Electronic Journal of Statistics
Vol. 10 (2016) 90–120
ISSN: 1935-7524
DOI: 10.1214/15-EJS1083

# Oracle estimation of parametric transformation models*

**Yair Goldberg**

*Department of Statistics*
*The University of Haifa*
*Mount Carmel, Haifa 31905, Israel*
*e-mail:* ygoldberg@stat.haifa.ac.il

**Wenbin Lu**

*Department of Statistics*
*North Carolina State University*
*Stinson Dr, Raleigh, North Carolina 27607, U.S.A.*
*e-mail:* lu@stat.ncsu.edu

**and**

**Jason Fine**

*Department of Biostatistics*
*University of North Carolina at Chapel Hill*
*3101 McGavran-Greenberg Hall*
*Chapel Hill, North Carolina 27599, U.S.A.*
*e-mail:* jfine@email.unc.edu

**Abstract:** Transformation models, like the Box-Cox transformation, are widely used in regression to reduce non-additivity, non-normality, and heteroscedasticity. The question of whether one may or may not treat the estimated transformation parameter as fixed in inference about other model parameters has a long and controversial history (Bickel and Doksum, 1981, Hinkley and Runger, 1984). While the frequentist wisdom is that uncertainty regarding the true value of the transformation parameter cannot be ignored, in practice, difficulties in interpretation arise if the transformation is regarded as random and not fixed. In this paper, we suggest a golden mean methodology which attempts to reconcile these philosophies. Penalized estimation yields oracle estimates of transformations that enable treating the transformation parameter as known when the data indicate one of a prespecified set of transformations of scientific interest. When the true transformation is outside this set, rigorous frequentist inference is still achieved. The methodology permits multiple candidate values for the transformation, as is common in applications, as well as simultaneously accommodating variable selection in regression model. Theoretical issues, such as selection consistency and the oracle property, are rigorously established. Numerical studies, including extensive simulation studies and real data examples, illustrate the practical utility of the proposed methods.

**Keywords and phrases:** Box-Cox transformation, maximum likelihood estimation, oracle transformation, shrinkage estimation.

Received July 2015.

*The first author was funded in part by ISF grant 1308/12. The authors are grateful to anonymous reviewers for their helpful suggestions and comments.

## 1. Introduction

In regression analysis, it is sometimes worthwhile to transform the response variable $Y$, the explanatory vector $X$, or both, in order to reveal some basic properties of the data (Tukey, 1977, page 93). Using such transformations, one hopes to achieve the following three goals. Firstly, one may obtain a linear model in which the mean response is a known function of a linear transformation of the explanatory variables. Secondly, one may remove heteroscedasticity from the residual error. Thirdly, one may isolate a normal or nearly normal error distribution.

Parametric transformation models were pioneered in the early work of Box and Cox (1964) on the power transformation of the response in the linear model, where

$$Z \equiv h(Y, \lambda_0) = X^T \boldsymbol{\beta}_0 + \varepsilon\,, \tag{1}$$

for some $\lambda_0$ and regression parameter $\boldsymbol{\beta}_0$, the covariate $X$ includes the constant 1, and

$$h(Y, \lambda) = \begin{cases} \frac{Y^\lambda - 1}{\lambda} & \lambda \neq 0\,, Y > 0 \\ \log Y & \lambda = 0\,, Y > 0 \end{cases} \tag{2}$$

is a power transformation indexed by $\lambda$. The distribution of the residual $\varepsilon$ is often assumed to be a mean zero normal random variable with variance $\sigma^2$ independent of $X$. Parameter estimation with a sample of independent and identically distributed data, denoted by $(Y_i, X_i), i = 1, \ldots, n$, has been well studied. In practice, since $\lambda_0$ is unknown, it is estimated by $\hat{\lambda}$, the maximizer of the profile likelihood. After $\lambda_0$ is estimated, the parameters $\boldsymbol{\beta}_0$ and $\sigma_0$ may be estimated via least squares with the transformed responses $h(Y_i, \hat{\lambda})$.

Inference for the regression coefficient vector $\boldsymbol{\beta}_0$ in the transformed model is challenging. As pointed out by Bickel and Doksum (1981), the estimation of $\boldsymbol{\beta}_0$ depends on the estimation of $\lambda_0$. The problem is that the transformation parameter is not generally orthogonal to other model parameters. There is substantial empirical evidence demonstrating the potential for rather large variance inflation associated with the estimation of $\lambda_0$. Hence, inference for $\boldsymbol{\beta}_0$ needs to take into account the uncertainty regarding the true value of the parameter $\lambda_0$. More formally, the construction of confidence intervals for $\boldsymbol{\beta}_0$ requires use of the adjusted information matrix which reflects the decrease in information due to estimation of $\lambda_0$.

In practice, the uncertainty regarding $\lambda_0$ is rarely taken into account. It is common to treat $\hat{\lambda}$ as fixed (Hinkley and Runger, 1984, Carroll and Ruppert, 1988), or to restrict $\lambda_0$ to some finite set. For example, it has been recommended that $\lambda_0 \in \{0, \pm 1/2, \pm 1, \pm 2\}$ (see Carroll, 1982, for analysis of estimators restricted to a finite set). This may be justified as in Hinkley and Runger (1984), who explain that the regression parameters have meaning only with reference to particular scales or at least give a partial explanation on general scales (Brillinger, 1982, Stoker, 1986). For the Box-Cox transformation model,

estimating $\lambda_0$ is akin to determining the right scale for the data. As argued by Hinkley and Runger (1984), "This leads to the conclusion that when inference about parameters refers to specified scales of measurement (as must usually be the case), no allowance need be made for selecting those scales with the aid of the data". These principles lead Hinkley and Runger (1984) to a rejection of Bickel and Doksum (1981) regarding the need to account for the uncertainty regarding $\lambda_0$.

In this paper, we attempt to reconcile these conflicting philosophies. A golden mean methodology is presented which provides a theoretically justified framework in which $\hat{\lambda}$ can be regarded as fixed when the data indicate that $\lambda_0$ belongs to some finite candidate set, but otherwise takes into account the uncertainty regarding $\hat{\lambda}$. We propose a regularization procedure that maximizes a penalized version of the log likelihood with respect to $\boldsymbol{\beta}$, $\lambda$ and $\sigma$. The penalization consists of a weighted sum of the $\ell_1$ distance of $\lambda$ from a prespecified set of values, with the weights calculated from the data. The procedure is shown to correctly shrink $\hat{\lambda}$ to the true value when the value is in the set, with the resulting inferences for the other model parameters adaptive to whether or not the true $\lambda_0$ is contained in the candidate set. When $\lambda_0$ is in the set, the limit distribution is equivalent to that for an oracle estimator where $\lambda_0$ is known a priori. This theoretical finding supports treating $\hat{\lambda}$ as fixed, as advocated by Hinkley and Runger (1984). When $\lambda_0$ is not in the set, the joint estimator of $\lambda_0$ and the other model parameters is asymptotically equivalent to the unpenalized estimator, with inferences corresponding to those of Bickel and Doksum (1981) which account for uncertainty in $\hat{\lambda}$.

The approach we take here is sensible in practice since typically one has a small set of candidate transformations of interest, where the transformations are inherently meaningful and yield straightforward model interpretation. If one uses standard model selection criterion, like AIC, restricting transformations to this set, then one is implicitly finding the best fitting model amongst those models. This best fitting model may of course be misspecified, if the true model is not contained in that set. Our approach is conceptually different in that it will not select a model in the finite set unless it is the true model. Thus, our approach might be viewed as providing a goodness-of-fit assessment for procedures which restrict models to the finite set.

The issues discussed above for the Box-Cox response transformation model occur quite generally in regression models involving transformations of either the response $Y$ or the covariates $X$. A comprehensive overview is given in the definitive text of Carroll and Ruppert (1988). In Section 2, we formulate a unifying model for the mean of the response in which there may potentially exist multiple transformation parameters influencing the response, the covariates, and the relationship of the mean of the response to the covariates. This includes model (1) as a special case, as well as permitting generalized linear models for categorical outcomes where the link function is specified up to an unknown parameter $\lambda_0$. A broadly applicable shrinkage approach is discussed, in which each transformation parameter is shrunk towards values in a candidate set, with the

level of shrinkage determined by a weight calculated from the data. The approach further allows shrinkage of regression parameters, enabling simultaneous variable and transformation selection. The theoretical results described above for the Box-Cox model are demonstrated to be valid in the unifying model. If the true parameter lies in the candidate set, then with probability that converges to one, the estimated parameter will equal this value in finite samples and when it does, the corresponding inferences may regard the estimated value as fixed. On the other hand, if the true value does not lie in the set, then the usual asymptotic results for joint estimation of all parameters applies. We refer to this adaptation as the oracle properties of the estimator.

The oracle properties discussed above are valid for fixed values of the parameters. The construction of confidence regions for parameters typically demand stronger results such as uniform convergence. As in previous theoretical work on variable selection with oracle properties, the convergence is not uniform over the parameter space (Pötscher and Leeb, 2009, Pötscher and Schneider, 2010). While uniform convergence does not hold, we prove that uniform convergence over arbitrarily large subsets of the parameter space does hold. Here, arbitrarily large means that the subset for which uniform convergence does not hold can be chosen to have arbitrarily small Lebesgue measure. For Box-Cox transformation models, the subsets for which uniform convergence does not hold consist only of points that are very close to candidate transformations, and thus by the Hinkley and Runger (1984) paradigm, no allowance is needed for selecting those scales. In this work, we construct confidence regions which asymptotically achieve the nominal coverage level over arbitrary large subsets of the parameter space. Furthermore, we show that for any continuous and bounded prior on the parameter space these confidence regions asymptotically attain the desired coverage level.

One may view the proposed penalization methods in the spirit of earlier work on variable selection, inspired by the seminal lasso paper (Tibshirani, 1996). With a suitable choice of tuning parameter, one may consistently select important covariates by shrinking the coefficients of unimportant covariates to zero, with the coefficient estimates for the important covariates having asymptotic distribution which is equivalent to an oracle estimator with the unimportant covariates known a priori. Penalty functions yielding such estimators include the adaptive lasso (Zou, 2006), SCAD, the smoothly clipped absolute deviation (Fan and Li, 2001), the minimum convex penalty (Zhang, 2010), the smooth integration of counting and absolute deviation (Lv and Fan, 2009) and the log penalty (Friedman, 2012). Our penalization strategy adapts that in Zou (2006), with unpenalized estimates of the unifying model providing the necessary weights for penalized estimation. A major technical innovation is that we permit simultaneous shrinkage to multiple values of interest, as needed with transformation models. In addition, we present results which allow the size of the value-of-interest set to converge to infinity as the size of the sample grows. As in the variable selection setting, in our approach, if one constructs the weights to be large when the transformation is close to the candidate values, then the penalty enforces shrinkage to those candidate values, with the tuning parameter providing the necessary counterbalance.

In Section 2, we present our unifying model and penalization method, along with the associated theoretical results. Section 3 presents their specialization to the Box-Cox model (1). Simulations studies are discussed in Section 4, with real data examples used to illustrate the methods in Section 5. Some remarks conclude in Section 6. Detailed proofs and simulation results may be found in the Appendix. The R code for algorithm can be found in Goldberg et al. (2016).

## 2. Penalized likelihood estimation with multiple candidate values

### 2.1. Data and model

Let $V_1, \ldots, V_n$ be independent $d$-dimensional random vectors with distributed function $G(\boldsymbol{v})$. The observations $V_i$ can be pairs $(X_i, Y_i)$ of explanatory and response variables but are not limited to this setting. Let the parameter space $\Theta$ be a compact subset of a $\mathbb{R}^p$. Let $\{F(\boldsymbol{v};\boldsymbol{\theta}), \boldsymbol{\theta} \in \Theta\}$ be a family of distributions. Denote the density of $F(\boldsymbol{v};\boldsymbol{\theta})$, with respect to some dominating measure $\nu$, as $f(\boldsymbol{v};\boldsymbol{\theta})$. Define the log-likelihood of the data with respect to the family of distributions $\mathcal{F} = \{F(\boldsymbol{v};\boldsymbol{\theta}), \boldsymbol{\theta} \in \Theta\}$ as

$$L_n(V_1, \ldots, V_n; \boldsymbol{\theta}) = \sum_{i=1}^{n} \ell(V_i; \boldsymbol{\theta}),$$

where $\ell(\boldsymbol{v};\boldsymbol{\theta}) = \log f(\boldsymbol{v};\boldsymbol{\theta})$. Let $\widetilde{\boldsymbol{\theta}}_n$ denote the maximizer of $L_n$. Define $\dot{\ell}(\boldsymbol{v};\boldsymbol{\theta}) = \partial\ell(\boldsymbol{v};\boldsymbol{\theta})/\partial\boldsymbol{\theta}$, $\ddot{\ell}(\boldsymbol{v};\boldsymbol{\theta}) = \partial^2\ell(\boldsymbol{v};\boldsymbol{\theta})/(\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^T)$. Denote $\Gamma(\boldsymbol{\theta}) = E(\ddot{\ell}(\boldsymbol{v};\boldsymbol{\theta}))$, $\Delta(\boldsymbol{\theta}) = E(\dot{\ell}(\boldsymbol{v};\boldsymbol{\theta})\dot{\ell}(\boldsymbol{v};\boldsymbol{\theta})^T)$, and $\Lambda(\boldsymbol{\theta}) = \Gamma(\boldsymbol{\theta})^{-1}\Delta(\boldsymbol{\theta})\Gamma(\boldsymbol{\theta})^{-1}$. This setting allows for model misspecification, such that the true distribution $G$ needs not to be in $\mathcal{F}$ (see White, 1982, for discussion).

We assume the following conditions.

(A1) $G$ has a density $g$ with respect to the dominating measure $\nu$.
(A2) $\boldsymbol{\theta}_0$ is an inner point of $\Theta$ and is the unique maximizer of $E\{f(V,\boldsymbol{\theta})\}$.
(A3) $\ell(\boldsymbol{v};\boldsymbol{\theta})$ is twice continuously differentiable with respect to $\boldsymbol{\theta}$ for all $\boldsymbol{\theta} \in \Theta$ and $\boldsymbol{v}$, and the components of $|\ell(\boldsymbol{v};\boldsymbol{\theta})|$, $|\dot{\ell}(\boldsymbol{v};\boldsymbol{\theta})\dot{\ell}(\boldsymbol{v};\boldsymbol{\theta})^T|$, and $|\ddot{\ell}(\boldsymbol{v};\boldsymbol{\theta})|$ are dominated by integrable functions.
(A4) $E\{\dot{\ell}(V;\boldsymbol{\theta})\} = 0$ for all $\boldsymbol{\theta}$, $\Gamma(\boldsymbol{\theta}_0)$ is a negative definite matrix, and $\Delta(\boldsymbol{\theta}_0)$ is invertible.

The conditions above ensure consistency of $\widetilde{\boldsymbol{\theta}}_n$ for $\boldsymbol{\theta}_0$ and that $\sqrt{n}(\widetilde{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0) \to_d N(0, \Lambda(\boldsymbol{\theta}_0))$ (White, 1982, Theorem 3.1). When $G(\boldsymbol{v})$ is in the family of distributions $\mathcal{F}$, then we return to the usual setting of maximum likelihood estimation. We then have that $G(\boldsymbol{v}) \equiv F(\boldsymbol{v};\boldsymbol{\theta}_0)$ and $\Lambda(\boldsymbol{\theta}_0) = I(\boldsymbol{\theta}_0)^{-1}$ where $I(\boldsymbol{\theta})$ is the information matrix at $\boldsymbol{\theta}$. However, the above conditions do not require that the assumed model is correctly specified, permitting model misspecification, similarly to Lu et al. (2012), who studied likelihood based variable selection under misspecification. For some misspecified models, the parameter $\boldsymbol{\theta}_0$ being estimated may still be meaningful. This generality is important for the transformation model, as discussed in Section 3.

### *2.2. Penalized estimators*

Let $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_p)^T$ be the vector of parameters. For each component of $\boldsymbol{\theta}$, one can select a set of values of interest. We refer to the sets of values of interest as candidate sets. In the variable selection problem (see, for example, Fan and Li, 2001), for each component, the value of interest is zero, and thus the candidate set is of size one. In power transformations of the response or of both-sides (Carroll and Ruppert, 1988), typical values of interest for the power parameter $\lambda_0$ are $0, \pm 1/2, \pm 1, \pm 2$. Thus the candidate set for $\lambda_0$ is $\{0, \pm 1/2, \pm 1, \pm 2\}$. In power transformation models, there may not be interest in shrinking the estimator of the regression parameter $\beta_0$, in which case the size of the component of $\boldsymbol{\beta}_0$ is 0. Of course, the candidate set may be different for each parameter, as would occur, for example, when performing variable selection together with power transformation (see discussion on this model in Yeo, 2005). In this case, one can simultaneously penalize the power transformation parameter in the Box-Cox response model to a finite set of values and perform variable selection in which each regression parameter is shrunk to zero.

Let $\{\theta_j^1, \ldots, \theta_j^{k_j}\}$ be the candidate set for the $j$th component of the parameter vector, $k_j \in \{0, 1, 2, \ldots\}$. Here we let $k_j$ be equal to 0 to allow no values of interest for some of the components.

We define the penalized log-likelihood function

$$\Phi_n(\boldsymbol{\theta}) \equiv L_n(V_1, \ldots, V_n; \boldsymbol{\theta}) - \sum_{j=1}^{p} a_{nj} \sum_{k=1}^{k_j} \hat{w}_j^k |\theta_j - \theta_j^k|, \tag{3}$$

where $\hat{w}_j^k = |\widetilde{\theta}_{nj} - \theta_j^k|^{-\gamma}$ for some $\gamma > 0$ are weights, and $\boldsymbol{a}_n = (a_{n1}, \ldots, a_{np})^T$ is a vector of tuning parameters with positive components. Let $\hat{\boldsymbol{\theta}}_n$ denote the maximizer of $\Phi_n(\boldsymbol{\theta})$. In the next subsection, we demonstrate that these penalized estimators are selection consistent in the sense that with probability that converges to one, a particular parameter estimator will equal a candidate value for finite $n$ if that candidate value is the true value. This generalizes earlier work on variable selection (Fan and Li, 2001, Zou, 2006, Lv and Fan, 2009, Zou and Zhang, 2009, Friedman, 2012), where with probability that converges to one, a particular regression parameter estimator will equal 0 for finite $n$ if the corresponding covariate is unimportant. Moreover, the resulting estimators are oracle, having a limiting normal distribution whose variance equals that of an unpenalized estimator in which it is known a priori which of the candidate values are the true parameter values.

### *2.3. Theoretical properties*

Let $\mathcal{A}$ and $\mathcal{A}^C$ be sets of indices defined as

$$\mathcal{A} = \left\{j : \theta_{0j} \neq \theta_j^k, (k = 1, \ldots, k_j) \text{ or } k_j = 0\right\},$$
$$\mathcal{A}^C = \left\{j : \theta_{0j} = \theta_j^k \text{ for some } k \in \{1, \ldots, k_j\}\right\}.$$

Without loss of generality, we assume that $\mathcal{A} = \{1, \ldots, p_1\}$, $\mathcal{A}^C = \{p_1 + 1, \ldots, p\}$, and that for all $j \in \mathcal{A}^C$, $\theta_j^1 = \theta_{0j}$. Write $\boldsymbol{\theta}_0 = (\boldsymbol{\theta}_{01}^T, \boldsymbol{\theta}_{02}^T)^T$, where $\boldsymbol{\theta}_{01}$ is a $p_1$-dimensional vector of parameters, corresponding to the indices in $\mathcal{A}$, and $\boldsymbol{\theta}_{02}$ is a $p_2 = (p - p_1)$-dimensional vector, corresponding to the indices in $\mathcal{A}^C$. Accordingly, we write $\hat{\boldsymbol{\theta}}_n = (\hat{\boldsymbol{\theta}}_{n1}^T, \hat{\boldsymbol{\theta}}_{n2}^T)^T$.

In the following we present the main theoretical results of the paper. We first show that asymptotically, one can estimate $\boldsymbol{\theta}_0$ as if the components of $\boldsymbol{\theta}_{02}$ were known. In other words, with probability that tends toward 1, $\hat{\boldsymbol{\theta}}_{n2} = \boldsymbol{\theta}_{02}$, and the asymptotic variance matrix of $\hat{\boldsymbol{\theta}}_{n1}$ achieves the information bound of the estimation problem in which $\boldsymbol{\theta}_{02}$ is known. We then use this result to derive pointwise asymptotically-consistent confidence regions for $\boldsymbol{\theta}_0$.

**Theorem 1.** *Assume that conditions (A1)–(A4) hold, that* $a_{nj} n^{-1/2} \to 0$ *and that* $a_{nj} n^{(\gamma-1)/2} \to \infty$ *as* $n \to \infty$. *Then, for each fixed* $\boldsymbol{\theta}_0$

$$\hat{\theta}_{nj} \to_P \theta_{0j} \,.$$

*For all* $j \in \mathcal{A}$ *and for all* $k = 1, \ldots, k_j$ *such that* $k_j \geq 1$

$$P(\hat{\theta}_{nj} \neq \theta_j^k) \to 1 \,.$$

*If* $\mathcal{A}^C$ *is not empty, then for all* $j \in \mathcal{A}^C$

$$P(\hat{\theta}_{nj} = \theta_{0j}) \to 1 \,.$$

*Moreover,*

$$n^{1/2}(\hat{\boldsymbol{\theta}}_{n1} - \boldsymbol{\theta}_{01}) \to_d N\left(0, \Lambda_{11}(\boldsymbol{\theta}_0)^{-1}\right) \,,$$

*where* $\Lambda_{11}(\boldsymbol{\theta}_0)$ *is the upper-left* $p_1 \times p_1$ *submatrix of* $\Lambda(\boldsymbol{\theta}_0)$.

The proof of the theorem appears in Appendix A.1. Such results do not require that the model is correctly specified, with the limiting variance $\Lambda_{11}$ being robust to model misspecification. In addition, as noted previously, these asymptotic results are pointwise and may not hold for contiguous sequences converging to parameter values which include candidate shrinkage points. The implications of this fact for the theoretical properties of the resulting confidence intervals and regions are discussed in the next subsection.

In general one can have a different regularization constant for each parameter that has a non-empty candidate set. For the transformation model in Section 3, there is only one candidate set; therefore only one $a_n$ is used, with tuning parameter selection using standard methods. In the more general case with multiple parameters being shrunk, as in variable selection, typically one uses the same regularization constant for all candidate sets (see Zou, 2006, for adaptive lasso, as well as other key papers in penalization). This is in part because the use of multiple tuning parameters tends to complicate the analysis, owing to the need to simultaneously select the multiple tuning parameters, and in part because there is little evidence in the literature to suggest that this yields marked improvements in the empirical performance of the penalization methods.

### *2.4. Inference*

We now derive general pointwise asymptotically valid confidence intervals and confidence regions for the components of $\boldsymbol{\theta}_0$. First, we need some definitions. Define the sets

$$\begin{aligned} \mathcal{A}_n &= \left\{ j : \hat{\theta}_{nj} \neq \theta_j^k \text{ for all } k \in \{1, \ldots, k_j\} \text{ or } k_j = 0 \right\}, \\ \mathcal{A}_n^C &= \left\{ j : \hat{\theta}_{nj} = \theta_j^k \text{ for some } k \in \{1, \ldots, k_j\} \right\}. \end{aligned} \tag{4}$$

Let $J = \{j_1, \ldots, j_r\}$ be a set of indices, $J \subset \{1, \ldots, p\}$ for some $1 \leq r \leq p$. For any $p \times 1$ vector $v$, define $v^{(J)}$ to be the $r$th length vector with components $(v_{j_1}, \ldots, v_{j_r})^T$. Similarly, for a $p$ by $p$ matrix $A$, define $A^{(J)}$ to be the $r$ by $r$ matrix with entries $A_{st}^{(J)} = A_{j_s j_t}$, $(s, t = 1, \ldots, r)$. Finally, for a set $C$, define $C^{(J)} = \{\phi^{(J)} : \phi \in C\}$. Let $J_1 = \{j_{1,1}, \ldots, j_{1,r_1}\}$ and $J_2 = \{j_{2,1}, \ldots, j_{2,r_2}\}$, such that $J_2 \subset J_1$. With some abuse of notation, for any $p \times 1$ vector $v$, we define

$$\left(v^{(J_1)}\right)^{(J_2)} \equiv \left(v_{j_{2,1}}, \ldots, v_{j_{2,r_2}}\right)^T,$$

and similarly for matrices and sets. In the following we discuss confidence regions for a subset of parameters in which one treats parameters having been shrunk to a candidate value as fixed. The confidence regions may then be constructed using standard methods for maximum likelihood estimators.

**Theorem 2.** *Let $J = \{j_1, \ldots, j_r\}$ be a set of indices. Define $J_{n1} = J \cap \mathcal{A}_n$ and $J_{n2} = J \cap \mathcal{A}_n^C$. Let $\Lambda_n$ be a consistent estimator of $\Lambda(\boldsymbol{\theta}_0)$. For each $s = 1, \ldots, r$ choose a set $D_s$ such that $P(Z_s \in D_s) = 1 - \alpha$, where $Z_s$ is a Gaussian random vector with mean 0 and identity variance matrix of dimension $s$. For fixed $\boldsymbol{\theta}_0$ with parameter subset ${\boldsymbol{\theta}_0}^{(J)}$, define the set $C_n$ as*

$$\left\{ \boldsymbol{\theta}^{(J)} : \boldsymbol{\theta} \in \Theta, \boldsymbol{\theta}^{(J_{n1})} \in \left\{ \left( \left( n \Lambda_n^{(\mathcal{A}_n)} \right)^{-1/2} \right)^{(J_{n1})} D_s + \hat{\boldsymbol{\theta}}_n^{(J_{n1})} \right\}, \boldsymbol{\theta}^{(J_{n2})} = \hat{\boldsymbol{\theta}}_n^{(J_{n2})} \right\}, \tag{5}$$

*where $s$ is the cardinality of $J_{n1}$. Then*

$$\liminf_{n \to \infty} P\left( {\boldsymbol{\theta}_0}^{(J)} \in C_n \right) \geq 1 - \alpha.$$

See proof in Appendix A.2.

The above theorem can be simplified when there is only one component of the vector $\boldsymbol{\theta}_0$ for which there are values of interest. This result is useful in the Box-Cox model (1), when the regression parameters have no candidate values.

**Corollary 1.** *For fixed $\boldsymbol{\theta}_0$, let $k_j = 0$ for $j = 1, \ldots, p-1$ and $k_p > 0$; where $k_j$ is the number of values of interest for the $j$th component of $\boldsymbol{\theta}_0$.*
*For every $j = 1, \ldots, p-1$, define*

$$C_{nj} = \begin{cases} \left[ \hat{\theta}_{nj} - \frac{z_{1-\alpha/2}}{n^{1/2}} \left(\Sigma_{jj}\right)^{1/2}, \hat{\theta}_{nj} + \frac{z_{1-\alpha/2}}{n^{1/2}} \left(\Sigma_{jj}\right)^{1/2} \right] & \hat{\theta}_{np} \in \{\theta_p^1, \ldots, \theta_p^{k_p}\}, \\ \left[ \hat{\theta}_{nj} - \frac{z_{1-\alpha/2}}{n^{1/2}} (\Lambda_n^{-1/2})_{jj}, \hat{\theta}_{nj} + \frac{z_{1-\alpha/2}}{n^{1/2}} (\Lambda_n^{-1/2})_{jj} \right] & \hat{\theta}_{np} \notin \{\theta_p^1, \ldots, \theta_p^{k_p}\}, \end{cases}$$

*where* $\Sigma$ *is the inverse of the* $(p-1)\times(p-1)$ *upper-left submatrix of* $\Lambda_n$*, i.e.,* $\Sigma = (\Lambda_n^{(\{1,\ldots,p-1\})})^{-1}$*. Then, for every* $j = 1,\ldots,p-1$

$$\liminf_{n\to\infty} P\left(\theta_{0j} \in C_{nj}\right) \geq 1-\alpha\,,$$

The above result provide guarantees regarding the coverage probabilities of confidence intervals and regions for fixed $\boldsymbol{\theta}_0$. These guarantees are not uniform in $\boldsymbol{\theta}_0$, owing to the lack of uniform (in $\theta$) convergence of the limit distributions of estimators based on penalization procedures. The difficulties occur for points in the parameter space which are arbitrarily close to shrinkage values. Hence, assuming our model is correctly specified, the conventional definition of a confidence region, that is,

$$\liminf_{n\to\infty} \inf_{\boldsymbol{\theta}\in\Theta} P_{\boldsymbol{\theta}}\left(\boldsymbol{\theta}^{(J)} \in C_n\right) \geq 1-\alpha\,. \tag{6}$$

cannot be satisfied. This raises fundamental questions concerning the practical utility of the pointwise confidence regions in Theorem 2. In particular, one may not know a priori whether the true parameter value is sufficiently separated from shrinkage points to yield valid inferences.

To address this issue, we now investigate the extent to which poor performance of the confidence regions may occur under weak a priori assumptions on the true parameter values with a correctly specified model. We say that a sequence of (potentially random) sets $C_n$ is an asymptotically almost-everywhere confidence set for $\boldsymbol{\theta}_0$ in $\Theta$ if there is a sequence of parameter subspaces $\Theta_n \subset \Theta$ such that the Lebesgue measure of the sets $\Theta/\Theta_n$ converges to zero as $n\to\infty$ and

$$\liminf_{n\to\infty} \inf_{\boldsymbol{\theta}\in\Theta_n} P_{\boldsymbol{\theta}}\left(\boldsymbol{\theta}^{(J)} \in C_n\right) \geq 1-\alpha\,. \tag{7}$$

We have the following result.

**Theorem 3.** *Let* $J = \{j_1,\ldots,j_r\}$ *be a set of indices. Let* $C_n$ *be defined as in* (5)*. Assume (A3) and that (A4) holds for all inner points of* $\Theta$*. Then the sequence* $C_n$ *is asymptotically almost-everywhere confidence sets in* $\Theta$*.*

See proof in Appendix A.3. The above result shows that the confidence regions defined in Theorem 2 guarantee asymptotic coverage probabilities of $1-\alpha$ uniformly on a large subset of $\Theta$. Here, a large subset means that the Lebesgue measure of difference set of $\Theta$ and the subset can be arbitrary small. Moreover, the following corollary shows that for any continuous and bounded prior on $\Theta$, the probability of $\theta$ being in the confidence region achieves asymptotically the nominal coverage rate $1-\alpha$.

**Corollary 2.** *Let* $C_n$ *be defined as in* (5)*. Assume (A3) and that (A4) holds for all inner points of* $\Theta$*. Assume that* $\theta$ *is a random vector with bounded density* $\pi_\theta(\vartheta)$*,* $\vartheta\in\Theta$*. Then*

$$\liminf_{n\to\infty} P\left(\theta \in C_n\right) \geq 1-\alpha.$$

The proof appears in Appendix A.4.

### *2.5. Generalization to growing number of shrinkage points*

We now consider the case that the number of shrinkage points for each component of $\theta$ may grow with the sample size. This permits scenarios with infinite shrinkage points, a set-up which to our knowledge has not been considered previously in the penalization literature. Let $\Theta_j^{(n)} = \{\theta_{j,1}^{(n)}, \ldots, \theta_{j,k_{nj}}^{(n)}\}$ be the candidate set for the $j$th component of the parameter vector, $k_{nj} \in \{0, 1, 2, \ldots\}$. The size of $\Theta_j^{(n)}$ can change with the sample size $n$. We assume that for all $n$, $\Theta_j^{(n)} \subseteq \Theta_j^{(n+1)}$.

We define the penalized log-likelihood function

$$\Phi_n^{(n)}(\boldsymbol{\theta}) \equiv L_n(V_1, \ldots, V_n; \boldsymbol{\theta}) - \sum_{j=1}^{p} a_{nj} \sum_{k=1}^{k_{nj}} \hat{w}_j^k |\theta_j - \theta_{j,k}^{(n)}|, \tag{8}$$

where $\hat{w}_1, \ldots, \hat{w}_p$ and $\boldsymbol{a}_n$ are defined as before. Let $\hat{\boldsymbol{\theta}}_n$ denote the maximizer of $\Phi_n^{(n)}(\boldsymbol{\theta})$.

Let $\mathcal{B}$ and $\mathcal{B}^C$ be sets of indices defined as

$$\mathcal{B} = \left\{ j : \theta_{0j} \notin \Theta_j^{(n)} \text{ for all } n = 1, 2, \ldots \right\},$$
$$\mathcal{B}^C = \left\{ j : \text{ there is } N_{0,j} \text{ such that } \theta_{0j} \in \Theta_j^{(n)} \text{ for all } n \geq N_{0,j} \right\}.$$

Without loss of generality, we assume that $\mathcal{B} = \{1, \ldots, p_1\}$, $\mathcal{B}^C = \{p_1+1, \ldots, p\}$, and that for all $j \in \mathcal{B}^C$, $\theta_{j,1}^{(n)} = \theta_{0j}$ for all $n \geq N_0 \equiv \max_{j \in \mathcal{B}^C} N_{0,j}$. Write $\boldsymbol{\theta}_0 = (\boldsymbol{\theta}_{01}^T, \boldsymbol{\theta}_{02}^T)^T$, where $\boldsymbol{\theta}_{01}$ is a $p_1$-dimensional vector of parameters, corresponding to the indices in $\mathcal{B}$, and $\boldsymbol{\theta}_{02}$ is a $p_2 = (p - p_1)$-dimensional vector, corresponding to the indices in $\mathcal{B}^C$. Accordingly, we write $\hat{\boldsymbol{\theta}}_n = (\hat{\boldsymbol{\theta}}_{n1}^T, \hat{\boldsymbol{\theta}}_{n2}^T)^T$.

We need the following definitions regarding the size and denseness of the sets $\Theta_j^{(n)}$. Define

$$\delta_n = \min_{j \in \{1, \ldots, p\}} \min_{k : \theta_{0j} \neq \theta_{j,k}^{(n)}} |\theta_{0j} - \theta_{j,k}^{(n)}|,$$

if there exists a pair $(j, k)$ such that $\theta_{0j} \neq \theta_{j,k}^{(n)}$, and $\delta_n = 1$ otherwise. Let $\eta_n = \sum_{j=1}^{p} k_{nj}$.

**Theorem 4.** *Assume that conditions (A1)–(A4) hold, that* $n^{-1/2} a_{nj} \eta_n \delta_n^{-\gamma} = o(1)$, $\delta_n^{-1} = o_P(n^{1/2})$, *and* $a_{nj} n^{(\gamma-1)/2} \to \infty$ *as* $n \to \infty$. *Then, for all* $j \in \mathcal{B}$

$$P\left(\hat{\theta}_{nj} \notin \Theta_j^{(n)}\right) \to 1.$$

*For all* $j \in \mathcal{B}^C$

$$P(\hat{\theta}_{nj} = \theta_{j0}) \to 1.$$

*Moreover*

$$n^{1/2}(\hat{\boldsymbol{\theta}}_{n1} - \boldsymbol{\theta}_{01}) \to_d N\left(0, \Lambda_{11}(\boldsymbol{\theta}_0)^{-1}\right) ,$$

*where* $\Lambda_{11}(\boldsymbol{\theta}_0)$ *is the upper-left* $p_1 \times p_1$ *submatrix of* $\Lambda(\boldsymbol{\theta}_0)$.

The proof appears in Appendix A.5.

Theorem 4 shows that one can obtain an asymptotically oracle estimator of $\boldsymbol{\theta}_0$, even when the candidate set size tends to infinity, or when the limit of the candidate set is a dense in $\Theta$, or both. Specifically, the candidate set size can grow to infinity at a rate of up to $n^{1/2-\varepsilon}$ for an arbitrary small $\varepsilon > 0$. The candidate set can also converge to a dense set such that in the limit, for each given $j$, there are points arbitrarily close to $\theta_{0j}$. The distance between $\theta_{0j}$ and its neighboring points can decrease at a rate not less than $n^{-1/2+\varepsilon}$ for an arbitrary small $\varepsilon > 0$. For the case that both size and denseness enlarge with $n$, the results of Theorem 4 hold when choosing $\gamma = 1$, $a_{nj} \asymp n^{\alpha_1}$ for some $\alpha_1 > 0$, $\eta_n \asymp n^{\alpha_2}$, for some $\alpha_2 > 0$ such that $\alpha_1 + \alpha_2 < \frac{1}{2}$, and $\delta_{nj} \asymp n^{\frac{1}{2}-\alpha_3}$ for $\alpha_1 + \alpha_2 < \alpha_3 < \frac{1}{2}$.

## 3. Application to transformation models

In this section, we discuss application of the general results in Section 2 to the Box-Cox transformation model. Let $(X_1, Y_1), \ldots, (X_n, Y_n)$ be $n$ independent and identically distributed observations, where $X_i$ are random vectors and $Y_i$ are positive random variables.

Assuming the usual Box-Cox transformation model defined (1)–(2), one can write the log likelihood for this model by

$$\ell(X, Y; \boldsymbol{\theta}) = -\frac{\varepsilon^2}{2} - \log\sigma + (\lambda - 1)\log(Y) + C ,$$

where $\boldsymbol{\theta} = (\boldsymbol{\beta}^T, \sigma, \lambda)^T$ and $\varepsilon = (h(Y, \lambda) - \boldsymbol{\beta}^T X)/\sigma$. Following Hernandez and Johnson (1980), we note that this model may not be correctly specified but that conditions (A1)–(A4) may be satisfied (see also Bickel and Doksum, 1981, Yeo and Johnson, 2000). The parameter being estimated corresponds to the maximizer of the expectation of the likelihood function $\ell$ with respect to the true underlying distribution of $(X, Y)$. Such parameter may still have a meaningful interpretation in the regression context.

Let $\widetilde{\boldsymbol{\theta}}_n$ be the solution the estimation equation $\sum_{i=1}^n \dot{\ell}(X_i, Y_i; \boldsymbol{\theta}) = 0$ where the score function $\dot{\ell}$ is

$$\dot{\ell}(X, Y; \boldsymbol{\theta}) = \left\{ \frac{\varepsilon X^T}{\sigma}, \frac{\varepsilon^2 - 1}{\sigma}, \log Y - \frac{\varepsilon}{\sigma}\frac{\partial h(Y, \lambda)}{\partial \lambda} \right\}^T , \tag{9}$$

where

$$\frac{\partial h(y, \lambda)}{\partial \lambda} = \frac{y^\lambda \log y - h(y, \lambda)}{\lambda} .$$

Write

$$\ddot{\ell}(X,Y;\boldsymbol{\theta}) = \frac{1}{\sigma^2}\begin{Bmatrix} -XX^T, & -\frac{2\varepsilon}{\sigma}X^T & -\frac{\partial h(Y,\lambda)}{\partial\lambda}X^T \\ & 1-3\varepsilon^2 & 2\varepsilon\frac{\partial h(Y,\lambda)}{\partial\lambda} \\ & & -\left(\frac{\partial h(Y,\lambda)}{\partial\lambda}\right)^2 - \sigma\varepsilon\frac{\partial^2 h(Y,\lambda)}{\partial\lambda^2} \end{Bmatrix} \quad (10)$$

The matrix $\Lambda(\boldsymbol{\theta}) = \Gamma(\boldsymbol{\theta})^{-1}\Delta(\boldsymbol{\theta})\Gamma(\boldsymbol{\theta})^{-1}$ can be consistently estimated by $\hat{\Lambda}_n = (\hat{\Gamma}_n)^{-1}\hat{\Delta}_n(\hat{\Gamma}_n)^{-1}$, where

$$\hat{\Gamma}_n = \frac{1}{n}\sum_{i=1}^{n}\ddot{\ell}(X_i,Y_i;\widetilde{\boldsymbol{\theta}}_n)$$
$$\hat{\Delta}_n = \frac{1}{n}\sum_{i=1}^{n}\dot{\ell}(X_i,Y_i;\widetilde{\boldsymbol{\theta}}_n)\dot{\ell}(X_i,Y_i;\widetilde{\boldsymbol{\theta}}_n)^T.$$

The goal is estimation of the regression parameters which is adaptive to the unknown power transformation. Denote the candidate set for lambda by $\{\lambda^1,\ldots,\lambda^k\}$ and define the penalized log-likelihood function

$$\Phi_n(\boldsymbol{\theta}) \equiv \sum_{i=1}^{n}\ell(X_i,Y_i;\boldsymbol{\theta}) - a_n\sum_{j=1}^{k}\hat{w}_n^j|\lambda-\lambda^j|, \quad (11)$$

where $\boldsymbol{\theta} = (\boldsymbol{\beta}^T,\sigma,\lambda)^T \in \mathbb{R}^p$, $a_n$ is a regularization constant, and $\hat{w}_n^j = |\tilde{\lambda}_n - \lambda^j|^{-\gamma}$ where $\tilde{\lambda}_n$ is the maximum likelihood estimator of $\lambda_0$ without adaptive selection. Let $\hat{\boldsymbol{\theta}}_n = (\hat{\boldsymbol{\beta}}_n^T,\hat{\sigma}_n,\hat{\lambda}_n)^T$ be the maximizer of (11).

**Lemma 1.** *Assume conditions (A1)–(A4), that* $a_n/n^{1/2} \to 0$ *and that* $a_n n^{(\gamma-1)/2} \to \infty$ *as* $n\to\infty$. *Assume also that* $E\{h(Y,\lambda_0) \mid X\} = \boldsymbol{\beta}_0{}^T X$ *and* $\mathrm{var}\{h(Y,\lambda_0)\mid X\} = \sigma_0^2$. *Then*

$$n^{1/2}(\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0) \to_d \begin{cases} N\left(0,{\sigma_0}^2\{E(XX^T)\}^{-1}\right) & \lambda_0 \in \{\lambda^1,\ldots,\lambda^k\} \\ N\left(0,\Lambda(\boldsymbol{\theta}_0)^{(\{1,\ldots,p-2\})}\right) & \textit{otherwise} \end{cases}.$$

*Moreover, the following holds*
*(i) When* $\lambda_0 \in \{\lambda^1,\ldots,\lambda^k\}$, *with probability that tends to* 1, $\hat{\lambda}_n = \lambda_0$.
*(ii) When* $\lambda_0 \notin \{\lambda^1,\ldots,\lambda^k\}$, *with probability that tends to* 1, $\hat{\lambda}_n \notin \{\lambda^1,\ldots,\lambda^k\}$.

This lemma follows from Theorem 1, and direct computation of $\Lambda(\boldsymbol{\theta}_0)$ using (9)–(10).

To obtain confidence intervals for the model parameters, we apply Corollary 1. When $\hat{\lambda}_n$ is in the candidate set, i.e., $\lambda \in \{\lambda^1,\ldots,\lambda^k\}$, an asymptotically almost-everywhere confidence interval for $\beta_{0j}$, $j=1,\ldots,p-2$, with confidence level of $100(1-\alpha)\%$, is

$$\beta_{0j} \in \left[\hat{\beta}_{nj} - \frac{z_{1-\alpha/2}}{n^{1/2}}(C_{jj})^{1/2},\ \hat{\beta}_{nj} + \frac{z_{1-\alpha/2}}{n^{1/2}}(C_{jj})^{1/2}\right].$$

Here, the matrix $C = \hat{\sigma}_n^2(\sum_{i=1}^n X_i X_i^T)^{-1}$ and $z_\alpha$ is the $\alpha$-quantile of the standard normal distribution. When $\hat{\lambda}_n$ is not in the candidate set, an asymptotically almost-everywhere confidence interval for $\beta_{0j}$ $j = 1, \ldots, p-2$, with confidence level of $100(1-\alpha)\%$, is

$$\beta_{0j} \in \left[\hat{\beta}_{nj} - \frac{z_{1-\alpha/2}}{n^{1/2}} \{(\Lambda_n)_{jj}\}^{1/2} \, , \, \hat{\beta}_{nj} + \frac{z_{1-\alpha/2}}{n^{1/2}} \{(\Lambda_n)_{jj}\}^{1/2}\right] ;$$

and an asymptotically almost-everywhere confidence interval for $\lambda_0$ is

$$\lambda_0 \in \left[\hat{\lambda}_n - \frac{z_{1-\alpha/2}}{n^{1/2}} \{(\Lambda_n)_{pp}\}^{1/2} \, , \, \hat{\lambda}_n + \frac{z_{1-\alpha/2}}{n^{1/2}} \{(\Lambda_n)_{pp}\}^{1/2}\right] .$$

## 4. Simulation study

We conducted simulations to evaluate the performance of the penalized estimators of the power transformation model. The data were generated from the Box-Cox model (1), under transformation (2). We considered a one-dimensional covariate vector that was generated from the standard normal distribution. The error terms were generated from the standard normal distribution. We considered five values for the true transformation parameter: $\lambda_0 = 0, 1/2, -1/2, 1, -1$. The candidate set $\mathcal{A}_\lambda$ equals $\{0, 1/2, -1/2, 2, -2\}$ for the transformation parameter. Note that $\mathcal{A}_\lambda$ does not include $1, -1$.

For our method, we first computed the maximum likelihood estimators for $\lambda_0$ and $\boldsymbol{\beta}_0$ without penalization, denoted by $\tilde{\lambda}_n$ and $\tilde{\boldsymbol{\beta}}_n$, respectively. Then we obtained our proposed estimators as defined in (11). The tuning parameter $a_n$ was selected using the 5-fold cross-validation. The weights' parameter $\gamma$ was set to 1. For comparison, we computed results for the unpenalized estimator and the oracle estimator, in which the maximum likelihood estimator is calculated with the true value of the transformation parameter.

For each method, we computed the bias and the median absolute deviation of the estimates divided by 0.6745. The reason we computed the median absolute deviation instead of the sample standard deviation of the estimates is that the maximum likelihood estimator and penalized estimators may have a few outliers due to the instability of the estimation. In addition, for the maximum likelihood estimator and penalized estimators, we computed the median of estimated standard errors and the empirical coverage probability of Wald-type 95% confidence intervals. The standard error of the maximum likelihood estimator was estimated using the standard likelihood theory. The standard error of the adaptive estimator was estimated based on the asymptotic results established in Lemma 1. The empirical coverage probabilities for 95% confidence intervals based on the model-based standard errors is provided for the maximum likelihood estimator and adaptive estimators. For the maximum likelihood estimator, we also computed the coverage probabilities when $\hat{\lambda}$ is treated as fixed. Finally, for the adaptive estimator, we include the selection frequency of the true power transformation parameter when it is contained in the candidate set and the selection frequency of the values in the candidate set when the true power

transformation parameter is not contained in the candidate set. We report these results in Appendix B, Tables 3–5.

Based on the simulation results, we make the following observations: (i) The oracle estimators for the regression parameters generally have much smaller median absolute deviation compared to the maximum likelihood estimator. (ii) When the true power transformation parameter is contained in the candidate set, the adaptive and oracle estimators show very comparable performance in terms of both bias and median absolute deviation. In addition, the median of estimated standard errors are all close to the median absolute deviation with the empirical coverage probability close to the nominal level. The adaptive method selects the true power transformation parameter with a high frequency and the selection performance improves as the sample size increases. (iii) When the true power transformation parameter is not contained in the candidate set, the adaptive and the maximum likelihood estimator estimators show comparable performance, and the estimation performance improves as the sample size increases, as expected. In addition, for the adaptive method, the selection frequency of the values in the candidate set is low and decreases to 0 as the sample size increases. (iv) When using maximum likelihood estimator and treating $\hat{\lambda}$ as fixed, the confidence intervals may severely undercover. The findings (i)–(iv) support the theoretical results in Sections 2 and 3.

Next, we conducted simulations to evaluate the performance of the proposed adaptive estimator when varying the signal-to-noise ratio. In particular, we study the inflation in the standard error estimates of the maximum likelihood estimator compared with the adaptive estimator and the power of Wald test for testing $\beta_{02} = 0$ based on the adaptive estimator, as the signal-to-noise ratio varies. We consider the same simulation settings with $\lambda_0 = 0, 1/2$, and $-1/2$, and set $\beta_{02} = 0.05, 0.1, 0.25, 0.5, 0.75$ and 1.0. For each setting, we conducted 500 runs with sample size $n = 100$. The true value $\lambda_0$ is contained in the candidate set. Therefore, the standard error estimates of the adaptive estimator are obtained as if $\lambda_0$ was known as long as the corresponding estimator is shrunk to a value in the candidate set. The mean standard error ratios of the maximum likelihood estimator over the adaptive estimator for $\beta_{02}$ are plotted in the upper-left panel of Figure 1, while the power of Wald test for testing $\beta_{02} = 0$ based on the adaptive estimator is plotted in other panels of Figure 1. We observe that comparing with the adaptive estimator, the maximum likelihood estimator shows larger standard error inflation as the signal-to-noise ratio increases, and when the signal-to-noise ratio is close to 0, there is almost no inflation. These agree with the findings in Bickel and Doksum (1981) and Doksum and Wong (1983). Furthermore, the power of Wald test increases as the signal-to-noise ratio increases as expected, and the adaptive estimator has significantly improved power compared with the maximum likelihood estimator for most $\beta_{02}$ values under all scenarios.

Finally, we conducted simulations to evaluate the performance of the proposed adaptive estimator in terms of prediction interval and compare it with the maximum likelihood estimator. Here the prediction interval is constructed

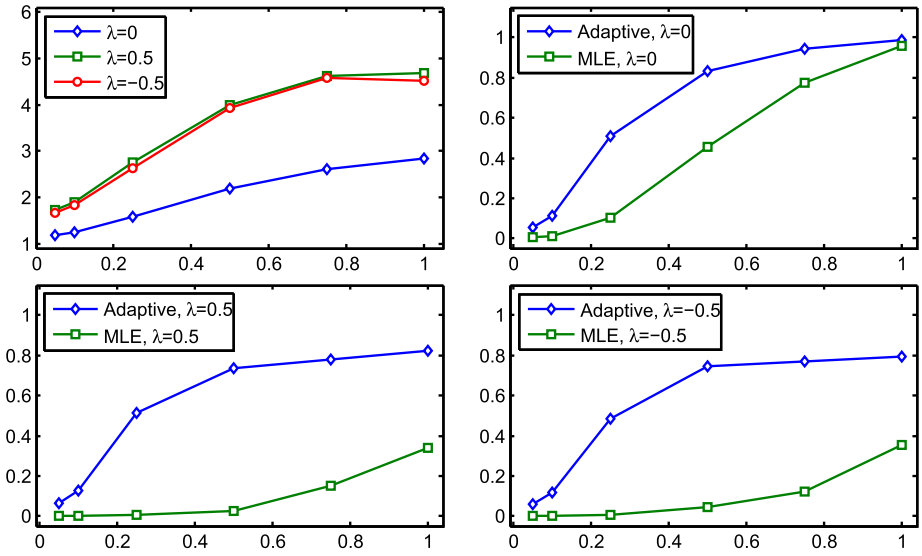


FIG 1. *Upper-left panel: ratio between the mean standard error of the maximum likelihood estimator (MLE) and the adaptive estimator (Adaptive) for* $\lambda = 0,\ 0.5,$ *and* $-0.5$. *Other panels: the power function of the Wald test for testing* $\beta_{02} = 0$, *for* $\lambda = 0,\ 0.5,$ *and* $-0.5$.

following the method of Cho et al. (2001). The coverage probabilities of corresponding 95% prediction intervals and their standard errors are reported in Table 6 in the Appendix. Based on simulation results, we observe that the prediction intervals constructed by both estimates give reasonable coverage probabilities. In addition, the coverage probabilities of the adaptive method tend to have smaller standard deviations compared with the maximum likelihood estimator method when $\lambda_0 = 0, 1/2,$ and $-1/2$, i.e. $\lambda_0$ is contained in the candidate set, while they are comparable when $\lambda_0 = 1, -1$, i.e., $\lambda_0$ is not contained in the candidate set.

## 5. Data examples

In this section, we apply the proposed adaptive selection method to two datasets studied in Box and Cox (1964) and Hinkley and Runger (1984). The R code for algorithm can be found in Goldberg et al. (2016).

For the textile example, the response is cycles to failure. There are three explanatory variables, $v_1$, $v_2$, $v_3$, denoting the factor levels. A linear regression model was considered for the Box-Cox transformation of the response. The sample size is $n = 27$. The analysis results for the maximum likelihood estimator and the proposed adaptive estimation method are given in Table 1. The maximum likelihood estimator gives $\hat{\lambda} = -0.06$ and the results of the maximum likelihood estimator are the same as those reported in Table 5 of Hinkley and Runger (1984). For the proposed adaptive estimation method, we considered the

TABLE 1
*Analysis results for the textile example*

| method | $\lambda$ | $\sigma$ | intercept | | $v_1$ | | $v_2$ | | $v_3$ | |
|---|---|---|---|---|---|---|---|---|---|---|
| | Est. | Est. | Est. | SE | Est. | SE | Est. | SE | Est. | SE |
| MLE | -0.06 | 0.12 | 5.25 | 1.51 | 0.57 | 0.35 | -0.43 | 0.27 | -0.27 | 0.17 |
| adaptive | 0.00 | 0.17 | 6.33 | 0.03 | 0.83 | 0.04 | -0.63 | 0.04 | -0.39 | 0.04 |

$\lambda$, the power transformation parameter; $\sigma$, the standard deviation of the normal error; intercept, $v_1$, $v_2$, and $v_3$ are regression parameters; Est., the estimates; SE, the estimated standard errors of the estimates.

TABLE 2
*Analysis results for the biology example*

| $\lambda$ | $\sigma$ | $v_0$ | $v_1$ | $v_2$ | $v_3$ | $v_4$ | $v_5$ |
|---|---|---|---|---|---|---|---|
| -0.82 | 0.36 | -1.35 | -0.64 | -1.84 | 1.19 | 0.98 | -0.45 |
| | | $v_6$ | $v_7$ | $v_8$ | $v_9$ | $v_{10}$ | $v_{11}$ |
| | | 0.55 | -0.19 | -1.44 | 0.70 | 0.70 | -0.51 |

$\lambda$, the power transformation parameter; $\sigma$, the standard deviation of the normal error; Est., the estimates; $v_0$, the intercept; $v_1$-$v_{11}$, the 11 dummy variables. The estimated standard error of the estimates $v_0$-$v_{11}$ is between 0.27 and 0.98.

candidate set $\mathcal{A}_\lambda = \{0, 1/2, -1/2, 1, -1\}$ as suggested in Hinkley and Runger (1984). Moreover, we used 3-fold cross-validation for choosing the tuning parameter. The adaptive oracle estimator is $\hat{\lambda} = 0$. It is noted that the estimated standard errors of the adaptive estimates for regression parameters are much smaller than those of the maximum likelihood estimator estimates. This agrees with our simulation findings, since when $\hat{\lambda}$ is in the candidate set, it is taken as known when making inference for the estimates of the regression coefficient vector.

Next, we consider the biological example from Hinkley and Runger (1984). The data consists of survival times from animals in a $3 \times 4$ factorial experiment. In this experiment, one unit of time equals ten hours. The two factors are treatment with four levels (A–D), and poison with three levels (I,II,III). There are 48 subjects, with four subjects in each one of the twelve treatment/poison combinations. As in Hinkley and Runger (1984), we consider a saturated model with 11 dummy variables indicating the 11 combinations. The baseline was taken as the combination of treatment A and poison I. The maximum likelihood estimator gives $\hat{\lambda} = -0.82$, which agrees with the results in Hinkley and Runger (1984). The results of the maximum likelihood estimator estimates are given in Table 2. For the proposed adaptive estimation method, we used the same candidate set as in the textile example with 4-fold cross-validation for choosing the tuning parameter. The penalization method also gives $\hat{\lambda} = -0.82$. That is, $\hat{\lambda}$ is not in the candidate set. Interestingly, in this example, the penalized and unpenalized estimates of the parameters are the same, indicating that there was no shrinkage of the transformation parameter. Therefore, the other estimates from our method are all the same as the maximum likelihood estimator estimates.

These two examples show the adaptivity of the proposed method for estimating the power transformation parameter given a candidate set, and then estimating the regression parameters and making the inference accordingly. The textile example provides rigorous treatment for scenarios where the estimated transformation is shrunk to a candidate value and may be treated as fixed, while the biological example evidences correct inferences in scenarios where the uncertainty in transformation estimation must be addressed via standard likelihood inference.

## 6. Concluding remarks

We have assumed throughout that the number of parameters in the model is finite, although we allowed the size of the candidate set for each of these parameters to grow to infinity as a function of the sample size. It is interesting to consider the asymptotic behavior and oracle properties of the proposed estimators for a model in which one allows the number of parameters to grow as a function of the number of observations. Similar research questions were investigated by Fan and Li (2001), Zou and Zhang (2009), and others, in the context of variable selection. It seems that for finite candidate values greater than one such proofs might be adapted, with this being a topic for future research.

The proposed penalized estimators are nonregular estimators. The irregularity arises because the estimators behave differently for parameters that are in the candidate set and close-by points. The irregularity poses challenges when constructing confidence regions to the parameters of interest. Discussion can be found in Pötscher and Leeb (2009), Pötscher and Schneider (2010), among others, for the variable-selection setting where each parameter has at most a single candidate point. In this work, we proposed a definition for an asymptotically almost everywhere confidence region and showed that our penalization based procedure yields inferences satisfying this definition. That is, the confidence region holds except on a set of parameter values close to the shrinkage points having asymptotically measure zero. Additional investigation is needed for understanding the properties of these nonregular estimators under general parametric models involving transformation, regression, and scale parameters when multiple candidate values are considered.

The proposed approach seems to simplify the predictions on the original untransformed scale. Typically, when fitting transformation models, one generally needs to account for estimation of both the regression parameter and the transformation parameter when making inference on the original scale. Such backtransformation procedures are greatly complicated by estimation of the transformation parameter. With our approach, one may ignore estimation of the transformation parameter when it is shrunk to one of the candidate values; otherwise, one must account for the estimation, similarly to other methods. The study of prediction on the original scale is thus an interesting topic for future research.

## Appendix A: Proofs

### *A.1. Proof of Theorem 1*

Define

$$\begin{aligned}
\Psi_n(u) \equiv & \Phi_n\left(\boldsymbol{\theta}_0 + \frac{\boldsymbol{u}}{\sqrt{n}}\right) - \Phi_n(\boldsymbol{\theta}_0) \\
= & L_n\left(V_1, \ldots, V_n; \boldsymbol{\theta}_0 + \frac{\boldsymbol{u}}{\sqrt{n}}\right) - L_n\left(V_1, \ldots, V_n; \boldsymbol{\theta}_0\right) \\
& - \sum_{j=1}^{p} a_{nj} \sum_{k=1}^{k_j} \hat{w}_j^k \left(\left|\theta_{0j} + \frac{u_j}{\sqrt{n}} - \theta_j^k\right| - \left|\theta_{0j} + \frac{u_j}{\sqrt{n}} - \theta_j^k\right|\right) \\
= & \frac{\boldsymbol{u}^T}{\sqrt{n}} \sum_{i=1}^{n} \dot{\ell}\left(V_i; \boldsymbol{\theta}_0\right) + \frac{1}{2} \frac{\boldsymbol{u}^T}{\sqrt{n}} \sum_{i=1}^{n} \ddot{\ell}\left(V_i; \check{\boldsymbol{\theta}}\right) \frac{\boldsymbol{u}}{\sqrt{n}} \\
& - \sum_{j=1}^{p} a_{nj} \sum_{k=1}^{k_j} \hat{w}_j^k \left(\left|\theta_{0j} + \frac{u_j}{\sqrt{n}} - \theta_j^k\right| - \left|\theta_{0j} + \frac{u_j}{\sqrt{n}} - \theta_j^k\right|\right) \\
\equiv & T_1^{(n)}(u) + T_2^{(n)}(u) - T_3^{(n)}(u)\,,
\end{aligned} \tag{12}$$

where $\check{\boldsymbol{\theta}}$ is between $\boldsymbol{\theta}_0$ and $\boldsymbol{\theta}_0 + \boldsymbol{u}/\sqrt{n}$.

By the central limit theorem, $T_1^{(n)}(u) \to_d u^T N(0, \Delta(\boldsymbol{\theta}_0))$. By the law of large numbers, Assumptions (A3) and (A4), and the continuous mapping theorem, $T_2^{(n)}(u) \to_d -u^T \Gamma(\boldsymbol{\theta}_0) \boldsymbol{u}$. Consider now the limiting behaviour of $T_3^{(n)}$. Recall that $\tilde{\theta}_{nj}$ is consistent for $\theta_{0j}$ for every $j \in \{1, \ldots, p\}$ (see Section 2.1). Hence, for every $k = 1, \ldots, k_j$, for $j \in \mathcal{A}$, and $k = 2, \ldots, k_j$, for $j \in \mathcal{A}^C$,

$$\hat{w}_j^k \equiv |\tilde{\theta}_{nj} - \theta_j^k|^{-\gamma} \to_P |\theta_{0j} - \theta_j^k|^{-\gamma} > 0\,.$$

Also,

$$\sqrt{n}\left(\left|\theta_{0j} + \frac{u_j}{\sqrt{n}} - \theta_j^k\right| - \left|\theta_{0j} - \theta_j^k\right|\right) \to u_j \cdot \operatorname{sign}(\theta_{0j} - \theta_j^k)\,.$$

Since $\frac{a_{nj}}{\sqrt{n}} \to 0$, we conclude that

$$\frac{a_{nj}}{\sqrt{n}} \hat{w}_j^k \sqrt{n}\left(\left|\theta_{0j} + \frac{u_j}{\sqrt{n}} - \theta_j^k\right| - \left|\theta_{0j} - \theta_j^k\right|\right) \to_P 0\,. \tag{13}$$

Recall that for $j \in \mathcal{A}^C$, $\theta_j^1 = \theta_{0j}$. Hence

$$\sqrt{n}\left(\left|\theta_{0j} + \frac{u_j}{\sqrt{n}} - \theta_j^1\right| - \left|\theta_{0j} - \theta_j^1\right|\right) = |u_j|\,,$$

and $n^{-1/2}a_{nj}\hat{w}_j^1 = a_{nj}n^{(\gamma-1)/2}|n^{1/2}(\tilde{\theta}_{nj}-\theta_{0j})|^{-\gamma}$ where $n^{1/2}(\tilde{\theta}_{nj}-\theta_{0j}) = O_P(1)$ (see Section 2.1). Thus, we obtain

$$a_{nj}\hat{w}_j^k\left(\left|\theta_{0j}+\frac{u_j}{\sqrt{n}}-\theta_j^k\right|-\left|\theta_{0j}-\theta_j^k\right|\right)\xrightarrow{\mathrm{P}}\begin{cases}0 & \text{if } j\in\mathcal{A},\\ 0 & \text{if } j\in\mathcal{A}^C \text{ and } k\geq 2\\ 0 & \text{if } j\in\mathcal{A}^C, k=1, \text{ and } u_j=0\\ \infty & \text{if } j\in\mathcal{A}^C, k=1, \text{ and } u_j\neq 0\end{cases} \tag{14}$$

We conclude that $\Psi_n(u)\to_d\Psi(u)$, where

$$\Psi(u)=\begin{cases}u_1^TW-\frac{1}{2}u_1^T\Gamma_{11}(\boldsymbol{\theta}_0)u_1 & \text{if } u_j=0, j\in\mathcal{A}^C,\\ -\infty & \text{otherwise}\end{cases} \tag{15}$$

where $W\sim N(\mathbf{0},\Delta_{11}(\boldsymbol{\theta}_0))$. Simple algebra shows that the maximizer of $\Psi(\boldsymbol{u})$ is $\hat{\boldsymbol{u}}=(\hat{\boldsymbol{u}}_1,\mathbf{0})$, where $\hat{\boldsymbol{u}}_1=\Gamma_{11}(\boldsymbol{\theta}_0)^{-1}W$.

Let $\hat{\boldsymbol{\theta}}_n=\operatorname{argmax}\Phi(\boldsymbol{\theta})$. Define $\hat{\boldsymbol{u}}_n=\operatorname{argmax}_{\boldsymbol{u}}\Psi_n(\boldsymbol{u})$; then $\hat{\boldsymbol{u}}_n=n^{1/2}(\hat{\boldsymbol{\theta}}_n-\boldsymbol{\theta}_0)$. We would like to show that $\hat{\boldsymbol{u}}_n\to_d\hat{\boldsymbol{u}}$. Note that $\Psi_n(\boldsymbol{u})$ for all $n\geq 1$, and $\Psi(\boldsymbol{u})$ are stochastic processes indexed by $\mathbb{R}^p$. The sample paths of $\Psi$ are upper semicontinuous and posses a unique maximum at $\hat{\boldsymbol{u}}$. Note that the inverse of $\Gamma_{11}(\boldsymbol{\theta}_0)$ is well defined by Assumption (A4). We would like to show that $\{\hat{\boldsymbol{u}}_n\}_n=O_P(1)$. To see that, we will show that in a probability that tends towards one, there is a local maximizer $\hat{\boldsymbol{u}}$ of $\Phi_n(\boldsymbol{u})$ such that $n^{-1/2}\hat{\boldsymbol{u}}_n=\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}_0=O_P(n^{-1/2})$. More specifically, we show that or any given $\varepsilon>0$, there exists a constant $C$ such that

$$P\left(\sup_{\|\boldsymbol{u}\|=C}\Phi_n(\boldsymbol{\theta}_0+n^{-1/2}\boldsymbol{u})<\Phi_n(\boldsymbol{\theta}_0)\right)\geq 1-\varepsilon\,. \tag{16}$$

By the Taylor expansion of (12), we obtain

$$\Phi_n(\boldsymbol{\theta}_0+n^{-1/2}\boldsymbol{u})-\Phi_n(\boldsymbol{\theta}_0)=T_1^{(n)}(\boldsymbol{u})+T_2^{(n)}(\boldsymbol{u})-T_3^{(n)}(\boldsymbol{u})\,.$$

Note that $T_1^{(n)}(\boldsymbol{u})=O_P(1)$, $T_2^{(n)}(\boldsymbol{u})=-\boldsymbol{u}^T\Gamma(\boldsymbol{\theta}_0)\boldsymbol{u}(1+o_P(1))$,

$$\begin{aligned}T_3^{(n)}(\boldsymbol{u})\geq&\sum_{j=1}^{p_1}\frac{a_{nj}}{n^{1/2}}\sum_{k=1}^{k_j}\hat{w}_j^k\sqrt{n}\left(\left|\theta_{0j}+\frac{u_j}{\sqrt{n}}-\theta_j^k\right|-\left|\theta_{0j}-\theta_j^k\right|\right)\\&+\sum_{j=p_1+1}^{p}\frac{a_{nj}}{n^{1/2}}\sum_{k=2}^{k_j}\hat{w}_j^k\sqrt{n}\left(\left|\theta_{0j}+\frac{u_j}{\sqrt{n}}-\theta_j^k\right|-\left|\theta_{0j}-\theta_j^k\right|\right)\end{aligned}$$

which is $o_P(1)$ by (13). Since by Assumption (A4), $\Gamma(\boldsymbol{\theta}_0)$ is negative definite, taking $C$ large enough, $T_2^{(n)}(\boldsymbol{u})$ dominates the other two terms, and thus (16) holds and $\{\hat{\boldsymbol{u}}_n\}_n$ is uniformly tight. Hence, all the conditions of the Argmax Theorem (Kosorok, 2008, Theorem 14.1) hold, and consequently we proved that $\hat{\boldsymbol{u}}_n\to_d\hat{\boldsymbol{u}}$. Summarizing, we have

$$\hat{\boldsymbol{u}}_{n1}\to_d\Gamma_{11}(\boldsymbol{\theta}_0)^{-1}W,\quad \hat{\boldsymbol{u}}_{n2}\to_d 0\,.$$

In other words, $n^{1/2}(\hat{\boldsymbol{\theta}}_{n1} - \boldsymbol{\theta}_{01}) \to_d N(\mathbf{0}, \Lambda_{11}(\boldsymbol{\theta}_0))$ and the normality part is proven.

We now move to prove the sparsity property. For all $j \in \mathcal{A}$, the asymptotically normality indicates that $P(\hat{\theta}_{nj} \neq \theta_{0j}, j = 1, \ldots, k_j) \to 1$. Thus, it suffice to show that for every $j \in \mathcal{A}^C$, $P(\hat{\theta}_{nj} = \theta_{0j}) \to 1$. It is sufficient to show that for any sequence of $\boldsymbol{\theta}_n$, satisfying $\|\boldsymbol{\theta}_{n1} - \boldsymbol{\theta}_{01}\| = O_P(n^{-1/2})$, and for any constant $C > 0$,

$$\Phi_n(\boldsymbol{\theta}_{n1}, \boldsymbol{\theta}_{02}) = \max_{\{\boldsymbol{\theta}_2 : \|\boldsymbol{\theta}_2 - \boldsymbol{\theta}_{02}\| \leq Cn^{-1/2}\}} \Phi_n(\boldsymbol{\theta}_{n1}, \boldsymbol{\theta}_2)\,. \tag{17}$$

For all $j \in \mathcal{A}^C$, $\partial\Phi_n(\boldsymbol{\theta})/\partial\theta_j$ exists for all $\boldsymbol{\theta}$, such that $\theta_j \neq \theta_j^k$ for some $k \in \{1, \ldots, k_j\}$. Hence, for any fixed constant $C$, and all $n$ large enough, $\partial\Phi_n(\boldsymbol{\theta})/\partial\theta_j$ exists for all $\boldsymbol{\theta}$ such that $\|\boldsymbol{\theta}_2 - \boldsymbol{\theta}_{02}\| \leq Cn^{-1/2}$ for which $\theta_j \neq \theta_{0j}$. Thus, in order to show (17), it is enough to show that with probability tending towards 1, $\partial\Phi_n(\boldsymbol{\theta}_n)/\partial\theta_j$ is positive for $\boldsymbol{\theta}_{nj} < \boldsymbol{\theta}_{0j}$ and negative for $\boldsymbol{\theta}_{nj} > \boldsymbol{\theta}_{0j}$ when $|\boldsymbol{\theta}_{nj} - \boldsymbol{\theta}_{0j}| < Cn^{-1/2}$ for all $j \in \mathcal{A}^C$.

By (3), the derivative

$$n^{-\frac{1}{2}} \frac{\partial\Phi_n(\boldsymbol{\theta}_{n1}, \boldsymbol{\theta}_2)}{\partial\theta_j} = n^{-\frac{1}{2}} \sum_{i=1}^{n} \frac{\partial\ell(V_i; \boldsymbol{\theta}_n)}{\partial\theta_j} - \frac{a_{nj}}{\sqrt{n}} \sum_{k=1}^{k_j} \hat{w}_j^k \operatorname{sign}(\theta_{nj} - \theta_j^k) \tag{18}$$

By Assumption (A3),

$$n^{-\frac{1}{2}} \sum_{i=1}^{n} \frac{\partial\ell(V_i; \boldsymbol{\theta}_n)}{\partial\theta_j} = n^{-\frac{1}{2}} \sum_{i=1}^{n} \frac{\partial\ell(V_i; \boldsymbol{\theta}_0)}{\partial\theta_j} + \frac{\sqrt{n}((\boldsymbol{\theta}_{n1}, \boldsymbol{\theta}_2) - \boldsymbol{\theta}_0)}{n} \sum_{i=1}^{n} \frac{\partial\ell(V_i; \check{\boldsymbol{\theta}})}{\partial\theta_j}$$

which equals $O_P(1)$, where $\check{\boldsymbol{\theta}}$ is between $(\boldsymbol{\theta}_{n1}, \boldsymbol{\theta}_2)$ and $\boldsymbol{\theta}_0$, and the last assertion follows since $\sqrt{n}((\boldsymbol{\theta}_{n1}, \boldsymbol{\theta}_2) - \boldsymbol{\theta}_0) = O_P(1)$ for all $\boldsymbol{\theta}_2$ such that $\|\boldsymbol{\theta}_2 - \boldsymbol{\theta}_{02}\| \leq Cn^{-1/2}$. As for the second expression in (18)

$$\begin{aligned}
\frac{a_{nj}}{\sqrt{n}} \sum_{k=1}^{k_j} \hat{w}_j^k \operatorname{sign}(\theta_j - \theta_j^k) =& a_{nj} n^{(\gamma-1)/2} |n^{1/2}(\tilde{\theta}_{nj} - \theta_{0j})|^{-\gamma} \operatorname{sign}(\theta_{nj} - \theta_{0j}) \\
&+ \frac{a_{nj}}{\sqrt{n}} \sum_{k=2}^{k_j} \left(|\theta_{0j} - \theta_j^k|^{-\gamma} + o_P(1)\right) \operatorname{sign}(\theta_{nj} - \theta_j^k) \\
&\to_P \operatorname{sign}(\theta_{nj} - \theta_{0j}) \cdot \infty\,,
\end{aligned}$$

since $a_{nj} n^{(\gamma-1)/2} \to \infty$, $|n^{1/2}(\tilde{\theta}_{nj} - \theta_{0j})| = O_P(1)$, and $n^{-1/2} a_{nj} \to 0$. Summarizing,

$$n^{-\frac{1}{2}} \frac{\partial\Phi_n(\boldsymbol{\theta}_{n1}, \boldsymbol{\theta}_2)}{\partial\theta_j} \to -\operatorname{sign}(\theta_{nj} - \theta_{0j}) \infty\,,$$

which proves that for all $n$ large enough, $\frac{\partial\Phi_n(\boldsymbol{\theta}_n)}{\partial\theta_j}$ is positive for $\boldsymbol{\theta}_{nj} < \boldsymbol{\theta}_{0j}$ and negative for $\boldsymbol{\theta}_{nj} > \boldsymbol{\theta}_{0j}$.

### A.2. Proof of Theorem 2

We need to prove that $\liminf P({\boldsymbol\theta_0}^{(J)} \in C_n) \geq 1-\alpha$. First, in order to ${\boldsymbol\theta_0}^{(J)} \in C_n$ we need $J_{n2} = J_2$ where $J_2 = J \cap \mathcal{A}^C$. Hence, we can write

$$\begin{aligned}
&\liminf P({\boldsymbol\theta_0}^{(J)} \in C_n) \\
&\quad = \liminf P({\boldsymbol\theta_0}^{(J_1)} \in C_n^{(J_1)}, J_{n2} = J_2) \\
&\quad \geq \liminf P({\boldsymbol\theta_0}^{(J_1)} \in C_n^{(J_1)} \mid \{J_{n2} = J_2\}) \liminf P(J_{n2} = J_2) \qquad (19)\\
&\quad \geq \liminf P({\boldsymbol\theta_0}^{(J_1)} \in C_n^{(J_1)} \mid \{\mathcal{A}_n = \mathcal{A}\}) \liminf P(\mathcal{A}_n^C = \mathcal{A}^C) \\
&\quad = \liminf P({\boldsymbol\theta_0}^{(J_1)} \in C_n^{(J_1)} \mid \{\mathcal{A}_n = \mathcal{A}\}),
\end{aligned}$$

where the one before last inequality holds since when $\mathcal{A}_n = \mathcal{A}$, we also have $J_{n2} = J_2$; and where the last inequality follows from Theorem 1, since $\liminf P(\{\mathcal{A}_n^C = \mathcal{A}^C) = 1$. When $\mathcal{A}_n = \mathcal{A}$ we have

$$\begin{aligned}
&{\boldsymbol\theta_0}^{(J_1)} \in C_n^{(J_1)} \\
\Leftrightarrow &\left((n\Lambda_n^{(\mathcal{A})})^{1/2}\right)^{(J_1)}({\boldsymbol\theta_0}^{(J_1)} - \hat{\boldsymbol\theta}_n^{(J_1)}) \in \left\{\left((n\Lambda_n^{(\mathcal{A})})^{1/2}\right)^{(J_1)}(C_n^{(J_1)} - \hat{\boldsymbol\theta}_n^{(J_1)})\right\} \\
\Leftrightarrow &\left((n\Lambda_n^{(\mathcal{A})})^{1/2}\right)^{(J_1)}({\boldsymbol\theta_0}^{(J_1)} - \hat{\boldsymbol\theta}_n^{(J_1)}) \in D_s\,.
\end{aligned}$$

Recall that by Theorem 1, $\left((n\Lambda_n^{(\mathcal{A})})^{1/2}\right)^{(J_1)}({\boldsymbol\theta_0}^{(J_1)} - \hat{\boldsymbol\theta}_n^{(J_1)})$ weakly converges to is a Gaussian random vector $Z$ with mean $\mathbf{0}$ and identity variance matrix of dimension equals to the cardinality of $J_1$. It thus follows from the Portmanteau Lemma (van der Vaart, 2000, Lemma 2.2) that

$$\begin{aligned}
&\liminf P\left({\boldsymbol\theta_0}^{(J_1)} \in C_n^{(J_1)} \mid \{\mathcal{A}_n = \mathcal{A}\}\right) \\
&\quad = \liminf P\left(({\boldsymbol\theta_0}^{(J_1)} - \hat{\boldsymbol\theta}_n^{(J_1)}) \in \{C_n^{(J_1)} - \hat{\boldsymbol\theta}_n^{(J_1)}\} \mid \{\mathcal{A}_n = \mathcal{A}\}\right) \\
&\quad = \liminf P\left(\left((n\Lambda_n^{(\mathcal{A})})^{1/2}\right)^{(J_1)}({\boldsymbol\theta_0}^{(J_1)} - \hat{\boldsymbol\theta}_n^{(J_1)}) \in D_s \mid \{\mathcal{A}_n = \mathcal{A}\}\right) \geq 1-\alpha\,, \qquad (20)
\end{aligned}$$

since $P(Z \in D_s) = 1-\alpha$. Substituting (20) in (19) and the result follows.

### A.3. Proof of Theorem 3

We define the sets $\Theta_n$ as follows. For each pair $(j,k)$, $j \in 1,\ldots,p$ and $k \in 1,\ldots,k_j$ and $\varepsilon > 0$, define the sets $G_{j,k}^{\varepsilon} \equiv \{\boldsymbol\theta \in \Theta : \theta_j^k - \varepsilon < \theta_j < \theta_j^k + \varepsilon\}$. Define the sets $H^{\varepsilon} \equiv \{\boldsymbol\theta \in \Theta : d(\boldsymbol\theta, \text{boundary}(\Theta)) < \varepsilon\}$ where $d$ is the Euclidean distance function. Define

$$\Theta_n \equiv \Theta / \left\{\left(\cup G_{j,k}^{n^{-1/2+\delta}}\right) \cup H^{n^{-1/2+\delta}}\right\}. \qquad (21)$$

for some fixed $0 < \delta < \frac{1}{2}$. Clearly, the Lebesgue measure of the sets $\Theta/\Theta_n$ converges to zero as $n \to \infty$. We now need to show that (7) holds.

It is enough to show that for every sequence $\boldsymbol{\theta}_n \in \Theta_n$ that converges to some $\boldsymbol{\theta}_0 \in \Theta$, we have that

$$\liminf_{n\to\infty} P_{\boldsymbol{\theta}_n}\left(C_n \in \boldsymbol{\theta}_n\right) \geq 1-\alpha\,. \tag{22}$$

Indeed, if (7) does not hold, then there is a sequence $\{\boldsymbol{\theta}_n\}$, $\boldsymbol{\theta}_n \in \Theta_n$, such that

$$\liminf_{n\to\infty} P_{\boldsymbol{\theta}_n}\left(\boldsymbol{\theta}_n \in C_n\right) = 1-\alpha-\varepsilon\,, \tag{23}$$

for some $\varepsilon > 0$. By the compactness of $\Theta$, this subsequence has a subsequence $\boldsymbol{\theta}_{n_k}$ that converges to some limit $\boldsymbol{\theta}_0$ for which (23) holds, and therefore (22) will not hold. Fix a sequence $\boldsymbol{\theta}_n \in \Theta_n$ that converges to some $\boldsymbol{\theta}_0 \in \Theta$. Note that

$$\begin{aligned} &L_n\left(V_1,\ldots,V_n;\boldsymbol{\theta}_n+\frac{\boldsymbol{u}}{\sqrt{n}}\right) - L_n\left(V_1,\ldots,V_n;\boldsymbol{\theta}_n\right) \\ &\quad= \frac{\boldsymbol{u}^T}{\sqrt{n}}\sum_{i=1}^{n}\dot{\ell}\left(V_i;\boldsymbol{\theta}_n\right) + \frac{1}{2}\frac{\boldsymbol{u}^T}{\sqrt{n}}\sum_{i=1}^{n}\ddot{\ell}\left(V_i;\check{\boldsymbol{\theta}}\right)\frac{\boldsymbol{u}}{\sqrt{n}} \equiv T_1^{(n)}(u)+T_2^{(n)}(u)\,, \end{aligned} \tag{24}$$

where $\check{\boldsymbol{\theta}}$ is between $\boldsymbol{\theta}_n$ and $\boldsymbol{\theta}_n + \boldsymbol{u}/\sqrt{n}$.

By the Lindeberg-Feller central limit theorem,

$$\Delta^{-1/2}(\boldsymbol{\theta}_n)T_1^{(n)}(u) \underset{\mathrm{P}_{\boldsymbol{\theta}_n}}{\to}{}_d\, u^T N(0,\Delta(\boldsymbol{\theta}_0))\,.$$

By the law of large numbers, Assumptions (A3) and (A4), and the continuous mapping theorem, $T_2^{(n)}(u) \to_{P_{\boldsymbol{\theta}_n}} -u^T\Gamma(\boldsymbol{\theta}_0)\boldsymbol{u}$. Hence, since $\boldsymbol{\theta}_n \to \boldsymbol{\theta}_0$, there exists a constant $C$ such that for all $n$ large enough

$$P_{\boldsymbol{\theta}_n}\left(\sup_{\|\boldsymbol{u}\|=C} L_n(\boldsymbol{\theta}_n+n^{-1/2}\boldsymbol{u}) < L_n(\boldsymbol{\theta}_n)\right) \geq 1-\varepsilon\,.$$

Since $\tilde{\boldsymbol{\theta}}_n$ is consistent to $\boldsymbol{\theta}_n$, we obtain that

$$P_{\boldsymbol{\theta}_n}\left(\left|\sqrt{n}(\tilde{\boldsymbol{\theta}}_n-\boldsymbol{\theta}_n)\right| < C\right) \geq 1-\varepsilon\,. \tag{25}$$

Define

$$\Psi_n(\boldsymbol{u}) \equiv \Phi_n\left(\boldsymbol{\theta}_n+\frac{\boldsymbol{u}}{\sqrt{n}}\right) - \Phi_n(\boldsymbol{\theta}_n) \equiv T_1^{(n)}(u)+T_2^{(n)}(u)-T_3^{(n)}(u)\,,$$

where $\check{\boldsymbol{\theta}}$ is between $\boldsymbol{\theta}_n$ and $\boldsymbol{\theta}_n + \boldsymbol{u}/\sqrt{n}$ and

$$T_3^{(n)} = \sum_{j=1}^{p} a_{nj}\sum_{k=1}^{k_j}\hat{w}_j^k\left(\left|\theta_{nj}+\frac{u_j}{\sqrt{n}}-\theta_j^k\right| - \left|\theta_{nj}+\frac{u_j}{\sqrt{n}}-\theta_j^k\right|\right)$$

The limiting distribution of $T_1^{(n)}$ and $T_2^{(n)}$ was discussed above. Consider the limiting behavior of $T_3^{(n)}$. Recall that $\hat{w}_j^k \equiv |\tilde{\theta}_{nj} - \theta_j^k|^{-\gamma}$. By (25), for all $n$ large enough, and for every $k = 1, \ldots, k_j$, and $j \in 1, \ldots, p$,

$$2|\theta_{0j} - \theta_j^k|^{-\gamma} > \left(|\theta_{nj} - \theta_j^k| - \frac{C}{\sqrt{n}}\right)^{-\gamma} > \hat{w}_j^k > \left(|\theta_{nj} - \theta_j^k| + \frac{C}{\sqrt{n}}\right)^{-\gamma} > 0\,.$$

where the left inequality follows since by the definition of $\Theta_n$, $(1 - 2^{-1/\gamma})|\theta_{nj} - \theta_j^k| > \frac{C}{\sqrt{n}}$ for all $n$ large enough.

Also, for all $n$ large enough, the sign of $\left|\theta_{nj} + \frac{u_j}{\sqrt{n}} - \theta_j^k\right|$ is constant and hence

$$\sqrt{n}\left(\left|\theta_{nj} + \frac{u_j}{\sqrt{n}} - \theta_j^k\right| - \left|\theta_{nj} - \theta_j^k\right|\right) \to u_j \cdot \operatorname{sign}(\theta_{0j} - \theta_j^k)\,.$$

Since $\frac{a_{nj}}{\sqrt{n}} \to 0$, we conclude that

$$\frac{a_{nj}}{\sqrt{n}}\hat{w}_j^k\sqrt{n}\left(\left|\theta_{0j} + \frac{u_j}{\sqrt{n}} - \theta_j^k\right| - \left|\theta_{0j} - \theta_j^k\right|\right) \underset{P_{\boldsymbol{\theta}_n}}{\to}{}_d\, 0\,.$$

We conclude that

$$\Psi_n(u) \underset{P_{\boldsymbol{\theta}_n}}{\to}{}_d\, \Psi(u) \equiv u^T W - \frac{1}{2}u^T\Gamma(\boldsymbol{\theta}_0)u\,,$$

where $W \sim N(\mathbf{0}, \Delta(\boldsymbol{\theta}_0))$. Simple algebra shows that the maximizer of $\Psi(\boldsymbol{u})$ is $\hat{\boldsymbol{u}} = \Gamma(\boldsymbol{\theta}_0)^{-1}W$. Using the same arguments as in the proof of Theorem 1, we can show that all the conditions of the Argmax Theorem (Kosorok, 2008, Theorem 14.1) hold, and thus

$$n^{1/2}(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_n) \underset{P_{\boldsymbol{\theta}_n}}{\to}{}_d\, N(\mathbf{0}, \Lambda(\boldsymbol{\theta}_0)^{-1})\,.$$

Similarly to the proof of Theorem 2, we can now show that that

$$\begin{aligned}
&\liminf P_{\boldsymbol{\theta}_n}\left({\boldsymbol{\theta}_0}^{(J)} \in C_n^{(J)} \mid \{\mathcal{A}_n = \{1, \ldots, p\}\}\right)\\
&\quad = \liminf P_{\boldsymbol{\theta}_n}\left((\boldsymbol{\theta}_0{}^{(J)} - \hat{\boldsymbol{\theta}}_n^{(J)}) \in \{C_n^{(J)} - \hat{\boldsymbol{\theta}}_n^{(J)}\} \mid \{\mathcal{A}_n = \{1, \ldots, p\}\}\right)\\
&\quad = \liminf P_{\boldsymbol{\theta}_n}\left(\left((n\Lambda_n)^{1/2}\right)^{(J)}(\boldsymbol{\theta}_0{}^{(J)} - \hat{\boldsymbol{\theta}}_n^{(J)}) \in D_p \mid \{\mathcal{A}_n = \{1, \ldots, p\}\}\right)\\
&\quad \geq 1 - \alpha\,,
\end{aligned}$$

by the Portmanteau Lemma (van der Vaart, 2000, Lemma 2.2), where by construction $P(Z \in D_p) = 1 - \alpha$.

### *A.4. Proof of Corollary 2*

*Proof.* Write

$$P\left(\theta \in C_n\right) = \int_\Theta P\left(\theta \in C_n \mid \theta = \vartheta\right) \mathrm{d}\vartheta$$

$$= \int_{\Theta_n} P\left(\vartheta \in C_n \mid \theta = \vartheta\right) \pi(\vartheta) \mathrm{d}\vartheta + \int_{\Theta/\Theta_n} P\left(\vartheta \in C_n \mid \theta = \vartheta\right) \pi_\theta(\vartheta) \mathrm{d}\vartheta$$

$$\geq \inf_{\vartheta \in \Theta_n} P(\vartheta \in C_n \mid \theta = \vartheta) \int_{\Theta_n} \pi_\theta(\vartheta) \mathrm{d}\vartheta + \int_{\Theta/\Theta_n} P\left(\vartheta \in C_n \mid \theta = \vartheta\right) \pi_\theta(\vartheta) \mathrm{d}\vartheta \,.$$

By Theorem 3 and the definition of $\Theta_n$ in (7),

$$\liminf_{n\to\infty} \inf_{\vartheta \in \Theta_n} P(\vartheta \in C_n \mid \theta = \vartheta) \int_{\Theta_n} \pi_\theta(\vartheta) \mathrm{d}\vartheta \geq 1 - \alpha \,.$$

Since $\pi_\theta(\vartheta)$ is bounded and the Lebesgue measure of $\Theta_n$ converges to zero, we have that

$$\int_{\Theta/\Theta_n} P\left(\vartheta \in C_n \mid \theta = \vartheta\right) \pi_\theta(\vartheta) \mathrm{d}\vartheta \to 0$$

and the result follows. □

### *A.5. Proof of Theorem 4*

Define

$$\Psi_n^{(n)}(u) \equiv \Phi_n\left(\boldsymbol{\theta}_0 + \frac{\boldsymbol{u}}{\sqrt{n}}\right) - \Phi_n(\boldsymbol{\theta}_0) = T_1^{(n)}(u) + T_2^{(n)}(u) - T_{3,n}^{(n)}(u) \tag{26}$$

where $T_1^{(n)}$ and $T_2^{(n)}$ are defined in (12) and where

$$T_{3,n}^{(n)}(u) \equiv \sum_{j=1}^{p} a_{nj} \sum_{k=1}^{k_{nj}} \hat{w}_j^k \left( \left| \theta_{0j} + \frac{u_j}{\sqrt{n}} - \theta_{j,k}^{(n)} \right| - \left| \theta_{0j} + \frac{u_j}{\sqrt{n}} - \theta_j^k \right| \right)$$

where $\check{\boldsymbol{\theta}}$ is between $\boldsymbol{\theta}_0$ and $\boldsymbol{\theta}_0 + \boldsymbol{u}/\sqrt{n}$.

The only difference between $\Psi_n^{(n)}$ and $\Psi_n$ defined in (12), is in the that the term $T_{3,n}^{(n)}$ replaces $T_3^{(n)}$.

Consider the limiting behavior of $T_{3,n}^{(n)}$. Recall that $\tilde{\theta}_{nj}$ is consistent for $\theta_{0j}$ for every $j \in \{1, \ldots, p\}$. First, by the inverse triangle inequality for every $k = 1, \ldots, k_{nj}$, and for every $j \in \mathcal{B}$, and $k = 2, \ldots, k_{nj}$, for $j \in \mathcal{B}^C$,

$$\begin{aligned} \hat{w}_j^k \equiv |\tilde{\theta}_{nj} - \theta_{j,k}^{(n)}|^{-\gamma} &= |\tilde{\theta}_{nj} - \theta_{0j} + \theta_{0j} - \theta_{j,k}^{(n)}|^{-\gamma} \\ &\leq \left( \left| |\theta_{0j} - \theta_{j,k}^{(n)}| - |\tilde{\theta}_{nj} - \theta_{0j}| \right| \right)^{-\gamma} \end{aligned}$$

Since $|\theta_{0j} - \theta_{j,k}^{(n)}| \geq \delta_n$, where $\delta_n^{-1} = o_P(n^{1/2})$, and $\tilde{\theta}_{nj} - \theta_{0j} = O_P(n^{-1/2})$,

$$\max_{k \in \{1, \ldots, k_{nj}\}} \hat{w}_j^k \leq \delta_n^{-\gamma} + o_P(1) \,.$$

Also, for fixed $u_j$, there exists $N_{0,j}$ such that for all $n \geq N_{0,j}$, $|u_j/\sqrt{n}| \leq |\theta_{0j} - \theta_{j,k}^{(n)}|$ for every $k = 1, \ldots, k_{nj}$, and for every $j \in \mathcal{B}$, and $k = 2, \ldots, k_{nj}$, for $j \in \mathcal{B}^C$. Hence, for all $n \geq N_{0,j}$,

$$\max_{k \in \{1,\ldots,k_{nj}\}} \sqrt{n} \left| \theta_{0j} + \frac{u_j}{\sqrt{n}} - \theta_{j,k}^{(n)} \right| - \left| \theta_{0j} - \theta_{j,k}^{(n)} \right| = u_j \cdot \operatorname{sign}(\theta_{0j} - \theta_{j,k}^{(n)}) \,.$$

Thus we conclude that for all $n \geq N_{0,j}$

$$\left| a_{nj} \sum_{k:\theta_{0j} \neq \theta_{j,k}^{(n)}} \hat{w}_j^k \left( \left| \theta_{0j} + \frac{u_j}{\sqrt{n}} - \theta_{j,k}^{(n)} \right| - \left| \theta_{0j} + \frac{u_j}{\sqrt{n}} - \theta_j^k \right| \right) \right|$$
$$\leq n^{-1/2} a_{nj} \eta_n |u_j| \max \hat{w}_j^k = n^{-1/2} a_{nj} \eta_n |u_j| (\delta_n^{-\gamma} + o_P(1)) \to_P 0 \,. \tag{27}$$

Recall that for $j \in \mathcal{B}^C$, $\theta_j^1 = \theta_{0j}$. Hence

$$\sqrt{n} \left( \left| \theta_{0j} + \frac{u_j}{\sqrt{n}} - \theta_j^1 \right| - \left| \theta_{0j} - \theta_j^1 \right| \right) = |u_j| \,,$$

and $\frac{a_{nj}}{n^{1/2}} \hat{w}_j^1 = a_{nj} n^{(\gamma-1)/2} |n^{1/2}(\tilde{\theta}_{nj} - \theta_{0j})|^{-\gamma}$ where $n^{1/2}(\tilde{\theta}_{nj} - \theta_{0j}) = O_P(1)$ (see Section 2.1). Thus, we obtain

$$a_{nj} \sum_{k=1}^{k_{nj}} \hat{w}_j^k \left( \left| \theta_{0j} + \frac{u_j}{\sqrt{n}} - \theta_{j,k}^{(n)} \right| - \left| \theta_{0j} + \frac{u_j}{\sqrt{n}} - \theta_j^k \right| \right) \xrightarrow{\mathrm{P}} \begin{cases} 0 & j \in \mathcal{B}, \\ 0 & j \in \mathcal{B}^C, u_j = 0 \\ \infty & j \in \mathcal{B}^C, u_j \neq 0 \end{cases} \tag{28}$$

We conclude that $\Psi_n^{(n)}(u) \to_d \Psi(u)$, where $\Psi(u)$ is defined in (15). The fact that $T_{3,n}(\boldsymbol{u}) \geq o_P(1)$ follows from (27). Thus, using similar arguments to those in the proof of Theorem 1, one can show that all the conditions of the Argmax Theorem (Kosorok, 2008, Theorem 14.1) hold, and consequently $\hat{\boldsymbol{u}}_n \to_d \hat{\boldsymbol{u}}$ where $\hat{\boldsymbol{u}}_n = \operatorname{argmax}_{\boldsymbol{u}} \Psi_n^{(n)}(\boldsymbol{u}) = \sqrt{n}(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0)$. In other words, $n^{1/2}(\hat{\boldsymbol{\theta}}_{n1} - \boldsymbol{\theta}_{01}) \to_d N(\mathbf{0}, \Lambda_{11}(\boldsymbol{\theta}_0))$ which proves the normality part.

The sparsity property can be proved similarly to the proof of Theorem 1. The main difference in the proof is that the expression for derivative of $\Phi_n^{(n)}$ is

$$n^{-\frac{1}{2}} \frac{\partial \Phi_n^n(\boldsymbol{\theta}_{n1}, \boldsymbol{\theta}_2)}{\partial \theta_j} = n^{-\frac{1}{2}} \sum_{i=1}^{n} \frac{\partial \ell(V_i; \boldsymbol{\theta}_n)}{\partial \theta_j} - \frac{a_{nj}}{\sqrt{n}} \sum_{k=1}^{k_{nj}} \hat{w}_j^k \operatorname{sign}(\theta_{nj} - \theta_{j,k}^{(n)}) \,, \tag{29}$$

Using (27), one can show that

$$n^{-\frac{1}{2}} \frac{\partial \Phi_n^{(n)}(\boldsymbol{\theta}_{n1}, \boldsymbol{\theta}_2)}{\partial \theta_j} \to -\operatorname{sign}(\theta_{nj} - \theta_{0j}) \infty \,,$$

which proves that for all $n$ large enough, $\frac{\partial \Phi_n^{(n)}(\boldsymbol{\theta}_n)}{\partial \theta_j}$ is positive for $\boldsymbol{\theta}_{nj} < \boldsymbol{\theta}_{0j}$ and negative for $\boldsymbol{\theta}_{nj} > \boldsymbol{\theta}_{0j}$ when $|\boldsymbol{\theta}_{nj} - \boldsymbol{\theta}_{0j}| < Cn^{-1/2}$ for all $j \in \mathcal{B}^C$.

TABLE 3
*Simulation results for* $\lambda_0 = 0$

| method | $\beta_{01}$ | | | | $\beta_{02}$ | | | | $\lambda_0$ | |
|---|---|---|---|---|---|---|---|---|---|---|
| | Bias | MAD | ESE | CP | Bias | MAD | ESE | CP | Bias | Freq. |
| | $\lambda_0 = 0$, $n = 100$ | | | | | | | | | |
| MLE | 0.10 | 0.66 | 0.89 | 0.97 | 0.03 | 0.26 | 0.35 | 0.96 | 0.00 | |
| Fixed | | | (0.10) | (0.22) | (0.10) | (0.56) | | | | |
| adaptive | 0.03 | 0.15 | 0.10 | 0.96 | 0.00 | 0.12 | 0.11 | 0.94 | 0.00 | 0.71 |
| oracle | -0.00 | 0.10 | | | -0.01 | 0.10 | | | | |
| | $\lambda_0 = 0$, $n = 400$ | | | | | | | | | |
| MLE | -0.02 | 0.33 | 0.37 | 0.96 | -0.01 | 0.13 | 0.14 | 0.94 | -0.00 | |
| Fixed | | | (0.05) | (0.22) | | | (0.05) | (0.55) | | |
| adaptive | 0.00 | 0.06 | 0.05 | 0.95 | 0.00 | 0.06 | 0.05 | 0.93 | 0.00 | 0.89 |
| oracle | -0.00 | 0.05 | | | -0.00 | 0.05 | | | | |
| | $\lambda_0 = 0$, $n = 800$ | | | | | | | | | |
| MLE | -0.00 | 0.22 | 0.25 | 0.94 | 0.00 | 0.09 | 0.10 | 0.94 | -0.00 | |
| Fixed | | | (0.04) | (0.23) | | | (0.04) | (0.56) | | |
| adaptive | -0.00 | 0.04 | 0.04 | 0.95 | -0.00 | 0.04 | 0.04 | 0.93 | 0.00 | 0.93 |
| oracle | 0.00 | 0.04 | | | 0.00 | 0.04 | | | | |
| | $\lambda_0 = 0$, $n = 1600$ | | | | | | | | | |
| MLE | -0.00 | 0.19 | 0.18 | 0.96 | 0.00 | 0.07 | 0.07 | 0.96 | 0.00 | |
| Fixed | | | (0.03) | (0.28) | | | (0.03) | (0.52) | | |
| adaptive | -0.00 | 0.03 | 0.03 | 0.92 | -0.00 | 0.03 | 0.03 | 0.93 | 0.00 | 0.94 |
| oracle | -0.00 | 0.03 | | | 0.00 | 0.02 | | | | |

MAD, the median absolute deviation of the estimates divided by 0.6745; ESE, the median of the estimated standard errors; CP, the empirical coverage probabilities of the Wald-type 95% confidence interval; Freq., the selection frequency of the true power transformation parameter over 500 runs.

## Appendix B: Simulation results

As described in Section 4, we conducted the following simulations. First, we considered five values for the true transformation parameter: $\lambda_0 = 0, 1/2, -1/2, 1, -1$. The regression parameters were set as $\boldsymbol{\beta}_0 = (\beta_{01}, 1)^T$, where $\beta_{01}$ is the intercept parameter. For $\lambda_0 = 0, 1/2$ we chose $\beta_{01} = 5$; for $\lambda_0 = 1$, $\beta_{01} = 8$; for $\lambda_0 = -1/2$, $\beta_{01} = -5$; and for $\lambda_0 = -1$, $\beta_{01} = -8$. The different choices of $\beta_{01}$ were chosen to ensure positive response. The candidate set $\mathcal{A}_\lambda$ equals $\{0, 1/2, -1/2, 2, -2\}$ for the transformation parameter. Note that $\mathcal{A}_\lambda$ does not include $1, -1$. For each setting, we conducted 500 simulation runs with sample sizes of $n = 100, 400, 800$ and 1600. The simulation results for $\lambda_0 = 0, 1/2$, and $-1/2$ are summarized in Tables 3 and 4. For such cases, $\lambda_0$ is contained in the candidate set $\mathcal{A}_\lambda$. The simulation results for $\lambda_0 = 1, -1$ are summarized in Table 5. For such cases, $\lambda_0$ is not contained in the candidate set $\mathcal{A}_\lambda$.

We also conducted a comparison between the prediction intervals based on the proposed adaptive method and the method of Cho et al. (2001). We considered the same simulation settings as above with the sample size of $n = 100$, 400. Five different prediction points are used, which are $x_0 = (1, 0)$, $(1, 1)$, $(1, -1)$, $(1, 2)$ and $(1, -2)$ with the first component being the intercept. The coverage probabilities of corresponding 95% prediction intervals and their standard errors are reported in Table 6

TABLE 4
*Simulation results for* $\lambda_0 = 1/2$ *and* $-1/2$

| method | $\beta_{01}$ | | | | $\beta_{02}$ | | | | $\lambda_0$ | |
|---|---|---|---|---|---|---|---|---|---|---|
| | Bias | MAD | ESE | CP | Bias | MAD | ESE | CP | Bias | Freq. |
| | | | | $\lambda_0 = 0.5$, $n = 100$ | | | | | | |
| MLE | 0.16 | 1.23 | 1.71 | 0.95 | 0.06 | 0.40 | 0.54 | 0.93 | -0.01 | |
| Fixed | | | (0.10) | (0.12) | | | (0.10) | (0.35) | | |
| adaptive | -0.17 | 0.17 | 0.10 | 0.91 | -0.05 | 0.14 | 0.11 | 0.88 | -0.04 | 0.67 |
| oracle | -0.00 | 0.10 | | | -0.01 | 0.10 | | | | |
| | | | | $\lambda_0 = 0.5$, $n = 400$ | | | | | | |
| -0.05 | 0.62 | 0.69 | 0.95 | -0.01 | 0.19 | 0.22 | 0.93 | -0.01 | | |
| Fixed | | | (0.05) | (0.14) | | | (0.05) | (0.38) | | |
| adaptive | -0.04 | 0.06 | 0.05 | 0.93 | -0.01 | 0.06 | 0.05 | 0.91 | -0.01 | 0.84 |
| oracle | -0.00 | 0.05 | | | -0.00 | 0.05 | | | | |
| | | | | $\lambda_0 = 0.5$, $n = 800$ | | | | | | |
| MLE | -0.01 | 0.44 | 0.49 | 0.94 | -0.00 | 0.14 | 0.16 | 0.94 | -0.01 | |
| Fixed | | | (0.04) | (0.13) | | | (0.04) | (0.36) | | |
| adaptive | -0.02 | 0.04 | 0.04 | 0.94 | -0.01 | 0.04 | 0.04 | 0.92 | -0.00 | 0.91 |
| oracle | 0.00 | 0.04 | | | 0.00 | 0.04 | | | | |
| | | | | $\lambda_0 = 0.5$, $n = 1600$ | | | | | | |
| MLE | -0.00 | 0.33 | 0.34 | 0.96 | 0.00 | 0.11 | 0.11 | 0.96 | -0.00 | |
| Fixed | | | (0.03) | (0.11) | | | (0.03) | (0.35) | | |
| adaptive | -0.01 | 0.03 | 0.03 | 0.93 | -0.00 | 0.03 | 0.03 | 0.94 | -0.00 | 0.92 |
| oracle | -0.00 | 0.03 | | | 0.00 | 0.02 | | | | |
| | | | | $\lambda_0 = -0.5$, $n = 100$ | | | | | | |
| MLE | -0.05 | 1.18 | 1.64 | 0.96 | 0.02 | 0.38 | 0.51 | 0.95 | 0.02 | |
| Fixed | | | (0.09) | (0.13) | | | (0.10) | (0.38) | | |
| adaptive | 0.19 | 0.17 | 0.10 | 0.90 | -0.05 | 0.14 | 0.11 | 0.88 | 0.04 | 0.64 |
| oracle | -0.00 | 0.10 | | | -0.01 | 0.10 | | | | |
| | | | | $\lambda_0 = -0.5$, $n = 400$ | | | | | | |
| MLE | -0.08 | 0.67 | 0.72 | 0.95 | 0.03 | 0.21 | 0.23 | 0.95 | -0.00 | |
| Fixed | | | (0.05) | (0.12) | | | (0.05) | (0.36) | | |
| adaptive | 0.03 | 0.06 | 0.05 | 0.94 | -0.01 | 0.06 | 0.05 | 0.92 | 0.01 | 0.88 |
| oracle | -0.00 | 0.05 | | | -0.00 | 0.05 | | | | |
| | | | | $\lambda_0 = -0.5$, $n = 800$ | | | | | | |
| MLE | -0.02 | 0.44 | 0.49 | 0.96 | 0.01 | 0.14 | 0.16 | 0.96 | 0.00 | |
| Fixed | | | (0.04) | (0.11) | | | (0.04) | (0.36) | | |
| adaptive | -0.02 | 0.04 | 0.04 | 0.95 | 0.00 | 0.04 | 0.04 | 0.94 | -0.00 | 0.92 |
| oracle | 0.00 | 0.04 | | | 0.00 | 0.04 | | | | |
| | | | | $\lambda_0 = -0.5$, $n = 1600$ | | | | | | |
| MLE | -0.02 | 0.35 | 0.34 | 0.96 | 0.01 | 0.11 | 0.11 | 0.95 | -0.00 | |
| Fixed | | | (0.03) | (0.11) | | | (0.03) | (0.34) | | |
| adaptive | -0.01 | 0.03 | 0.03 | 0.93 | 0.00 | 0.03 | 0.03 | 0.94 | -0.00 | 0.93 |
| oracle | -0.00 | 0.03 | | | 0.00 | 0.02 | | | | |

MAD, the median absolute deviation of the estimates divided by 0.6745; ESE, the median of the estimated standard errors; CP, the empirical coverage probabilities of the Wald-type 95% confidence interval; Freq., the selection frequency of the true power transformation parameter over 500 runs.

TABLE 5
*Simulation results for $\lambda_0 = 1$ and $-1$*

| method | $\beta_{01}$ | | | | $\beta_{02}$ | | | | $\lambda_0$ | |
|---|---|---|---|---|---|---|---|---|---|---|
| | Bias | MAD | ESE | CP | Bias | MAD | ESE | CP | Bias | Freq. |
| | | | | $\lambda_0 = 1$, $n = 100$ | | | | | | |
| MLE | 2.01 | 4.72 | 6.80 | 0.92 | 0.47 | 0.85 | 1.26 | 0.89 | -0.01 | |
| Fixed | | | (0.11) | (0.04) | | | (0.10) | (0.18) | | |
| adaptive | 1.82 | 5.40 | 5.82 | 0.79 | 0.45 | 0.86 | 1.07 | 0.77 | -0.04 | 0.12 |
| oracle | -0.00 | 0.10 | | | -0.01 | 0.10 | | | | |
| | | | | $\lambda_0 = 1$, $n = 400$ | | | | | | |
| MLE | 0.16 | 2.31 | 2.68 | 0.92 | 0.05 | 0.42 | 0.50 | 0.91 | -0.03 | |
| Fixed | | | (0.05) | (0.03) | | | (0.05) | (0.16) | | |
| adaptive | -0.05 | 2.42 | 2.64 | 0.88 | 0.02 | 0.43 | 0.48 | 0.87 | -0.05 | 0.08 |
| oracle | -0.00 | 0.05 | | | -0.00 | 0.05 | | | | |
| | | | | $\lambda_0 = 1$, $n = 800$ | | | | | | |
| MLE | 0.10 | 1.73 | 1.93 | 0.93 | 0.03 | 0.33 | 0.36 | 0.93 | -0.01 | |
| Fixed | | | (0.04) | (0.03) | | | (0.04) | (0.20) | | |
| adaptive | -0.06 | 1.71 | 1.89 | 0.92 | 0.00 | 0.32 | 0.35 | 0.91 | -0.03 | 0.02 |
| oracle | 0.00 | 0.04 | | | 0.00 | 0.04 | | | | |
| | | | | $\lambda_0 = 1$, $n = 1600$ | | | | | | |
| MLE | 0.04 | 1.29 | 1.33 | 0.95 | 0.01 | 0.24 | 0.24 | 0.95 | -0.01 | |
| Fixed | | | (0.03) | (0.02) | | | (0.00) | (0.15) | | |
| adaptive | -0.05 | 1.29 | 1.31 | 0.94 | -0.00 | 0.24 | 0.24 | 0.94 | -0.01 | 0.00 |
| oracle | -0.00 | 0.03 | | | 0.00 | 0.02 | | | | |
| | | | | $\lambda_0 = -1$, $n = 100$ | | | | | | |
| MLE | -1.49 | 4.23 | 6.03 | 0.92 | 0.38 | 0.72 | 1.09 | 0.88 | 0.05 | |
| Fixed | | | (0.09) | (0.04) | | | (0.09) | (0.17) | | |
| adaptive | -1.31 | 4.58 | 5.17 | 0.75 | 0.36 | 0.75 | 0.93 | 0.73 | 0.09 | 0.15 |
| oracle | -0.00 | 0.10 | | | -0.01 | 0.10 | | | | |
| | | | | $\lambda_0 = -1$, $n = 400$ | | | | | | |
| MLE | -0.70 | 2.68 | 2.91 | 0.95 | 0.15 | 0.49 | 0.53 | 0.94 | -0.01 | |
| Fixed | | | (0.05) | (0.04) | | | (0.05) | (0.17) | | |
| adaptive | -0.49 | 2.61 | 2.86 | 0.90 | 0.12 | 0.49 | 0.52 | 0.89 | 0.01 | 0.06 |
| oracle | -0.00 | 0.05 | | | -0.00 | 0.05 | | | | |
| | | | | $\lambda_0 = -1$, $n = 800$ | | | | | | |
| MLE | -0.25 | 1.69 | 1.94 | 0.95 | 0.06 | 0.31 | 0.36 | 0.95 | 0.00 | |
| Fixed | | | (0.04) | (0.03) | | | (0.04) | (0.15) | | |
| adaptive | -0.10 | 1.74 | 1.87 | 0.95 | 0.03 | 0.31 | 0.35 | 0.94 | 0.01 | 0.01 |
| oracle | 0.00 | 0.04 | | | 0.00 | 0.04 | | | | |
| | | | | $\lambda_0 = -1$, $n = 1600$ | | | | | | |
| MLE | -0.16 | 1.34 | 1.35 | 0.96 | 0.04 | 0.24 | 0.25 | 0.95 | -0.00 | |
| Fixed | | | (0.03) | (0.04) | | | (0.03) | (0.15) | | |
| adaptive | -0.08 | 1.33 | 1.34 | 0.95 | 0.02 | 0.24 | 0.24 | 0.94 | 0.00 | 0.00 |
| oracle | -0.00 | 0.03 | | | 0.00 | 0.02 | | | | |

MAD, the median absolute deviation of the estimates divided by 0.6745; ESE, the median of the estimated standard errors; CP, the empirical coverage probabilities of the Wald-type 95% confidence interval; Freq., the selection frequency of the true power transformation parameter over 500 runs.

TABLE 6
*Simulation results for coverage probabilities of prediction intervals and their standard deviations*

| $\lambda_0$ | $n$ | method | $x_0 = (1, 0)$ | $x_0 = (1, 1)$ | $x_0 = (1, -1)$ | $x_0 = (1, 2)$ | $x_0 = (1, -2)$ |
|---|---|---|---|---|---|---|---|
| 0.0 | 100 | MLE | 0.945 (18) | 0.943 (21) | 0.945 (22) | 0.940 (32) | 0.942 (33) |
| | | adaptive | 0.945 (17) | 0.945 (19) | 0.945 (20) | 0.943 (25) | 0.944 (28) |
| | 400 | MLE | 0.948 ( 8) | 0.949 (10) | 0.947 (10) | 0.948 (15) | 0.946 (15) |
| | | adaptive | 0.948 ( 8) | 0.948 ( 9) | 0.948 ( 9) | 0.948 (11) | 0.948 (11) |
| 0.5 | 100 | MLE | 0.945 (18) | 0.944 (21) | 0.943 (23) | 0.942 (29) | 0.939 (37) |
| | | adaptive | 0.945 (17) | 0.947 (18) | 0.942 (22) | 0.948 (22) | 0.936 (37) |
| | 400 | MLE | 0.948 ( 8) | 0.949 (10) | 0.947 (10) | 0.949 (14) | 0.945 (17) |
| | | adaptive | 0.948 ( 8) | 0.949 ( 9) | 0.948 ( 9) | 0.949 (10) | 0.947 (13) |
| -0.5 | 100 | MLE | 0.945 (18) | 0.942 (22) | 0.946 (22) | 0.936 (36) | 0.944 (30) |
| | | adaptive | 0.945 (17) | 0.941 (22) | 0.947 (20) | 0.934 (38) | 0.948 (25) |
| | 400 | MLE | 0.948 ( 8) | 0.948 (10) | 0.948 (10) | 0.947 (16) | 0.947 (14) |
| | | adaptive | 0.948 ( 8) | 0.948 ( 9) | 0.949 ( 9) | 0.947 (12) | 0.949 (10) |
| 1.0 | 100 | MLE | 0.945 (18) | 0.944 (21) | 0.944 (22) | 0.942 (29) | 0.940 (36) |
| | | adaptive | 0.945 (18) | 0.945 (21) | 0.943 (23) | 0.943 (30) | 0.938 (37) |
| | 400 | MLE | 0.948 ( 8) | 0.949 (10) | 0.947 (10) | 0.949 (14) | 0.945 (17) |
| | | adaptive | 0.948 ( 8) | 0.949 (10) | 0.947 (11) | 0.950 (14) | 0.944 (18) |
| -1.0 | 100 | MLE | 0.945 (18) | 0.943 (22) | 0.945 (21) | 0.937 (35) | 0.944 (30) |
| | | adaptive | 0.945 (18) | 0.942 (22) | 0.946 (22) | 0.935 (37) | 0.945 (30) |
| | 400 | MLE | 0.948 ( 8) | 0.948 (10) | 0.948 (10) | 0.948 (16) | 0.947 (14) |
| | | adaptive | 0.948 ( 8) | 0.948 (11) | 0.948 (10) | 0.946 (17) | 0.947 (15) |

The numbers in the parenthesis are the standard deviations of the coverage probabilities$\times 10^3$.

## Supplementary Material

### R Code

(doi: 10.1214/15-EJS1083SUPP; .zip). The R files that contains the code for the case studies analysis.

## References

P. J. BICKEL and K. A. DOKSUM. An analysis of transformations revisited. *Journal of the American Statistical Association*, 76(374):296–311, 1981. MR0624332

G. E. P. BOX and D. R. COX. An analysis of transformations. *Journal of the Royal Statistical Society. Series B*, 26(2):211–252, 1964. MR0192611

D. R. BRILLINGER. *A Festschrift For Erich L. Lehmann*, chapter, A Generalized Linear Model With "Gaussian" Regressor Variables, pages 97–114. Chapman and Hall, 1982.

R. J. CARROLL. Prediction and power transformations when the choice of power is restricted to a finite set. *Journal of the American Statistical Association*, 77(380):908–915, 1982. MR0686417

R. J. CARROLL and D. RUPPERT. *Transformation and weighting in regression*. Chapman and Hall, 1988. MR1014890

K. Cho, I. K. Yeo, R. A. Johnson, and W. Y. Loh. Prediction interval estimation in transformed linear models. *Statistics and Probability Letter*, 51:345–350, 2001. MR1820792

K. A. Doksum and C.-W. Wong. Statistical tests based on transformed data. *Journal of the American Statistical Association*, 78(382):411–417, 1983.

J. Fan and R. Li. Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association*, 96(456):1348–1360, 2001. MR1946581

J. H. Friedman. Fast sparse regression and classification. *International Journal of Forecasting*, 28(3):722–738, 2012.

Y. Goldberg, W. Lu, and J. Fine. Supplement to "Oracle estimation of parametric transformation models.". DOI:10.1214/15-EJS1083SUPP, 2016.

F. Hernandez and R. A. Johnson. The large-sample behavior of transformations to normality. *Journal of the American Statistical Association*, 75(372):855–861, 1980. MR0600967

D. V. Hinkley and G. Runger. The analysis of transformed data. *Journal of the American Statistical Association*, 79(386):302–309, 1984. MR0755087

M. R. Kosorok. *Introduction to Empirical Processes and Semiparametric Inference*. Springer, 2008. MR2724368

W. Lu, Y. Goldberg, and J. P. Fine. On the robustness of the adaptive lasso to model misspecification. *Biometrika*, 99:717–731, 2012. MR2966780

J. Lv and Y. Fan. A unified approach to model selection and sparse recovery using regularized least squares. *The Annals of Statistics*, 37(6A):3498–3528, 2009. MR2549567

B. M. Pötscher and H. Leeb. On the distribution of penalized maximum likelihood estimators: The Lasso, SCAD, and thresholding. *Journal of Multivariate Analysis*, 100(9):2065–2082, 2009. MR2543087

B. M. Pötscher and U. Schneider. Confidence sets based on penalized maximum likelihood estimators in Gaussian regression. *Electronic Journal of Statistics*, 4:334–360, 2010. MR2645488

T. M. Stoker. Consistent estimation of scaled coefficients. *Econometrica*, 54(6):1461–1481, 1986. MR0868152

R. Tibshirani. Regression shrinkage and selection via the Lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, 58(1):267–288, 1996. MR1379242

J. W. Tukey. *Exploratory Data Analysis*. Addison-Wesley, 1977.

A. W. van der Vaart. *Asymptotic Statistics*. Cambridge University Press, 2000. MR1652247

H. White. Maximum likelihood estimation of misspecified models. *Econometrica*, 50(1):1–25, 1982. MR0640163

I.-K. Yeo. Variable selection and transformation in linear regression models. *Statistics & Probability Letters*, 72(3):219–226, 2005. MR2137164

I. K. Yeo and R. A. Johnson. A new family of power transformations to improve normality or symmetry. *Biometrika*, 87(4):954–959, 2000. MR1813988

C. H. Zhang. Nearly unbiased variable selection under minimax concave penalty. *The Annals of Statistics*, 38(2):894–942, 2010. MR2604701

H. Zou. The adaptive Lasso and its oracle properties. *Journal of the American Statistical Association*, 101(476):1418–1429, 2006. MR2279469

H. Zou and H. H. Zhang. On the adaptive elastic-net with a diverging number of parameters. *The Annals of Statistics*, 37(4):1733–1751, 2009. MR2533470