entropy

MDPI

Article

Sparse Density Estimation with Measurement Errors

Xiaowei Yang [1,†], Huiming Zhang [2,3,*,†], Haoyu Wei [4,†] and Shouzheng Zhang [5]

[1] School of Mathematics and Statistics, Chaohu University, Hefei 238000, China; yxw8290@163.com
[2] Department of Mathematics, Faculty of Science and Technology, University of Macau, Macau 999078, China
[3] UMacau Zhuhai Research Institute, Zhuhai 519031, China
[4] Department of Statistics, North Carolina State University, Raleigh, NC 27695, USA; hwei4@ncsu.edu
[5] Graduate School of Arts and Science, Yale University, New Haven, CT 06510-8034, USA; shouzheng.zhang@yale.edu
* Correspondence: huimingzhang@um.edu.mo; Tel.: +86-152-7191-5190
† Xiaowei Yang, Huiming Zhang and Haoyu Wei are co-first authors which contributes equally to this work.

Abstract: This paper aims to estimate an unknown density of the data with measurement errors as a linear combination of functions from a dictionary. The main novelty is the proposal and investigation of the corrected sparse density estimator (CSDE). Inspired by the penalization approach, we propose the weighted Elastic-net penalized minimal ℓ_2-distance method for sparse coefficients estimation, where the adaptive weights come from sharp concentration inequalities. The first-order conditions holding a high probability obtain the optimal weighted tuning parameters. Under local coherence or minimal eigenvalue assumptions, non-asymptotic oracle inequalities are derived. These theoretical results are transposed to obtain the support recovery with a high probability. Some numerical experiments for discrete and continuous distributions confirm the significant improvement obtained by our procedure when compared with other conventional approaches. Finally, the application is performed in a meteorology dataset. It shows that our method has potency and superiority in detecting multi-mode density shapes compared with other conventional approaches.

Keywords: density estimation; Elastic-net; measurement errors; support recovery; multi-mode data

Citation: Yang, X.; Zhang, H.; Wei, H.; Zhang, S. Sparse Density Estimation with Measurement Errors. *Entropy* **2022**, *24*, 30. https://doi.org/10.3390/e24010030

Academic Editors: Andrea Prati, Carlos A. Iglesias, Luis Javier García Villalba and Vincent A. Cicirello

Received: 16 November 2021
Accepted: 15 December 2021
Published: 24 December 2021

1. Introduction

Over the years, the mixture models have been extensively applied to model unknown distributional shapes in astronomy, biology, economics, and genomics (see [1] and references therein). The distributions of real data involving potential complex variables often show multi-mode and heterogeneity. Due to the flexibility, it also appears in various distribution-based statistical techniques, such as cluster analysis, discriminant analysis, survival analysis, and empirical Bayesian inference. Flexible mixture models can naturally represent how the data are generated as mathematical artifacts. Theoretical results show that the mixture can approximate any density in the Euclidean space well, and the amount of the mixture can also be finite (for example, a mixture of several Gaussian distributions). Although the mixture model is inherently attractive to the statistical modeling, it is well-known that it is difficult to infer (see [2,3]). From the computational aspect, the optimization problems of mixture models are non-convex. Although existing computational methods, such as EM and various MCMC algorithms, can make the mixture model fit the data relatively easily. It should be emphasized that the mixture problems are essentially challenging, even unrecognizable, and the number of components (says, the order selection) is hard to determine (see [4]). There is a large amount of literature on its approximation theory, and various methods have been proposed to estimate the components (see [5] and references therein).

The nonparametric method and combinatorial method in density estimation were well studied in [6,7], as well as [8]. These can consistently estimate the number of the mixture's components when the components have known functional forms. When the number

of candidate components is large, the non-parametric method becomes computationally infeasible. Fortunately, the high-dimensional inference would compensate for this gap and guarantee the corrected identification of the mixture components with a probability tending to one. With the advancement of technology, high-dimensional problems have been applied at the forefront of statistical research. The high-dimensional inference method has been applied to the infinite mixture models with a sparse mixture of $p \to \infty$ components, which is an interesting and challenging problem (see [9,10]). We propose an improvement of the sparse estimation strategy proposed in [9], in which Bunea et al. propose a ℓ_1-type penalty [11] to obtain a sparse density estimate (SPADES). At the same time, we add a ℓ_2-type penalty and extend the oracle-inequality results to our new estimator.

In the real data, we often encounter the situation that the i.i.d. samples $X_i = Z_i + \varepsilon_i$ are contained by some zero-mean measurement errors $\{\varepsilon_i\}_{i=1}^n$; see [12–16]. For density estimation of $\{Z_i\}_{i=1}^n$, if there exists orthogonal basis functions, the estimation method is quite easy. In the measurement errors setting, however, finding an orthogonal-based density function is not easy (see [17]). Ref. [17] suggests the assumption that the conditional distribution function of X_i given Z_i is known. This condition is somewhat strong since most conditional distributions cannot obtain the explicit formula (except the Gaussian distribution). To address this predicament, particularly with nonorthogonal base functions, the SPADES model is attractive and makes the situation easier to deal with. Based on the SPADES method, our approach is an Elastic-net calibration approach, which is simpler and more interpreted than the conditional inference procedure proposed by [17]. In this paper, we proposed the corrected loss function to debiasing the measurement errors, and this is motivated by [18]. The main problem of considering measurement errors in various statistical models is that they are responsible for the bias of the classical statistical estimates; this is true, e.g., in linear regression, which has traditionally been the main focus of studies related to measurement errors. Debiasing represents an important task in various statistical models. In linear regression, it can be performed in the basic measurement errors (ME) model, which is also denoted as the Errors-in-variables (EIV or EV) model, if it is possible to estimate the variability of measurement errors (see [19,20]). We derive the real variable selection consistency based on weighted ℓ_1+ ℓ_2 penalty [21]. At the same time, some theoretical results of SPADES only contain the situation of the equal weights setting, which is not plausible in the sense of adaptive (data-dependent) penalized estimation. Moreover, we perform the Poisson mixture model to approximate the complex discrete distribution in the simulation part, while existing papers only emphasize the performance of continuous distribution models. Note that the multivariate kernel density estimator can only deal with a continuous distribution, and it requires a multivariate bandwidths section, while our method is dimensional-free (the number of the required tuning parameters is only two). There has been quite a lot of work in this area, starting with [22].

There are several differences between our article and [9]. The first point is that the upper bound of non-asymptotic oracle inequality in in our Theorems 1 and 2 is tighter than Theorems 1 and 2 in [9], and the optimal weighted tuning parameters are derived. The second point is that the ℓ_1-penalized techniques are applied in [9] to estimate the sparse density. Still, this paper considers the estimation of density functions in the presence of a classical measurement error. We opt to use an Elastic-net criterion function to estimate the density, which is taken to be approximated by a series of basis functions. The third point is that the tuning parameters are chosen by the coordinate descent algorithm in [9], and the mixture weights are calculated by the generalized bisection method (GBM). However, this paper directly calculates the optimal weights, so our algorithm is more accessible to implement than [9].

This paper is presented as follows. Section 2 introduces the density estimator, which can deal with measurement errors. This section introduces data-dependent weights for the Lasso penalty, and the weights are derived by the event of KKT conditions holding a high probability. In Section 3, we give a condition that can accurately estimate the weights of the mixture, with a probability tending to 1. We show that, in an increasing dimensional

mixture model under the local coherence assumption, if the tuning parameter is higher than the noise level, the recovery of the mixture components can hold with a high probability. In Section 4, we study the performance of our approach on artificial data generated from mixture Gaussian or Poisson distributions compared with other conventional methods, which indeed shows the improvement by employing our procedures. Moreover, the simulation also demonstrates that our method is better than the traditional EM algorithm, even under a low-dimensional model. Considering the multi-modal density aspect of the meteorology dataset, our proposed estimator has a stronger ability to detect multiple modes for the underlying distribution compared with other methods, such as SPADES or un-weighted Elastic-net estimator. Section 5 is the summary, and the proof of theoretical results is delivered in the Appendix A.

2. Density Estimation

2.1. Mixture Models

Suppose that $\{Z_i\}_{i=1}^n \in \mathbb{R}^d$ are independent random variables with a common unknown density h. However, the observations are contaminated with measurement errors $\{\varepsilon_i\}_{i=1}^n$ as latent variables, the observed data are actually $X_i = Z_i + \varepsilon_i$. Let $\{h_j\}_{j=1}^W$ be a series of density functions (such as Gaussian density or Poisson mass function), and $\{h_j\}_{j=1}^W$ are also called basis functions. Assume that the h belongs to the linear combination of $\{h_j\}_{j=1}^W$. The $W := W_n$ is a function of n, which is particularly intriguing for us since there may be $W \gg n$ (the high-dimensional setting). Let $\beta^* := (\beta_1^*, \cdots, \beta_W^*) \in \mathbb{R}^W$ be the unknown true parameter. Assume that

- (H.1): the $h := h_{\beta^*}$ is defined as

$$Z \sim h(z) := h_{\beta^*}(z) = \sum_{j=1}^{W} \beta_j^* h_j(z), \quad \text{with} \sum_{j=1}^{W} \beta_j^* = 1. \tag{1}$$

If the base is orthogonal and there are no measurement errors, a perfectly natural method is to estimate h by an orthogonal series of estimators in the form of $h_{\tilde{\beta}}$, where $\tilde{\beta}$ has the coordinates $\tilde{\beta}_j = \frac{1}{n}\sum_{i=1}^n h_j(X_i)$ (see [17]). However, this estimator depends on the choice of W, and a data-driven selection of W or the threshold needs to be adaptive. This estimator can only be applied to $W \leq n$. Nevertheless, we want to solve more general problems for $W > n$, and the base functions $\{h_j\}_{j=1}^W$ may not orthogonal.

We aim to achieve the best convergence for the estimator when the W is not necessarily less than n. Theorem 33.2 in [5] states that any smooth density can be well-approximated by a finite mixture of some continuous functions. However, Theorem 33.2 in [5] does not confirm how many components W are required for the mixture. Thus, the hypothesis of the increasing-dimensional W is reasonable. For discrete distributions, there is also a similar mixture density approximation—see Remark of Theorem 33.2 in [5].

2.2. The Density Estimation with Measurement Errors

This subsection aims to construct a sparse estimator for the density $h(z) := h_{\beta^*}(z)$ as a linear combination of known densities.

Recall the definition of the $L_2(\mathbb{R}^d)$ norm $\|f\| = \left(\int_{\mathbb{R}^d} f^2(x)dx\right)^{\frac{1}{2}}$. For $f, g \in L_2(\mathbb{R}^d)$, let $< f, g >= \int_{\mathbb{R}^d} f(x)g(x)dx$ be the inner product. If two functions f and g satisfy $< f, g >= 0$, then we call these two functions are orthogonal. Note that if the density $h(z)$ belongs to $L_2(\mathbb{R}^d)$ and assume that $\{X_i\}_{i=1}^n$ has the same distribution X, for any $f \in L_2$, we have $< f, h >= \int_{\mathbb{R}^d} f(x)h(x)dx = Ef(X)$. If $h(x)$ is the density function for a discrete distribution, the integral is replaced by summation, and we can define the inner product as $< f, h >:= \sum_{k\in\mathbb{Z}^d} f(k)h(k)$.

For true observations $\{Z_i\}_{i=1}^n$, we minimize the $\|h_\beta - h\|^2$ on $\beta \in \mathbb{R}^W$ to obtain the estimate of $h(z) := h_{\beta^*}(z)$, i.e., minimizing

$$\|h_\beta - h\|^2 = \|h\|^2 + \|h_\beta\|^2 - 2 < h_\beta, h >= \|h\|^2 + \|h_\beta\|^2 - 2Eh_\beta(Z) \propto -2Eh_\beta(Z) + \|h_\beta\|^2,$$

which implies that minimizing the $\|h_\beta - h\|^2$ is equivalent to minimizing

$$-2Eh_\beta(Z) + \|h_\beta\|^2 \approx -\frac{2}{n}\sum_{i=1}^{n} h_\beta(Z_i) + \|h_\beta\|^2. \tag{2}$$

It is plausible to assign more constrains for the candidate set of β in the optimization, for example, the ℓ_1 constrains $\|\beta\|_1 \le a$, where a is the tuning parameter. More adaptively, we prefer to use the weighted ℓ_1 restriction $\sum_{j=1}^W \omega_j|\beta_j| \le a$, where the weights ω_j's are data-dependent and will be specified later. From [23], we add Elastic-net penality $2\sum_{j=1}^W \omega_j|\beta_j| + c\sum_{j=1}^W \beta_j^2$ with tuning parameter c, which is in regards to the measurement errors (see [24,25]) for a similar purpose. We would have $c = 0$ in the situation without measurement errors. The c indeed becomes larger if the measurement errors become more serious, i.e., we can say that the c is proportional to the increasing variability of the measurement errors. It is different from SPADES since adjusting for the measurement errors is important for accurately describing the relationship between the observed varables and the outcomes of interest.

From the discussion above, now we propose the following *Corrected Sparse Density Estimator* (CSDE):

$$\hat{\beta} := \hat{\beta}(\omega_1, \cdots, \omega_W) = \arg\min_{\beta \in \mathbb{R}^W} \left\{ -\frac{2}{n}\sum_{i=1}^{n} h_\beta(X_i) + \|h_\beta\|^2 + 2\sum_{j=1}^{W} \omega_j|\beta_j| + c\sum_{j=1}^{W} \beta_j^2 \right\} \tag{3}$$

where the c is the tuning parameter for ℓ_2-penalty, and the c also presents the correction for adjusting the measurement errors in our observations.

For CSDE, if $\{h_j\}_{j=1}^W$ is an orthogonal system, it can be clearly seen that the CSDE estimator is consistent with the soft thresholding estimator, and the explicit solution is $\hat{\beta}_j = \frac{(1-\omega_j/|\tilde{\beta}_j|)_+\tilde{\beta}_j}{1+c}$, where $\tilde{\beta}_j = \frac{1}{n}\sum_{i=1}^n h_j(X_i)$ and $x_+ = \max(0, x)$. In this case, we can see that ω_j is the threshold of the j-th component of the simplest mean estimator $\tilde{\beta} = (\tilde{\beta}_1, \cdots, \tilde{\beta}_W)$.

From the sub-differential of the convex optimization, the corresponding Karush–Kuhn–Tucker conditions (necessary and sufficient first-order condition) for the minimizer in Equation (3) is

Lemma 1 (KKT conditions, Lemma 4.2 of [26]). *Let $k \in \{1, 2, \cdots, W\}$ and $c > 0$. Then, a necessary and sufficient condition for CSDE to be a solution of Equation* (3)

1. $\hat{\beta}_k :\neq 0$ *if* $\frac{1}{n}\sum_{i=1}^n h_k(X_i) - \sum_{j=1}^W \hat{\beta}_j < h_j, h_k > -c\hat{\beta}_k = w_k \text{sign}(\hat{\beta}_k)$.
2. $\hat{\beta}_k = 0$ *if* $\left|\frac{1}{n}\sum_{i=1}^n h_k(X_i) - \sum_{j=1}^W \hat{\beta}_j < h_j, h_k > -c\hat{\beta}_k\right| \le w_k$.

Since all values of β_j^* are non-negative, when conducting minimization in Equation (3), we have to put a non-negative restriction for optimizing Equation (3).

Due to the computational feasibility and optimal first-order conditions, we prefer an adaptively weighted Lasso penalty as a convex adaptive ℓ_1 penalization. We require that the larger weights are assigned to the coefficients of unimportant covariates, while significant covariates accompany the smaller weights. So, the weights represent the importance of the covariates. The larger (smaller) weights shrink to zero more easily (difficultly) than the unweighted Lasso, with appropriate or even optimal weights, leading to less bias and more efficient variable selection. The derivation of the weights will be given in Section 3.1.

In the end of this part, we will illustrate that in the mixture models, even without measurement errors, Equation (1) cannot be partially transformed into the linear model $Y = \mathrm{X}^T\beta + \varepsilon$, where Y is the n-dimensional response variables, X is the $W \times n$-dimensional fixed design matrix, β is a W-dimensional vector of model parameters, the ε is a $n \times 1$-dimensional vector for random error terms with zero mean and finite variance. Consider the least square objective function $U(\beta)$ for estimating β, $U(\beta) = (Y - \mathrm{X}^T\beta)^T(Y - \mathrm{X}^T\beta) = -2Y^T\mathrm{X}^T\beta + \beta^T\mathrm{X}\mathrm{X}^T\beta + Y^TY$. Minimizing $U(\beta)$ is equivalent to minimizing $U^*(\beta)$ in Formula (4)

$$U^*(\beta) = -2Y^T\mathrm{X}^T\beta + \beta^T\mathrm{X}\mathrm{X}^T\beta. \tag{4}$$

Comparing the objective function in Formula (4) with Equation (2), it is easy to obtain $Y = (\frac{1}{n}, \frac{1}{n}, \cdots, \frac{1}{n})^T$, $\beta = (\beta_1, \beta_2, \cdots, \beta_W)^T$, $\mathrm{X} = \begin{pmatrix} h_1(X_1) & \cdots & h_1(X_n) \\ \vdots & \ddots & \vdots \\ h_W(X_1) & \cdots & h_W(X_n) \end{pmatrix}$. Substituting Y, X and β into a linear regression model, we obtain

$$\begin{pmatrix} \frac{1}{n} \\ \vdots \\ \frac{1}{n} \end{pmatrix}_{n\times 1} = \begin{pmatrix} h_1(X_1) & \cdots & h_W(X_1) \\ \vdots & \ddots & \vdots \\ h_1(X_n) & \cdots & h_W(X_n) \end{pmatrix}_{n\times W} \begin{pmatrix} \beta_1 \\ \vdots \\ \beta_W \end{pmatrix}_{W\times 1} + \begin{pmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{pmatrix}_{n\times 1}.$$

Then,

$$\varepsilon_i = \frac{1}{n} - \sum_{j=1}^{W} \beta_j h_j(X_i),\ i = 1, 2, \cdots, n. \tag{5}$$

It can be seen from Equation (5) that the value of ε_i is no longer random if X was the fixed design matrix. Furthermore, even for a random design X, take the expectation on both sides of Equation (5), and one can find that the left side is not equal to the right side, that is, $E(\varepsilon_i) = 0 = \frac{1}{n} - \sum_{j=1}^{W} \beta_j E h_j(X_i)$. It leads to an additional requirement $\sum_{j=1}^{W} \beta_j E h_j(X_i) = \frac{1}{n} \to 0$, which is meaningless as $n \to \infty$, since all β_j and h_j are positive. This is a contradiction to $\sum_{j=1}^{W} \beta_j E h_j(X_i) > 0$ for all n.

Both of the two situations above contradict the definition of the assumed linear regression model. Hence, we cannot convert the estimation of Equation (1) into the estimation problem of linear models. Thus, the existing oracle inequalities are not applicable anymore, and we will propose new ones later. However, we can transform the mixture models to a corrected score Dantzig selector, such as in [27]. Although [10] studies the oracle inequalities for adaptive Dantzig density estimation, their study does not contain the error-in-variables framework and the support recovery content.

3. Sparse Mixture Density Estimation

In this section, we will present the oracle inequalities for estimators $\hat{\beta}$ and $h_{\hat{\beta}}$. The core of this section consists of five main results corresponding to the oracle inequalities for estimated density (Theorems 1 and 2), upper bounds on ℓ_1-estimation error (Corollaries 1 and 2) and support consistency (Theorem 3) as the byproduct of Corollary 2.

3.1. Data-Dependent Weights

The weights ω_j's are chosen adequately such that the KKT conditions for stochastic optimization problems have a high probability of being satisfied.

As mentioned before, the weights in Equation (3) rely on the observed data since we calculate the weights, ensuring the KKT conditions hold with a high probability. The weighted Lasso estimates could have less ℓ_1 estimation error than Lasso estimates (see the simulation part and [28]). Next, we need to consider what kind of data-dependent weight configuration can enable the KKT conditions to be satisfied with a high probability. A way to obtain data-dependent weights is to apply a concentration inequality for a weighted sum

of independent random variables. Moreover, the weights should be a known data function without any unknown parameters. A criterion can help obtain the weights grounded on McDiarmid's inequality (see [29] for more details).

Lemma 2. *Suppose $X_1, \cdots, X_n$ are independent random variables, and all values belong to a set A. Let $f : A^n \to \mathbb{R}$ be a function and satisfy the bounded difference conditions*

$$\sup_{x_1,\cdots,x_n,x_s' \in A} |f(x_1, \cdots, x_n) - f(x_1, \cdots, x_{s-1}, x_s', x_{s+1}, \cdots, x_n)| \le C_s,$$

then for all $t > 0$, $P\{|f(X_1, \cdots, X_n) - Ef(X_1, \cdots, X_n)| \ge t\} \le 2\exp\left\{-\frac{2t^2}{\sum_{s=1}^n C_s^2}\right\}$.

We define the KKT conditions of optimization evaluated at β^* (it is from the sub-gradient of the optimization function evaluated at β^*) by the events below:

$$\mathcal{F}_k(\omega_k) := \left\{ \left| \frac{1}{n}\sum_{i=1}^n h_k(X_i) - \sum_{j=1}^W \beta_j^* < h_j, h_k > -c\beta_k^* \right| \le \omega_k \right\}, k = 1, 2, \cdots, W.$$

Assume that

- (H.2): $\exists\, L_k > 0$ s.t. $\|h_k\|_\infty = \max\limits_{1\le i\le n} |h_k(X_i)| \le 2L_k$;
- (H.3): $0 < \max\limits_{1\le j\le W} |\beta_j^*| \le B$.

(H.2) is an assumption in sparse ℓ_1 estimation, and the assumption (H.3) is a classical compact parameter space assumption in sparse high-dimensional regressions (see [9,25]).

Next, we check that the event $\mathcal{F}_k(\omega_k)$ is hold with high probability. Note that $Eh_k(X_i) = \sum_{j=1}^W \beta_j^* < h_j, h_k >$ (which is free of X_i), we find

$$\frac{1}{n}\left|\sum_{i=1}^n h_k(X_i) - (\sum_{i\ne s}^n h_k(X_i) + h_k(X_s'))\right| = \frac{1}{n}\left|h_k(X_s) - h_k(X_s')\right| \le \frac{1}{n}(|h_k(X_s)| + |h_k(X_s')|) \le \frac{4L_k}{n},$$

where the last inequality is due to $|h_k(X_i)| \le \max\limits_{1\le i\le n} |h_k(X_i)| \le 2L_k$.

Next, we apply the McDiarmid's inequality on the event $\mathcal{F}_k^c(\omega_k)$ by (H.3). Then

$$\begin{aligned} P(\mathcal{F}_k^c(\omega_k)) &= P\left\{ \left| \frac{1}{n}\sum_{i=1}^n h_k(X_i) - \sum_{j=1}^W \beta_j^* < h_j, h_k > -c\beta_k^* \right| \ge \omega_k \right\} \\ \text{(by (H.3))} \ &\le P\left\{ \left| \frac{1}{n}\sum_{i=1}^n h_k(X_i) - Eh_k(X_i) \right| \ge \omega_k - cB \right\} \\ \text{(define } \tilde{\omega}_k := \omega_k - cB > 0) \ &\le 2\exp\left\{ -\frac{2\tilde{\omega}_k^2}{16L_k^2/n} \right\} = 2\exp\left\{ -\frac{n\tilde{\omega}_k^2}{8L_k^2} \right\} =: \frac{\delta}{W}, \quad 0 < \delta < 1. \end{aligned}$$

Considering the previous line,

$$\omega_k := 2\sqrt{2}L_k\sqrt{\frac{1}{n}\log\frac{2W}{\delta}} + cB =: 2\sqrt{2}L_k v(\delta/2) + cB, \text{ where } v = v(\delta) := \sqrt{\frac{1}{n}\log\frac{W}{\delta}}. \quad (6)$$

The weight ω_k in our paper is different from [9], which gives the un-shift version ($\check{\omega}_k = 4L_k\sqrt{\frac{1}{n}\log\frac{W}{\delta/2}}$), due to the Elastic-net penalty. Define the modified KKT conditions:

$$\mathcal{K}_k(\omega_k) := \{|\frac{1}{n}\sum_{i=1}^n h_k(X_i) - \sum_{j=1}^W \beta_j^* < h_j, h_k > | \le \tilde{\omega}_k\}, k = 1, 2, \cdots, W \quad (7)$$

which hold with probability of at least $1 - 2\exp\left\{-\frac{n\tilde{\omega}_k^2}{8L_k^2}\right\}$.

3.2. Non-Asymptotic Oracle Inequalities

Introduced by [30], oracle inequality is a powerful non-asymptotic and analytical tool that seeks to provide the distance between the obtained estimator and a true estimator. The sharp oracle inequality connects the optimal convergence of an obtained estimator compared with the true parameter (see [31,32]).

For $\forall \beta \in \mathbb{R}^W$, let $I(\beta) = \{j \in \{1, \cdots, W\} : \beta_j \neq 0\}$ be the indices corresponding to the non-zero components of the vector β, i.e., the support in mathematical jargon. If there is no ambiguity, we would like to write $I(\beta^*)$ as I_* for simplicity. Define $W(\beta) = \sum_{j=1}^{W} I(\beta_j \neq 0)$ as the number of its non-zero components, where $I(\cdot)$ represents the indicative function. Let $\sigma_j^2 = Var(h_j(X_1))$, $1 \leq j \leq W$.

Below, we will state the non-asymptotic oracle inequalities for $h_{\hat{\beta}}$ (with the probability at least$1 - \delta(W, n)$ for any integer W and n), which measures the L_2 distance between $h_{\hat{\beta}}$ and h. For $\beta \in \mathbb{R}^W$, define the correlation for the two base densities: h_i and h_j, $\rho_W(i, j) = \frac{<h_i, h_j>}{\|h_i\|\|h_j\|}$, $i, j = 1, \cdots, W$. Our results will be established under the local coherence condition, and we define the maximal local coherence as:

$$\rho(\beta) = \max_{i \in I(\beta)} \max_{j \neq i} |\rho_W(i, j)|.$$

It is easy to see that $\rho(\beta)$ measures the separation of the variables in the set $I(\beta)$ from one another and the rest. The degree of separation is measured in terms of the size of the correlation coefficients. However, the regular condition introduced by this coherence may be too strong. It may exclude cases that the "correlation" can be relatively significant for a small number of pairs (i, j) and almost zero otherwise. Thus, we consider the definition of the cumulative local coherence given by [9]: $\rho_*(\beta) = \sum_{i \in I(\beta)} \sum_{j > i} |\rho_W(i, j)|$. Define $H(\beta) = \max_{j \in I(\beta)} \frac{\omega_j}{v(\delta/2)\|h_j\|}$, $F = \max_{1 \leq j \leq W} \frac{v(\delta/2)\|h_j\|}{\tilde{\omega}_j} = \max_{1 \leq j \leq W} \frac{\|h_j\|}{2\sqrt{2}L_j}$, where $\tilde{\omega}_j := 2\sqrt{2}L_j v(\delta/2)$.

By using the definition of $\rho_*(\beta)$ and the notations above, we present the main results of this paper, which lays the foundation for the oracle inequality of the estimated mixture coefficients.

Theorem 1. *Under (H.1)–(H.3), let* $c = \frac{\min_{1 \leq j \leq W}\{\tilde{\omega}_j\}}{B}$ *and a given constant* $0 < \gamma \leq 1$. *If the true base functions* $\{h_j\}_{j=1}^W$ *conform to the cumulative local coherence assumption for all* $\beta \in \mathbb{R}^W$,

$$12FH(\beta)\rho_*(\beta)\sqrt{W(\beta)} \leq \gamma, \tag{8}$$

then the $\hat{\beta}$ *of the optimization problem in Equation (3) has the following oracle inequality with a probability at least* $1 - \delta$,

$$\begin{aligned} &\|h_{\hat{\beta}} - h\|^2 + \frac{\alpha_{opt1}(1-\gamma)}{(\alpha_{opt1}-1)} \sum_{j=1}^{W} \tilde{\omega}_j |\hat{\beta}_j - \beta_j| + \frac{\alpha_{opt1}}{\alpha_{opt1}-1} \sum_{j=1}^{W} c(\hat{\beta}_j - \beta_j)^2 \\ &\leq \frac{\alpha_{opt1}+1}{\alpha_{opt1}-1}\|h_\beta - h\|^2 + \frac{18\alpha_{opt1}^2}{\alpha_{opt1}-1} H^2(\beta) v^2(\delta/2) W(\beta), \end{aligned}$$

where $\alpha_{opt1} = 1 + \sqrt{1 + \frac{\|h_\beta - h\|^2}{9H^2(\beta)v^2(\delta/2)W(\beta)}}$.

It is worthy to note that here we use $\sqrt{W(\beta)}$ instead of $W(\beta)$, and the latter is used in [9]. The upper bound of the oracle inequality by Theorem 1 is sharper than the upper bound of Theorem 1 in [9]. Further, we give the value of the optimal α_{opt1}, but [9] did not give it. The reason for this phenomenon is quite clean actually: from the proof, it is due to ineuqality (A5). Now, let us address the sparse Gram matrix $\Psi_W = (< h_i, h_j >)_{1 \leq i,j \leq W}$ with a small number of non-zero elements in off-diagonal positions, define $\psi_W(i, j)$ as the

element (i,j)-th of position ψ_W. Condition (8) in Theorem 1 can be transformed to the condition

$$12SH(\beta)\sqrt{W(\beta)} \leq \gamma,$$

where the number S is called the sparse index of matrix Ψ_W, which is defined as $S = |\{(i,j) : i,j \in \{1,\cdots,W\}, i > j \text{ and } \psi_W(i,j) \neq 0\}|$, where $|A|$ is the number of elements of set A.

Sometimes the assumption in Condition (8) does not imply the positive definiteness of Ψ_W. Next, we give a similar oracle inequality that is valid under the hypothesis that the Gram matrix Ψ_W is positive definite.

Theorem 2. *Under the assumption of (H.1)–(H.3) and that the Gram matrix Ψ_W is positive definite with a minimum eigenvalue greater than or equal to $\lambda_W > 0$. For all $\beta \in \mathbb{R}^W$, the $\hat{\beta}$ of the optimization problem in Equation (3) has the following oracle inequality with probability at least $1-\delta$,*

$$\begin{aligned}&\|h_{\hat{\beta}} - h\|^2 + \frac{\alpha_{opt2}}{\alpha_{opt2}-1}\sum_{j=1}^{W}\tilde{\omega}_j|\hat{\beta}_j - \beta_j| + \frac{\alpha_{opt2}}{\alpha_{opt2}-1}\sum_{j=1}^{W}c(\hat{\beta}_j - \beta_j)^2\\ &\leq \frac{\alpha_{opt2}+1}{\alpha_{opt2}-1}\|h_\beta - h\|^2 + \frac{576\alpha_{opt2}^2}{\alpha_{opt2}-1}\frac{G}{\lambda_W}v^2(\delta/2),\end{aligned}$$

where $G = G(\beta) := \sum_{j\in I(\beta)} L_j^2$ and $\alpha_{opt2} = 1 + \sqrt{1 + \frac{\|h_\beta - h\|^2}{288\frac{G}{\lambda_W}v^2(\delta/2)}}$.

Remark 1. *The argument and result of Theorem 1 in this paper is more refined than the conclusion of Theorem 1 in [9] for Lasso by letting $\gamma = 1/2$ and $c = 0$. In addition, Theorems 1 and 2 of this paper, respectively, give the optimal α value of the density estimation oracle inequalities, namely α_{opt1}, α_{opt2}. It provides a potentially sharper bound for the ℓ_1-estimation error bound.*

Next, we will present the ℓ_1-estimation error for the estimator $\hat{\beta}$ by Equation (3), and the weights are defined by Equation (6). For the technical point, we consider that $\|h_j\| = 1$ for all j in Equation (3), i.e., the base functions are normalized. This normalization mimics the covariates' standardization procedure when doing penalized estimations in generalized linear models. For simplicity, we put $L := \max_{1\leq j\leq W} L_j$.

For any other choice of $v(\delta/2)$ greater than or equal to $\sqrt{\frac{1}{n}\log\frac{2W}{\delta}}$, the conclusions of Section 3 are valid with a high probability. It imposes a restriction on the predictive performance of CSDE. As pointed out in [33], for the ℓ_1-penalty in the regression, the adjusted sequence ω_j required for the corrected selection is usually larger than the adjusted sequence ω_j that produces a good prediction. The selection of the mixture density shown below is also true. Specifically, we will take the value $\beta = \beta^*$ and $v = v(\delta/2W) = \sqrt{\frac{\log(2W^2/\delta)}{n}}$, then $\alpha_{opt1}, \alpha_{opt2} = 2$. Below, we give the Corollaries of Theorems 1 and 2.

Corollary 1. *Given the same conditions as Theorem 1 with $\|h_j\| = 1$ for all j, let $\alpha_{opt1} = 2$, then we have the following ℓ_1-estimation error oracle inequality:*

$$\sum_{j=1}^{W}|\hat{\beta}_j - \beta_j^*| \leq \frac{72\sqrt{2}v(\delta/2W)W(\beta^*)}{1-\gamma}\frac{(L+L_{\min})^2}{L_{\min}} \tag{9}$$

with probability at least $1-\delta/W$, where $L_{\min} = \min_{1\leq j\leq W} L_j$.

Corollary 2. *Given the same conditions as Theorem 2 with* $\|h_j\| = 1$ *for all* j*, let* $\alpha_{opt2} = 2$*, then we have the following oracle inequality, with probability at least* $1 - \delta/W$*,*

$$\sum_{j=1}^{W} |\hat{\beta}_j - \beta_j^*| \leq \frac{288\sqrt{2}v(\delta/2W)G^*}{L_{\min}\lambda_W}, \text{ where } G^* = \sum_{j\in I_*} L_j^2.$$

If the number $W(\beta^*)$ of the mixture indicator elements is much smaller than $\sqrt{n}$, then inequality (9) guarantees that the estimated $\hat{\beta}$ is close to the true β^*, and the ℓ_1-estimation error will be presented in the numerical simulation in Section 4. Our results of Corollaries 1 and 2 are non-asymptotic for any W and n. The oracle inequalities are guiders for us to find an optimal tuning parameter with order $O(\sqrt{\frac{\log W}{n}})$ for a sharper estimation error and better prediction performance. This is also an intermediate and crucial result, which leads to the main results of correctly identifying the mixture components in Section 3.3. In the following section, we turn to cope with the identification of I_*. Corrected components are selected by the proposed oracle inequality for the weighted ℓ_1+ ℓ_2 penalty.

3.3. Corrected Support Identification of Mixture Models

In this section, we will study the results of the support recovery of our CSDE estimator. There are few results on support recovery, while most of the results are the consistency of the ℓ_1-error and prediction errors. Here, we borrow the framework of [25,33]. They give many proof techniques to deal with the corrected support identification in linear models by $\ell_1 + \ell_2$ regularization. Let $\hat{I}$ be the set of indicators consisting of non-zero elements of $\hat{\beta}$ in the given Equation (3). In other words, $\hat{I}$ is an estimate of the true support set $I(\beta^*) := I_*$. We will study $P(\hat{I} = I(\beta^*)) \geq 1 - \varepsilon$ for a given $0 < \varepsilon < 1$ under some mild conditions.

To identify the I_* consistently, we need more assumptions about some special correlation conditions than ℓ_1-error consistency.

Condition (A): $\rho_*(\beta^*) \leq \frac{LL_{\min}\lambda_W}{288G^*}$.

Moreover, we need an additional condition that the minimal signal should be higher than a threshold level and quantified by order of tuning parameter. Therefore, we state it as follows:

Condition (B): $\min_{j\in I^*} |\beta_j^*| \geq 4\sqrt{2}v(\frac{\delta}{2W})L$, where $v(\frac{\delta}{2W}) := \sqrt{\frac{1}{n}\log\frac{2W^2}{\delta}}$.

When performing simulation, Condition (B) is the theoretical guarantee that the minor magnitude of β_j must be greater than a threshold value as a minimal signal condition. It is also called the Beta-min condition (see [26]).

Theorem 3. *Let* $0 < \delta < \frac{1}{2}$ *and define* $\epsilon_k := |E[h_k(X_1)] - E[h_k(Z_1)]|$*. Assume that both conditions (A) and (B) are true and give the same conditions as Corollary 2, then*

$$P(\hat{I} = I_*) \geq 1 - \left(4W\left(\frac{\delta}{2W^2}\right)^{(1-\epsilon_k^*)^2} + 2\delta\right), \text{ where } \epsilon_k^* = \epsilon_k/\sqrt{2}v(\delta/2W)L.$$

Under the Beta-min condition, the support estimation is very close to the true support of β_j^*. The probability of the event $\{\hat{I} = I_*\}$ is high when W is growing. The $\hat{\beta}$ recovers the corrected support with probability at least $1 - (4W(\frac{\delta}{2W^2})^{(1-\epsilon_k^*)^2} + 2\delta)$. The result is non-asymptotic and it is true for any fixed W and n. There is a similar conclusion about support consistency (see Theorem 6 of [25]).

4. Simulation and Real Data Analysis

Ref. [9] proposes the SPADES estimation to deal with the samples for sparse mixture density, and they also derive an algorithm from complementing their theoretical result. Their findings successfully handle the high-dimensional adaptive density estimation to some degree. However, their algorithm is costly and unstable. In this section, we deal with the tuning parameter directly and compare our CSDE method with the SPADES method in [9] and other similar methods. In all cases here, we fix $n = 100$ for $W = 81, 131, 211, 321$, which is known as the dimension of the unknown parameter β^*. The performance of each estimator is evaluated by the ℓ_1-estimation error and the total variation (TV) distance between the estimator and the true value of β^*. The total variation (TV) error is defined as: $\mathrm{TV}(h_{\beta^*}, h_{\hat{\beta}}) = \int |h_{\beta^*}(x) - h_{\hat{\beta}}(x)|\, dx$.

4.1. Tuning Parameter Selection

In [9], the λ_1 is chosen by the coordinate descent method, while the mixture weights are detected by GBM. However, in our article, the optimal weights can be computed directly. Thus, it is much easier to carry out than [9]. The ℓ_1-penalty term $\sum_{j=1}^{W} \omega_j |\beta_j|$ with optimal weights are defined by $\omega_k := 2\sqrt{2} L_k v(\delta/2) + cB$, where $L_j = \|h_j\|_\infty$, which usually can be computed easily for a continuous h_j.

For a discrete base density $\{h_j\}_{j=1}^{W}$, it can be estimated as the following approximation by using concentration inequalities from Exercise 4.3.3 of [34]: $|\mathrm{med}(X) - E(X)| \leq \sqrt{2Var(X)}$, $\bar{x} \approx x_{med}\big(1 + O(n^{-1})\big) \approx h^{-1}(L_j)\big(1 + O(n^{-1})\big)$, where $\bar{x}$ and x_{med} represent the sample mean and sample median, respectively, in each simulation, then we only need to select the λ_1 and $c = \lambda_2$, and they can be detected by the nesting coordinate descent method. Moreover, the precision level is assigned as $\xi = 0.001$ in our simulation.

4.2. Multi-Modal Distributions

First, we examine our method in a multi-modal Gaussian model that is similar to the first model in [9]. However, our mixture Gaussian has a different variance, which leads the meaningful weights to our estimation. The density function for the i.i.d. sample $Z_1, \ldots, Z_n$ is assigned as follows:

$$h^*_\beta(x) = \sum_{j=1}^{W} \beta^*_j \phi(x|aj, \sigma_j), \tag{10}$$

where $\phi(x|aj, \sigma_j)$ is the density of $N(aj, \sigma_j^2)$. However, to estimate β^*, we only observe i.i.d. data $X_1, \ldots, X_n$ with density $g_{\beta^*_j}(x) = \sum_{j=1}^{W} \beta^*_j \phi(x|aj, \sqrt{1.1}\sigma_j)$. Put $a = 0.5, n = 100$ and

$$\beta^* = \Big(\mathbf{0}_8^T, 0.2, \mathbf{0}_{10}^T, 0.1, \mathbf{0}_5^T, 0.1, \mathbf{0}_{10}^T, 0.1, \mathbf{0}_{10}^T, 0.1, \mathbf{0}_5^T, 0.15, \mathbf{0}_{10}^T, 0.15, \mathbf{0}_{10}^T, 0.1, \mathbf{0}_{W-76}^T\Big)^T, \tag{11}$$

with $\sigma = \big(\mathbf{1}_{20}^T, \mathbf{0.8}_6^T, \mathbf{0.6}_{11}^T, \mathbf{0.4}_{11}^T, \mathbf{0.6}_6^T, \mathbf{0.8}_{11}^T, \mathbf{1.2}_{W-76}^T\big)^T$.

We replicate the simulation $N = 100$ times. Simulation results are presented in Table 1, from which we can see our method has more and more excellent performances as the W increases, which matches the non-asymptotic results in the previous section. The best performance is far better than the other three methods when $W = 321$. It is worthy to note that the better approximation follows the increase in W, matching Equation (7) and Theorem 3 in our previous section.

We plot the solution path to compare the performance of the four estimators in $\beta_j \in I(\beta)$ for every W in Figure 1 (the result of Elastic-net in $W = 321$ is not be shown due to its poor performance.). These figures also provide strong support for the above analysis. Meanwhile, we plot the probability densities of the several estimators and the true density to complement the visual sensory of the advantage in our method in Figure 2. The robust competency of detecting the multi-mode is shown (whereas other methods only find the most strongest signal, ignoring other meaningful but relatively weak signals).

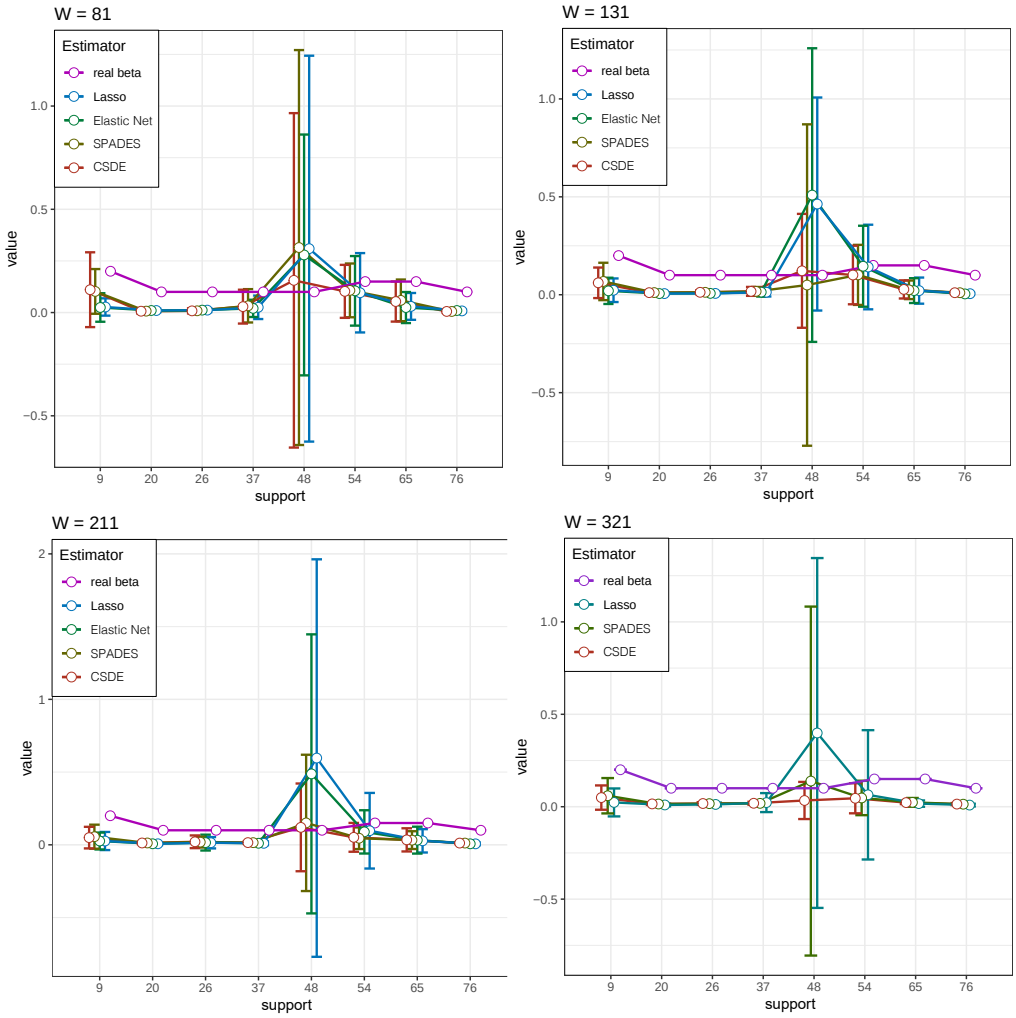

Figure 1. The simulation result in Section 4.2. The estimated support of β^* by the four types of estimators, and the W is varying. The circles represent the means of the estimators under the four specific approaches, while the half vertical lines mean the standard deviations.

Table 1. The simulation results in Section 4.2. The mean and standard deviation of the errors in the four estimators of β^* under $N = 100$ simulations, with $n = 100$. The quasi-optimal λ_2 is $c = 0.002$ for Elastic-net, while $c = 0.027$ is for the CSDE.

	W	λ_1	L_1 Error	TV Error
Lasso	81	0.065	2.133 (2.467)	1.137 (1.115)
Elastic-net			2.061 (1.439)	1.114 (0.805)
SPADES		0.053	1.922 (2.211)	1.258 (1.296)
CSDE			2.191 (4.812)	1.405 (2.329)
Lasso	131	0.068	2.032 (0.985)	1.352 (0.712)
Elastic-net			2.236 (2.498)	1.409 (1.056)
SPADES		0.056	1.880 (2.644)	0.972 (1.204)
CSDE			1.635 (0.342)	0.863 (0.402)

Table 1. *Cont.*

	W	λ_1	L_1 Error	TV Error
Lasso	211	0.071	2.572 (4.187)	1.605 (2.702)
Elastic-net			2.061 (1.883)	1.353 (1.516)
SPADES		0.058	1.764 (1.041)	0.832 (0.610)
CSDE			1.648 (0.168)	0.791 (0.415)
Lasso	321	0.074	2.120 (2.842)	1.146 (1.115)
Elastic-net			10.173 (82.753)	7.839 (67.887)
SPADES		0.061	2.106 (4.816)	0.818 (1.565)
CSDE			1.623 (0.085)	0.634 (0.199)

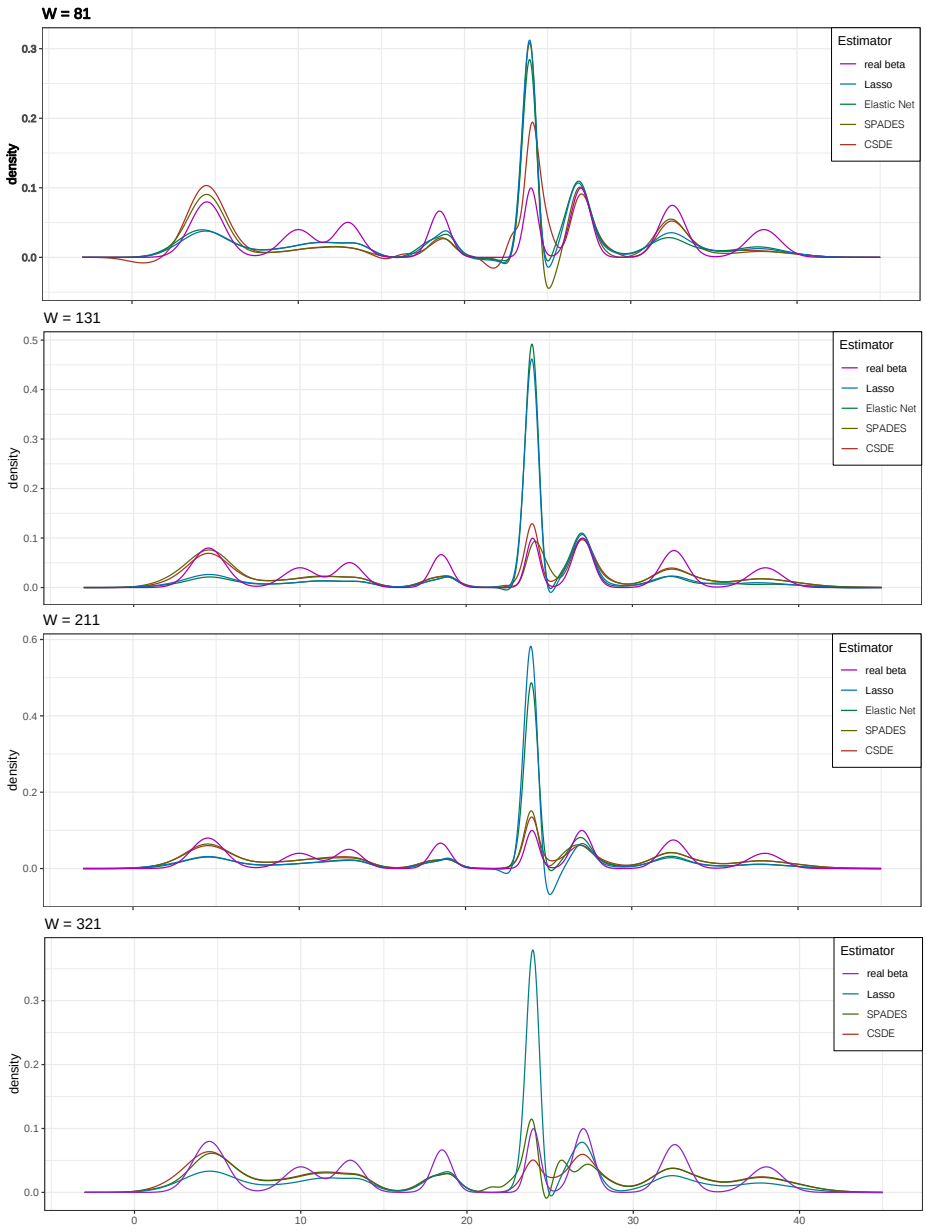

Figure 2. The simulation results in Section 4.2. The density map of the four estimators. The result of Elastic-net in $W = 321$ is not be shown due to its poor performance.

4.3. Mixture of Poisson Distributions

We study the mixture of discrete distribution: the mixture Poisson distribution

$$h_{\beta^*}(x) = \sum_{j=1}^{W} \beta_j^* p\left(x|\lambda_j = a \cdot j\right), \tag{12}$$

where $p\left(x|\lambda_j = a \cdot j\right)$ is the probability mass function (p.m.f.) of the Poisson distribution with mean λ_j. We set $a = 0.1$ and

$$\beta^* = \left(\mathbf{0}_8^T, 0.2, \mathbf{0}_{10}^T, 0.1, \mathbf{0}_5^T, 0.1, \mathbf{0}_{10}^T, 0.1, \mathbf{0}_{10}^T, 0.1, \mathbf{0}_5^T, 0.15, \mathbf{0}_{10}^T, 0.15, \mathbf{0}_{10}^T, 0.1, \mathbf{0}_{W-75}^T\right)^T. \tag{13}$$

The adjusted weights are calculated by Equation (3), and in discrete distributions, we define $\langle f, g\rangle = \sum_{k=1}^{\infty} f(k)g(k)$. Meanwhile, the Poisson random variable with measurement errors can be treated as a negative binomial random variable. Let $n(x|\lambda_j, r)$ be the p.m.f. of the Poisson distribution with the mean λ_j and dispersion parameter r. Suppose the observed data with sample size $n = 100$ has the p.m.f.

$$g_{\beta^*}(x) = \sum_{j=1}^{W} \beta_j^* n\left(x|\lambda_j = a \cdot j, r\right), \tag{14}$$

where $r = 6$, which leads to an increment of variance from Poisson to the negative binomial distribution. Similarly, we replicate each simulation to estimate the parameter $N = 100$ times with the sample from the mixture negative binomial distribution above. The result is shown in Table 2. The result is actually akin to that in the previous mixture Gaussian distribution, while the strong performance of our method is shown clearly when W is considerable.

Table 2. The simulation result in Section 4.3. The mean and standard deviation of the errors in the four estimators of β^* under $N = 100$ simulations. The λ_2 is chosen as $c = 0.005$ for Elastic-net, while $c = 0.203$ for the CSDE.

	W	λ_1	L_1 Error	TV Error
Lasso	81	0.048	1.796 (0.006)	0.002 (0.001)
Elastic-net			1.796 (0.006)	0.002 (0.001)
SPADES		0.138	1.811 (0.013)	0.002 (0.005)
CSDE			1.806 (0.008)	0.003 (0.005)
Lasso	131	0.051	1.828 (0.006)	0.003 (0.001)
Elastic-net			1.830 (0.009)	0.004 (0.002)
SPADES		0.145	1.880 (0.006)	0.002 (0.005)
CSDE			1.854 (0.006)	0.002 (0.004)
Lasso	211	0.053	1.935 (0.010)	0.005 (0.003)
Elastic-net			2.061 (0.014)	0.007 (0.008)
SPADES		0.152	1.935 (0.008)	0.005 (0.003)
CSDE			1.861 (0.005)	0.003 (0.002)
Lasso	321	0.055	1.927 (0.031)	0.005 (0.002)
Elastic-net			2.123 (0.026)	0.009 (0.009)
SPADES		0.158	1.938 (0.008)	0.005 (0.003)
CSDE			1.852 (0.002)	0.002 (0.001)

4.4. Low-Dimensional Mixture Model

Surprisingly, our method has more competitive efficacy than some popular methods (such as EM algorithm), even the dimension W is relatively small. To see this, we introduce the following numerical experiments to estimate the weights of the low-dimensional Gaussian mixture model: the samples $X_1, \cdots, X_n$ come from the model: $h_{\beta^*}(x) = \sum_{j=1}^{W} \beta_j^* \phi(x|\mu_j, \sigma_j)$. The updated equation for the EM algorithm in t-th step is:

$$\omega_{ij}^{(t)} = \frac{p_j^{(t)} \phi(x_i; \mu_t, \sigma_t)}{\sum_{s=1}^{W} p_s^{(t)} \phi(x_i; \mu_s, \sigma_s)}, \quad \beta_j^{(t+1)} = \frac{\sum_{i=1}^{W} \omega_{ij}^{(t)}}{\sum_{i=1}^{n} \sum_{j=1}^{W} \omega_{ij}^{(t)}}.$$

Here, we consider two scenarios:

(1) $W = 6, \beta = (0.3, 0, 0, 0.3, 0, 0.4)^T, \mu = (0, 10, 20, 30, 40, 50)^T, \sigma = (1, 2, 3, 4, 5, 6)^T$;
(2) $W = 7, \beta = (0.1, 0, 0, 0.8, 0, 0, 0.1)^T, \mu = (0, 1, 2, 3, 4, 5, 6)^T$,
$\sigma = (0.3, 0.2, 0.2, 0.1, 0.2, 0.2, 0.3)^T$.

For each scenario $n = 50$, and the fitter levels (cessation level) in the EM approach and our method are both $\xi = 10^{-4}$. A well-advised initial value in the EM approach is the equal weight.

We replicate the simulation $N = 100$ times, and the optimal tuning parameters stem from the cross-validation (CV). Thus, under each simulation, they are not the same, albeit they are very close to each other. The result can be seen in Table 3.

Table 3. The low-dimensional simulation result in Section 4.4.

		L_1 **Error**	TV **Error**
Scenario 1	EM	0.255 (0.122)	0.205 (0.098)
	CSDE	0.206 (0.145)	0.185 (0.104)
Scenario 2	EM	0.111 (0.055)	0.111 (0.055)
	CSDE	0.109 (0.037)	0.108 (0.037)

4.5. Real Data Examples

Practically, we consider using our method to estimate some densities in the environmental science field. Wind, which is mercurial, has been an advisable object to study for a long time in meteorology. Please note that the wind's speed at one specific location may not be diverse so we will use the wind's azimuth angle with a more sparse density at two sites in China. Many types of research about the estimated density for wind exist, so there is a possibility of using our approach to cope with some difficulties in meteorology science.

There have been some very credible meteorological dataset. We used the ERA5 hourly data in [35] to continue our analysis. We want chose a continental area and a coastal area in China, so we chose Beijing Nongzhanguan and Qingdao Coast. The locations of these two areas are: (116.3125° E, 116.4375° E) × (39.8125° N , 39.8125° N). Take notice that the wind in one day may be highly correlated. Therefore, using the data at a specific time point of each day in a consecutive period as i.i.d. samples is reasonable. The sample histograms at 6 am in Beijing Nongzhanguan and at midnight on the Qingdao Coast are shown in Figure 3. Here, we used the data from 1 January 2013 to 12 December 2015.

As we can see, their density does multi-peak (we used 1095 samples). Now, we can use our approach to estimate the multi-mode densities based on a relatively small size of samples, which is only a tiny part of the whole data from 1 January 2013 to 12 December 2015. Because one year has about 360 days, we may assume that every day is a latent factor that forms the base density. Thus, the model is designed as $h_{\beta^*}(x) = \sum_{j=1}^{360} \beta_j^* \phi(x|\mu_j, \sigma_j)$ with the mean and variance parameters $\mu = (1, 2, \ldots, 360)^T$, $\sigma = t \cdot \mathbf{1}_{360}^T$,, where t is the

bandwidth (or tuning parameter). With the different sub-samples, the computed values are different.

Another critical issue is how to choose the tuning parameters λ_1 and λ_2. Then, we apply the cross-validation criterion, namely choosing λ_i, to minimize the difference between the two estimators derived from the separated samples in a random dichotomy.

Now, start to construct the samples for the estimating procedure. Assume that an observatory wants to figure out some information about the two areas' wind. However, it does not have intact data due to the limited budget at its inception. The only samples it has are several days' information each month for the two areas, and these days scatter randomly. Furthermore, sample size $n = 168$ exactly. These imperfect data increase the challenge of estimating a trustworthy density. We compared our method with other previous methods, in which appraising the difference between the complete data sample histogram and the estimated density under each method is for the evaluation. Notice that the samples are only a tiny part of the data, so the $n = 168 < M(= 360)$ is relatively small. The small sample and large dimension setting coincide with the non-asymptotic theory provided in the previous section. The estimated density has been shown in Figure 4.

In this practical application, our method vindicates its more efficient estimating performance and stability from its propinquity of the complete sample histogram, namely the productive capacity of detecting the shape of the multi-mode density and the stronger inclination to bear a resemblance to each other sub-sample (although some subtle nuances do exist because of the different sub-sample). An alternative approach can be to consider principles and tools of circular statistics, which has been reviewed in [36].

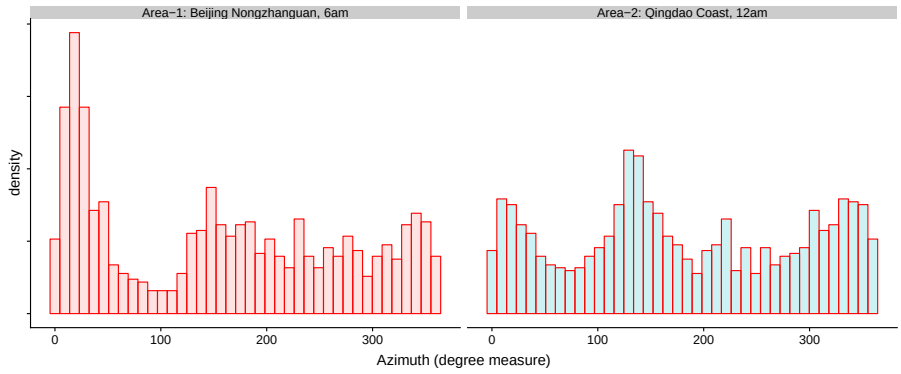

Figure 3. The sample histogram of the azimuth in Beijing Nongzhanguan at 6am and Qingdao Coast at 12 am.

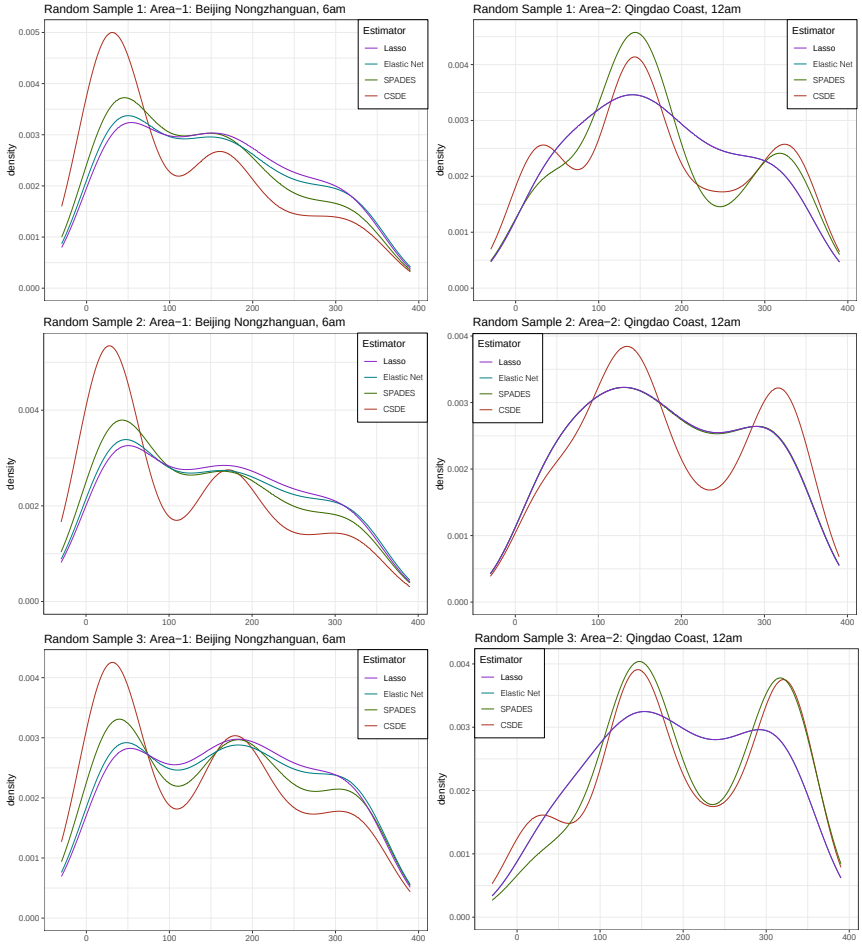

Figure 4. The density map of the four estimators' approaches for the three random sub-samples from the real-world data in Section 4.5.

5. Summary and Discussions

The paper deals with the deconvolution problem using Lasso-type methods: the observations $X_1, \cdots, X_n$ are independent and generated from $X_i = Z_i + \varepsilon_i$, and the goal is to estimate the unknown density h of the Z_i. We assume that the function h can be written as $h(\cdot) = h_{\beta^*}(\cdot) = \sum_{j=1}^{W} \beta_j^* h_j(\cdot)$ based on some functions $\{h_j\}_{j=1}^{W}$ from a specific dictionary and propose estimating the coefficients of this decomposition with the Elastic-net method. For this estimator, we show that under some classical assumptions of the model, such as coherence of the Gram matrix, finite sample bounds for the estimation and the prediction errors valid with a relatively high probability can be obtained. Moreover, we prove a variable selection consistency result under a beta-min condition and conduct an extensive numerical study. The following estimation problem is also similar to the CSDE.

For future study, it is also interesting and meaningful to do hypothesis testing about the coefficients $\beta^* \in \mathbb{R}^W$ in sparse mixture models. For a general function $h : \mathbb{R}^W \mapsto \mathbb{R}^m$ and a nonempty closed set $\Omega \in \mathbb{R}^m$, we can consider

$$H_0 : h(\beta^*) \in \Omega \text{ vs. } H_1 : h(\beta^*) \notin \Omega.$$

It is possible to use [37] as a general approach to hypothesis testing within models with measurement errors.

Author Contributions: Conceptualization, H.Z.; Data curation, X.Y.; Formal analysis, X.Y. and S.Z.; Funding acquisition, X.Y.; Investigation, X.Y. and S.Z.; Methodology, X.Y., H.Z., H.W. and S.Z.; Project administration, H.Z.; Resources, H.W.; Software, H.W.; Supervision, X.Y. and H.Z.; Visualization, H.W.; Writing—original draft, X.Y., H.Z., H.W. and S.Z.; Writing—review and editing, H.Z. and H.W. All authors have read and agreed to the published version of the manuscript.

Funding: Xiaowei Yang is supported in part by the General Research Project of Chaohu University (XLY-201906), Chaohu University Applied Curriculum Development Project (ch19yykc21), Key Project of Natural Science Foundation of Anhui Province Colleges and Universities (KJ2019A0683), Key Scientific Research Project of Chaohu University (XLZ-202105). Huiming Zhang is supported in part by the University of Macau under UM Macao Talent Program (UMMTP-2020-01). This work also is supported in part by the National Natural Science Foundation of China (Grant No. 11701109, 11901124) and the Guangxi Science Foundation (Grant No. 2018GXNSFAA138164).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The authors would like to thank Song Xi Chen's Group https://songxichen.com/(accessed on 20 December 2021) for sharing the meteorological dataset.

Acknowledgments: The Appendix includes the proofs of the lemmas, corollaries and theorems in the main body.

Conflicts of Interest: The authors declare no conflict of interest.

Appendix A

For convenience, we first give a preliminary lemma and proof. Define the random variables $M_j = \frac{1}{n}\sum_{i=1}^{n}\{h_j(X_i) - Eh_j(X_i)\}$. Consider the event $\mathcal{E}$ by $\mathcal{E} = \bigcap_{j=1}^{W}\{2|M_j| \le \tilde{\omega}_j\}$, where $\tilde{\omega}_k := 2\sqrt{2}L_k\sqrt{\frac{1}{n}\log\frac{W}{\delta/2}} =: 2\sqrt{2}L_k v(\delta/2)$. Then, we have the following lemma, which is cornerstone for the proofs in below.

Lemma A1. *Suppose* $\max_{1\le j\le W} L_j < \infty$ *and* $c = \frac{\min_{1\le j\le W}\{\tilde{\omega}_j\}}{B}$, *for any* $\beta \in \mathbb{R}^W$ *on the event* $\mathcal{E}$, *we have* $\|h_{\hat{\beta}} - h\|^2 + \sum_{j=1}^{W}\tilde{\omega}_j|\hat{\beta}_j - \beta_j| + \sum_{j=1}^{W} c(\hat{\beta}_j - \beta_j)^2 \le \|h_\beta - h\|^2 + 6\sum_{j\in I(\beta)}\omega_j|\hat{\beta}_j - \beta_j|$.

Appendix A.1. Proof of Lemma A1

According to the definition of $\hat{\beta}$, for any $\beta \in \mathbb{R}^W$, we find $-\frac{2}{n}\sum_{i=1}^{n} h_{\hat{\beta}}(X_i) + \|h_{\hat{\beta}}\|^2 + 2\sum_{j=1}^{W}\omega_j|\hat{\beta}_j| + c\sum_{j=1}^{W}\hat{\beta}_j^2 \le -\frac{2}{n}\sum_{i=1}^{n} h_\beta(X_i) + \|h_\beta\|^2 + 2\sum_{j=1}^{W}\omega_j|\beta_j| + c\sum_{j=1}^{W}\beta_j^2$. Then

$$\|h_{\hat{\beta}}\|^2 - \|h_\beta\|^2 \le \frac{2}{n}\sum_{i=1}^{n} h_{\hat{\beta}}(X_i) - \frac{2}{n}\sum_{i=1}^{n} h_\beta(X_i) + 2\sum_{j=1}^{W}\omega_j|\beta_j| - 2\sum_{j=1}^{W}\omega_j|\hat{\beta}_j| + c\sum_{j=1}^{W}\beta_j^2 - c\sum_{j=1}^{W}\hat{\beta}_j^2.$$

Note that

$$\begin{aligned}\|h_{\hat{\beta}} - h\|^2 &= \|h_{\hat{\beta}} - h_\beta + h_\beta - h\|^2 = \|h_{\hat{\beta}} - h_\beta\|^2 + \|h_\beta - h\|^2 + 2 < h_\beta - h, h_{\hat{\beta}} - h_\beta > \\ &= \|h_\beta - h\|^2 - 2 < h, h_{\hat{\beta}} - h_\beta > +2 < h_\beta, h_{\hat{\beta}} - h_\beta > + \|h_{\hat{\beta}} - h_\beta\|^2 \\ &= \|h_\beta - h\|^2 - 2 < h, h_{\hat{\beta}} - h_\beta > + \|h_{\hat{\beta}}\|^2 - \|h_\beta\|^2.\end{aligned}$$

Combining the two result above, we obtain

$$\begin{aligned}\|h_{\hat{\beta}} - h\|^2 \le \|h_\beta - h\|^2 &+ 2\sum_{j=1}^{W}\omega_j|\beta_j| - 2\sum_{j=1}^{W}\omega_j|\hat{\beta}_j| + c\sum_{j=1}^{W}\beta_j^2 - c\sum_{j=1}^{W}\hat{\beta}_j^2 \\ &- 2 < h, h_{\hat{\beta}} - h_\beta > + \frac{2}{n}\sum_{i=1}^{n} h_{\hat{\beta}}(X_i) - \frac{2}{n}\sum_{i=1}^{n} h_\beta(X_i). \qquad \text{(A1)}\end{aligned}$$

According to the definition of $h_\beta(x)$, it gives $h_\beta(x) = \sum_{j=1}^{W} \beta_j h_j(x)$ with $\beta = (\beta_1, \cdots, \beta_W)$. For the three terms in Equation (A1), we have

$$\begin{aligned}
&-2 < h, h_{\hat{\beta}} - h_\beta > + \frac{2}{n}\sum_{i=1}^{n} h_{\hat{\beta}}(X_i) - \frac{2}{n}\sum_{i=1}^{n} h_\beta(X_i) \\
&= 2 \cdot \frac{1}{n}\sum_{i=1}^{n}\left(\sum_{j=1}^{W}\hat{\beta}_j h_j(X_i) - \sum_{j=1}^{W}\beta_j h_j(X_i)\right) - 2E(h_{\beta'} - h_\beta)(X_i)|_{\beta'=\hat{\beta}} \\
&= 2\sum_{j=1}^{W}\frac{1}{n}\sum_{i=1}^{n} h_j(X_i)(\hat{\beta}_j - \beta_j) - 2\sum_{j=1}^{W} E[h_j(X_i)](\hat{\beta}_j - \beta_j) \\
&= 2\sum_{j=1}^{W}\left(\frac{1}{n}\sum_{i=1}^{n} h_j(X_i) - E[h_j(X_i)]\right)(\hat{\beta}_j - \beta_j).
\end{aligned}$$

Then

$$\begin{aligned}
\|h_{\hat{\beta}} - h\|^2 \le \|h_\beta - h\|^2 &+ 2\sum_{j=1}^{W}\left(\frac{1}{n}\sum_{i=1}^{n} h_j(X_i) - E[h_j(X_i)]\right)(\hat{\beta}_j - \beta_j) \\
&+ 2\sum_{j=1}^{W}\omega_j|\beta_j| - 2\sum_{j=1}^{W}\omega_j|\hat{\beta}_j| + c\sum_{j=1}^{W}\beta_j^2 - c\sum_{j=1}^{W}\hat{\beta}_j^2.
\end{aligned}$$

Conditioning on $\mathcal{E}$, we have $\|h_{\hat{\beta}} - h\|^2 \le \|h_\beta - h\|^2 + \sum_{j=1}^{W}\tilde{\omega}_j|\hat{\beta}_j - \beta_j| + 2\sum_{j=1}^{W}\omega_j(|\beta_j| - |\hat{\beta}_j|) + c\sum_{j=1}^{W}(\beta_j^2 - \hat{\beta}_j^2)$. We add $\sum_{j=1}^{W}\tilde{\omega}_j|\hat{\beta}_j - \beta_j| + c\sum_{j=1}^{W}(\beta_j - \hat{\beta}_j)^2$ to both sides of the inequality, it gives

$$\begin{aligned}
&\|h_{\hat{\beta}} - h\|^2 + \sum_{j=1}^{W}\tilde{\omega}_j|\hat{\beta}_j - \beta_j| + c\sum_{j=1}^{W}(\beta_j - \hat{\beta}_j)^2 \\
&\le \|h_\beta - h\|^2 + 2\sum_{j=1}^{W}\tilde{\omega}_j|\hat{\beta}_j - \beta_j| + 2\sum_{j=1}^{W}\omega_j(|\beta_j| - |\hat{\beta}_j|) + c\sum_{j=1}^{W}(\beta_j^2 - \hat{\beta}_j^2) + c\sum_{j=1}^{W}(\beta_j - \hat{\beta}_j)^2.
\end{aligned}$$

Note that

$$\begin{aligned}
&c[\sum_{j=1}^{W}(\beta_j^2 - \hat{\beta}_j^2) + \sum_{j=1}^{W}(\beta_j - \hat{\beta}_j)^2] = c[\sum_{j=1}^{W}(\beta_j^2 - \hat{\beta}_j^2 + \beta_j^2 - 2\beta_j\hat{\beta}_j + \hat{\beta}_j^2)] \\
&= 2c\sum_{j=1}^{W}\beta_j(\beta_j - \hat{\beta}_j) = 2c\sum_{j\in I(\beta)}\beta_j(\beta_j - \hat{\beta}_j) \le 2cB\sum_{j\in I(\beta)}|\beta_j - \hat{\beta}_j| \le 2\sum_{j\in I(\beta)}\tilde{\omega}_j|\beta_j - \hat{\beta}_j|,
\end{aligned}$$

where the last inequality is due to the assumption $c = \frac{\min_{1\le j\le W}\{\tilde{\omega}_j\}}{B} \le \frac{\tilde{\omega}_j}{B}$. Thus, we obtain

$$\begin{aligned}
&\|h_{\hat{\beta}} - h\|^2 + \sum_{j=1}^{W}\tilde{\omega}_j|\hat{\beta}_j - \beta_j| + c\sum_{j=1}^{W}(\hat{\beta}_j - \beta_j)^2 \\
&\le \|h_\beta - h\|^2 + 2\sum_{j=1}^{W}\tilde{\omega}_j|\hat{\beta}_j - \beta_j| + 2\sum_{j=1}^{W}\omega_j(|\beta_j| - |\hat{\beta}_j|) + 2\sum_{j\in I(\beta)}\tilde{\omega}_j|\hat{\beta}_j - \beta_j| \\
&\le \|h_\beta - h\|^2 + 2\sum_{j=1}^{W}\omega_j|\hat{\beta}_j - \beta_j| + 2\sum_{j=1}^{W}\omega_j(|\beta_j| - |\hat{\beta}_j|) + 2\sum_{j\in I(\beta)}\omega_j|\hat{\beta}_j - \beta_j|,
\end{aligned}$$

where the last inequality follows from $\tilde{\omega}_j \le \omega_j$ for all j.

We know $\beta_j \neq 0$ if $j \in I(\beta)$, and $\beta_j = 0$ if $j \notin I(\beta)$. Considering $|\beta_j| - |\hat{\beta}_j| \leq |\hat{\beta}_j - \beta_j|$ for all j, we have $2\sum_{j=1}^{W} \omega_j |\hat{\beta}_j - \beta_j| + 2\sum_{j=1}^{W} \omega_j (|\beta_j| - |\hat{\beta}_j|) \leq 4\sum_{j\in I(\beta)} \omega_j |\hat{\beta}_j - \beta_j|$. Then

$$\begin{aligned}
&\|h_{\hat{\beta}} - h\|^2 + \sum_{j=1}^{W} \tilde{\omega}_j |\hat{\beta}_j - \beta_j| + c\sum_{j=1}^{W} (\hat{\beta}_j - \beta_j)^2 \\
&\leq \|h_\beta - h\|^2 + 4\sum_{j\in I(\beta)} \omega_j |\hat{\beta}_j - \beta_j| + 2\sum_{j\in I(\beta)} \omega_j |\hat{\beta}_j - \beta_j| \\
&= \|h_\beta - h\|^2 + 6\sum_{j\in I(\beta)} \omega_j |\hat{\beta}_j - \beta_j|.
\end{aligned}$$

Appendix A.2. Proof of Theorems

According to $\tilde{\omega}_j = 2\sqrt{2}L_j\sqrt{\frac{1}{n}\log\frac{2W}{\delta}}$ in Equation (6), the sum of the independent random variables $\zeta_{ij} = h_j(X_i) - Eh_j(X_i)$ is determined by Hoeffding's inequality, and $|h_j(X_i)| \leq 2L_j$. We obtain

$$\begin{aligned}
P(\mathcal{E}^c) = P\left(\bigcup_{j=1}^{W} \{2|M_j| > \tilde{\omega}_j\}\right) &\leq \sum_{j=1}^{W} P(2|M_j| > \tilde{\omega}_j) \leq 2\sum_{j=1}^{W} \exp\left(-\frac{2n^2 \cdot \tilde{\omega}_j^2/4}{4nL_j^2}\right) \\
&= 2\sum_{j=1}^{W} \exp\left(-\log\frac{2W}{\delta}\right) = 2W \cdot \frac{\delta}{2W} = \delta.
\end{aligned}$$

Appendix A.3. Proof of Theorem 1

By Lemma A1, we need an upper bound on $\sum_{j\in I(\beta)} \omega_j |\hat{\beta}_j - \beta_j|$. For easy notation, let $q_j = \hat{\beta}_j - \beta_j$, $Q(\beta) = \sum_{j\in I(\beta)} |q_j| \|h_j\|$, $Q = \sum_{j=1}^{W} |q_j| \|h_j\|$. According to the definition of $H(\beta)$, that is, $H(\beta) = \max_{j\in I(\beta)} \frac{\omega_j}{v(\delta/2)\|h_j\|}$, we have

$$\sum_{j\in I(\beta)} \omega_j |\hat{\beta}_j - \beta_j| \leq v(\delta/2) H(\beta) Q(\beta). \tag{A2}$$

Let $Q_*(\beta) := \sqrt{\sum_{j\in I(\beta)} q_j^2 \|h_j\|^2}$. Using the definition of $h_\beta(x)$, we obtain $Q_*^2(\beta) = \sum_{j\in I(\beta)} q_j^2 \|h_j\|^2 = \|h_{\hat{\beta}} - h_\beta\|^2 - \sum_{i,j\notin I(\beta)} q_i q_j < h_i, h_j > -(2\sum_{i\notin I(\beta)} \sum_{j\in I(\beta)} q_i q_j < h_i, h_j > + \sum\sum_{i,j\in I(\beta), i\neq j} q_i q_j < h_i, h_j >)$. As $i, j \notin I(\beta)$, $\beta_i = \beta_j = 0$, it is easy to see $\sum\sum_{i,j\notin I(\beta)} < h_i, h_j > q_i q_j \geq 0$. Observe that

$$\begin{aligned}
&2\sum_{i\notin I(\beta)} \sum_{j\in I(\beta)} q_i q_j < h_i, h_j > + \sum\sum_{i,j\in I(\beta), i\neq j} q_i q_j < h_i, h_j > \\
&= 2\sum_{i\notin I(\beta)} \sum_{j\in I(\beta)} q_i q_j < h_i, h_j > + 2\sum\sum_{i,j\in I(\beta), j>i} q_i q_j < h_i, h_j > = 2\sum\sum_{i\in I(\beta), j>i} q_i q_j < h_i, h_j > .
\end{aligned}$$

By the definitions of $\rho_W(i,j)$ and $\rho_*(\beta)$, then

$$\begin{aligned}
Q_*^2(\beta) &\leq \|h_{\hat{\beta}} - h_\beta\|^2 + 2\sum\sum_{i\in I(\beta), j>i} |q_i||q_j| \|h_i\| \|h_j\| \frac{< h_i, h_j >}{\|h_i\| \|h_j\|} \\
&\leq \|h_{\hat{\beta}} - h_\beta\|^2 + 2\rho_*(\beta) \max_{i\in I(\beta), j>i} |q_i| \|h_i\| |q_j| \|h_j\|.
\end{aligned}$$

By $\max_{i\in I(\beta)}|q_i|\|h_i\| \le \sqrt{\sum_{j\in I(\beta)} q_j^2\|h_j\|^2} = Q_*(\beta)$, $\max_{i\in I(\beta),j>i}|q_j|\|h_j\| \le \sum_{j=1}^{W}|q_j|\|h_j\|$,

$$Q_*^2(\beta) \le \|h_{\hat{\beta}} - h_\beta\|^2 + 2\rho_*(\beta)Q_*(\beta)\sum_{j=1}^{W}|q_j|\|h_j\| = \|h_{\hat{\beta}} - h_\beta\|^2 + 2\rho_*(\beta)Q_*(\beta)Q. \quad \text{(A3)}$$

By Equation (A3), we can obtain $Q_*^2(\beta) - 2\rho_*(\beta)Q_*(\beta)Q - \|h_{\hat{\beta}} - h_\beta\|^2 \le 0$.

To find the upper bound of $Q_*(\beta)$, applying the properties of the quadratic inequality to the above formula, we obtain that

$$\begin{aligned} Q_*(\beta) &\le \rho_*(\beta)Q + \sqrt{\rho_*^2(\beta)Q^2 + \|h_{\hat{\beta}} - h_\beta\|^2} \le \rho_*(\beta)Q + [\rho_*(\beta)Q + \|h_{\hat{\beta}} - h_\beta\|] \\ &\le 2\rho_*(\beta)Q + \|h_{\hat{\beta}} - h_\beta\|. \end{aligned} \quad \text{(A4)}$$

Note that $W(\beta) = |I(\beta)| = \sum_{j=1}^{W} I(\beta_j \neq 0)$, employing the Cauchy–Schwarz inequalities, we have

$$\begin{aligned} W(\beta)\sum_{j\in I(\beta)}|q_j|^2\|h_j\|^2 &= \sum_{j\in I(\beta)} I^2(j\in I(\beta))\sum_{j\in I(\beta)}|q_j|^2\|h_j\|^2 \\ &\ge (\sum_{j\in I(\beta)} I(\{j\in I(\beta))|q_j|\|h_j\|)^2 = Q^2(\beta). \end{aligned}$$

Then, $Q_*^2(\beta) = \sum_{j\in I(\beta)}|q_j|^2\|h_j\|^2 \ge Q^2(\beta)/W(\beta)$. In combination with Equation (A4), we can obtain $Q(\beta)/\sqrt{W(\beta)} \le Q_*(\beta) \le 2\rho_*(\beta)Q + \|h_{\hat{\beta}} - h_\beta\|$. Therefore,

$$Q(\beta) \le 2\rho_*(\beta)\sqrt{W(\beta)}Q + \sqrt{W(\beta)}\|h_{\hat{\beta}} - h_\beta\|. \quad \text{(A5)}$$

By Lemma A1 and Equation (A2), we have the following inequality with probability exceeding $1-\delta$,

$$\begin{aligned} &\|h_{\hat{\beta}} - h\|^2 + \sum_{j=1}^{W}\tilde{\omega}_j|\hat{\beta}_j - \beta_j| + \sum_{j=1}^{W}c(\hat{\beta}_j - \beta_j)^2 \le \|h_\beta - h\|^2 + 6\sum_{j\in I(\beta)}\omega_j|\hat{\beta}_j - \beta_j| \\ &\le \|h_\beta - h\|^2 + 6v(\delta/2)H(\beta)Q(\beta) \\ &\le \|h_\beta - h\|^2 + 6v(\delta/2)H(\beta)[2\rho_*(\beta)\sqrt{W(\beta)}\sum_{j=1}^{W}|q_j|\|h_j\| + \sqrt{W(\beta)}\|h_{\hat{\beta}} - h_\beta\|] \ \text{(by (A5))} \\ &= \|h_\beta - h\|^2 + 12v(\delta/2)H(\beta)\rho_*(\beta)\sqrt{W(\beta)}\sum_{j=1}^{W}\tilde{\omega}_j|\hat{\beta}_j - \beta_j|\frac{\|h_j\|}{\tilde{\omega}_j} \\ &+ 6v(\delta/2)H(\beta)\sqrt{W(\beta)}\|h_{\hat{\beta}} - h_\beta\| \\ &\le \|h_\beta - h\|^2 + 12FH(\beta)\rho_*(\beta)\sqrt{W(\beta)}\sum_{j=1}^{W}\tilde{\omega}_j|\hat{\beta}_j - \beta_j| + 6v(\delta/2)H(\beta)\sqrt{W(\beta)}\|h_{\hat{\beta}} - h_\beta\| \\ &\le \|h_\beta - h\|^2 + \gamma\sum_{j=1}^{W}\tilde{\omega}_j|\hat{\beta}_j - \beta_j| + 6v(\delta/2)H(\beta)\sqrt{W(\beta)}\|h_{\hat{\beta}} - h_\beta\|, \end{aligned}$$

where the second last inequality follows from the definition of $F := \max_{1\le j\le W}\frac{v(\delta/2)\|h_j\|}{\tilde{\omega}_j}$, and the last inequality is derived by the assumption $12FH(\beta)\rho_*(\beta)\sqrt{W(\beta)} \le \gamma, (0 < \gamma \le 1)$.

Further, we can find that, with probability at least $1-\delta$,

$$\begin{aligned}
&\|h_{\hat{\beta}}-h\|^2+(1-\gamma)\sum_{j=1}^{W}\tilde{\omega}_j|\hat{\beta}_j-\beta_j|+\sum_{j=1}^{W}c(\hat{\beta}_j-\beta_j)^2\\
&\leq \|h_{\beta}-h\|^2+6v(\delta/2)H(\beta)\sqrt{W(\beta)}\|h_{\hat{\beta}}-h_{\beta}\|\\
&= \|h_{\beta}-h\|^2+6v(\delta/2)H(\beta)\sqrt{W(\beta)}\|h_{\hat{\beta}}-h+h-h_{\beta}\|\\
&\leq \|h_{\beta}-h\|^2+6v(\delta/2)H(\beta)\sqrt{W(\beta)}\|h_{\hat{\beta}}-h\|+6v(\delta/2)H(\beta)\sqrt{W(\beta)}\|h-h_{\beta}\|.
\end{aligned}$$

Using the elementary inequality $2st\leq s^2/\alpha+\alpha t^2$ $(s,t\in\mathbb{R},\alpha>1)$ to the last two terms of the above inequality, it yields

$$\begin{aligned}
2\{3v(\delta/2)H(\beta)\sqrt{W(\beta)}\}\|h_{\hat{\beta}}-h\| &\leq \alpha\cdot 9v^2(\delta/2)H^2(\beta)W(\beta)+\|h_{\hat{\beta}}-h\|^2/\alpha,\\
2\{3v(\delta/2)H(\beta)\sqrt{W(\beta)}\}\|h_{\beta}-h\| &\leq \alpha\cdot 9v^2(\delta/2)H^2(\beta)W(\beta)+\|h_{\beta}-h\|^2/\alpha.
\end{aligned}$$

Thus,

$$\begin{aligned}
&\|h_{\hat{\beta}}-h\|^2+(1-\gamma)\sum_{j=1}^{W}\tilde{\omega}_j|\hat{\beta}_j-\beta_j|+\sum_{j=1}^{W}c(\hat{\beta}_j-\beta_j)^2\\
&\leq \|h_{\beta}-h\|^2+18\alpha v^2(\delta/2)H^2(\beta)W(\beta)+\|h_{\hat{\beta}}-h\|^2/\alpha+\|h_{\beta}-h\|^2/\alpha.
\end{aligned}$$

Simplifying, we have

$$\begin{aligned}
&\|h_{\hat{\beta}}-h\|^2+\frac{\alpha(1-\gamma)}{(\alpha-1)}\sum_{j=1}^{W}\tilde{\omega}_j|\hat{\beta}_j-\beta_j|+\frac{\alpha}{\alpha-1}\sum_{j=1}^{W}c(\hat{\beta}_j-\beta_j)^2\\
&\leq \frac{\alpha+1}{\alpha-1}\|h_{\beta}-h\|^2+\frac{18\alpha^2}{\alpha-1}H^2(\beta)v^2(\delta/2)W(\beta),\ \alpha>1,0<\gamma\leq 1. \qquad \text{(A6)}
\end{aligned}$$

Optimizing α to obtain the sharp upper bounds for the above oracle inequality

$$\begin{aligned}
\alpha_{opt1} :&= \arg\min_{\alpha>1}\left\{\frac{\alpha+1}{\alpha-1}\|h_{\beta}-h\|^2+\frac{18\alpha^2}{\alpha-1}H^2(\beta)v^2(\delta/2)W(\beta)\right\}\\
&= 1+\sqrt{1+\frac{\|h_{\beta}-h\|^2}{9H^2(\beta)v^2(\delta/2)W(\beta)}}
\end{aligned}$$

by the first order condition. To date, Theorem 1 is proved by substituting α_{opt1} into Equation (A6).

Appendix A.4. Proof of Theorem 2

By the minimal eigenvalue assumption for ψ_W, we have

$$\|h_{\beta}\|^2=\|\sum_{j=1}^{W}\beta_j h_j(x)\|^2=\beta^T\psi_W\beta\geq\lambda_W\|\beta\|^2\geq\lambda_W\sum_{j\in I(\beta)}\beta_j^2. \qquad \text{(A7)}$$

Using the definition of ω_j and assumption $L_{\min}:=\min_{1\leq j\leq W}L_j>0$,

$$\omega_j=2L_j\left(\sqrt{\frac{2\log(2W/\delta)}{n}}+\frac{cB}{2L_j}\right)\leq 2L_j\left(\sqrt{\frac{2\log(2W/\delta)}{n}}+\frac{cB}{2L_{\min}}\right).$$

Since $cB = \tilde{\omega}_{\min} = 2\sqrt{2}L_{\min}v(\delta/2)$ and $v(\delta/2) = \sqrt{\frac{\log(2W/\delta)}{n}}$, we have $\omega_j \le 4\sqrt{2}L_j v(\delta/2)$.

Let $G(\beta) = \sum_{j\in I(\beta)} L_j^2$, by the Cauchy–Schwartz inequality, we obtain

$$\begin{aligned} 6\sum_{j\in I(\beta)} \omega_j|\hat{\beta}_j - \beta_j| &\le 24\sqrt{2}v(\delta/2)\sum_{j\in I(\beta)} L_j|\hat{\beta}_j - \beta_j| \\ &\le 24\sqrt{2}v(\delta/2)\sqrt{\sum_{j\in I(\beta)} L_j^2}\sqrt{\sum_{j\in I(\beta)}(\hat{\beta}_j - \beta_j)^2} \le 24\sqrt{2}v(\delta/2)\sqrt{\frac{G(\beta)}{\lambda_W}}\|h_{\hat{\beta}} - h_\beta\|, \end{aligned} \tag{A8}$$

where the last inequality above is from Equation (A7) due to

$$\|h_{\hat{\beta}} - h_\beta\|^2 = \sum_{1\le i,j\le W}\sum (\hat{\beta}_i - \beta_i)(\hat{\beta}_j - \beta_j) < h_i, h_j > \ge \lambda_W \sum_{j\in I(\beta)} (\hat{\beta}_j - \beta_j)^2.$$

Let $b(\beta) := 12\sqrt{2}v(\delta/2)\sqrt{\frac{G(\beta)}{\lambda_W}}$, Lemma 2 implies

$$\begin{aligned} &\|h_{\hat{\beta}} - h\|^2 + \sum_{j=1}^{W}\tilde{\omega}_j|\hat{\beta}_j - \beta_j| + \sum_{j=1}^{W} c(\hat{\beta}_j - \beta_j)^2 \le \|h_\beta - h\|^2 + 2b(\beta)\|h_{\hat{\beta}} - h_\beta\| \\ &= \|h_\beta - h\|^2 + 2b(\beta)(\|h_{\hat{\beta}} - h + h - h_\beta\|) \le \|h_\beta - h\|^2 + 2b(\beta)\|h_{\hat{\beta}} - h\| + 2b(\beta)\|h_\beta - h\|. \end{aligned}$$

Using the inequality $2st \le s^2/\alpha + \alpha t^2$ $(s, t \in R, \alpha > 1)$ for the last two terms on the right side of the above inequality, we find

$$\begin{aligned} 2b(\beta)\|h_{\hat{\beta}} - h\| + 2b(\beta)\|h_\beta - h\| &\le \|h_{\hat{\beta}} - h\|^2/\alpha + b^2(\beta)\alpha + \|h_\beta - h\|^2/\alpha + b^2(\beta)\alpha \\ &= \|h_{\hat{\beta}} - h\|^2/\alpha + \|h_\beta - h\|^2/\alpha + 2b^2(\beta)\alpha. \end{aligned}$$

Thus,

$$\begin{aligned} \|h_{\hat{\beta}} - h\|^2 + \sum_{j=1}^{W}\tilde{\omega}_j|\hat{\beta}_j - \beta_j| + \sum_{j=1}^{W} c(\hat{\beta}_j - \beta_j)^2 &\le \|h_\beta - h\|^2 + \|h_{\hat{\beta}} - h\|^2/\alpha \\ &\quad + \|h_\beta - h\|^2/\alpha + 2b^2(\beta)\alpha \end{aligned}$$

gives $\frac{\alpha-1}{\alpha}\|h_{\hat{\beta}} - h\|^2 + \sum_{j=1}^{W}\tilde{\omega}_j|\hat{\beta}_j - \beta_j| + \sum_{j=1}^{W} c(\hat{\beta}_j - \beta_j)^2 \le \frac{\alpha+1}{\alpha}\|h_\beta - h\|^2 + 2\alpha b^2(\beta)$. Therefore,

$$\begin{aligned} &\|h_{\hat{\beta}} - h\|^2 + \frac{\alpha}{\alpha-1}\sum_{j=1}^{W}\tilde{\omega}_j|\hat{\beta}_j - \beta_j| + \frac{\alpha}{\alpha-1}\sum_{j=1}^{W} c(\hat{\beta}_j - \beta_j)^2 \\ &\le \frac{\alpha+1}{\alpha-1}\|h_\beta - h\|^2 + \frac{2\alpha^2}{\alpha-1}b^2(\beta) \\ &= \frac{\alpha+1}{\alpha-1}\|h_\beta - h\|^2 + \frac{576\alpha^2}{\alpha-1}\frac{G(\beta)}{\lambda_W}v^2(\delta/2). \end{aligned}$$

To obtain the sharp upper bounds for the above oracle inequality, we optimize α

$$\alpha_{opt2} := \arg\min_{\alpha>1}\left\{\frac{\alpha+1}{\alpha-1}\|h_\beta - h\|^2 + \frac{576\alpha^2}{\alpha-1}\frac{G(\beta)}{\lambda_W}v^2(\delta/2)\right\} = 1 + \sqrt{1 + \frac{\|h_\beta - h\|^2}{288\frac{G(\beta)}{\lambda_W}v^2(\delta/2)}},$$

by the first-order condition. This completes the proof of Theorem 2.

Appendix A.5. Proof of Corollory 1

Let $\tilde{\omega}_{\min} := \min_{1\le j\le W}\tilde{\omega}_j$. We replace $v(\delta/2)$ in Theorem 1 by the larger value $v(\delta/2W)$. Substituting $\beta = \beta^*$ in Theorem 1, we have

$$\frac{\alpha_{opt1}(1-\gamma)}{\alpha_{opt1}-1}\sum_{j=1}^{W}\tilde{\omega}_j|\hat{\beta}_j-\beta_j^*| \le \frac{18\alpha_{opt1}^2}{\alpha_{opt1}-1}H^2(\beta^*)v^2(\delta/2W)W(\beta^*)$$

by $h = h_{\beta^*}$. Since $\tilde{\omega}_j \ge \tilde{\omega}_{\min}$ for all j, we obtain

$$\sum_{j=1}^{W}|\hat{\beta}_j-\beta_j^*| \le \frac{18\alpha_{opt1}}{1-\gamma}\cdot\frac{1}{\tilde{\omega}_{\min}}\cdot\max_{j\in I(\beta)}\frac{\omega_j^2}{\|h_j\|^2}\cdot W(\beta^*).$$

In this case, $\alpha_{opt1} = 2$, and $\|h_j\| = 1$; thus,

$$\begin{aligned}\|\hat{\beta}-\beta^*\| &\le \frac{36}{1-\gamma}\cdot\max_{j\in I(\beta)}\frac{\omega_j^2}{\tilde{\omega}_{\min}}\cdot W(\beta^*)\\ &= \frac{72\sqrt{2}v(\delta/2W)W(\beta^*)}{1-\gamma}\max_{j\in I(\beta)}\frac{(L_j+L_{\min})^2}{L_{\min}} \le \frac{72\sqrt{2}v(\delta/2W)W(\beta^*)}{1-\gamma}\frac{(L+L_{\min})^2}{L_{\min}}\end{aligned}$$

from $\tilde{\omega}_{\min} = 2\sqrt{2}v(\delta/2W)L_{\min}$ and

$$\omega_j^2 = [2\sqrt{2}v(\delta/2W)]^2\Big[L_j+\tfrac{\tilde{\omega}_{\min}}{2\sqrt{2}v(\delta/2W)}\Big]^2 = [2\sqrt{2}v(\delta/2W)]^2[L_j+L_{\min}]^2.$$

This completes the proof of Corollary 1.

Appendix A.6. Proof of Corollary 2

Let $\beta = \beta^*$ in Theorem 2, with $\alpha_{opt2} = 2$, we replace $v(\delta/2)$ in Theorem 2 by the larger value $v(\delta/2W)$, then $\sum_{j=1}^{W}\tilde{\omega}_{\min}|\hat{\beta}_j-\beta_j^*| \le \sum_{j=1}^{W}\tilde{\omega}_j|\hat{\beta}_j-\beta_j^*| \le \frac{576\alpha_{opt2}G^*}{\lambda_W}v^2(\delta/2W)$. By the definition of $\tilde{\omega}_{\min}$, we can obtain

$$\sum_{j=1}^{W}|\hat{\beta}_j-\beta_j^*| \le \frac{576\alpha_{opt2}G^*v^2(\delta/2)}{\tilde{\omega}_{\min}\lambda_W} = \frac{576\cdot 2G^*v^2(\delta/2W)}{2\sqrt{2}v(\delta/2W)L_{\min}\lambda_W} = \frac{288\sqrt{2}G^*v(\delta/2W)}{L_{\min}\lambda_W}.$$

This concludes the proof of Corollary 2.

Appendix A.7. Proof of Theorem 3

The following lemma is by virtue of the KKT conditions. It derives a bound of $P(I_* \nsubseteq \hat{I})$, which is easily analyzed.

Lemma A2 (Proposition 3.3 in [33]). $P(I_* \nsubseteq \hat{I}) \le W(\beta^*)\max_{k\in I_*}P(\hat{\beta}_k = 0 \text{ and } \beta_k^* \ne 0)$.

To present the proof of Theorem 3, we first notice that $P(\hat{I} \ne I_*) \le P(I_* \nsubseteq \hat{I}) + P(\hat{I} \nsubseteq I_*)$. Next, we control the probability on the right side of the above inequality.

For the control of $P(I_* \nsubseteq \hat{I})$, by Lemma A2, it remains to bound $P(\hat{\beta}_k = 0 \text{ and } \beta_k^* \ne 0)$.

Below, we will use the conclusion of Lemma 2 (KKT conditions). Recall that $E[h_k(Z_1)] = \sum_{j\in I_*} \beta_j^* < h_k, h_j >= \sum_{j=1}^W \beta_j^* < h_k, h_j >$. Since we assume that the density of Z_1 is the mixture density $h_{\beta^*} = \sum_{j\in I_*} \beta_j^* h_j$. Therefore, for $k \in I_*$, we have,

$$
\begin{aligned}
&P(\hat{\beta}_k = 0 \text{ and } \beta_k^* \neq 0) = P\left(\left|\frac{1}{n}\sum_{i=1}^n h_k(X_i) - \sum_{j=1}^W \hat{\beta}_j < h_j, h_k >\right| \leq 2\sqrt{2}v(\delta/2W)L_k; \beta_k^* \neq 0\right) \\
&= P\left(\left|\frac{1}{n}\sum_{i=1}^n h_k(X_i) - E[h_k(Z_1)] + E[h_k(Z_1)] - \sum_{j=1}^W \hat{\beta}_j < h_j, h_k >\right| \leq 2\sqrt{2}v(\delta/2W)L_k; \beta_k^* \neq 0\right) \\
&= P\left(\left|\frac{1}{n}\sum_{i=1}^n h_k(X_i) - E[h_k(Z_1)] - \sum_{j=1}^W (\hat{\beta}_j - \beta_j^*) < h_j, h_k >\right| \leq 2\sqrt{2}v(\delta/2W)L_k; \beta_k^* \neq 0\right) \\
&= P\left(\left|\frac{1}{n}\sum_{i=1}^n h_k(X_i) - E[h_k(Z_1)] - \sum_{j\neq k}^W (\hat{\beta}_j - \beta_j^*) < h_j, h_k > +\beta_k^*\|h_k\|^2\right| \leq 2\sqrt{2}v(\delta/2W)L_k\right) \\
&\leq P\left(|\beta_k^*|\|h_k\|^2 - 2\sqrt{2}v(\delta/2W)L_k \leq \left|\frac{1}{n}\sum_{i=1}^n h_k(X_i) - E[h_k(Z_1)]\right| + \left|\sum_{j\neq k}^W (\hat{\beta}_j - \beta_j^*) < h_j, h_k >\right|\right) \\
&\leq P\left(\left|\frac{1}{n}\sum_{i=1}^n h_k(X_i) - E[h_k(Z_1)]\right| \geq \frac{|\beta_k^*|\|h_k\|^2}{2} - \sqrt{2}v(\delta/2W)L_k\right) \qquad \text{(A9)} \\
&\quad + P\left(\left|\sum_{j\neq k}^W (\hat{\beta}_j - \beta_j^*) < h_j, h_k >\right| \geq \frac{|\beta_k^*|\|h_k\|^2}{2} - \sqrt{2}v(\delta/2W)L_k\right). \qquad \text{(A10)}
\end{aligned}
$$

Similar to Lemma 2, for Equation (A9), we use Hoeffding's inequality. Since $\|h_k\| = 1$ for all k. Put $\epsilon_k := |E[h_k(X_1)] - E[h_k(Z_1)]|$. Consider Condition (B), $\min_{k\in I_*} |\beta_k^*| \geq 4\sqrt{2}v(\delta/2W)L$ and $L \geq \max_{1\leq k\leq W} L_k$, then we have

$$
\begin{aligned}
&P\left(\left|\frac{1}{n}\sum_{i=1}^n h_k(X_i) - E[h_k(Z_1)]\right| \geq \frac{|\beta_k^*|\|h_k\|^2}{2} - \sqrt{2}v(\delta/2W)L_k\right) \\
&= P\left(\left|\frac{1}{n}\sum_{i=1}^n h_k(X_i) - E[h_k(X_1)] + E[h_k(X_1)] - E[h_k(Z_1)]\right| \geq \frac{|\beta_k^*|\|h_k\|^2}{2} - \sqrt{2}v(\delta/2W)L_k\right) \\
&\leq P\left(\left|\frac{1}{n}\sum_{i=1}^n h_k(X_i) - E[h_k(X_1)]\right| \geq \frac{|\beta_k^*|}{2} - \sqrt{2}v(\delta/2W)L - \epsilon_k\right) \\
&\leq P\left(\left|\frac{1}{n}\sum_{i=1}^n h_k(X_i) - E[h_k(X_1)]\right| \geq 2\sqrt{2}v(\delta/2W)L - \sqrt{2}v(\delta/2W)L - \epsilon_k\right) \\
&= P\left(\left|\frac{1}{n}\sum_{i=1}^n h_k(X_i) - E[h_k(X_1)]\right| \geq \sqrt{2}v(\delta/2W)L - \epsilon_k\right) \\
&= P\left(\left|\frac{1}{n}\sum_{i=1}^n h_k(X_i) - E[h_k(X_1)]\right| \geq \sqrt{2}v(\delta/2W)L(1-\epsilon_k^*)\right) \quad (\text{let } \epsilon_k^* = \epsilon_k/\sqrt{2}v(\delta/2W)L) \\
&\leq 2\exp\left\{-\frac{4n^2v^2(\delta/2W)L^2(1-\epsilon_k^*)^2}{4nL^2}\right\} \\
&= 2\exp\left\{-n(1-\epsilon_k^*)^2\frac{\log(2W^2/\delta)}{n}\right\} = 2\left(\frac{\delta}{2W^2}\right)^{(1-\epsilon_k^*)^2}. \qquad \text{(A11)}
\end{aligned}
$$

For the upper bound of Equation (A10), using Condition (A) and Condition (B), by the definitions of $\rho_*(\beta^*)$ and $W(\beta^*)$, we obtain

$$\begin{aligned}
&P\left(\left|\sum_{j\neq k}^{W}(\hat{\beta}_j-\beta_j^*)<h_j,h_k>\right|\geq\frac{|\beta_k^*|\|h_k\|^2}{2}-\sqrt{2}v(\delta/2W)L_k\right)\\
&=P\left(\left|\sum_{j\neq k}^{W}(\hat{\beta}_j-\beta_j^*)<h_j,h_k>\right|\geq\frac{|\beta_k^*|}{2}-\sqrt{2}v(\delta/2W)L_k\right)\\
&\leq P\left(\left|\sum_{j\neq k}^{W}(\hat{\beta}_j-\beta_j^*)\frac{<h_j,h_k>}{\|h_j\|\|h_k\|}\cdot\|h_j\|\|h_k\|\right|\geq 2\sqrt{2}v(\delta/2W)L-\sqrt{2}v(\delta/2W)L\right)\\
&\leq P\left(\rho_*(\beta^*)\sum_{j\neq k}^{W}\left|\hat{\beta}_j-\beta_j^*\right|\geq\sqrt{2}v(\delta/2W)L\right)\leq P\left(\sum_{j=1}^{W}\left|\hat{\beta}_j-\beta_j^*\right|\geq\frac{\sqrt{2}v(\delta/2W)L}{\rho_*(\beta^*)}\right)\\
&\leq P\left(\sum_{j=1}^{W}\left|\hat{\beta}_j-\beta_j^*\right|\geq\frac{288\sqrt{2}G^*v(\delta/2W)}{L_{\min}\lambda_W}\right)\leq\frac{\delta}{W}.
\end{aligned}$$

where the second last inequality is by Condition (A), and the last inequality above is by using the ℓ_1-estimation oracle inequality in Corollary 2.

Therefore, by the definition of $W(\beta^*)$, $W(\beta^*)=|I_*|\leq W$, we find

$$\begin{aligned}
P(I_*\nsubseteq\hat{I})\leq W(\beta^*)\max_{k\in I_*}P(\hat{\beta}_k=0)&\leq W(\beta^*)2\left(\frac{\delta}{2W^2}\right)^{(1-\epsilon_k^*)^2}+W(\beta^*)\frac{\delta}{W}\\
&\leq 2W\left(\frac{\delta}{2W^2}\right)^{(1-\epsilon_k^*)^2}+W\frac{\delta}{W}=2W\left(\frac{\delta}{2W^2}\right)^{(1-\epsilon_k^*)^2}+\delta.
\end{aligned}$$

For the control of $P(\hat{I}\nsubseteq I_*)$, let

$$\tilde{\eta}=\underset{\eta\in\mathbb{R}^{W(\beta^*)}}{\arg\min}z(\eta), \tag{A12}$$

where $z(\eta)=-\frac{2}{n}\sum_{i=1}^{n}\sum_{j\in I_*}\eta_j h_j(X_i)+\|\sum_{j\in I_*}\eta_j h_j\|^2+\sum_{j\in I_*}(4\sqrt{2}v(\delta/2)L_j+2cB)|\eta_j|+c\sum_{j\in I_*}\eta_j^2$. Consider the following random event

$$\begin{aligned}
&\bigcap_{k\notin I_*}\left\{\left|-\frac{1}{n}\sum_{i=1}^{n}h_k(X_i)+\sum_{j\in I_*}\tilde{\eta}_j<h_j,h_k>\right|\leq 2\sqrt{2}v(\delta/2)L_k\right\}\\
&\quad\subseteq\bigcap_{k\notin I_*}\left\{\left|-\frac{1}{n}\sum_{i=1}^{n}h_k(X_i)+\sum_{j\in I_*}\tilde{\eta}_j<h_j,h_k>\right|\leq 2\sqrt{2}v(\delta/2W)L\right\}:=\Psi.
\end{aligned}\tag{A13}$$

Let $\bar{\eta}\in\mathbb{R}^W$ be a vector corresponding to the component of the index set I_* having $\tilde{\eta}$ given by Equation (A12), and the component at other corresponding positions is 0. By Lemma 1, we know that $\bar{\eta}\in\mathbb{R}^W$ is a solution of Equation (3) on the event Ψ. It is recalled that $\hat{\beta}\in\mathbb{R}^W$, which is also a solution of Equation (3). Through the definition of the indicator set $\hat{I}$, we have $\hat{\beta}_k\neq 0$ for $k\in\hat{I}$. By construction, we obtain $\tilde{\eta}_k\neq 0$ for some subset $T\subseteq I_*$. The

KKT conditions indicate that any two solutions have non-zero components at the same positions. Therefore, $\hat{I} = T \subseteqq I_*$ on the event Ψ. Further, we can write

$$
\begin{aligned}
P(\hat{I} \not\subseteqq I_*) \le P(\Psi^c) &= P\left(\bigcup_{k \notin I_*} \left\{ \left| -\frac{1}{n}\sum_{i=1}^{n} h_k(X_i) + \sum_{j \in I_*} \tilde{\eta}_j < h_j, h_k > \right| \ge 2\sqrt{2} v(\delta/2W) L \right\}\right) \\
&\le \sum_{k \notin I_*} P\left\{ \left| -\frac{1}{n}\sum_{i=1}^{n} h_k(X_i) + \sum_{j \in I_*} \tilde{\eta}_j < h_j, h_k > \right| \ge 2\sqrt{2} v(\delta/2W) L \right\} \\
&= \sum_{k \notin I_*} P\left\{ \left| -\frac{1}{n}\sum_{i=1}^{n} h_k(X_i) + E[h_k(Z_1)] - E[h_k(Z_1)] + \sum_{j \in I_*} \tilde{\eta}_j < h_j, h_k > \right| \ge 2\sqrt{2} v(\delta/2W) L \right\} \\
&= \sum_{k \notin I_*} P\left\{ \left| \frac{1}{n}\sum_{i=1}^{n} h_k(X_i) - E[h_k(Z_1)] - \sum_{j \in I_*} (\tilde{\eta}_j - \beta_j^*) < h_j, h_k > \right| \ge 2\sqrt{2} v(\delta/2W) L \right\} \\
&\le \sum_{k \notin I_*} P\left\{ \left| \frac{1}{n}\sum_{i=1}^{n} h_k(X_i) - E[h_k(Z_1)] \right| \ge \sqrt{2} v(\delta/2W) L \right\} \qquad \text{(A14)} \\
&\quad + \sum_{k \notin I_*} P\left\{ \sum_{j \in I_*} |\tilde{\eta}_j - \beta_j^*| |< h_j, h_k >| \ge \sqrt{2} v(\delta/2W) L \right\}. \qquad \text{(A15)}
\end{aligned}
$$

According to the previously proven Formula (A11), we find

$$
\begin{aligned}
&\sum_{k \notin I_*} P\left\{ \left| \frac{1}{n}\sum_{i=1}^{n} h_k(Z_i) - E h_k(Z_1) \right| \ge \sqrt{2} v(\delta/2W) L \right\} \\
&\le \sum_{k \notin I_*} P\left\{ \left| \frac{1}{n}\sum_{i=1}^{n} h_k(X_i) - E h_k(X_1) \right| \ge \sqrt{2} v(\delta/2W) L - \epsilon_k \right\} \\
&= \sum_{k=1}^{W} P\left\{ \left| \frac{1}{n}\sum_{i=1}^{n} h_k(X_i) - E h_k(X_1) \right| \ge \sqrt{2} v(\delta/2W) L (1 - \epsilon_k^*) \right\} \le 2W \left(\frac{\delta}{2W^2}\right)^{(1-\epsilon_k^*)^2}.
\end{aligned}
$$

For the upper bound of Equation (A15), observe Theorem 2, we can use a larger $v(\delta/2W)$ instead of $v(\delta/2)$. Consider the construction of $\tilde{\eta}$ in Equation (A12), we obtain

$$
P\left(\sum_{j \in I_*} |\tilde{\eta}_j - \beta_j^*| \ge \frac{288\sqrt{2} G^* v(\delta/2W)}{L_{\min} \lambda_W} \right) \le \frac{\delta}{W}.
$$

Similarly, we have

$$
\begin{aligned}
&\sum_{k \notin I_*} P\left\{ \sum_{j \in I_*} |\tilde{\eta}_j - \beta_j^*| |< h_j, h_k >| \ge \sqrt{2} v(\delta/2W) L \right\} \\
&\le \sum_{k=1}^{W} P\left\{ \sum_{j \in I_*} |\tilde{\eta}_j - \beta_j^*| \left| \frac{< h_j, h_k >}{\|h_j\| \|h_k\|} \|h_j\| \|h_k\| \right| \ge \sqrt{2} v(\delta/2W) L \right\} \\
&\le \sum_{k=1}^{W} P\left\{ \sum_{j \in I_*} |\tilde{\eta}_j - \beta_j^*| \rho_*(\beta^*) \ge \sqrt{2} v(\delta/2W) L \right\} \\
&= \sum_{k=1}^{W} P\left\{ \sum_{j \in I_*} |\tilde{\eta}_j - \beta_j^*| \ge \frac{\sqrt{2} v(\delta/2W) L}{\rho_*(\beta^*)} \right\} \\
\text{(using Condition (A))} \quad &\le \sum_{k=1}^{W} P\left\{ \sum_{j \in I_*} |\tilde{\eta}_j - \beta_j^*| \ge \frac{288\sqrt{2} G^* v(\delta/2W)}{L_{\min} \lambda_W} \right\} \le \sum_{k=1}^{W} \frac{\delta}{W} = \delta.
\end{aligned}
$$

Combining all the bounds above, we can obtain

$$P(\hat{I} \neq I_*) \leq P(I_* \nsubseteq \hat{I}) + P(\hat{I} \nsubseteq I_*) \leq 2W\left(\frac{\delta}{2W^2}\right)^{(1-\epsilon_k^*)^2} + \delta + 2W\left(\frac{\delta}{2W^2}\right)^{(1-\epsilon_k^*)^2} + \delta$$
$$= 4W\left(\frac{\delta}{2W^2}\right)^{(1-\epsilon_k^*)^2} + 2\delta.$$

This completes the proof of Theorem 3.

References

1. McLachlan, G.J.; Lee, S.X.; Rathnayake, S.I. Finite mixture models. *Ann. Rev. Stat. Appl.* **2019**, *6*, 355–378. [CrossRef]
2. Balakrishnan, S.; Wainwright, M.J.; Yu, B. Statistical guarantees for the EM algorithm: From population to sample-based analysis. *Ann. Stat.* **2017**, *45*, 77–120. [CrossRef]
3. Wu, Y.; Zhou, H.H. Randomly initialized EM algorithm for two-component Gaussian mixture achieves near optimality in $O(\sqrt{n})$ iterations. *arXiv* **2019**, arXiv:1908.10935.
4. Chen, J.; Khalili, A. Order selection in finite mixture models with a nonsmooth penalty. *J. Am. Stat. Assoc.* **2008**, *103*, 1674–1683. [CrossRef]
5. DasGupta, A. *Asymptotic Theory of Statistics and Probability*; Springer: New York, NY, USA, 2008.
6. Devroye, L.; Lugosi, G. *Combinatorial Methods in Density Estimation*; Springer: New York, NY, USA, 2001.
7. Biau, G.; Devroye, L. Density estimation by the penalized combinatorial method. *J. Multivar. Anal.* **2005**, *94*, 196–208. [CrossRef]
8. Martin, R. Fast Nonparametric Estimation of a Mixing Distribution with Application to High Dimensional Inference. Ph.D. Thesis, Purdue University, West Lafayette, IN, USA, 2009.
9. Bunea, F.; Tsybakov, A.B.; Wegkamp, M.H.; Barbu, A. Spades and mixture models. *Ann. Stat.* **2010**, *38*, 2525–2558. [CrossRef]
10. Bertin, K.; Le Pennec, E.; Rivoirard, V. Adaptive Dantzig density estimation. *Annales de l'IHP Probabilités et Statistiques* **2011**, *47*, 43–74 . [CrossRef]
11. Tibshirani, R. Regression shrinkage and selection via the lasso. *J. R. Stat. Soc. Ser. B (Methodological)* **1996**, *58*, 267–288. [CrossRef]
12. Hall, P.; Lahiri, S.N. Estimation of distributions, moments and quantiles in deconvolution problems. *Ann. Stat.* **2008**, *36*, 2110–2134. [CrossRef]
13. Meister, A. Density estimation with normal measurement error with unknown variance. *Stat. Sinica* **2006**, *16*, 195–211.
14. Cheng, C.L; van Ness, J.W. *Statistical Regression with Measurement Error*; Wiley: New York, NY, USA, 1999.
15. Zhu, H.; Zhang, R.; Zhu, G. Estimation and Inference in Semi-Functional Partially Linear Measurement Error Models. *J. Syst. Sci. Complex.* **2020**, *33*, 1179–1199. [CrossRef]
16. Zhu, H.; Zhang, R.; Yu, Z.; Lian, H.; Liu, Y. Estimation and testing for partially functional linear errors-in-variables models. *J. Multivar. Anal.* **2019**, *170*, 296–314. [CrossRef]
17. Bonhomme, S. Penalized Least Squares Methods for Latent Variables Models. In *Advances in Economics and Econometrics: Volume 3, Econometrics: Tenth World Congress*; Cambridge University Press: Cambridge, UK, 2013; Volume 51, p. 338.
18. Nakamura, T. Corrected score function for errors-in-variables models: Methodology and application to generalized linear models. *Biometrika* **1990**, *77*, 127–137. [CrossRef]
19. Buonaccorsi, J.P. Measurement error. In *Models, Methods, and Applications*; Chapman & Hall/CRC: Boca Raton, FL, USA, 2010.
20. Carroll, R.J.; Ruppert, D.; Stefanski, L.A.; Crainiceanu, C.M. Measurement error in nonlinear models. In *A Modern Perspective*, 2nd ed.; Chapman & Hall/CRC: Boca Raton, FL, USA, 2006.
21. Zou, H.; Zhang, H. On the adaptive elastic-net with a diverging number of parameters. *Ann. Stat.* **2009**, *37*, 1733–1751. [CrossRef]
22. Aitchison, J.; Aitken, C.G. Multivariate binary discrimination by the kernel method. *Biometrika* **1976**, *63*, 413–420. [CrossRef]
23. Zou, H.; Hastie, T. Regularization and variable selection via the elastic net. *J. R. Stat. Soc. Ser. B (Stat. Methodol.)* **2005**, *67*, 301–320. [CrossRef]
24. Rosenbaum, M.; Tsybakov, A.B. Sparse recovery under matrix uncertainty. *Ann. Stat.* **2010**, *38*, 2620–2651. [CrossRef]
25. Zhang, H.; Jia, J. Elastic-net regularized high-dimensional negative binomial regression: Consistency and weak signals detection. *Stat. Sinica* **2022**, 32. [CrossRef]
26. Buhlmann, P.; van de Geer, S. *Statistics for High-Dimensional Data: Methods, Theory and Applications*; Springer: New York, NY, USA, 2011.
27. Belloni, A.; Rosenbaum, M.; Tsybakov, A.B. Linear and conic programming estimators in high dimensional errors-in-variables models. *J. R. Stat. Soc. Series B (Stat. Methodol.)* **2017**, *79*, 939–956. [CrossRef]
28. Huang, H.; Gao, Y.; Zhang, H.; Li, B. Weighted Lasso estimates for sparse logistic regression: Non-asymptotic properties with measurement errors. *Acta Math. Sci.* **2021**, *41*, 207–230. [CrossRef]
29. Zhang, H.; Chen, S.X. Concentration Inequalities for Statistical Inference. *Commun. Math. Res.* **2021**, *37*, 1–85.
30. Donoho, D.L.; Johnstone, J.M. Ideal spatial adaptation by wavelet shrinkage. *Biometrika* **1994**, *81*, 425–455. [CrossRef]
31. Deng, H.; Chen, J.; Song, B.; Pan, Z. Error bound of mode-based additive models. *Entropy* **2021**, *23*, 651. [CrossRef]

32. Bickel, P.J.; Ritov, Y.A.; Tsybakov, A.B. Simultaneous analysis of Lasso and Dantzig selector. *Ann. Stat.* **2009**, *37*, 1705–1732. [CrossRef]
33. Bunea, F. Honest variable selection in linear and logistic regression models via ℓ_1 and $\ell_1 + \ell_2$ penalization. *Electron. J. Stat.* **2008**, *2*, 1153–1194. [CrossRef]
34. Chow, Y.S.; Teicher, H. *Probability Theory: Independence, Interchangeability, Martingales*, 3rd ed.; Springer: New York, NY, USA, 2003.
35. Hersbach, H.; de Rosnay, P.; Bell, B.; Schepers, D.; Simmons, A.; Soci, C.; Abdalla, S.; Alonso-Balmaseda, M.; Balsamo, G.; Bechtold, P.; et al. *Operational Global Reanalysis: Progress, Future Directions and Synergies with NWP*; European Centre for Medium Range Weather Forecasts: Reading, UK, 2018
36. Fisher, N.I. *Statistical Analysis of Circular Data*; Cambridge University Press: Cambridge, UK, 1995.
37. Broniatowski, M.; Jureckova, J.; Kalina, J. Likelihood ratio testing under measurement errors. *Entropy* **2018**, *20*, 966. [CrossRef]