{\beta}_n^L = \operatorname{argmin}_{\beta} Z_n^L(\beta; \lambda_n) = \operatorname{argmin}_{\beta} (Z_n^L(\beta; \lambda_n) - Z_n^L(\beta^*; \lambda_n))$$

$$(74) \quad = \operatorname{argmin}_{\beta} V_n(u) = \beta^* + \frac{1}{\sqrt{n}} \operatorname{argmin}_u V_n(u).$$

Therefore, we have $\operatorname{argmin}_u V_n(u) = \sqrt{n}(\hat{\beta}_n^L - \beta^*)$ and

$$(75) \quad \sqrt{n}(\hat{\beta}_n^L - \beta^*) \xrightarrow{d} \operatorname{argmin}_u V(u).$$

B.1.3 Proof of Lemma 2.6

Let $u^* := \operatorname{argmin}_u V(u)$ where $V(u)$ is defined in (70). Note that $\hat{S}_n = S$ implies $\hat{\beta}_j = 0 \forall j \in S^c$ and $\sqrt{n}\hat{\beta}_{S^c} \xrightarrow{d} u_{S^c}^*$. By the weak convergence result, we have

$$(76) \quad P(\hat{S}_n = S) \leq P(\sqrt{n}\hat{\beta}_j = 0 \forall j \in S^c),$$

and

$$(77) \quad \limsup_n P(\hat{S}_n = S) \leq \limsup_n P(\sqrt{n}\hat{\beta}_j = 0 \forall j \in S^c) \leq P(u_j^* = 0 \forall j \in S^c) =: c.$$

If $\lambda_0 = 0$, then we have

$$(78) \quad u^* = C^{-1}W \sim \mathcal{N}(0, \sigma^2 C^{-1}),$$

resulting in $c = 0$. If $\lambda_0 > 0$, then the KKT condition yields

$$(79) \quad \begin{cases} -2W_S + 2(Cu^*)_S + \lambda_0 \operatorname{sgn}(\beta_S^*) = 0 \\ |-2W_{S^c} + 2(Cu^*)_{S^c}| \leq \lambda_0. \end{cases}$$

When $u_{S^c}^* = 0$, the above conditions indicate

$$(80) \quad |-2W_{S^c} + C_{S^c S} C_{SS}^{-1}(2W_S - \lambda_0 \operatorname{sgn}(\beta_S^*)| \leq \lambda_0,$$

thus we have

$$(81) \quad c \leq P(|-2W_{S^c} + C_{S^c S} C_{SS}^{-1}(2W_S - \lambda_0 \operatorname{sgn}(\beta_S^*))| \leq \lambda_0) < 1.$$

B.1.4 Proof of Lemma 2.7

Let $u := (n/\lambda_n)(\beta - \beta^*)$. We have

$$(82) \quad Z_n^{\mathcal{L}}(\beta) := \frac{1}{n} \|y - X\beta\|_2^2 + \frac{\lambda_n}{n} \sum_j |\beta_j|$$

$$(83) \quad = \frac{1}{n} \left\| \varepsilon - \frac{\lambda_n}{n} Xu \right\|_2^2 + \frac{\lambda_n}{n} \sum_j \left| \beta_j^* + \frac{\lambda_n}{n} u_j \right|,$$

and

$$(84)$$

$$\begin{aligned} & Z_n^{\mathcal{L}}(\beta) - Z_n^{\mathcal{L}}(\beta^*) \\ (85) \quad &= \frac{1}{n} \left\| \varepsilon - \frac{\lambda_n}{n} Xu \right\|_2^2 - \frac{1}{n} \|\varepsilon\|_2^2 + \frac{\lambda_n}{n} \sum_j \left(\left| \beta_j^* + \frac{\lambda_n}{n} u_j \right| - |\beta_j^*| \right) \\ (86) \quad &= \underbrace{\frac{\lambda_n^2}{n^2} \left\{ u^\top \left(\frac{1}{n} X^\top X \right) u - 2 \frac{\sqrt{n}}{\lambda_n} u^\top \left(\frac{1}{\sqrt{n}} X^\top \varepsilon \right) + \sum_j \left(\left| u_j + \frac{n}{\lambda_n} \beta_j^* \right| - \left| \frac{n}{\lambda_n} \beta_j^* \right| \right) \right\}}_{=: V_n(u)}. \end{aligned}$$

Assuming $\lambda_n/n \rightarrow 0$ and $\lambda_n/\sqrt{n} \rightarrow \infty$, we obtain

$$(87) \quad V_n(u) \xrightarrow{d} u^\top Cu + \sum_{j=1}^p (u_j \operatorname{sgn}(\beta_j^*) I(\beta_j^* \neq 0) + |u_j| I(\beta_j^* = 0)) =: V(u).$$

Since $V_n(u)$ is convex and $V(u)$ has a unique minimum, based on [7], we have

$$(88) \quad \underset{u}{\operatorname{argmin}} V_n(u) \xrightarrow{d} \underset{u}{\operatorname{argmin}} V(u).$$

On the other hand, we have

$$(89) \quad \hat{\beta}_n^{\mathcal{L}} = \underset{\beta}{\operatorname{argmin}} Z_n^{\mathcal{L}}(\beta; \lambda_n) = \underset{\beta}{\operatorname{argmin}} (Z_n^{\mathcal{L}}(\beta; \lambda_n) - Z_n^{\mathcal{L}}(\beta^*; \lambda_n))$$

$$(90) \quad = \underset{\beta}{\operatorname{argmin}} V_n(u) = \beta^* + \frac{\lambda_n}{n} \underset{u}{\operatorname{argmin}} V_n(u).$$

Therefore, we have $\underset{u}{\operatorname{argmin}} V_n(u) = (n/\lambda_n)(\hat{\beta}_n^{\mathcal{L}} - \beta^*)$ and

$$(91) \quad \frac{n}{\lambda_n} (\hat{\beta}_n^{\mathcal{L}} - \beta^*) \xrightarrow{d} \underset{u}{\operatorname{argmin}} V(u).$$

B.1.5 Proof of Lemma 2.9

Asymptotic Normality Part: Let $u := \sqrt{n}(\beta - \beta^*)$. We have

$$(92) \quad Z_n^{\mathcal{A}}(\beta) := \frac{1}{n} \|y - X\beta\|_2^2 + \frac{\lambda_n}{n} \sum_j w_j |\beta_j|$$

$$(93) \quad = \frac{1}{n} \left\| \varepsilon - \frac{1}{\sqrt{n}} Xu \right\|_2^2 + \frac{\lambda_n}{n} \sum_j w_j \left| \beta_j^* + \frac{u_j}{\sqrt{n}} \right|$$

and

$$(94)$$

$$\begin{aligned} & Z_n^{\mathcal{A}}(\beta) - Z_n^{\mathcal{A}}(\beta^*) \\ (95) \quad &= \frac{1}{n} \left\| \varepsilon - \frac{1}{\sqrt{n}} Xu \right\|_2^2 - \frac{1}{n} \|\varepsilon\|_2^2 + \frac{\lambda_n}{n} \sum_j w_j \left(\left| \beta_j^* + \frac{u_j}{\sqrt{n}} \right| - |\beta_j^*| \right) \\ (96) \quad &= \frac{1}{n} \underbrace{\left\{ u^\top \left(\frac{1}{n} X^\top X \right) u - 2u^\top \left(\frac{1}{\sqrt{n}} X^\top \varepsilon \right) + \frac{\lambda_n}{\sqrt{n}} \sum_j w_j (|u_j + \sqrt{n}\beta_j^*| - |\sqrt{n}\beta_j^*|) \right\}}_{=:V_n(u)}. \end{aligned}$$

We know that

$$(97) \quad -2u^\top \left(\frac{1}{\sqrt{n}} X^\top \varepsilon \right) \xrightarrow{d} -2u^\top W, \quad W \sim \mathcal{N}(0, \sigma^2 C)$$

$$(98) \quad u^\top \left(\frac{1}{n} X^\top X \right) u \rightarrow u^\top C u.$$

Now we consider the last term of (96). If $\beta_j^* \neq 0$, then $w_j \rightarrow 1/|\beta_j^*|^\gamma$ and $|u_j + \sqrt{n}\beta_j^*| - |\sqrt{n}\beta_j^*| \rightarrow u_j \operatorname{sgn}(\beta_j^*)$, thus by Slutsky's theorem,

$$(99) \quad \frac{\lambda_n}{\sqrt{n}} w_j (|u_j + \sqrt{n}\beta_j^*| - |\sqrt{n}\beta_j^*|) \xrightarrow{p} 0,$$

as $\lambda_n/\sqrt{n} \rightarrow 0$. If $\beta_j^* = 0$, then $\lambda_n w_j / \sqrt{n} = \lambda_n n^{(\gamma-1)/2} |\sqrt{n}\tilde{\beta}_j|^{-\gamma}$ and $|u_j + \sqrt{n}\beta_j^*| - |\sqrt{n}\beta_j^*| = |u_j|$, thus by Slutsky's theorem,

$$(100) \quad \frac{\lambda_n}{\sqrt{n}} w_j (|u_j + \sqrt{n}\beta_j^*| - |\sqrt{n}\beta_j^*|) \xrightarrow{d} \begin{cases} 0 & \text{if } u_j = 0 \\ \infty & \text{otherwise,} \end{cases}$$

as $\lambda_n n^{(\gamma-1)/2} \rightarrow \infty$. Therefore, we have for every u ,

$$(101) \quad V_n(u) \xrightarrow{d} V(u) := \begin{cases} -2u_S^\top W_S + u_S^\top C_{SS} u_S & \text{if } u_j = 0 \ \forall j \in S^c \\ \infty & \text{otherwise.} \end{cases}$$

Since $V_n(u)$ is convex and $V(u)$ has a unique minimum of $(C_{SS}^{-1} W_S, 0)^\top$, based on [6], we have

$$(102) \quad \operatorname{argmin}_u V_n(u) \xrightarrow{d} \operatorname{argmin}_u V(u) = (C_{SS}^{-1} W_S, 0)^\top.$$

On the other hand,

$$(103) \quad \hat{\beta}_n^{\mathcal{L}} = \operatorname{argmin}_{\beta} Z_n(\beta; \lambda_n) = \operatorname{argmin}_{\beta} (Z_n(\beta; \lambda_n) - Z_n(\beta^*; \lambda_n))$$

$$(104) \quad = \operatorname{argmin}_{\beta} V_n(u) = \beta^* + \frac{1}{\sqrt{n}} \operatorname{argmin}_u V_n(u).$$

Therefore, we have $\operatorname{argmin}_u V_n(u) = \sqrt{n}(\hat{\beta}_n - \beta^*)$ and

$$(105) \quad \sqrt{n}(\hat{\beta}_S - \beta_S^*) \xrightarrow{d} \mathcal{N}(0, \sigma^2 C_{SS}^{-1})^\top, \quad \sqrt{n}\hat{\beta}_{S^c} \xrightarrow{d} 0.$$

Variable Selection Consistency Part: Asymptotic normality indicates that $\hat{\beta} \xrightarrow{p} \beta^*$, thus

$$(106) \quad \forall j \in S, \quad P(j \in \operatorname{supp}(\hat{\beta})) \rightarrow 1.$$

Now, we consider the event $j \in S^c$ and $j \in \text{supp}(\hat{\beta})$. By the KKT conditions, we have

$$(107) \quad 2\mathbf{x}_j^\top(y - X\hat{\beta}) + \lambda_n w_j \text{sgn}(\hat{\beta}_j) = 0.$$

This yields

$$(108) \quad 2\left(\frac{1}{n}\mathbf{x}_j^\top X\right)\sqrt{n}(\beta^* - \hat{\beta}) + 2\frac{1}{\sqrt{n}}\mathbf{x}_j^\top \varepsilon + \lambda_n n^{(\gamma-1)/2}|\sqrt{n}\tilde{\beta}_j|^{-\gamma} \text{sgn}(\hat{\beta}_j) = 0$$

By Slutsky's theorem, the first and second terms on the left-hand side converge to some normal distribution, but the last term on the left-hand side diverges to infinity. Therefore, we have for $\forall j \in S^c$,

$$(109) \quad P(j \in \text{supp}(\hat{\beta})) \leq P\left(2\mathbf{x}_j^\top(y - X\hat{\beta}) + \lambda_n w_j \text{sgn}(\hat{\beta}_j) = 0\right) \rightarrow 0.$$

B.2 Proofs of Transfer Lasso

B.2.1 Proof of Lemma 3.2

The proof is similar to that of Lemma 2.9. If $\tilde{\beta}$ is a \sqrt{m} -consistent initial estimator and $\lambda_n \sqrt{m^\gamma/n} \rightarrow \infty$, (100) reduces to

$$(110) \quad \lambda_n w_j \left(\left| \beta_j^* + \frac{u_j}{\sqrt{n}} \right| - |\beta_j^*| \right) = \lambda_n \sqrt{\frac{m^\gamma}{n}} |\sqrt{m}\tilde{\beta}_j|^{-\gamma} |u_j| \xrightarrow{d} \begin{cases} 0 & \text{if } u_j = 0 \\ \infty & \text{otherwise} \end{cases}$$

and (108) reduces to

$$(111) \quad 2\left(\frac{1}{n}\mathbf{x}_j^\top X\right)\sqrt{n}(\beta^* - \hat{\beta}) + 2\frac{1}{\sqrt{n}}\mathbf{x}_j^\top \varepsilon + \lambda_n \sqrt{\frac{m^\gamma}{n}} |\sqrt{m}\tilde{\beta}_j|^{-\gamma} \text{sgn}(\hat{\beta}_j) = 0.$$

These modifications does not affect the remaining proofs.

B.2.2 Proof of Theorem 3.5 and Theorem A.6

Let $u := l(\beta - \beta^*)$ where $l = l(n, m, \lambda_n)$ is a certain function as defined later. Let $z := \sqrt{m}(\tilde{\beta} - \beta^*)$. Since $\tilde{\beta}$ is a \sqrt{m} -consistent estimator, z follows

some distribution. Suppose that $n/m \rightarrow r_0 \geq 0$. The objective function for Transfer Lasso is

$$(112) \quad Z_n^T(\beta) := \frac{1}{n} \|y - X\beta\|_2^2 + \frac{\lambda_n}{n} \sum_j |\beta_j| + \frac{\eta_n}{n} \sum_j |\beta_j - \tilde{\beta}_j|$$

$$(113) \quad = \frac{1}{n} \left\| \varepsilon - \frac{1}{l} Xu \right\|_2^2 + \frac{\lambda_n}{n} \sum_j \left| \beta_j^* + \frac{u_j}{l} \right| + \frac{\eta_n}{n} \sum_j \left| u_j - \frac{z_j}{\sqrt{m}} \right|$$

$$(114) \quad = \frac{1}{l^2} u^\top \left(\frac{1}{n} X^\top X \right) u - \frac{2}{\sqrt{nl}} u^\top \left(\frac{1}{\sqrt{n}} X^\top \varepsilon \right) + \frac{1}{n} \|\varepsilon\|_2^2$$

$$(115) \quad + \frac{\lambda_n}{nl} \sum_j |u_j + l\beta_j^*| + \frac{\eta_n}{nl} \sum_j \left| u_j - \frac{l}{\sqrt{m}} z_j \right|,$$

and we have

$$(116) \quad Z_n^T(\beta) - Z_n^T(\beta^*) = \frac{1}{l^2} u^\top \left(\frac{1}{n} X^\top X \right) u - \frac{2}{\sqrt{nl}} u^\top \left(\frac{1}{\sqrt{n}} X^\top \varepsilon \right) \\ + \frac{\lambda_n}{nl} \sum_j (|u_j + l\beta_j^*| - |l\beta_j^*|) + \frac{\eta_n}{nl} \sum_j \left(\left| u_j - \frac{l}{\sqrt{m}} z_j \right| - \left| \frac{l}{\sqrt{m}} z_j \right| \right).$$

We divide the case into three cases: $l = \sqrt{m}$ (Case I), \sqrt{n} (Case II), and n/λ_n (Case III).

Case I. Let $l = \sqrt{m}$. Then, (116) reduces to

(117)

$$Z_n^T(\beta) - Z_n^T(\beta^*) = \frac{1}{m} u^\top \left(\frac{1}{n} X^\top X \right) u - \frac{2}{\sqrt{nm}} u^\top \left(\frac{1}{\sqrt{n}} X^\top \varepsilon \right) \\ + \frac{\lambda_n}{n\sqrt{m}} \sum_j (|u_j + \sqrt{m}\beta_j^*| - |\sqrt{m}\beta_j^*|) + \frac{\eta_n}{n\sqrt{m}} \sum_j (|u_j - z_j| - |z_j|).$$

Let $V_n(u) := (n\sqrt{m}/\eta_n)(Z_n^T(\beta) - Z_n^T(\beta^*)) + \sum_j |z_j|$, then we have

$$(119) \quad V_n(u) = \frac{n}{\sqrt{m}\eta_n} u^\top \left(\frac{1}{n} X^\top X \right) u - \frac{2\sqrt{n}}{\eta_n} u^\top \left(\frac{1}{\sqrt{n}} X^\top \varepsilon \right)$$

$$(120) \quad + \frac{\lambda_n}{\eta_n} \sum_j (|u_j + \sqrt{m}\beta_j^*| - |\sqrt{m}\beta_j^*|) + \sum_j |u_j - z_j|.$$

On the other hand, we have

(121)

$$\hat{\beta}_n^T = \operatorname{argmin}_{\beta} Z_n^T(\beta) = \operatorname{argmin}_{\beta} (Z_n^T(\beta) - Z_n^T(\beta^*)) = \beta^* + \frac{1}{\sqrt{m}} \operatorname{argmin}_u V_n(u),$$

and hence

$$(122) \quad \sqrt{m}(\hat{\beta}_n - \beta^*) = \operatorname{argmin}_u V_n(u).$$

Consider the case (i) where $\eta_n/\sqrt{n} \rightarrow \infty$, $\lambda_n/\eta_n \rightarrow \rho_0$, and $0 \leq \rho_0 < 1$. Let $V(u) := \lim_{n \rightarrow \infty} V_n(u)$, then we have

$$(123) \quad V(u) = \sum_j ((\rho_0 u_j \operatorname{sgn}(\beta_j^*) + |u_j - z_j|) I(\beta_j^* \neq 0) + (\rho_0 |u_j| + |u_j - z_j|) I(\beta_j^* = 0)).$$

Because $V_n(u)$ is convex and $V(u)$ has a unique minimum, we obtain

$$(124) \quad \sqrt{m}(\hat{\beta}_n - \beta^*) \rightarrow \operatorname{argmin}_u V(u) = z.$$

Case II. Let $l = \sqrt{n}$. Then, (116) reduces to

$$(125)$$

$$(126) \quad \begin{aligned} & Z_n^T(\beta) - Z_n^T(\beta^*) \\ &= \frac{1}{n} u^\top \left(\frac{1}{n} X^\top X \right) u - \frac{2}{n} u^\top \left(\frac{1}{\sqrt{n}} X^\top \varepsilon \right) \end{aligned}$$

$$(127)$$

$$+ \frac{\lambda_n}{n\sqrt{n}} \sum_j (|u_j + \sqrt{n}\beta_j^*| - |\sqrt{n}\beta_j^*|) + \frac{\eta_n}{n\sqrt{n}} \sum_j \left(\left| u_j - \sqrt{\frac{n}{m}} z_j \right| - \left| \sqrt{\frac{n}{m}} z_j \right| \right).$$

Let $V_n(u) := n(Z_n^T(\beta) - Z_n^T(\beta^*))$, then we have

$$(128)$$

$$V_n(u) = u^\top \left(\frac{1}{n} X^\top X \right) u - 2u^\top \left(\frac{1}{\sqrt{n}} X^\top \varepsilon \right)$$

$$(129)$$

$$+ \frac{\lambda_n}{\sqrt{n}} \sum_j (|u_j + \sqrt{n}\beta_j^*| - |\sqrt{n}\beta_j^*|) + \frac{\eta_n}{\sqrt{n}} \sum_j \left(\left| u_j - \sqrt{\frac{n}{m}} z_j \right| - \left| \sqrt{\frac{n}{m}} z_j \right| \right).$$

Consider the case (ii) where $\lambda_n/\sqrt{n} \rightarrow \lambda_0 \geq 0$ and $\eta_n/\sqrt{n} \rightarrow \eta_0 \geq 0$.

Let $V(u) := \lim_{n \rightarrow \infty} V_n(u)$, then we have

$$(130) \quad V(u) = u^\top Cu - 2u^\top W$$

$$(131) \quad + \lambda_0 \sum_j (u_j \operatorname{sgn}(\beta_j^*) I(\beta_j^* \neq 0) + |u_j| I(\beta_j^* = 0)) + \eta_0 \sum_j (|u_j - \sqrt{r_0} z_j| - |\sqrt{r_0} z_j|).$$

On the other hand, we have

$$(132) \quad \hat{\beta}_n^T = \underset{\beta}{\operatorname{argmin}} Z_n^T(\beta) = \underset{\beta}{\operatorname{argmin}} (Z_n^T(\beta) - Z_n^T(\beta^*)) = \beta^* + \frac{1}{\sqrt{n}} \underset{u}{\operatorname{argmin}} V_n(u),$$

and hence

$$(133) \quad \sqrt{n}(\hat{\beta}_n - \beta^*) = \underset{u}{\operatorname{argmin}} V_n(u).$$

Because $V_n(u)$ is convex and $V(u)$ has a unique minimum, we obtain

$$(134) \quad \sqrt{n}(\hat{\beta}_n - \beta^*) \rightarrow \underset{u}{\operatorname{argmin}} \left\{ u^\top Cu - 2u^\top W \right.$$

$$(135) \quad \left. + \lambda_0 \sum_j (u_j \operatorname{sgn}(\beta_j^*) I(\beta_j^* \neq 0) + |u_j| I(\beta_j^* = 0)) + \eta_0 \sum_j (|u_j - \sqrt{r_0} z_j| - |\sqrt{r_0} z_j|) \right\}.$$

Case III. Let $l = n/\lambda_n (\rightarrow \infty)$. Then, (116) reduces to

$$(136) \quad Z_n^T(\beta) - Z_n^T(\beta^*)$$

$$(137) \quad = \frac{\lambda_n^2}{n^2} u^\top \left(\frac{1}{n} X^\top X \right) u - \frac{\lambda_n}{n\sqrt{n}} u^\top \left(\frac{1}{\sqrt{n}} X^\top \varepsilon \right)$$

$$(138) \quad + \frac{\lambda_n^2}{n^2} \sum_j \left(\left| u_j + \frac{n}{\lambda_n} \beta_j^* \right| - \left| \frac{n}{\lambda_n} \beta_j^* \right| \right) + \frac{\lambda_n \eta_n}{n^2} \sum_j \left(\left| u_j - \sqrt{\frac{n}{m}} \frac{\sqrt{n}}{\lambda_n} z_j \right| - \left| \sqrt{\frac{n}{m}} \frac{\sqrt{n}}{\lambda_n} z_j \right| \right)$$

Let $V_n(u) := (n^2/\lambda_n^2)(Z_n^T(\beta) - Z_n^T(\beta^*))$, then we have

(139)

$$V_n(u) = u^\top \left(\frac{1}{n} X^\top X \right) u - \frac{2\sqrt{n}}{\lambda_n} u^\top \left(\frac{1}{\sqrt{n}} X^\top \varepsilon \right)$$

(140)

$$+ \sum_j \left(\left| u_j + \frac{n}{\lambda_n} \beta_j^* \right| - \left| \frac{n}{\lambda_n} \beta_j^* \right| \right) + \frac{\eta_n}{\lambda_n} \sum_j \left(\left| u_j - \sqrt{\frac{n}{m}} \frac{\sqrt{n}}{\lambda_n} z_j \right| - \left| \sqrt{\frac{n}{m}} \frac{\sqrt{n}}{\lambda_n} z_j \right| \right).$$

Consider the case (iii) where $\lambda_n/\sqrt{n} \rightarrow \infty$, $\lambda_n/n \rightarrow 0$, and $\eta_n/\lambda_n \rightarrow \rho'_0 \geq 0$. Let $V(u) := \lim_{n \rightarrow \infty} V_n(u)$, then we have

(141)

$$V(u) = u^\top C u + \sum_j ((u_j \operatorname{sgn}(\beta_j^*) + \rho'_0 |u_j|) I(\beta_j^* \neq 0) + (1 + \rho'_0) |u_j| I(\beta_j^* = 0)).$$

On the other hand, we have

(142)

$$\hat{\beta}_n^T = \underset{\beta}{\operatorname{argmin}} Z_n^T(\beta) = \underset{\beta}{\operatorname{argmin}} (Z_n^T(\beta) - Z_n^T(\beta^*)) = \beta^* + \frac{\lambda_n}{n} \underset{u}{\operatorname{argmin}} V_n(u),$$

and hence

(143)

$$\frac{n}{\lambda_n} (\hat{\beta}_n - \beta^*) = \underset{u}{\operatorname{argmin}} V_n(u).$$

Because $V_n(u)$ is convex and $V(u)$ has a unique minimum, we obtain

(144)

$$\begin{aligned} & \frac{n}{\lambda_n} (\hat{\beta}_n - \beta^*) \\ & \rightarrow \underset{u}{\operatorname{argmin}} \left\{ u^\top C u + \sum_j ((u_j \operatorname{sgn}(\beta_j^*) + \rho'_0 |u_j|) I(\beta_j^* \neq 0) + (1 + \rho'_0) |u_j| I(\beta_j^* = 0)) \right\}. \end{aligned}$$

In addition, if $\rho'_0 \geq 1$, then the right-hand side of (144) reduces to 0.

B.2.3 Proof of Theorem 3.9

By Theorem 3.5 and Corollary 3.6, the Transfer Lasso estimator satisfies $\hat{\beta}_n^T \xrightarrow{P} \beta^*$, thus

$$(146) \quad \forall j \in S, \limsup_{n \rightarrow \infty} P(j \in \text{supp}(\hat{\beta}_n^T)) = 1.$$

Let $u^* := \arg\min_u V(u)$ where $V(u)$ is the asymptotic objective function (123) in the case (i) and (130) in the case (ii). By the weak convergence result, we have

$$(147) \quad \limsup_{n \rightarrow \infty} P(\hat{S}_n^T = S) = \limsup_{n \rightarrow \infty} P(\hat{\beta}_j^T \neq 0 \ \forall j \in S \text{ and } \hat{\beta}_j^T = 0 \ \forall j \in S^c)$$

$$(148) \quad \leq \limsup_{n \rightarrow \infty} P(l\hat{\beta}_j^T = 0 \ \forall j \in S^c)$$

$$(149) \quad \leq P(u_j^* = 0 \ \forall j \in S^c),$$

where l is \sqrt{m} in the case (i) and \sqrt{n} in the case (ii). We evaluate the probability of $u_{S^c}^* = 0$ in each case.

Consider the case (i) where $\eta_n/\sqrt{n} \rightarrow \infty$ and $\lambda_n/\eta_n \rightarrow \rho_0$ with $0 \leq \rho_0 < 1$. The asymptotic distribution of the Transfer Lasso is equal to that of the initial estimator z . Because we suppose that z is inconsistent in terms of variable selection, we obtain

$$(150) \quad P(u_j^* = 0 \ \forall j \in S^c) = P(z_j = 0 \ \forall j \in S^c) \leq c < 1,$$

hence $\hat{\beta}_n^T$ is inconsistent in terms of variable selection.

Consider the case (ii) where $\lambda_n/\sqrt{n} \rightarrow \lambda_0 \geq 0$ and $\eta_n/\sqrt{n} \rightarrow \eta_0 \geq 0$. Suppose that $u_{S^c}^* = 0$. Let $S_1 := \{j : j \in S \text{ and } u_j^* \neq \sqrt{r_0}z_j\}$, $S_2 := \{j : j \in S \text{ and } u_j^* = \sqrt{r_0}z_j\}$, $S_1^c := \{j : j \in S^c \text{ and } u_j^* \neq \sqrt{r_0}z_j\}$, and $S_2^c := \{j : j \in S^c \text{ and } u_j^* = \sqrt{r_0}z_j\}$. By the KKT conditions of $\arg\min_u V(u)$, we have

$$(151) \quad \begin{cases} 2C_{S_1 S_1} u_{S_1}^* + 2\sqrt{r_0} C_{S_1 S_2} z_{S_2} - 2W_{S_1} + \lambda_0 \text{sgn}(\beta_{S_1}^*) + \eta_0 \text{sgn}(u_{S_1}^* - \sqrt{r_0}z_{S_1}) = 0 \\ |2C_{S_2 S_1} u_{S_1}^* + 2\sqrt{r_0} C_{S_2 S_2} z_{S_2} - 2W_{S_2} + \lambda_0 \text{sgn}(\beta_{S_2}^*)| \leq \eta_0 \\ |2C_{S^c S_1} u_{S_1}^* + 2\sqrt{r_0} C_{S^c S_2} z_{S_2} - 2W_{S_1^c} + \eta_0 \text{sgn}(-\sqrt{r_0}z_{S_1^c})| \leq \lambda_0 \\ |2C_{S^c S_1} u_{S_1}^* + 2\sqrt{r_0} C_{S^c S_2} z_{S_2} - 2W_{S_2^c}| \leq \lambda_0 + \eta_0. \end{cases}$$

If $S_1 \neq \emptyset$, the first equation yields

$$(152) \quad u_{S_1}^* = C_{S_1 S_1}^{-1} \left(W_{S_1} - \sqrt{r_0} C_{S_1 S_2} z_{S_2} - \frac{\lambda_0}{2} \text{sgn}(\beta_{S_1}^*) - \frac{\eta_0}{2} \text{sgn}(u_{S_1}^* - \sqrt{r_0}z_{S_1}) \right).$$

Because W follows a Gaussian and z follows some distribution, the probability holding the second, third, and fourth inequalities in (151) is less than 1. This indicates inconsistent variable selection. If $S_1 = \emptyset$, by the KKT conditions, we have

$$(153) \quad \begin{cases} |2\sqrt{r_0}C_{S_2S_2}z_{S_2} - 2W_{S_2} + \lambda_0 \operatorname{sgn}(\beta_{S_2}^*)| \leq \eta_0 \\ |2\sqrt{r_0}C_{S_1^cS_2}z_{S_2} - 2W_{S_1^c} + \eta_0 \operatorname{sgn}(-\sqrt{r_0}z_{S_1^c})| \leq \lambda_0 \\ |2\sqrt{r_0}C_{S_2^cS_2}z_{S_2} - 2W_{S_2^c}| \leq \lambda_0 + \eta_0. \end{cases}$$

The probability holding these inequalities is less than 1. This indicates inconsistent variable selection.

B.2.4 Proof of Theorem 3.10

Consider the case (i) in Theorem 3.5. Consider the event where $\hat{\beta}_j \neq \tilde{\beta}_j$ for some $j \in S$. By the KKT conditions,

$$(154) \quad \begin{cases} 2 \left(\frac{1}{n} \mathbf{x}_j^\top X \right) \sqrt{\frac{n}{m}} \sqrt{m} (\hat{\beta} - \beta^*) - \frac{2}{\sqrt{n}} \mathbf{x}_j^\top \varepsilon + \frac{\lambda_n}{\sqrt{n}} \operatorname{sgn}(\hat{\beta}_j) + \frac{\eta_n}{\sqrt{n}} \operatorname{sgn}(\hat{\beta}_j - \tilde{\beta}_j) = 0, \\ \left| 2 \left(\frac{1}{n} \mathbf{x}_j^\top X \right) \sqrt{\frac{n}{m}} \sqrt{m} (\hat{\beta} - \beta^*) - \frac{2}{\sqrt{n}} \mathbf{x}_j^\top \varepsilon + \frac{\eta_n}{\sqrt{n}} \operatorname{sgn}(\hat{\beta}_j - \tilde{\beta}_j) \right| \leq \frac{\lambda_n}{\sqrt{n}}, \text{ for } \hat{\beta}_j = 0. \end{cases} \quad \text{for } \hat{\beta}_j \neq 0$$

The term including η_n/\sqrt{n} in (154) in both $\hat{\beta}_j \neq 0$ and $\hat{\beta}_j = 0$ cases diverge to infinity faster compared to the rest terms. Therefore, we have

$$(155) \quad \forall j \in S, \lim_{n \rightarrow \infty} P(\hat{\beta}_j \neq \tilde{\beta}_j) = 0.$$

This concludes

$$(156) \quad \lim_{n \rightarrow \infty} P(\hat{\beta}_S = \tilde{\beta}_S) = 1.$$

B.2.5 Proof of Theorem 3.11

By Theorem 3.5 and Corollary 3.6, the Transfer Lasso estimator satisfies $\hat{\beta}_n^T \xrightarrow{P} \beta^*$, thus

$$(157) \quad \forall j \in S, P(j \in \operatorname{supp}(\hat{\beta}_n^T)) \rightarrow 1.$$

Consider the case (ii) where $\lambda_n/\sqrt{n} \rightarrow \lambda_0 \geq 0$ and $\eta_n/\sqrt{n} \rightarrow \eta_0 \geq 0$. Let $u^* := \operatorname{argmin}_u V(u)$ where $V(u)$ is the asymptotic objective function (130) for the case (ii). By the weak convergence result, we have

$$(158) \quad \limsup_{n \rightarrow \infty} P(\hat{\beta}_S^T = \tilde{\beta}_S) \leq P(u_j^* = \sqrt{r_0}z_j \ \forall j \in S).$$

Suppose that $u_S^* = \sqrt{r_0}z_S$. Let $S_1 := \{j : j \in S \text{ and } u_j^* \neq 0\}$, $S_2 := \{j : j \in S \text{ and } u_j^* = 0\}$, $S_1^c := \{j : j \in S^c \text{ and } u_j^* \neq 0 \text{ and } u_j^* \neq \sqrt{r_0}z_j\}$, $S_2^c := \{j : j \in S^c \text{ and } u_j^* \neq 0 \text{ and } u_j^* = \sqrt{r_0}z_j\}$, $S_3^c := \{j : j \in S^c \text{ and } u_j^* = 0 \text{ and } u_j^* \neq \sqrt{r_0}z_j\}$, and $S_4^c := \{j : j \in S^c \text{ and } u_j^* = 0 \text{ and } u_j^* = \sqrt{r_0}z_j\}$. By the KKT conditions of $\operatorname{argmin}_u V(u)$, we have

$$(159) \quad \begin{cases} |2\sqrt{r_0}C_{SS}z_S + 2C_{SS^c}u_{S^c}^* - 2W_S + \lambda_0 \operatorname{sgn}(\beta_S^*)| \leq \eta_0 \\ 2\sqrt{r_0}C_{S_1^c}S^c z_S + 2C_{S_1^c S^c}u_{S^c}^* - 2W_{S_1^c} + \lambda_0 \operatorname{sgn}(u_{S_1^c}^*) + \eta_0 \operatorname{sgn}(u_{S_1^c}^* - \sqrt{r_0}z_{S_1^c}) = 0 \\ |2\sqrt{r_0}C_{S_2^c}S^c z_S + 2C_{S_2^c S^c}u_{S^c}^* - 2W_{S_2^c} + \lambda_0 \operatorname{sgn}(u_{S_2^c}^*)| \leq \eta_0 \\ |2\sqrt{r_0}C_{S_3^c}S^c z_S + 2C_{S_3^c S^c}u_{S^c}^* - 2W_{S_3^c} + \eta_0 \operatorname{sgn}(u_{S_3^c}^* - \sqrt{r_0}z_{S_3^c})| \leq \lambda_0 \\ |2\sqrt{r_0}C_{S_4^c}S^c z_S + 2C_{S_4^c S^c}u_{S^c}^* - 2W_{S_4^c}| \leq \lambda_0 + \eta_0. \end{cases}$$

Note that $u_{S_2^c}^* = \sqrt{r_0}z_{S_2^c}$, $u_{S_2^c}^* = 0$, and $u_{S_4^c}^* = \sqrt{r_0}z_{S_4^c} = 0$. If $S_1 \neq \emptyset$, the second equation yields

$$(160) \quad u_{S_1^c}^* = C_{S_1^c S_1}^{-1} \left(W_{S_1^c} - \sqrt{r_0}C_{S_1^c}S^c z_S - \sqrt{r_0}C_{S_1^c S_2^c}z_{S_2^c} \right.$$

$$(161) \quad \left. - \frac{\lambda_0}{2} \operatorname{sgn}(u_{S_1^c}^*) - \frac{\eta_0}{2} \operatorname{sgn}(u_{S_1^c}^* - \sqrt{r_0}z_{S_1^c}) \right).$$

Hence, we have

$$(162) \quad \begin{cases} |2\sqrt{r_0}C_{SS}z_S + 2C_{SS^c}u_{S^c}^* + 2\sqrt{r_0}C_{SS^c}z_{S^c} - 2W_S + \lambda_0 \operatorname{sgn}(\beta_S^*)| \leq \eta_0 \\ |2\sqrt{r_0}C_{S_2^c}S^c z_S + 2C_{S_2^c S^c}u_{S^c}^* + 2\sqrt{r_0}C_{S_2^c S_2^c}z_{S_2^c} - 2W_{S_2^c} + \lambda_0 \operatorname{sgn}(u_{S_2^c}^*)| \leq \eta_0 \\ |2\sqrt{r_0}C_{S_3^c}S^c z_S + 2C_{S_3^c S^c}u_{S^c}^* + 2\sqrt{r_0}C_{S_3^c S_2^c}z_{S_2^c} - 2W_{S_3^c} + \eta_0 \operatorname{sgn}(u_{S_3^c}^* - \sqrt{r_0}z_{S_3^c})| \leq \lambda_0 \\ |2\sqrt{r_0}C_{S_4^c}S^c z_S + 2C_{S_4^c S^c}u_{S^c}^* + 2\sqrt{r_0}C_{S_4^c S_2^c}z_{S_2^c} - 2W_{S_4^c}| \leq \lambda_0 + \eta_0. \end{cases}$$

Because W follows a Gaussian distribution and $u_{S_1^c}^*$ and z follow some distribution, the probability holding these inequalities is less than 1. This

indicates inconsistent invariant variable selection. If $S_1 = \emptyset$, by the KKT conditions, we have

$$(163) \quad \begin{cases} |2\sqrt{r_0}C_{SS}z_S + 2\sqrt{r_0}C_{SS_2^c}z_{S_2^c} - 2W_S + \lambda_0 \operatorname{sgn}(\beta_S^*)| \leq \eta_0 \\ |2\sqrt{r_0}C_{S_2^c}S_S z_S + 2\sqrt{r_0}C_{S_2^c}S_2^c z_{S_2^c} - 2W_{S_2^c} + \lambda_0 \operatorname{sgn}(u_{S_2^c})| \leq \eta_0 \\ |2\sqrt{r_0}C_{S_3^c}S_S z_S + 2\sqrt{r_0}C_{S_3^c}S_2^c z_{S_2^c} - 2W_{S_3^c} + \eta_0 \operatorname{sgn}(u_{S_3^c} - \sqrt{r_0}z_{S_3^c})| \leq \lambda_0 \\ |2\sqrt{r_0}C_{S_4^c}S_S z_S + 2\sqrt{r_0}C_{S_4^c}S_2^c z_{S_2^c} - 2W_{S_4^c}| \leq \lambda_0 + \eta_0. \end{cases}$$

Because W follows a Gaussian distribution and z follows some distribution, the probability holding these inequalities is less than 1. This indicates inconsistent variable selection.

B.3 Proofs of Adaptive Transfer Lasso

B.3.1 Proof of Theorem 4.1

Let $u := l(\beta - \beta^*)$ where $l = l(n, m, \lambda_n)$ is a certain function as defined later. Let $z := \sqrt{m}(\tilde{\beta} - \beta^*)$. Since $\tilde{\beta}$ is a \sqrt{m} -consistent estimator, z follows some distribution. We suppose that $n/m \rightarrow r_0 \geq 0$. The objective function for the Adaptive Transfer Lasso is

$$(164) \quad Z_n^\#(\beta) := \frac{1}{n} \|y - X\beta\|_2^2 + \frac{\lambda_n}{n} \sum_j \frac{|\beta_j|}{|\tilde{\beta}_j|^{\gamma_1}} + \frac{\eta_n}{n} \sum_j |\tilde{\beta}_j|^{\gamma_2} |\beta_j - \tilde{\beta}_j|$$

$$(165) \quad = \frac{1}{n} \left\| \varepsilon - \frac{1}{l} Xu \right\|_2^2 + \frac{\lambda_n}{n} \sum_j \frac{\left| \beta_j^* + \frac{u_j}{l} \right|}{\left| \frac{z_j}{\sqrt{m}} + \beta_j^* \right|^{\gamma_1}} + \frac{\eta_n}{n} \sum_j \left| \frac{z_j}{\sqrt{m}} + \beta_j^* \right|^{\gamma_2} \left| \frac{u_j}{l} - \frac{z_j}{\sqrt{m}} \right|$$

$$(166) \quad = \frac{1}{l^2} u^\top \left(\frac{1}{n} X^\top X \right) u - \frac{2}{\sqrt{nl}} u^\top \left(\frac{1}{\sqrt{n}} X^\top \varepsilon \right) + \frac{1}{n} \|\varepsilon\|_2^2$$

$$(167) \quad + \frac{\sqrt{m^{\gamma_1}} \lambda_n}{nl} \sum_j \frac{1}{\left| z_j + \sqrt{m}\beta_j^* \right|^{\gamma_1}} |u_j + l\beta_j^*|$$

$$(168) \quad + \frac{\eta_n}{n\sqrt{m^{\gamma_2}}l} \sum_j |z_j + \sqrt{m}\beta_j^*|^{\gamma_2} \left| u_j - \frac{l}{\sqrt{m}} z_j \right|,$$

and we have

$$(169) \quad \begin{aligned} Z_n^\#(\beta) - Z_n^\#(\beta^*) &= \frac{1}{l^2} u^\top \left(\frac{1}{n} X^\top X \right) u - \frac{2}{\sqrt{nl}} u^\top \left(\frac{1}{\sqrt{n}} X^\top \varepsilon \right) \\ &\quad + \frac{\sqrt{m^{\gamma_1}} \lambda_n}{nl} \sum_j \frac{1}{|z_j + \sqrt{m}\beta_j^*|^{\gamma_1}} (|u_j + l\beta_j^*| - |l\beta_j^*|) \\ &\quad + \frac{\eta_n}{n\sqrt{m^{\gamma_2}}} \sum_j |z_j + \sqrt{m}\beta_j^*|^{\gamma_2} \left(\left| u_j - \frac{l}{\sqrt{m}} z_j \right| - \left| \frac{l}{\sqrt{m}} z_j \right| \right). \end{aligned}$$

We divide the case into three cases: $l = \sqrt{m}$ (Case I), $l = \sqrt{n}$ (Case II), and $l = n/\lambda_n$ (Case III).

Case I. Let $l = \sqrt{m}$. Then, (169) reduces to

$$(170)$$

$$(171) \quad \begin{aligned} Z_n^\#(\beta) - Z_n^\#(\beta^*) &= \frac{1}{m} u^\top \left(\frac{1}{n} X^\top X \right) u - \frac{2}{\sqrt{nm}} u^\top \left(\frac{1}{\sqrt{n}} X^\top \varepsilon \right) \\ &\quad + \frac{\sqrt{m^{\gamma_1-1}} \lambda_n}{n} \sum_j \frac{1}{|z_j + \sqrt{m}\beta_j^*|^{\gamma_1}} (|u_j + \sqrt{m}\beta_j^*| - |\sqrt{m}\beta_j^*|) \\ &\quad + \frac{\eta_n}{n\sqrt{m^{\gamma_2+1}}} \sum_j |z_j + \sqrt{m}\beta_j^*|^{\gamma_2} (|u_j - z_j| - |z_j|) \end{aligned}$$

$$(172) \quad \begin{aligned} &= \frac{1}{m} u^\top \left(\frac{1}{n} X^\top X \right) u - \frac{2}{\sqrt{nm}} u^\top \left(\frac{1}{\sqrt{n}} X^\top \varepsilon \right) \\ &\quad + \frac{\lambda_n}{n\sqrt{m}} \sum_j \left(\frac{|u_j + \sqrt{m}\beta_j^*| - |\sqrt{m}\beta_j^*|}{|\beta_j^* + \frac{z_j}{\sqrt{m}}|^{\gamma_1}} I(\beta_j^* \neq 0) + \frac{\sqrt{m^{\gamma_1}} |u_j|}{|z_j|^{\gamma_1}} I(\beta_j^* = 0) \right) \\ &\quad + \frac{\eta_n}{n\sqrt{m}} \sum_j \left(\left| \beta_j^* + \frac{z_j}{\sqrt{m}} \right|^{\gamma_2} I(\beta_j^* \neq 0) + \frac{|z_j|^{\gamma_2}}{\sqrt{m^{\gamma_2}}} I(\beta_j^* = 0) \right) (|u_j - z_j| - |z_j|). \end{aligned}$$

Consider the case (i) where

$$(173) \quad \frac{\eta_n}{n\sqrt{m^{\gamma_2+1}}} \gg \frac{1}{\sqrt{nm}} \text{ and } \frac{\eta_n}{n\sqrt{m^{\gamma_2+1}}} \gg \frac{\sqrt{m^{\gamma_1-1}} \lambda_n}{n},$$

that is,

$$(174) \quad \frac{\eta_n}{\sqrt{nm^{\gamma_2}}} \rightarrow \infty \text{ and } \frac{\sqrt{m^{\gamma_1+\gamma_2}}\lambda_n}{\eta_n} \rightarrow 0.$$

Let $V_n(u) := (n\sqrt{m^{\gamma_2+1}}/\eta_n)(Z_n^\#(\beta) - Z_n^\#(\beta^*)) + (\text{term irrelevant to } \beta)$, then we have

$$(175) \quad \begin{aligned} V_n(u) = & \frac{\sqrt{nm^{\gamma_2}}}{\eta_n} \sqrt{\frac{n}{m}} u^\top \left(\frac{1}{n} X^\top X \right) u - \frac{2\sqrt{nm^{\gamma_2}}}{\eta_n} u^\top \left(\frac{1}{\sqrt{n}} X^\top \varepsilon \right) \\ & + \frac{\sqrt{m^{\gamma_1+\gamma_2}}\lambda_n}{\eta_n} \sum_j \left(\frac{|u_j + \sqrt{m}\beta_j^*| - |\sqrt{m}\beta_j^*|}{|z_j + \sqrt{m}\beta_j^*|^{\gamma_1}} I(\beta_j^* \neq 0) + \frac{|u_j|}{|z_j|^{\gamma_1}} I(\beta_j^* = 0) \right) \\ & + \sum_j (|z_j + \sqrt{m}\beta_j^*|^{\gamma_2} I(\beta_j^* \neq 0) + |z_j|^{\gamma_2} I(\beta_j^* = 0)) |u_j - z_j|. \end{aligned}$$

Let $V(u) := \lim_{n \rightarrow \infty} V_n(u)$, then we have

$$(176) \quad V(u) = \begin{cases} \sum_j |z_j|^{\gamma_2} |u_j - z_j| I(\beta_j^* = 0) & \text{for } u_S = z_S, \\ \infty & \text{otherwise.} \end{cases}$$

Therefore, we obtain

$$(177) \quad \sqrt{m}(\hat{\beta}_n^\# - \beta^*) = \operatorname{argmin}_u V_n(u) \xrightarrow{d} \operatorname{argmin}_u V(u) = z.$$

Consider the case (ii) where

$$(178) \quad \frac{\sqrt{m^{\gamma_1-1}}\lambda_n}{n} \gg \frac{1}{\sqrt{nm}}, \quad \frac{\sqrt{m^{\gamma_1-1}}\lambda_n}{n} \gg \frac{\eta_n}{n\sqrt{m^{\gamma_2+1}}},$$

$$(179) \quad \frac{\eta_n}{n\sqrt{m}} \gg \frac{1}{\sqrt{nm}}, \quad \frac{\eta_n}{n\sqrt{m}} \gg \frac{\lambda_n}{n\sqrt{m}},$$

that is,

$$(180) \quad \frac{\sqrt{m^{\gamma_1}}\lambda_n}{\sqrt{n}} \rightarrow \infty, \quad \frac{\eta_n}{\sqrt{n}} \rightarrow \infty, \quad \frac{\eta_n}{\lambda_n} \rightarrow \infty, \quad \frac{\eta_n}{\sqrt{m^{\gamma_1+\gamma_2}}\lambda_n} \rightarrow 0.$$

We divide into three cases:

$$(181) \quad (\text{ii-a}) \frac{\eta_n}{n\sqrt{m}} \gg \frac{\sqrt{m^{\gamma_1-1}}\lambda_n}{n}, \quad (\text{ii-b}) \frac{\eta_n}{n\sqrt{m}} \ll \frac{\sqrt{m^{\gamma_1-1}}\lambda_n}{n},$$

$$(182) \quad (\text{ii-c}) \frac{\eta_n}{n\sqrt{m}} \asymp \frac{\sqrt{m^{\gamma_1-1}}\lambda_n}{n},$$

that is,

$$(183) \quad (\text{ii-a}) \frac{\sqrt{m^{\gamma_1}}\lambda_n}{\eta_n} \rightarrow 0, \quad (\text{ii-b}) \frac{\sqrt{m^{\gamma_1}}\lambda_n}{\eta_n} \rightarrow \infty, \quad (\text{ii-c}) \frac{\sqrt{m^{\gamma_1}}\lambda_n}{\eta_n} \rightarrow \rho_0 > 0.$$

In the case (ii-a), let $V_n(u) := (n/\sqrt{m^{\gamma_1-1}}\lambda_n)(Z_n^\#(\beta) - Z_n^\#(\beta^*)) + (\text{term irrelevant to } \beta)$, then we have

$$(184) \quad V_n(u) = \frac{\sqrt{n}}{\sqrt{m^{\gamma_1}}\lambda_n} \sqrt{\frac{n}{m}} u^\top \left(\frac{1}{n} X^\top X \right) u - \frac{2\sqrt{n}}{\sqrt{m^{\gamma_1}}\lambda_n} u^\top \left(\frac{1}{\sqrt{n}} X^\top \varepsilon \right) \\ + \sum_j \left(\frac{|u_j + \sqrt{m}\beta_j^*| - |\sqrt{m}\beta_j^*|}{|z_j + \sqrt{m}\beta_j^*|^{\gamma_1}} I(\beta_j^* \neq 0) + \frac{|u_j|}{|z_j|^{\gamma_1}} I(\beta_j^* = 0) \right) \\ + \frac{\eta_n}{\sqrt{m^{\gamma_1}}\lambda_n} \sum_j \left(\left| \beta_j^* + \frac{z_j}{\sqrt{m}} \right|^{\gamma_2} I(\beta_j^* \neq 0) + \frac{|z_j|^{\gamma_2}}{\sqrt{m^{\gamma_2}}} I(\beta_j^* = 0) \right) |u_j - z_j|.$$

Let $V(u) := \lim_{n \rightarrow \infty} V_n(u)$, then we have

$$(185) \quad V(u) = \begin{cases} \sum_j \frac{|u_j|}{|z_j|^{\gamma_1}} I(\beta_j^* = 0) & \text{for } u_S = z_S, \\ \infty & \text{otherwise.} \end{cases}$$

Therefore, we obtain

$$(186) \quad \sqrt{m}(\hat{\beta}_n^\# - \beta^*) = \operatorname{argmin}_u V_n(u) \xrightarrow{d} \operatorname{argmin}_u V(u) = \begin{cases} 0 & \text{for } j \in S^c, \\ z_j & \text{for } j \in S. \end{cases}$$

In the case (ii-b), let $V_n(u) := (n\sqrt{m}/\eta_n)(Z_n^\#(\beta) - Z_n^\#(\beta^*)) + (\text{term irrelevant to } \beta)$,

then we have

(187)

$$\begin{aligned} V_n(u) = & \frac{\sqrt{n}}{\eta_n} \sqrt{\frac{n}{m}} u^\top \left(\frac{1}{n} X^\top X \right) u - \frac{2\sqrt{n}}{\eta_n} u^\top \left(\frac{1}{\sqrt{n}} X^\top \varepsilon \right) \\ & + \frac{\sqrt{m^{\gamma_1}} \lambda_n}{\eta_n} \sum_j \left(\frac{|u_j + \sqrt{m}\beta_j^*| - |\sqrt{m}\beta_j^*|}{|z_j + \sqrt{m}\beta_j^*|^{\gamma_1}} I(\beta_j^* \neq 0) + \frac{|u_j|}{|z_j|^{\gamma_1}} I(\beta_j^* = 0) \right) \\ & + \sum_j \left(\left| \beta_j^* + \frac{z_j}{\sqrt{m}} \right|^{\gamma_2} I(\beta_j^* \neq 0) + \frac{|z_j|^{\gamma_2}}{\sqrt{m^{\gamma_2}}} I(\beta_j^* = 0) \right) |u_j - z_j|. \end{aligned}$$

Let $V(u) := \lim_{n \rightarrow \infty} V_n(u)$, then we have

$$(188) \quad V(u) = \begin{cases} \sum_j |\beta_j^*|^{\gamma_2} |u_j - z_j| I(\beta_j^* \neq 0) & \text{for } u_{S^c} = 0, \\ \infty & \text{otherwise.} \end{cases}$$

Therefore, we obtain

$$(189) \quad \sqrt{m}(\hat{\beta}_n^\# - \beta^*) = \operatorname{argmin}_u V_n(u) \xrightarrow{d} \operatorname{argmin}_u V(u) = \begin{cases} 0 & \text{for } j \in S^c, \\ z_j & \text{for } j \in S. \end{cases}$$

In the case (ii-c), let $V_n(u) := (n\sqrt{m}/\eta_n)(Z_n^\#(\beta) - Z_n^\#(\beta^*)) + (\text{term irrelevant to } \beta)$, then we have

(190)

$$\begin{aligned} V_n(u) = & \frac{\sqrt{n}}{\eta_n} \sqrt{\frac{n}{m}} u^\top \left(\frac{1}{n} X^\top X \right) u - \frac{2\sqrt{n}}{\eta_n} u^\top \left(\frac{1}{\sqrt{n}} X^\top \varepsilon \right) \\ & + \frac{\sqrt{m^{\gamma_1}} \lambda_n}{\eta_n} \sum_j \left(\frac{|u_j + \sqrt{m}\beta_j^*| - |\sqrt{m}\beta_j^*|}{|z_j + \sqrt{m}\beta_j^*|^{\gamma_1}} I(\beta_j^* \neq 0) + \frac{|u_j|}{|z_j|^{\gamma_1}} I(\beta_j^* = 0) \right) \\ & + \sum_j \left(\left| \beta_j^* + \frac{z_j}{\sqrt{m}} \right|^{\gamma_2} I(\beta_j^* \neq 0) + \frac{|z_j|^{\gamma_2}}{\sqrt{m^{\gamma_2}}} I(\beta_j^* = 0) \right) |u_j - z_j|. \end{aligned}$$

Let $V(u) := \lim_{n \rightarrow \infty} V_n(u)$, then we have

$$(191) \quad V(u) = \sum_j \left(|\beta_j^*|^{\gamma_2} |u_j - z_j| I(\beta_j^* \neq 0) + \frac{\rho_0 |u_j|}{|z_j|^{\gamma_1}} I(\beta_j^* = 0) \right).$$

Therefore, we obtain

$$(192) \quad \sqrt{m}(\hat{\beta}_n^\# - \beta^*) = \underset{u}{\operatorname{argmin}} V_n(u) \xrightarrow{d} \underset{u}{\operatorname{argmin}} V(u) = \begin{cases} 0 & \text{for } j \in S^c, \\ z_j & \text{for } j \in S. \end{cases}$$

Case II. Let $l = \sqrt{n}$. Then, (169) reduces to

(193)

$$\begin{aligned} & Z_n^\#(\beta) - Z_n^\#(\beta^*) \\ (194) \quad &= \frac{1}{n} u^\top \left(\frac{1}{n} X^\top X \right) u - \frac{2}{n} u^\top \left(\frac{1}{\sqrt{n}} X^\top \varepsilon \right) \\ &+ \frac{\sqrt{m^{\gamma_1}} \lambda_n}{n \sqrt{n}} \sum_j \frac{1}{|z_j + \sqrt{m} \beta_j^*|^{\gamma_1}} (|u_j + \sqrt{n} \beta_j^*| - |\sqrt{n} \beta_j^*|) \\ &+ \frac{\eta_n}{n \sqrt{nm^{\gamma_2}}} \sum_j |z_j + \sqrt{m} \beta_j^*|^{\gamma_2} \left(\left| u_j - \sqrt{\frac{n}{m}} z_j \right| - \left| \sqrt{\frac{n}{m}} z_j \right| \right) \\ (195) \quad &= \frac{1}{n} u^\top \left(\frac{1}{n} X^\top X \right) u - \frac{2}{n} u^\top \left(\frac{1}{\sqrt{n}} X^\top \varepsilon \right) \\ &+ \frac{\lambda_n}{n \sqrt{n}} \sum_j \left(\frac{|u_j + \sqrt{n} \beta_j^*| - |\sqrt{n} \beta_j^*|}{|\beta_j^* + \frac{z_j}{\sqrt{m}}|^{\gamma_1}} I(\beta_j^* \neq 0) + \frac{\sqrt{m^{\gamma_1}} |u_j|}{|z_j|^{\gamma_1}} I(\beta_j^* = 0) \right) \\ &+ \frac{\eta_n}{n \sqrt{n}} \sum_j \left(\left| \beta_j^* + \frac{z_j}{\sqrt{m}} \right|^{\gamma_2} I(\beta_j^* \neq 0) + \frac{|z_j|^{\gamma_2}}{\sqrt{m^{\gamma_2}}} I(\beta_j^* = 0) \right) \left(\left| u_j - \sqrt{\frac{n}{m}} z_j \right| - \left| \sqrt{\frac{n}{m}} z_j \right| \right). \end{aligned}$$

Consider the case (iii) where

$$(196) \quad \frac{1}{n} \gtrsim \frac{\sqrt{m^{\gamma_1}} \lambda_n}{n \sqrt{n}} \text{ and } \frac{1}{n} \gtrsim \frac{\eta_n}{n \sqrt{n}},$$

that is,

$$(197) \quad \frac{\sqrt{m^{\gamma_1}} \lambda_n}{\sqrt{n}} \rightarrow \lambda_1 \geq 0 \text{ and } \frac{\eta_n}{\sqrt{n}} \rightarrow \eta_0 \geq 0.$$

Let $V_n(u) := n(Z_n^\#(\beta) - Z_n^\#(\beta^*))$, then we have

(198)

$$\begin{aligned} V_n(u) &= u^\top \left(\frac{1}{n} X^\top X \right) u - 2u^\top \left(\frac{1}{\sqrt{n}} X^\top \varepsilon \right) \\ &\quad + \frac{\lambda_n}{\sqrt{n}} \sum_j \left(\frac{|u_j + \sqrt{n}\beta_j^*| - |\sqrt{n}\beta_j^*|}{|\beta_j^* + \frac{z_j}{\sqrt{m}}|^{\gamma_1}} I(\beta_j^* \neq 0) + \frac{\sqrt{m^{\gamma_1}}|u_j|}{|z_j|^{\gamma_1}} I(\beta_j^* = 0) \right) \\ &\quad + \frac{\eta_n}{\sqrt{n}} \sum_j \left(\left| \beta_j^* + \frac{z_j}{\sqrt{m}} \right|^{\gamma_2} I(\beta_j^* \neq 0) + \frac{|z_j|^{\gamma_2}}{\sqrt{m^{\gamma_2}}} I(\beta_j^* = 0) \right) \left(\left| u_j - \sqrt{\frac{n}{m}} z_j \right| - \left| \sqrt{\frac{n}{m}} z_j \right| \right). \end{aligned}$$

Let $V(u) := \lim_{n \rightarrow \infty} V_n(u)$, then we have

(199)

$$V(u) = u^\top C u - 2u^\top W$$

(200)

$$+ \sum_j \left(\frac{\lambda_1 |u_j|}{|z_j|^{\gamma_1}} I(\beta_j^* = 0) + \eta_0 |\beta_j^*|^{\gamma_2} (|u_j - \sqrt{r_0} z_j| - |\sqrt{r_0} z_j|) I(\beta_j^* \neq 0) \right).$$

Therefore, we obtain

(201)

$$\begin{aligned} \sqrt{n}(\hat{\beta}_n^\# - \beta^*) &= \operatorname{argmin}_u V_n(u) \xrightarrow{d} \operatorname{argmin}_u V(u) \\ (202) \quad &= \operatorname{argmin}_u \left\{ u^\top C u - 2u^\top W + \sum_j \left(\frac{\lambda_1 |u_j|}{|z_j|^{\gamma_1}} I(\beta_j^* = 0) + \eta_0 |\beta_j^*|^{\gamma_2} (|u_j - \sqrt{r_0} z_j| - |\sqrt{r_0} z_j|) I(\beta_j^* \neq 0) \right) \right\}. \end{aligned}$$

Consider the case (iv) where

$$(203) \quad \frac{1}{n} \gtrsim \frac{\sqrt{m^{\gamma_1}} \lambda_n}{n \sqrt{n}} \text{ and } \frac{\eta_n}{n \sqrt{n}} \gg \frac{1}{n} \gtrsim \frac{\eta_n}{n \sqrt{nm^{\gamma_2}}},$$

that is,

$$(204) \quad \frac{\sqrt{m^{\gamma_1}} \lambda_n}{\sqrt{n}} \rightarrow \lambda_1 \geq 0, \quad \frac{\eta_n}{\sqrt{n}} \rightarrow \infty, \quad \text{and} \quad \frac{\eta_n}{\sqrt{nm^{\gamma_2}}} \rightarrow \eta_1 \geq 0.$$

Let $V_n(u) := n(Z_n^\#(\beta) - Z_n^\#(\beta^*)) + (\text{term irrelevant to } \beta)$ and $V(u) := \lim_{n \rightarrow \infty} V_n(u)$. Then, we have (198) and

$$(205) \quad V(u) = \begin{cases} u^\top Cu - 2u^\top W + \sum_j \left(\frac{\lambda_1}{|z_j|^{\gamma_1}} |u_j| + \eta_1 |z_j|^{\gamma_2} |u_j - \sqrt{r_0} z_j| \right) I(\beta_j^* = 0) & \text{for } u_S = r_0 z_S, \\ \infty & \text{otherwise.} \end{cases}$$

Therefore, we obtain

$$(206) \quad \sqrt{n}(\hat{\beta}_n^\# - \beta^*) = \operatorname{argmin}_u V_n(u)$$

$$(207) \quad \xrightarrow{d} \operatorname{argmin}_u V(u) = \operatorname{argmin}_{u \in \mathcal{U}} \left\{ u^\top Cu - 2u^\top W + \sum_j \left(\frac{\lambda_1}{|z_j|^{\gamma_1}} |u_j| + \eta_1 |z_j|^{\gamma_2} |u_j - \sqrt{r_0} z_j| \right) \right\},$$

$$(208) \quad \mathcal{U} := \{u \mid u_S = r_0 z_S\}.$$

Consider the case (v) where

$$(209) \quad \frac{\sqrt{m^{\gamma_1}} \lambda_n}{n\sqrt{n}} \gg \frac{1}{n} \gtrsim \frac{\lambda_n}{n\sqrt{n}} \text{ and } \frac{1}{n} \gtrsim \frac{\eta_n}{n\sqrt{n}},$$

that is,

$$(210) \quad \frac{\sqrt{m^{\gamma_1}} \lambda_n}{\sqrt{n}} \rightarrow \infty, \frac{\lambda_n}{\sqrt{n}} \rightarrow \lambda_0 \geq 0, \text{ and } \frac{\eta_n}{\sqrt{n}} \rightarrow \eta_0 \geq 0.$$

Let $V_n(u) := n(Z_n^\#(\beta) - Z_n^\#(\beta^*))$ and $V(u) := \lim_{n \rightarrow \infty} V_n(u)$. Then, we have (198) and

$$(211) \quad V(u) = \begin{cases} u^\top Cu - 2u^\top W + \sum_j (\lambda_0 u_j \operatorname{sgn}(\beta_j^*) / |\beta_j^*|^{\gamma_1} + \eta_0 |\beta_j^*|^{\gamma_2} |u_j - \sqrt{r_0} z_j|) & \text{for } u_{S^c} = 0, \\ \infty & \text{otherwise.} \end{cases}$$

Therefore, we obtain

(212)

$$\sqrt{n}(\hat{\beta}_n^\# - \beta^*) = \operatorname{argmin}_u V_n(u)$$

(213)

$$\xrightarrow{d} \operatorname{argmin}_u V(u) = \operatorname{argmin}_{u \in \mathcal{U}} \left\{ u^\top C u - 2u^\top W + \sum_j \left(\frac{\lambda_0 u_j \operatorname{sgn}(\beta_j^*)}{|\beta_j^*|^{\gamma_1}} + \eta_0 |\beta_j^*|^{\gamma_2} |u_j - \sqrt{r_0} z_j| \right) \right\},$$

(214) $\mathcal{U} := \{u \mid u_{S^c} = 0\}.$

Case III. Let $l = n/\lambda_n$. Suppose that $n/\lambda_n \rightarrow \infty$. Then, (169) reduces to

(215)

$$Z_n^\#(\beta) - Z_n^\#(\beta^*)$$

(216)

$$\begin{aligned} &= \frac{\lambda_n^2}{n^2} u^\top \left(\frac{1}{n} X^\top X \right) u - \frac{2\lambda_n}{n\sqrt{n}} u^\top \left(\frac{1}{\sqrt{n}} X^\top \varepsilon \right) \\ &+ \frac{\sqrt{m^{\gamma_1}} \lambda_n^2}{n^2} \sum_j \frac{1}{|z_j + \sqrt{m}\beta_j^*|^{\gamma_1}} \left(\left| u_j + \frac{n}{\lambda_n} \beta_j^* \right| - \left| \frac{n}{\lambda_n} \beta_j^* \right| \right) \\ &+ \frac{\lambda_n \eta_n}{n^2 \sqrt{m^{\gamma_2}}} \sum_j |z_j + \sqrt{m}\beta_j^*|^{\gamma_2} \left(\left| u_j - \frac{n}{\sqrt{m}\lambda_n} z_j \right| - \left| \frac{n}{\sqrt{m}\lambda_n} z_j \right| \right) \end{aligned}$$

(217)

$$\begin{aligned} &= \frac{\lambda_n^2}{n^2} u^\top \left(\frac{1}{n} X^\top X \right) u - \frac{2\lambda_n}{n\sqrt{n}} u^\top \left(\frac{1}{\sqrt{n}} X^\top \varepsilon \right) \\ &+ \frac{\lambda_n^2}{n^2} \sum_j \left(\frac{\left| u_j + \frac{n}{\lambda_n} \beta_j^* \right| - \left| \frac{n}{\lambda_n} \beta_j^* \right|}{\left| \beta_j^* + \frac{z_j}{\sqrt{m}} \right|^{\gamma_1}} I(\beta_j^* \neq 0) + \frac{\sqrt{m^{\gamma_1}} |u_j|}{|z_j|^{\gamma_1}} I(\beta_j^* = 0) \right) \\ &+ \frac{\lambda_n \eta_n}{n^2} \sum_j \left(\left| \beta_j^* + \frac{z_j}{\sqrt{m}} \right|^{\gamma_2} I(\beta_j^* \neq 0) + \frac{|z_j|^{\gamma_2}}{\sqrt{m^{\gamma_2}}} I(\beta_j^* = 0) \right) \left(\left| u_j - \frac{n}{\sqrt{m}\lambda_n} z_j \right| - \left| \frac{n}{\sqrt{m}\lambda_n} z_j \right| \right). \end{aligned}$$

Consider the case (vi) where

$$(218) \quad \frac{\lambda_n^2}{n^2} \gg \frac{\lambda_n}{n\sqrt{n}} \text{ and } \frac{\lambda_n^2}{n^2} \gg \frac{\lambda_n \eta_n}{n^2},$$

that is,

$$(219) \quad \frac{\lambda_n}{\sqrt{n}} \rightarrow \infty, \quad \frac{\lambda_n}{n} \rightarrow 0, \quad \text{and} \quad \frac{\lambda_n}{\eta_n} \rightarrow \infty.$$

Let $V_n(u) := (n^2/\lambda_n^2)(Z_n^\#(\beta) - Z_n^\#(\beta^*))$, then we have

(220)

$$\begin{aligned} V_n(u) = & u^\top \left(\frac{1}{n} X^\top X \right) u - \frac{2\sqrt{n}}{\lambda_n} u^\top \left(\frac{1}{\sqrt{n}} X^\top \varepsilon \right) \\ & + \sum_j \left(\frac{\left| u_j + \frac{n}{\lambda_n} \beta_j^* \right| - \left| \frac{n}{\lambda_n} \beta_j^* \right|}{\left| \beta_j^* + \frac{z_j}{\sqrt{m}} \right|^{\gamma_1}} I(\beta_j^* \neq 0) + \frac{\sqrt{m^{\gamma_1}} |u_j|}{|z_j|^{\gamma_1}} I(\beta_j^* = 0) \right) \\ & + \frac{\eta_n}{\lambda_n} \sum_j \left(\left| \beta_j^* + \frac{z_j}{\sqrt{m}} \right|^{\gamma_2} I(\beta_j^* \neq 0) + \frac{|z_j|^{\gamma_2}}{\sqrt{m^{\gamma_2}}} I(\beta_j^* = 0) \right) \left(\left| u_j - \frac{n}{\sqrt{m}\lambda_n} z_j \right| - \left| \frac{n}{\sqrt{m}\lambda_n} z_j \right| \right). \end{aligned}$$

Let $V(u) := \lim_{n \rightarrow \infty} V_n(u)$, then we have

$$(221) \quad V(u) = \begin{cases} u^\top C u + \sum_j \frac{u_j \operatorname{sgn}(\beta_j^*)}{|\beta_j^*|^{\gamma_1}} I(\beta_j^* \neq 0) & \text{for } u_{S^c} = 0, \\ \infty & \text{otherwise.} \end{cases}$$

Therefore, we obtain

$$(222) \quad \frac{n}{\lambda_n} (\hat{\beta}_n^\# - \beta^*) = \operatorname{argmin}_u V_n(u) \xrightarrow{d} \operatorname{argmin}_u V(u) = \operatorname{argmin}_{u \in \mathcal{U}} \left\{ u^\top C u + \sum_j \frac{\operatorname{sgn}(\beta_j^*)}{|\beta_j^*|^{\gamma_1}} u_j \right\},$$

$$(223) \quad \mathcal{U} := \{u \mid u_{S^c} = 0\}.$$

B.3.2 Proof of Theorem 4.3

Suppose that $\tilde{\beta}$ is a \sqrt{m} -consistent estimator. Suppose that $\hat{\beta}$ is l -consistent and let $\hat{u} := l(\hat{\beta} - \beta^*)$ where $l = l(n, m, \lambda_n)$ is a certain function as defined later. Then, we have

$$(224) \quad \forall j \in S, \quad P(j \in \operatorname{supp}(\hat{\beta})) \rightarrow 1.$$

Consider the event $\beta_j^* \in S^c$ and $\hat{\beta}_j \neq 0$. By the KKT conditions, for $\beta_j^* = 0$ and $\hat{\beta}_j \neq \tilde{\beta}_j$,

$$(225) \quad 2 \left(\frac{1}{n} \mathbf{x}_j^\top X \right) \frac{\sqrt{n}}{l} \hat{u} - \frac{2}{\sqrt{n}} \mathbf{x}_j^\top \varepsilon + \frac{\sqrt{m^{\gamma_1}} \lambda_n}{\sqrt{n}} \frac{\operatorname{sgn}(\hat{\beta}_j)}{|\sqrt{m}\tilde{\beta}_j|^{\gamma_1}} + \frac{\eta_n}{\sqrt{nm^{\gamma_2}}} \left| \sqrt{m}\tilde{\beta}_j \right|^{\gamma_2} \operatorname{sgn}(\hat{\beta}_j - \tilde{\beta}_j) = 0,$$

and for $\beta_j^* = 0$ and $\hat{\beta}_j = \tilde{\beta}_j$,

$$(226) \quad \left| 2 \left(\frac{1}{n} \mathbf{x}_j^\top X \right) \frac{\sqrt{n}}{l} \hat{u} - \frac{2}{\sqrt{n}} \mathbf{x}_j^\top \varepsilon + \frac{\sqrt{m^{\gamma_1}} \lambda_n}{\sqrt{n}} \frac{\operatorname{sgn}(\hat{\beta}_j)}{|\sqrt{m}\tilde{\beta}_j|^{\gamma_1}} \right| \leq \frac{\eta_n}{\sqrt{nm^{\gamma_2}}} \left| \sqrt{m}\tilde{\beta}_j \right|^{\gamma_2}.$$

Suppose that $\hat{\beta}$ is l -consistent and that $\sqrt{m^{\gamma_1}} \lambda_n / \sqrt{n} \gg 1$, $\sqrt{m^{\gamma_1}} \lambda_n / \sqrt{n} \gg \sqrt{n}/l$, and $\sqrt{m^{\gamma_1}} \lambda_n / \sqrt{n} \gg \eta_n / \sqrt{nm^{\gamma_2}}$. For $l = \sqrt{m}, \sqrt{n}, n/\lambda_n$, these conditions reduce to the conditions where $\hat{\beta}$ is l -consistent, $\sqrt{m^{\gamma_1}} \lambda_n / \sqrt{n} \rightarrow \infty$, and $\sqrt{m^{\gamma_1+\gamma_2}} \lambda_n / \eta_n \rightarrow \infty$. Then, in both $\hat{\beta}_j \neq \tilde{\beta}_j$ and $\hat{\beta}_j = \tilde{\beta}_j$ cases, the term including $\sqrt{m^{\gamma_1}} / n \lambda_n$ diverges to infinity faster compared to the rest terms. Therefore, we have

$$(227) \quad \forall j \in S^c, \lim_{n \rightarrow \infty} P(j \in \operatorname{supp}(\hat{\beta})) = 0.$$

This concludes

$$(228) \quad \lim_{n \rightarrow \infty} P(\hat{S}_n = S) = 1.$$

B.3.3 Proof of Theorem 4.4

Suppose that $\tilde{\beta}$ is a \sqrt{m} -consistent estimator. Consider the event where $\beta_j^* \neq 0$ and $\hat{\beta}_j \neq \tilde{\beta}_j$. Suppose that $\hat{\beta}$ is l -consistent and let $\hat{u} := l(\hat{\beta} - \beta^*)$ where $l = l(n, m, \lambda_n)$ is a certain function as defined later. By the KKT conditions, for $\beta_j^* \neq 0$ and $\hat{\beta}_j \neq 0$,

$$(229) \quad 2 \left(\frac{1}{n} \mathbf{x}_j^\top X \right) \frac{\sqrt{n}}{l} \hat{u} - \frac{2}{\sqrt{n}} \mathbf{x}_j^\top \varepsilon + \frac{\lambda_n}{\sqrt{n}} \frac{\operatorname{sgn}(\hat{\beta}_j)}{|\tilde{\beta}_j|^{\gamma_1}} + \frac{\eta_n}{\sqrt{n}} \left| \tilde{\beta}_j \right|^{\gamma_2} \operatorname{sgn}(\hat{\beta}_j - \tilde{\beta}_j) = 0,$$

and for $\beta_j^* \neq 0$ and $\hat{\beta}_j = 0$,

$$(230) \quad \left| 2 \left(\frac{1}{n} \mathbf{x}_j^\top X \right) \frac{\sqrt{n}}{l} \hat{u} - \frac{2}{\sqrt{n}} \mathbf{x}_j^\top \varepsilon + \frac{\eta_n}{\sqrt{n}} \left| \tilde{\beta}_j \right|^{\gamma_2} \operatorname{sgn}(\hat{\beta}_j - \tilde{\beta}_j) \right| \leq \frac{\lambda_n}{\sqrt{n}} \frac{1}{|\tilde{\beta}_j|^{\gamma_1}}.$$

Suppose that $\hat{\beta}$ is l -consistent and that $\eta_n/\sqrt{n} \gg 1$, $\eta_n/\sqrt{n} \gg \sqrt{n}/l$, and $\eta_n/\sqrt{n} \gg \lambda_n/\sqrt{n}$. For $l = \sqrt{m}, \sqrt{n}$, these conditions reduce to the conditions where $\hat{\beta}$ is l -consistent, $\eta_n/\sqrt{n} \rightarrow \infty$, and $\eta_n/\lambda_n \rightarrow \infty$. Then, in both $\hat{\beta}_j \neq 0$ and $\hat{\beta}_j = 0$ cases, the term including η_n/\sqrt{n} diverges to infinity faster compared to the rest terms. Therefore, we have

$$(231) \quad \forall j \in S, \lim_{n \rightarrow \infty} P(\hat{\beta}_j \neq \tilde{\beta}_j) = 0.$$

This concludes

$$(232) \quad \forall j \in S, \lim_{n \rightarrow \infty} P(\hat{\beta}_j = \tilde{\beta}_j) = 1.$$

B.4 Proofs for Transfer Lasso with Deterministic Source Parameters

B.4.1 Proof of Theorem A.3

For $\tilde{\beta}_j = 0$, we consider the event $\hat{\beta}_j \neq 0$. Then, we have by KKT conditions

$$(233) \quad 2 \left(\frac{1}{n} \mathbf{x}_j^\top X \right) \sqrt{n}(\beta^* - \hat{\beta}) + 2 \frac{1}{\sqrt{n}} \mathbf{x}_j^\top \varepsilon + \frac{\lambda_n + \eta_n}{\sqrt{n}} \operatorname{sgn}(\hat{\beta}_j) = 0$$

The first and second terms converge to some truncated Gaussian-mixture distribution, and the third term diverges to infinity as $(\lambda_n + \eta_n)/\sqrt{n} \rightarrow \infty$, which is a contradiction. Hence, we have for $\tilde{\beta}_j = 0$

$$(234) \quad \lim_{n \rightarrow \infty} P(\hat{\beta}_j = 0) = 1$$

For $\tilde{\beta}_j \neq 0$, we consider the event $\hat{\beta}_j \neq 0$ and $\hat{\beta}_j \neq \tilde{\beta}_j$. Then, we have by KKT conditions,

$$(235) \quad 2 \left(\frac{1}{n} \mathbf{x}_j^\top X \right) \sqrt{n}(\beta^* - \hat{\beta}) + 2 \frac{1}{\sqrt{n}} \mathbf{x}_j^\top \varepsilon + \frac{\lambda_n}{\sqrt{n}} \operatorname{sgn}(\hat{\beta}_j) + \frac{\eta_n}{\sqrt{n}} \operatorname{sgn}(\hat{\beta}_j - \tilde{\beta}_j) = 0$$

The third and fourth terms diverge to infinity as $(\lambda_n + \eta_n)/\sqrt{n} \rightarrow \infty$ if $\operatorname{sgn}(\hat{\beta}_j) = \operatorname{sgn}(\hat{\beta}_j - \tilde{\beta}_j)$, which induces a contradiction. Hence, we have for $\hat{\beta}_j \neq 0$

$$(236) \quad \lim_{n \rightarrow \infty} P(\min \{0, \tilde{\beta}_j\} \leq \hat{\beta}_j \leq \max \{0, \tilde{\beta}_j\}) = 1$$

B.4.2 Proof of Theorem A.4

Asymptotic distribution: Let $u := \sqrt{n}(\beta - \beta^*)$. We have

(237)

$$\begin{aligned} Z_n^T(\beta; \tilde{\beta}, \lambda_n, \eta_n) &:= \frac{1}{n} \|y - X\beta\|_2^2 + \frac{\lambda_n}{n} \sum_j |\beta_j| + \frac{\eta_n}{n} |\beta_j - \tilde{\beta}_j| \\ (238) \quad &= \frac{1}{n} \left\| \varepsilon - \frac{1}{\sqrt{n}} Xu \right\|_2^2 + \frac{\lambda_n}{n} \sum_j \left| \beta_j^* + \frac{u_j}{\sqrt{n}} \right| + \frac{\eta_n}{n} \sum_j \left| \beta_j^* - \tilde{\beta}_j + \frac{u_j}{\sqrt{n}} \right| \end{aligned}$$

and

(239)

$$\begin{aligned} Z_n^T(\beta; \tilde{\beta}, \lambda_n, \eta_n) - Z_n^T(\beta^*; \tilde{\beta}, \lambda_n, \eta_n) \\ (240) \quad = \left\| \varepsilon - \frac{1}{\sqrt{n}} Xu \right\|_2^2 - \|\varepsilon\|_2^2 \\ (241) \quad + \lambda_n \sum_j \left(\left| \beta_j^* + \frac{u_j}{\sqrt{n}} \right| - \left| \beta_j^* \right| \right) + \eta_n \sum_j \left(\left| \beta_j^* - \tilde{\beta}_j + \frac{u_j}{\sqrt{n}} \right| - \left| \beta_j^* - \tilde{\beta}_j \right| \right) \end{aligned}$$

(242)

$$= u^\top \left(\frac{1}{n} X^\top X \right) u - 2u^\top \left(\frac{1}{\sqrt{n}} X^\top \varepsilon \right)$$

(243)

$$+ \lambda_n \sum_j \left(\left| \beta_j^* + \frac{u_j}{\sqrt{n}} \right| - \left| \beta_j^* \right| \right) + \eta_n \sum_j \left(\left| \beta_j^* - \tilde{\beta}_j + \frac{u_j}{\sqrt{n}} \right| - \left| \beta_j^* - \tilde{\beta}_j \right| \right)$$

The first and second terms are the same as (97) and (98). We consider the third and fourth terms.

(244)

$$\lambda_n \left(\left| \beta_j^* + \frac{u_j}{\sqrt{n}} \right| - \left| \beta_j^* \right| \right) = \begin{cases} \frac{\lambda_n}{\sqrt{n}} |u_j| & \text{if } \beta_j^* = 0 \\ \frac{\lambda_n}{\sqrt{n}} u_j \operatorname{sgn}(\beta_j^*) & \text{if } \beta_j^* \neq 0 \text{ and } \beta_j^*(\beta_j^* + \frac{u_j}{\sqrt{n}}) \geq 0 \\ -\frac{\lambda_n}{\sqrt{n}} u_j \operatorname{sgn}(\beta_j^*) - 2\lambda_n |\beta_j^*| & \text{if } \beta_j^* \neq 0 \text{ and } \beta_j^*(\beta_j^* + \frac{u_j}{\sqrt{n}}) < 0 \end{cases}$$

(245)

$$\begin{aligned}
& \eta_n \left(\left| \beta_j^* - \tilde{\beta}_j + \frac{u_j}{\sqrt{n}} \right| - \left| \beta_j^* - \tilde{\beta}_j \right| \right) \\
(246) \quad &= \begin{cases} \frac{\eta_n}{\sqrt{n}} |u_j| & \text{if } \beta_j^* = \tilde{\beta}_j \\ \frac{\eta_n}{\sqrt{n}} u_j \operatorname{sgn}(\beta_j^* - \tilde{\beta}_j) & \text{if } \beta_j^* \neq \tilde{\beta}_j \text{ and } (\beta_j^* - \tilde{\beta}_j)(\beta_j^* - \tilde{\beta}_j + \frac{u_j}{\sqrt{n}}) \geq 0 \\ -\frac{\eta_n}{\sqrt{n}} u_j \operatorname{sgn}(\beta_j^* - \tilde{\beta}_j) - 2\eta_n |\beta_j^* - \tilde{\beta}_j| & \text{if } \beta_j^* \neq \tilde{\beta}_j \text{ and } (\beta_j^* - \tilde{\beta}_j)(\beta_j^* - \tilde{\beta}_j + \frac{u_j}{\sqrt{n}}) < 0 \end{cases}
\end{aligned}$$

For large n such that ($n > u_j^2/\beta_j^{*2}$ for $\forall j \in S$) and ($n > u_j^2/(\beta_j^* - \tilde{\beta}_j)^2$ for $\forall j$ such that $\beta_j^* \neq \tilde{\beta}_j$), we have

$$\begin{aligned}
(247) \quad & \mathcal{R}_j := \lambda_n \left(\left| \beta_j^* + \frac{u_j}{\sqrt{n}} \right| - \left| \beta_j^* \right| \right) + \eta_n \left(\left| \beta_j^* - \tilde{\beta}_j + \frac{u_j}{\sqrt{n}} \right| - \left| \beta_j^* - \tilde{\beta}_j \right| \right) \\
(248) \quad &= \begin{cases} \frac{\lambda_n + \eta_n}{\sqrt{n}} |u_j| & \text{if } \beta_j^* = 0, \tilde{\beta}_j = 0 \\ \frac{\lambda_n}{\sqrt{n}} |u_j| - \frac{\eta_n}{\sqrt{n}} u_j \operatorname{sgn}(\tilde{\beta}_j) & \text{if } \beta_j^* = 0, \tilde{\beta}_j \neq 0 \\ \frac{\lambda_n}{\sqrt{n}} u_j \operatorname{sgn}(\beta_j^*) + \frac{\eta_n}{\sqrt{n}} |u_j| & \text{if } \beta_j^* \neq 0, \tilde{\beta}_j = \beta_j^* \\ \frac{\lambda_n}{\sqrt{n}} u_j \operatorname{sgn}(\beta_j^*) + \frac{\eta_n}{\sqrt{n}} u_j \operatorname{sgn}(\beta_j^* - \tilde{\beta}_j) & \text{if } \beta_j^* \neq 0, \tilde{\beta}_j \neq \beta_j^* \end{cases}
\end{aligned}$$

Assume $(\lambda_n + \eta_n)/\sqrt{n} \rightarrow \delta_1$ and $(\lambda_n - \eta_n)/\sqrt{n} \rightarrow \delta_2$. For $\beta_j^* = 0$ and $\tilde{\beta}_j = 0$,

$$(249) \quad \mathcal{R}_j \rightarrow \begin{cases} 0 & \text{if } u_j = 0 \\ \delta_1 |u_j| & \text{if } u_j \neq 0 \end{cases}$$

For $\beta_j^* = 0, \tilde{\beta}_j \neq 0$,

$$(250) \quad \mathcal{R}_j \rightarrow \begin{cases} 0 & \text{if } u_j = 0 \\ \delta_2 & \text{if } u_j \tilde{\beta}_j > 0 \\ \delta_1 |u_j| & \text{if } u_j \tilde{\beta}_j < 0 \end{cases}$$

For $\beta_j^* \neq 0, \tilde{\beta}_j = \beta_j^*$,

$$(251) \quad \mathcal{R}_j \rightarrow \begin{cases} 0 & \text{if } u_j = 0 \\ \delta_1 & \text{if } u_j \beta_j^* > 0 \\ -\delta_2 |u_j| & \text{if } u_j \beta_j^* < 0 \end{cases}$$

For $\beta_j^* \neq 0, \tilde{\beta}_j \neq \beta_j^*$,

$$(252) \quad \mathcal{R}_j \rightarrow \begin{cases} 0 & \text{if } u_j = 0 \\ \delta_2|u_j| & \text{if } \operatorname{sgn}(\beta_j^*) \neq \operatorname{sgn}(\beta_j^* - \tilde{\beta}_j) \text{ and } u_j \operatorname{sgn}(\beta_j^*) > 0 \\ -\delta_2|u_j| & \text{if } \operatorname{sgn}(\beta_j^*) \neq \operatorname{sgn}(\beta_j^* - \tilde{\beta}_j) \text{ and } u_j \operatorname{sgn}(\beta_j^*) < 0 \\ \delta_1|u_j| & \text{if } \operatorname{sgn}(\beta_j^*) = \operatorname{sgn}(\beta_j^* - \tilde{\beta}_j) \text{ and } u_j \operatorname{sgn}(\beta_j^*) > 0 \\ -\delta_1|u_j| & \text{if } \operatorname{sgn}(\beta_j^*) = \operatorname{sgn}(\beta_j^* - \tilde{\beta}_j) \text{ and } u_j \operatorname{sgn}(\beta_j^*) < 0 \end{cases}$$

Furthermore, we assume $\delta_1 = \infty$ and $\delta_2 = 0$. Then, for $\beta_j^* = 0$ and $\tilde{\beta}_j = 0$,

$$(253) \quad \mathcal{R}_j \rightarrow \begin{cases} 0 & \text{if } u_j = 0 \\ \infty & \text{if } u_j \neq 0 \end{cases}$$

For $\beta_j^* = 0, \tilde{\beta}_j \neq 0$,

$$(254) \quad \mathcal{R}_j \rightarrow \begin{cases} 0 & \text{if } u_j \tilde{\beta}_j \geq 0 \\ \infty & \text{if } u_j \tilde{\beta}_j < 0 \end{cases}$$

For $\beta_j^* \neq 0, \tilde{\beta}_j = \beta_j^*$,

$$(255) \quad \mathcal{R}_j \rightarrow \begin{cases} 0 & \text{if } u_j \beta_j^* \leq 0 \\ \infty & \text{if } u_j \beta_j^* > 0 \end{cases}$$

For $\beta_j^* \neq 0, \tilde{\beta}_j \neq \beta_j^*$,

$$(256) \quad \mathcal{R}_j \rightarrow \begin{cases} 0 & \text{if } \operatorname{sgn}(\beta_j^*) \neq \operatorname{sgn}(\beta_j^* - \tilde{\beta}_j) \text{ or } u_j = 0 \\ \infty & \text{if } \operatorname{sgn}(\beta_j^*) = \operatorname{sgn}(\beta_j^* - \tilde{\beta}_j) \text{ and } u_j \operatorname{sgn}(\beta_j^*) > 0 \\ -\infty & \text{if } \operatorname{sgn}(\beta_j^*) = \operatorname{sgn}(\beta_j^* - \tilde{\beta}_j) \text{ and } u_j \operatorname{sgn}(\beta_j^*) < 0 \end{cases}$$

Overall, we have

$$(257) \quad \mathcal{R}_j \rightarrow \begin{cases} 0 & \text{if } (\beta_j^* = 0 \text{ and } \tilde{\beta}_j = 0 \text{ and } u_j = 0) \\ & \text{or } (\beta_j^* = 0 \text{ and } \tilde{\beta}_j \neq 0 \text{ and } u_j \tilde{\beta}_j \geq 0) \\ & \text{or } (\beta_j^* \neq 0 \text{ and } \tilde{\beta}_j = \beta_j^* \text{ and } u_j \beta_j^* \leq 0) \\ & \text{or } (\beta_j^* \neq 0 \text{ and } \tilde{\beta}_j \neq \beta_j^* \text{ and } (\operatorname{sgn}(\beta_j^*) \neq \operatorname{sgn}(\beta_j^* - \tilde{\beta}_j) \text{ or } u_j = 0)) \\ \infty & \text{if } (\beta_j^* = 0 \text{ and } \tilde{\beta}_j = 0 \text{ and } u_j \neq 0) \\ & \text{or } (\beta_j^* = 0 \text{ and } \tilde{\beta}_j \neq 0 \text{ and } u_j \tilde{\beta}_j < 0) \\ & \text{or } (\beta_j^* \neq 0 \text{ and } \tilde{\beta}_j = \beta_j^* \text{ and } u_j \beta_j^* > 0) \\ & \text{or } (\beta_j^* \neq 0 \text{ and } \tilde{\beta}_j \neq \beta_j^* \text{ and } \operatorname{sgn}(\beta_j^*) = \operatorname{sgn}(\beta_j^* - \tilde{\beta}_j) \text{ and } u_j \operatorname{sgn}(\beta_j^*) > 0) \\ -\infty & \text{if } (\beta_j^* \neq 0 \text{ and } \tilde{\beta}_j \neq \beta_j^* \text{ and } \operatorname{sgn}(\beta_j^*) = \operatorname{sgn}(\beta_j^* - \tilde{\beta}_j) \text{ and } u_j \operatorname{sgn}(\beta_j^*) < 0) \end{cases}$$

If we assume $\text{sgn}(\beta_j^*) \neq \text{sgn}(\beta_j^* - \tilde{\beta}_j)$ for $\forall j$ such that $\beta_j^* \neq 0$ and $\tilde{\beta}_j \neq \beta_j^*$, we exclude the case of $\mathcal{R} \rightarrow -\infty$, and

$$(258) \quad V_n(u) \rightarrow V(u) := \begin{cases} -2u^\top W + u^\top Cu & \text{if } (u_j = 0 \text{ for } \forall j \text{ s.t. } \beta_j^* = 0 \text{ and } \tilde{\beta}_j = 0) \\ & \text{and } (u_j \tilde{\beta}_j \geq 0 \text{ for } \forall j \text{ s.t. } \beta_j^* = 0 \text{ and } \tilde{\beta}_j \neq 0) \\ & \text{and } (u_j \beta_j^* \leq 0 \text{ for } \forall j \text{ s.t. } \beta_j^* \neq 0 \text{ and } \tilde{\beta}_j = \beta_j^*) \\ \infty & \text{otherwise.} \end{cases}$$

Since $V_n(u)$ is convex and $V(u)$ has a unique minimum, in the same manner of the proof of Adaptive Lasso, we have

$$(259) \quad \sqrt{n}(\hat{\beta} - \beta^*) \xrightarrow{d} \underset{u \in \mathcal{U}}{\operatorname{argmin}} -2u^\top W + u^\top Cu$$

$$(260) \quad \mathcal{U} := \left\{ u \in \mathbb{R}^p \mid \begin{array}{l} u_j = 0 \text{ for } \forall j \text{ s.t. } \beta_j^* = 0 \text{ and } \tilde{\beta}_j = 0 \\ u_j \tilde{\beta}_j \geq 0 \text{ for } \forall j \text{ s.t. } \beta_j^* = 0 \text{ and } \tilde{\beta}_j \neq 0 \\ u_j \beta_j^* \leq 0 \text{ for } \forall j \text{ s.t. } \beta_j^* \neq 0 \text{ and } \tilde{\beta}_j = \beta_j^* \end{array} \right\}$$

Next, we show active/varying variable consistency. We consider the cases of (i) $j \in S \cap T$, (ii) $j \in S \cap T^c$, (iii) $j \in S^c \cap T$, and (iv) $j \in S^c \cap T^c$.

(i) For $j \in S \cap T$, we have $\tilde{\beta}_j \neq 0$ because of $\text{sgn}(\tilde{\beta}_j - \beta_j^*) = \text{sgn}(\beta_j^*)$. By KKT condition, we have

$$(261) \quad 2 \left(\frac{1}{n} \mathbf{x}_j^\top X \right) \sqrt{n}(\beta^* - \hat{\beta}) + 2 \frac{1}{\sqrt{n}} \mathbf{x}_j^\top \varepsilon + \frac{\lambda_n}{\sqrt{n}} \text{sgn}(\hat{\beta}_j) + \frac{\eta_n}{\sqrt{n}} \text{sgn}(\hat{\beta}_j - \tilde{\beta}_j) = 0$$

and

$$(262) \quad \lim_{n \rightarrow \infty} P \left(\min \left\{ 0, \tilde{\beta}_j \right\} \leq \hat{\beta}_j \leq \max \left\{ 0, \tilde{\beta}_j \right\} \right) = 1.$$

On the other hand, $\sqrt{n}(\hat{\beta}_j - \beta_j^*)$ converges to some Gaussian-mixture distribution for $j \in S \cap T$. Hence, we have $\hat{\beta}_j \rightarrow \beta_j^*$ and

$$(263) \quad \lim_{n \rightarrow \infty} P \left(\hat{\beta}_j = 0 \text{ or } \hat{\beta}_j = \tilde{\beta}_j \right) = 0$$

This concludes

$$(264) \quad \lim_{n \rightarrow \infty} P \left(0 < \hat{\beta}_j < \tilde{\beta}_j \text{ or } \tilde{\beta}_j < \hat{\beta}_j < 0 \right) = 1 \text{ for } \forall j \in (S \cap T)$$

(ii) For $j \in S \cap T^c$, we have $\beta_j^* \neq 0$ and $\beta_j^* = \tilde{\beta}_j$. By KKT condition, we have

$$(265) \quad 2 \left(\frac{1}{n} \mathbf{x}_j^\top X \right) \sqrt{n}(\beta^* - \hat{\beta}) + 2 \frac{1}{\sqrt{n}} \mathbf{x}_j^\top \varepsilon + \frac{\lambda_n}{\sqrt{n}} \operatorname{sgn}(\hat{\beta}_j) + \frac{\eta_n}{\sqrt{n}} \operatorname{sgn}(\hat{\beta}_j - \beta_j^*) = 0$$

and

$$(266) \quad \lim_{n \rightarrow \infty} P \left(\min \{0, \beta_j^*\} \leq \hat{\beta}_j \leq \max \{0, \beta_j^*\} \right) = 1$$

On the other hand, $\sqrt{n}(\hat{\beta}_j - \beta_j^*)$ converges to some mixture distribution of truncated Gaussian distribution truncated at zero and delta distribution at zero for $j \in S \cap T^c$. Hence, we have $\hat{\beta}_j \rightarrow \beta_j^* \neq 0$ and

$$(267) \quad \lim_{n \rightarrow \infty} P \left(\hat{\beta}_j = 0 \right) = 0$$

This concludes

$$(268) \quad \lim_{n \rightarrow \infty} P \left(0 < \hat{\beta}_j \leq \tilde{\beta}_j \text{ or } \tilde{\beta}_j \leq \hat{\beta}_j < 0 \right) = 1 \text{ for } \forall j \in (S \cap T^c)$$

(iii) For $j \in S^c \cap T$, we have $\beta_j^* = 0$, $\beta_j^* \neq \tilde{\beta}_j$, and $\tilde{\beta}_j \neq 0$. By KKT condition, we have

$$(269) \quad 2 \left(\frac{1}{n} \mathbf{x}_j^\top X \right) \sqrt{n}(\beta^* - \hat{\beta}) + 2 \frac{1}{\sqrt{n}} \mathbf{x}_j^\top \varepsilon + \frac{\lambda_n}{\sqrt{n}} \operatorname{sgn}(\hat{\beta}_j) + \frac{\eta_n}{\sqrt{n}} \operatorname{sgn}(\hat{\beta}_j - \tilde{\beta}_j) = 0$$

and

$$(270) \quad \lim_{n \rightarrow \infty} P \left(\min \{0, \beta_j^*\} \leq \hat{\beta}_j \leq \max \{0, \beta_j^*\} \right) = 1$$

On the other hand, $\sqrt{n}\hat{\beta}_j$ converges to some mixture distribution of truncated Gaussian distribution truncated at zero and delta distribution at zero for $j \in S \cap T^c$. Hence, we have $\hat{\beta}_j \rightarrow 0 \neq \tilde{\beta}_j$ and

$$(271) \quad \lim_{n \rightarrow \infty} P \left(\hat{\beta}_j = \tilde{\beta}_j \right) = 0$$

This concludes

$$(272) \quad \lim_{n \rightarrow \infty} P \left(0 \leq \hat{\beta}_j < \tilde{\beta}_j \text{ or } \tilde{\beta}_j < \hat{\beta}_j \leq 0 \right) = 1 \text{ for } \forall j \in (S^c \cap T)$$

(iv) For $j \in S^c \cap T^c$, we have $\tilde{\beta}_j = \beta_j^* = 0$ and

$$(273) \quad 2 \left(\frac{1}{n} \mathbf{x}_j^\top X \right) \sqrt{n}(\beta^* - \hat{\beta}) + 2 \frac{1}{\sqrt{n}} \mathbf{x}_j^\top \varepsilon + \frac{\lambda_n + \eta_n}{\sqrt{n}} \operatorname{sgn}(\hat{\beta}_j) = 0$$

This implies that

$$(274) \quad \lim_{n \rightarrow \infty} P(\hat{\beta}_j = 0) = 1$$

B.4.3 Proof of Theorem A.5

We assume that $\tilde{\beta}_j = 0$ for $\forall j$ such that $\beta_j^* = 0$, and $\tilde{\beta}_j \neq \beta_j^*$ for $\forall j$ such that $\beta_j^* \neq 0$. Then, $\mathcal{U} = \{u \mid u_{S^c} = 0\}$, and we have asymptotic normality, i.e.,

$$(275) \quad \sqrt{n}(\hat{\beta}_S - \beta_S^*) \xrightarrow{d} \mathcal{N}(0, \sigma^2 C_{SS}^{-1})^\top, \quad \sqrt{n}\hat{\beta}_{S^c} \xrightarrow{d} 0.$$

On the other hand, this indicates that

$$(276) \quad \forall j \in S, \quad P(j \in \operatorname{supp}(\hat{\beta})) \rightarrow 1.$$

Now, we consider the event $j \in S^c$ and $j \in \operatorname{supp}(\hat{\beta})$. By the KKT conditions,

(277)

$$\begin{aligned} & 2\mathbf{x}_j^\top(y - X\hat{\beta}) + \lambda_n \operatorname{sgn}(\hat{\beta}_j) + \eta_n \operatorname{sgn}(\hat{\beta}_j - \tilde{\beta}_j) = 0 \\ (278) \quad & \Rightarrow \frac{2}{\sqrt{n}}\mathbf{x}_j^\top(y - X\hat{\beta}) + \frac{1}{\sqrt{n}}\lambda_n \operatorname{sgn}(\hat{\beta}_j) + \frac{\eta_n}{\sqrt{n}} \operatorname{sgn}(\hat{\beta}_j - \tilde{\beta}_j) = 0 \\ (279) \quad & \Rightarrow 2 \left(\frac{1}{n} \mathbf{x}_j^\top X \right) \sqrt{n}(\beta^* - \hat{\beta}) + 2 \frac{1}{\sqrt{n}} \mathbf{x}_j^\top \varepsilon + \frac{\lambda_n}{\sqrt{n}} \operatorname{sgn}(\hat{\beta}_j) + \frac{\eta_n}{\sqrt{n}} \operatorname{sgn}(\hat{\beta}_j - \tilde{\beta}_j) = 0 \end{aligned}$$

Since we assume $\tilde{\beta}_j = 0$ for $j \in S^c$, we have

$$(280) \quad 2 \left(\frac{1}{n} \mathbf{x}_j^\top X \right) \sqrt{n}(\beta^* - \hat{\beta}) + 2 \frac{1}{\sqrt{n}} \mathbf{x}_j^\top \varepsilon + \frac{\lambda_n + \eta_n}{\sqrt{n}} \operatorname{sgn}(\hat{\beta}_j) = 0$$

The first and second terms on the left-hand side converge to some normal distribution, but the last term on the left-hand side diverges to infinity if $(\lambda_n + \eta_n)/\sqrt{n} \rightarrow \infty$. Hence, we have for $\forall j \in S^c$,

$$(281) \quad \lim_{n \rightarrow \infty} P(j \in \operatorname{supp}(\hat{\beta})) = 0.$$

This concludes

$$(282) \quad \lim_{n \rightarrow \infty} P(\hat{S}_n = S) = 1$$

On the other hand, (4) is symmetric in terms of 0 and $\tilde{\beta}$. If we reparameterize parameters as

$$(283) \quad \beta^{*\prime} := \beta^* - \tilde{\beta}, \quad \tilde{\beta}' := -\tilde{\beta}, \quad \hat{\beta}'_n := \hat{\beta}_n - \tilde{\beta}, \quad y' := y - X\tilde{\beta},$$

$$(284) \quad \hat{S}'_n := \{j : \hat{\beta}'_j \neq 0\}, \quad S' := \{j : \beta_j^{*\prime} \neq 0\},$$

$$(285) \quad \lambda'_n := \eta_n, \quad \eta'_n := \lambda_n,$$

we have

$$(286)$$

$$\hat{\beta}'_n = \hat{\beta}_n - \tilde{\beta}$$

$$(287)$$

$$= \operatorname{argmin}_{\beta'} \frac{1}{n} \|y - X(\beta' + \tilde{\beta})\|_2^2 + \frac{\lambda_n}{n} \|\beta' + \tilde{\beta}\|_1 + \frac{\eta_n}{n} \|(\beta' + \tilde{\beta}) - \tilde{\beta}'\|_1$$

$$(288)$$

$$= \operatorname{argmin}_{\beta'} \frac{1}{n} \|y' - X\beta'\|_2^2 + \frac{\lambda'_n}{n} \|\beta'\|_1 + \frac{\eta'_n}{n} \|\beta' - \tilde{\beta}'\|_1$$

Hence, the property for $\{\hat{\beta}'_n, \beta^{*\prime}, \hat{S}'_n, S'\}$ is the same as that for $\{\hat{\beta}_n, \beta^*, \hat{S}_n, S\}$ in Theorem A.5. In addition, we have

$$(289)$$

$$\hat{S}'_n = \{j : \hat{\beta}'_j \neq 0\} = \{j : \hat{\beta}_j \neq \tilde{\beta}_j\} = \hat{T}_n, \quad S' = \{j : \beta_j^{*\prime} \neq 0\} = \{j : \beta_j^* \neq \tilde{\beta}_j\} = T.$$

This concludes

$$(290) \quad \lim_{n \rightarrow \infty} P(\hat{T}_n = T) = \lim_{n \rightarrow \infty} P(\hat{S}'_n = S') = 1$$

B.5 Proofs of Transfer Lasso with Initial Estimator in Boundary Region

B.5.1 Proofs of Theorem A.6

Let $u := \sqrt{n}(\beta - \beta^*)$, $V_n(u) := n(Z_n^T(\beta) - Z_n^T(\beta^*))$, and $V(u) := \lim_{n \rightarrow \infty} V_n(u)$. Consider the case where $\lambda_n/\sqrt{n} \rightarrow \infty$, $\lambda_n/\eta_n \rightarrow 1$, and $(\lambda_n - \eta_n)/\sqrt{n} \rightarrow \delta_0$.

Suppose that $n/m \rightarrow 0$. In the same way as in Case II of Proof B.2.2 (Theorem 3.5), we obtain $V_n(u)$ as (128). Because

$$(291) \quad (|u_j + \sqrt{m}\beta_j^*| - |\sqrt{m}\beta_j^*|) \rightarrow u_j \operatorname{sgn}(\beta_j^*) I(\beta_j^* \neq 0) + |u_j| I(\beta_j^* = 0),$$

$$(292) \quad \left(\left| u_j - \sqrt{\frac{n}{m}} z_j \right| - \left| \sqrt{\frac{n}{m}} z_j \right| \right) \rightarrow |u_j|,$$

and

$$(293) \quad \frac{\lambda_n}{\sqrt{n}} (u_j \operatorname{sgn}(\beta_j^*) I(\beta_j^* \neq 0) + |u_j| I(\beta_j^* = 0)) + \frac{\eta_n}{\sqrt{n}} |u_j|$$

$$(294) \quad = \left(\frac{\lambda_n}{\sqrt{n}} u_j \operatorname{sgn}(\beta_j^*) + \frac{\eta_n}{\sqrt{n}} |u_j| \right) I(\beta_j^* \neq 0) + \frac{\lambda_n + \eta_n}{\sqrt{n}} |u_j| I(\beta_j^* = 0),$$

we have

$$(295) \quad V(u) = \begin{cases} u^\top C u - 2u^\top W - \delta_0 \sum_j |u_j| I(\beta_j^* \neq 0), \\ \quad \text{if } \beta_j^* u_j \leq 0 \text{ for } \forall j \in S \text{ and } u_j = 0 \text{ for } \forall j \in S^c, \\ \infty, \text{ otherwise.} \end{cases}$$

Since $V_n(u)$ is convex and $V(u)$ has a unique minimum, we obtain

$$(296) \quad \sqrt{n} (\hat{\beta}_n - \beta^*) \rightarrow \operatorname{argmin}_{u \in \mathcal{U}} \left\{ u^\top C u - 2u^\top W - \delta_0 \sum_j |u_j| \right\},$$

$$(297) \quad \mathcal{U} := \left\{ u \in \mathbb{R}^p \mid \begin{array}{l} \beta_j^* u_j \leq 0 \text{ for } \forall j \in S, \\ u_j = 0 \text{ for } \forall j \in S^c \end{array} \right\}.$$

On the other hand, the Lagrangian function of $\operatorname{argmin}_{u \in \mathcal{U}} V(u)$ is

$$(298) \quad u_S^\top C_{SS} u_S - 2u_S^\top W_S - \delta_0 \sum_{j \in S} |u_j| + \sum_{j \in S} \mu_j \beta_j^* u_j + \sum_{j \in S^c} \mu_j u_j,$$

where $\mu \in \mathbb{R}^p$ is the Lagrangian multiplier. Suppose that $u_{S^c}^* = 0$. Let $S_1 := \{j : j \in S \text{ and } u_j^* \neq 0\}$ and $S_2 := \{j : j \in S \text{ and } u_j^* = 0\}$. By the

KKT conditions of $\operatorname{argmin}_{u \in \mathcal{U}} V(u)$, we have

$$(299) \quad \begin{cases} 2C_{S_1 S_1} u_{S_1}^* - 2W_{S_1} - \delta_0 \operatorname{sgn}(u_{S_1}^*) + \mu_{S_1} \beta_{S_1}^* = 0 \\ |2C_{S_2 S_1} u_{S_1}^* - 2W_{S_2} + \mu_{S_2} \beta_{S_2}^*| \leq \delta_0 \\ \beta_S^* u_S^* \leq 0 \\ \mu_S \geq 0 \\ \mu_S \beta_S^* u_S^* = 0 \\ \mu_{S^c} = u_{S^c}^* = 0 \end{cases}$$

If $S_1 \neq \emptyset$, then we have

$$(300) \quad u_{S_1}^* = C_{S_1 S_1}^{-1} (W_{S_1} + \delta_0 \operatorname{sgn}(u_{S_1}^*)).$$

The probability holding the third inequality in (299) is less than 1. This indicates inconsistent variable selection. If $S_1 = \emptyset$, then we have

$$(301) \quad |-2W_{S_2} + \mu_{S_2} \beta_{S_2}^*| \leq \delta_0.$$

The probability holding this inequality is less than 1. This indicates inconsistent variable selection.

C Additional Empirical Results

We describe additional empirical results.

C.1 Inconsistent Initial Estimator

In this simulation, we simulated inconsistent initial estimators. We generated target data by $\beta_{target}^* = [3, 1.5, 0, 0, 2, 0, 0, \dots, 0]^\top$. Then, we generated the source data with parameters different from those of the target data. We considered the following two cases:

Case A $\beta_{source}^* = [3, 1.5, 0, 0, 2, 2, 0, \dots, 0]^\top$ (non-zero \rightarrow zero change for $j = 6$)

Case B $\beta_{source}^* = [3, 1.5, 0, 0, 0, 0, 0, \dots, 0]^\top$ (zero \rightarrow non-zero change for $j = 5$)

Other simulation settings were the same as those in Simulation I.

The results are shown in Figures 10 and 11 for Case A, and Figures 12 and 13 for Case B. In Case A, the Transfer Lasso and the Adaptive Transfer

Lasso were superior in estimation error, and the Adaptive Lasso was slightly inferior to the Transfer Lasso and the Adaptive Transfer Lasso. Moreover, the Adaptive Lasso and the Adaptive Transfer Lasso were still superior in variable selection. In Case B, the Adaptive Lasso and the Adaptive Transfer Lasso performed worse in terms of estimation error. This may be because the initial estimator was incorrectly estimated close to zero, and regularization was strongly applied to it. The performance degradation is particularly significant for the Adaptive Lasso, but not so significant for Adaptive Transfer Lasso. The superiority of the Adaptive Lasso and the Adaptive Transfer Lasso also diminished in variable selection. Overall, the Adaptive Transfer Lasso performed comparatively well in estimation error and variable selection, but the inconsistent initial estimators reduced its performance.

C.2 Other Initial Estimators

We compared other initial estimators for simulations of a large amount of target data. The initial estimators included

- Ridge (Figure 14, 15)
- Ridgeless [2, 8]: minimum ℓ_2 -norm solution of least squares (Figure 16, 17)
- Lassoless [14, 12]: minimum ℓ_1 -norm solution of least squares (Figure 18, 19)

The results of Ridge initial estimators were similar to those of Lasso initial estimators. The results of Ridgeless and Lassoless initial estimators showed double descent phenomena, but they did not perform as well as Lasso and Ridge.

C.3 Other Metrics

We evaluated other metrics for Simulation I (Section 5.2). The metrics included

- RMSE for prediction evaluation (Figure 20)
- sensitivity ($= (\# \text{ of correct selected variables}) / (\# \text{ of true active variables})$) for feature selection evaluation (Figure 21)
- specificity ($= (\# \text{ of correctly not selected variables}) / (\# \text{ of true inactive variables})$) for feature selection evaluation (Figure 22)

- positive predictive value ($= (\# \text{ of correctly selected variables}) / (\# \text{ of selected variables})$) for feature selection evaluation (Figure 23)
- number of active variables for feature selection evaluation (Figure 24)

The results of RMSE were similar to those of ℓ_2 estimation errors, but the difference among methods got smaller. The results of 4 metrics for feature selection showed that Adaptive Lasso and Adaptive Transfer Lasso selected small number of variables and achieved superior performance especially on specificity and positive predictive value.

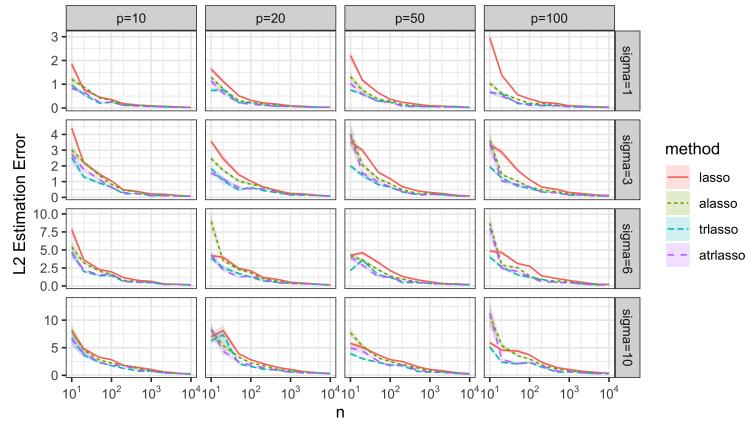


Figure 10: ℓ_2 estimation errors for inconsistent source data (Case A).

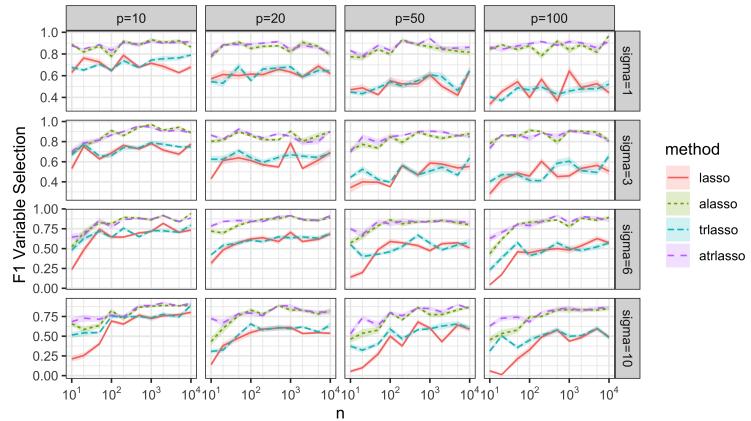


Figure 11: Variable selection F1-scores for inconsistent source data (Case A).

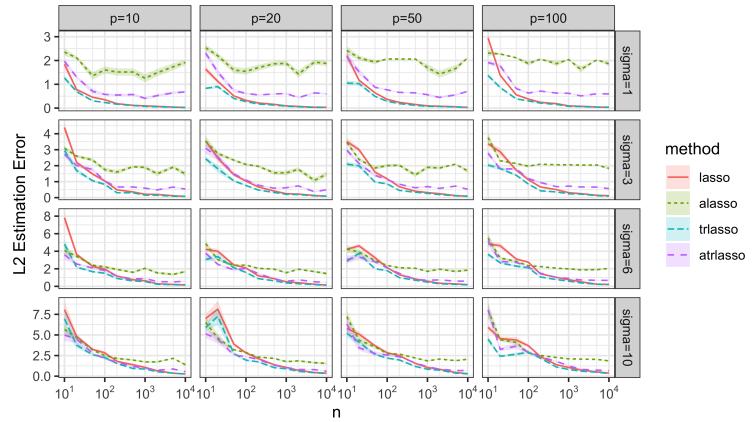


Figure 12: ℓ_2 estimation errors for inconsistent source data (Case B).

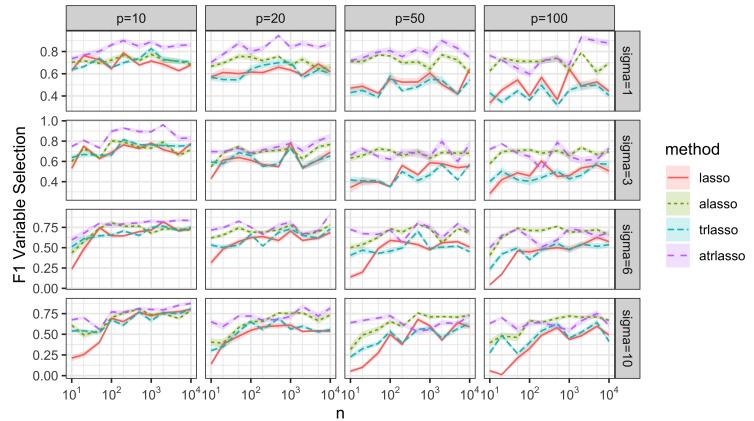


Figure 13: Variable selection F1-scores for inconsistent source data (Case B).

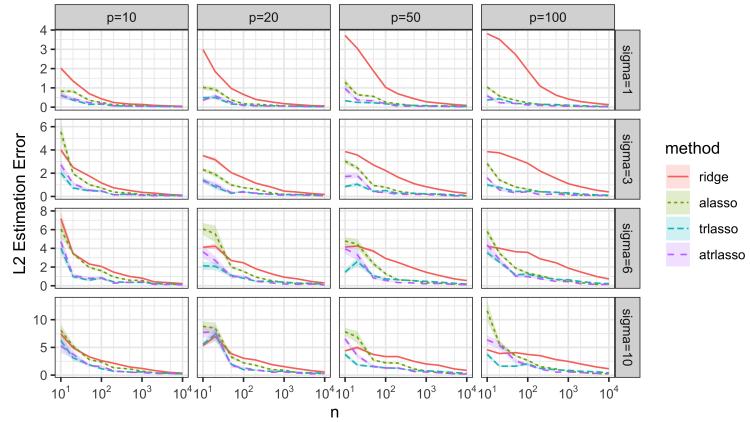


Figure 14: ℓ_2 estimation errors for a large amount of source data and Ridge initial estimators.

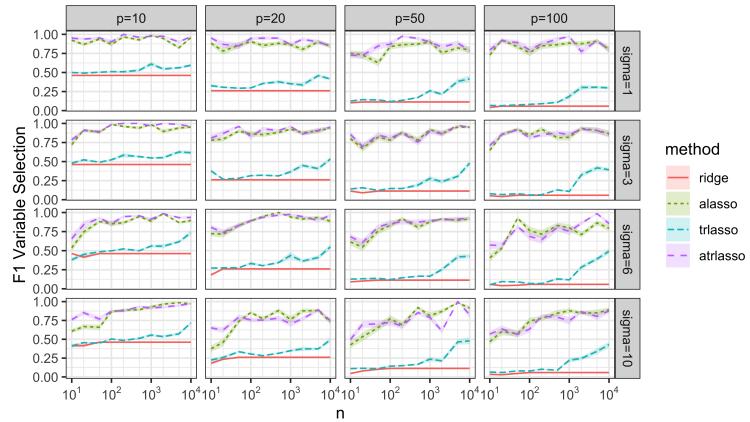


Figure 15: Variable selection F1-score for a large amount of source data and Ridge initial estimators.

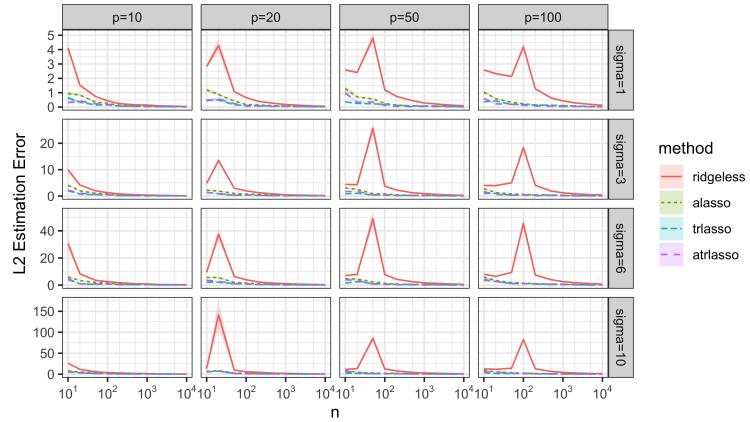


Figure 16: ℓ_2 estimation errors for a large amount of source data and Ridgeless initial estimators.

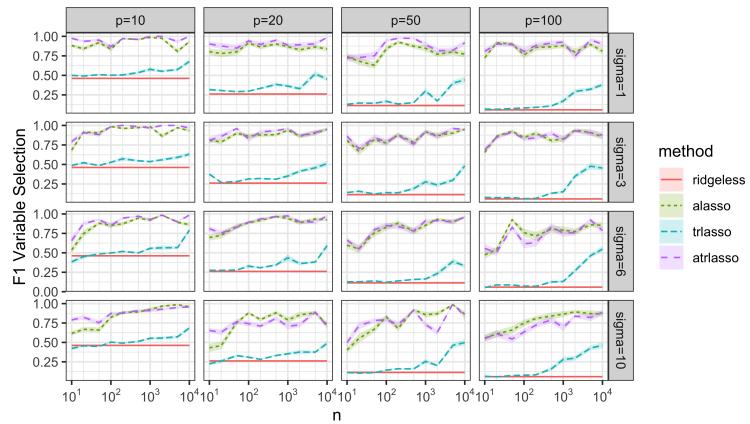


Figure 17: Variable selection F1-score for a large amount of source data and Ridgeless initial estimators.

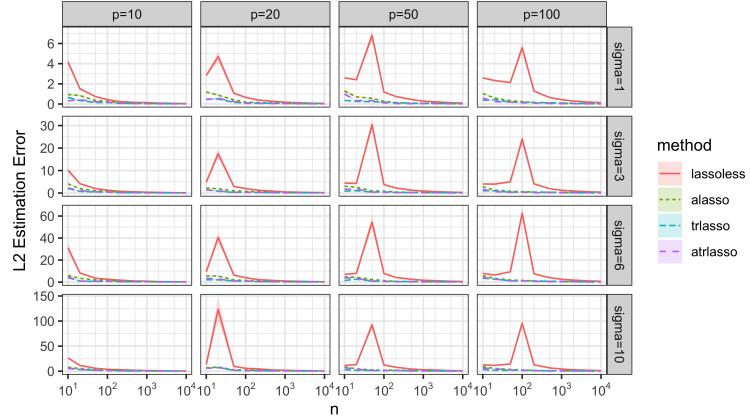


Figure 18: ℓ_2 estimation errors for a large amount of source data and Lassoless initial estimators.

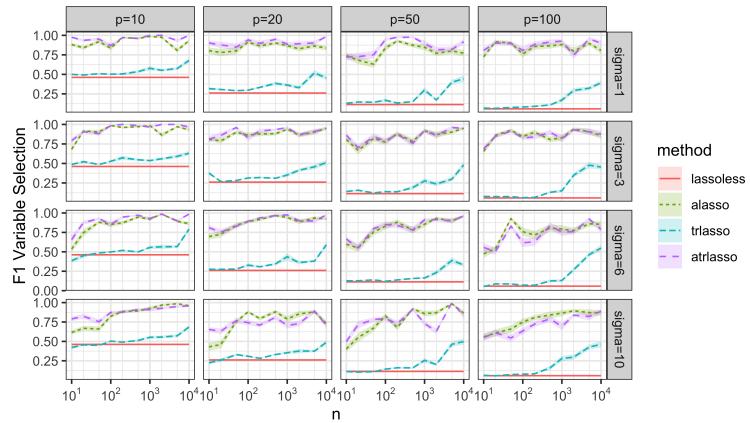


Figure 19: Variable selection F1-score for a large amount of source data and Lassoless initial estimators.

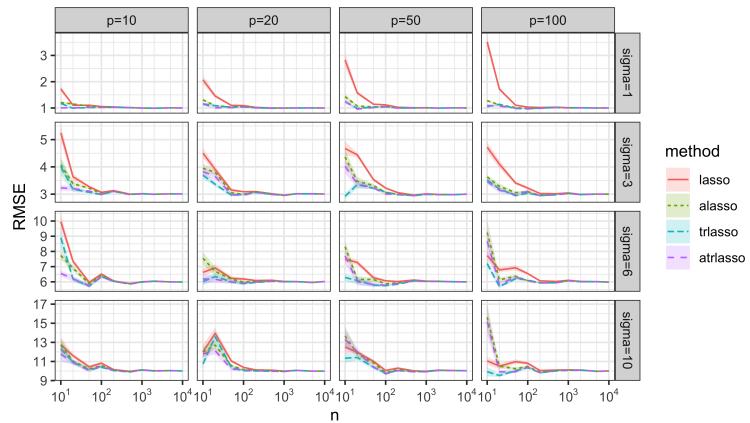


Figure 20: RMSE for a large amount of source data.

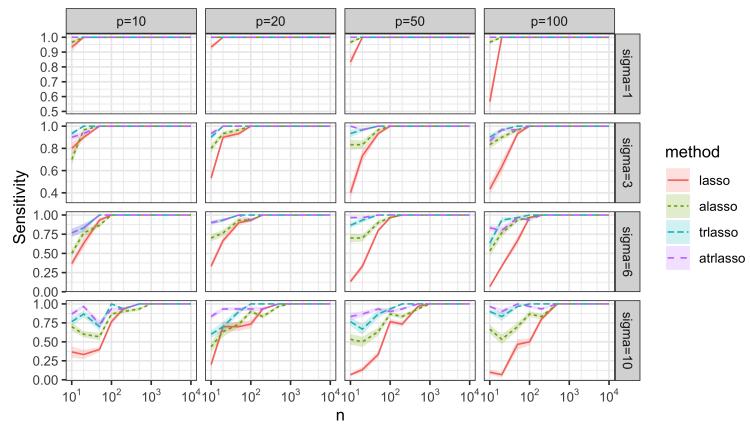


Figure 21: Sensitivity for a large amount of source data.

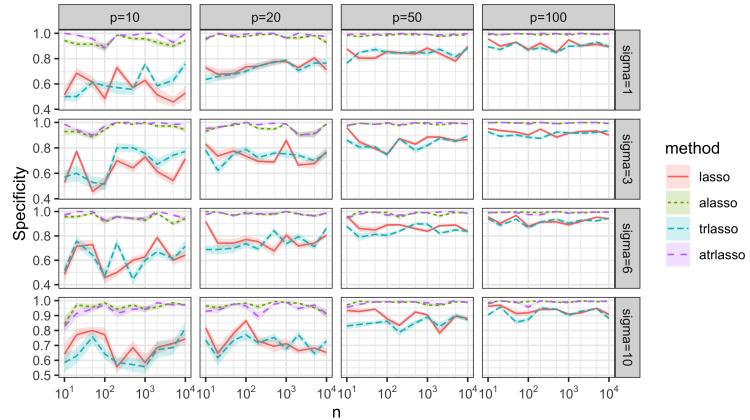


Figure 22: Specificity for a large amount of source data.

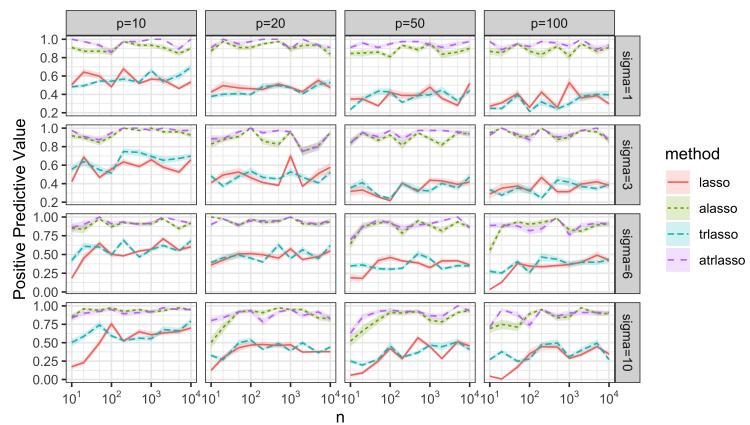


Figure 23: Positive predictive value for a large amount of source data.

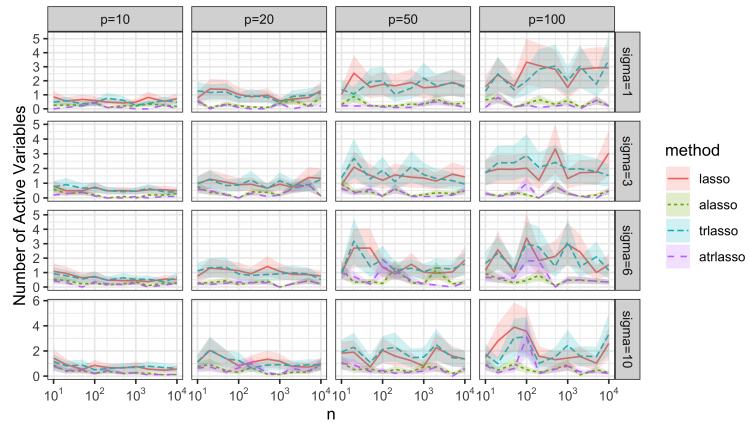


Figure 24: Number of active variables for a large amount of source data.