Consistence Condition of Kernel Selection In Regular Linear Kernel Regression

WangChao (WangChaoAnHui@126.com)
LiBo
LiuHuan
PengPai

Research Article

Keywords: Kernel Selection, Kernel Regression, Prediction Assessment, Covid-19

Posted Date: April 14th, 2022

DOI: https://doi.org/10.21203/rs.3.rs-1556479/v1

Consistence Condition of Kernel Selection In Regular Linear Kernel Regression

Wang Chao[1*], Li Bo[1†], Liu Huan[1†] and Peng Pai[1†]

[1*]Department of Mathematic and Statistics, Central China Normal University, Luoyu, Wuhan, 430071, Hubei, China.

*Corresponding author(s). E-mail(s): WangChaoAnHui@126.com;
Contributing authors: haoyoulibo@163.com; 1083236002@qq.com;
469555176@qq.com;
†These authors contributed equally to this work.

Abstract

Kernel regression is widely used in biology and economy, because it is more adaptable to complex laws than linear regression, and it has better interpretability than many methods in deep learning. In high-dimensional area, $\boldsymbol{l_1}$-norm penalization is a common method for variable selection, which may be derived from the excellent performance of the lasso algorithm. Although it seems natural to generalize from consistency in variable selection to consistency in kernel selection, there are still many details that need to be taken seriously, e.g. the lower eigenvalue condition. The consistence condition of kernel selection in $\boldsymbol{l_1}$-norm regular linear kernel regression is given, including the prediction error. In simulation study, the consistency of different levels of $\boldsymbol{\lambda_n}$ and dimension of features is carefully checked. Finally, the kernel selection method was applied to high-risk area exploration of Covid-19, with the dataset provided by the US Centers for Disease Control and Prevention(CDC). **Declarations of interest:**The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Keywords: Kernel Selection; Kernel Regression; Prediction Assessment; Covid-19

1 Introduction

In many real-world applications, high-dimensional data is often encountered, such as image processing, text classification and variable selection[1–3]. To handle high-dimensional data, regularization methods and kernel methods are two worthwhile approaches. For regularization methods, lasso, ridge, adaptive lasso and elastic net are very popular variable selection methods, among which lasso is especially widely used in practical, for its ability of compressing variable coefficients to zero[4–7]. Compared with the compression method of the regularization methods, the kernel methods are also effective methods through dimensionality reduction[8–10]. Kernel regression has more powerful learning ability than linear regression[11]. It is natural to consider combining kernel methods with regularization methods, especially for sparse kernel problems.

The innovation of combination of kernel methods with regularization methods actually comes from real-world applications, such as signal emission source detection and exploring sources of plague[12–14]. In these scenarios, the experimenter will sample different locations, including the characterization features of each observation location and irradiance of each location. We are subjected to finding out the true emission sources. In such case, the features of each location usually have no direct connection with the irradiance, which renders methods that try to mapping the features to the irradiance useless, including linear or nonlinear methods. Inspired by K Nearest Neighbors method[15], the irradiance of one location should be closely related with its neighborhoods. The difference is that the radiance of a location is no longer related to neighboring locations, but to the distance to the real sources. The real sources are mixed in some candidates locations. When the dimension of the characterization of one location is too large, how to define the distance between two sample points should be carefully checked, where Euclidean distance can not be suitable for the curse of dimensionality[16]. Cosine distance or cosine-like distance, e.g. gaussian kernel, are commonly used in such cases[17, 18]. The characteristics of this type of problem, including weak correlations between features and responses and large feature dimensions, all lead to the critical role of kernel methods.

In the signal emission sources detection problem, it is particularly important to explore the consistency conditions of kernel selection. As the response variable can be seen as the linear combination of the kernels, the true kernels actually stand for emission sources. Denote $K_\alpha(x) = K(\alpha, x)$ be a measure of the distance between α and x, where the Gaussian RBF(Radial Basis Function) Kernel and Linear Inner Product Kernel are most widely used[19–21]. In this paper, the kernel selection consistence condition of Regular Linear Kernel Regression was given, as a compliment theory for linear kernel regression. For a given group of kernels, $K_\alpha(\cdot), \alpha \in \bar{\alpha}$, where $\bar{\alpha}$ stands for an index set, Kernel Regression model has the form,

$$y = \sum_{\alpha \in \bar{\alpha}} \theta_\alpha K_\alpha(x) + w, \tag{1}$$

where, y stands for irradiance and w stands for observation noise. From the additivity of model, y is actually a linear combination of the characteristic Kernels at the point of x. We see a Kernel $K_\alpha(\cdot)$ as a influence function of a source area characterized by α. As stated, $K_\alpha(x)$ is a measure of the distance between α and x. For a long distance area x from source area α, the influence value $K_\alpha(x)$from α shall be small, which is coincident with the property of $K_\alpha(x)$. Upper to now, we explain the link between the spreading mode of emission sources and the Kernel Regression, that's emission sources are characterized by $K_\alpha(\cdot)$ in the model and the response in one area is actually the combination of the influences of those emission sources. Still, there are problems that must be check carefully. First, the number of true emission sources can be unlimited? For a given design matrix $\mathbf{X} \in \mathbb{R}^{n\times d}$, and a candidate set $Card(\bar{\alpha}) = m$, the number of true emission sources s must be less than or equal to $\min\{n, d, m\}$, otherwise the true emission sources will not be check out completely. Second, for a given estimation of $\{\hat{\theta}\}$ with a support set of $\hat{S}$, under which conditions the support set is consistent with the true support set S of true θ? This question is crucial, as it means whether we have got the true emission sources, as a non-zero θ_α means that the candidate area of $K_\alpha(\cdot)$ influences other areas, which suggests location characterized by α is an emission source.

The content of this article is arranged as follows. In the first section, the innovation of regular linear kernel regression was given, where the background scenarios all lead to the usage of regular linear kernel regression. In the second section, some notations was constructed and the explanation of the parameters of linear kernel regression is also here. In the third section, the consistence condition of kernel selection was given with two mild conditions. A corollary about the upper bound on the error of parameter estimation was given, which can be applied in practice. In the forth section, the risk assessment with regular linear kernel regression was discussed. Finally, simulation and real world experiments were designed for checking the change law of consistence. With large dimension of features and small candidate kernels, complete consistency is easy to be fulfilled, along with the no false inclusion. In real world dataset experiment, regular linear kernel regression method was applied to Covid-19 outbreaks of all states of United States[22], provided by the US Centers for Disease Control and Prevention(CDC). The kernel selection results reveal the sources of transmission and purification of the virus. The prediction results evaluate the risk level of each state.

2 Model Construction

Suppose we are given relevant features in observation regions, $x_i = (x_{i1}, x_{i2}, \cdots, x_{id})^T$ and corresponding response values y_i, for $i = 1, \cdots, n$, which stands for the most concerned variable, e.g. the number of epidemic outbreaks. Here, n stands for sample size. A significant research is to find possible emission sources from many candidate areas $\alpha_j = (\alpha_{j1}, \cdots, \alpha_{jd})^T$, for

$j = 1, \cdots, m$. The true emission sources α_S are supposed to be contained in those candidate areas $\bar{\alpha} = [\alpha_1^T, \cdots, \alpha_m^T]^T$ indexed by S and $S \subseteq [m]$ stands for the true support of emission sources. Here $[m]$ is a simplified notation of $\{1, \cdots, m\}$.

Intuitively, the response variable y at observation point x is the linear superimposed effects of these true emission sources, from the superimposed signals viewpoint[23, 24], that's

$$y = \sum_{j \in S} \theta_j < x, \alpha_j > + w, \tag{2}$$

where $S \subseteq \{1, \cdots, m\}$ stands for the supports of truth emission sources, and w stands for observation noise. $< \cdot, \cdot >$ stands for inner product. Since S is unknown in practice, the estimation problem of S will be more meaningful, except for the estimation of $\hat{\theta}_j, j \in S$. Actually if the support S of the truth emission sources was given, we can estimate $\hat{\theta}$ very well. With the true S unknown, there is no hope for obtaining an unique solution of θ_j, when $m > \min\{n, d\}$, from the truth $\text{Rank}(\mathbf{X}\alpha^T) \leq \min\{\text{Rank}(\mathbf{X}), \text{Rank}(\alpha)\}$. Without uniqueness, there is no way to talk about the Kernel Selection Consistence. Inspired by Lasso of l_1 norm regularization, consider the following optimization,

$$\hat{\theta} \leftarrow \arg\min \sum_{i=1}^{n} \frac{1}{2n} (y_i - \sum_{j \in [m]} \theta_j < x_i, \alpha_j >)^2 + \lambda_n \|\theta\|_1. \tag{3}$$

Here, $[m] = \{1, \cdots, m\}$ is a simple notation and λ_n is a given parameter. Denote $\mathbf{X} = [x_1, \cdots, x_n]^T \in \mathbb{R}^{n \times d}$, $\alpha = [\alpha_1, \cdots, \alpha_m]^T \in \mathbb{R}^{m \times d}$, $\theta = [\theta_1, \cdots, \theta_m]^T \in \mathbb{R}^m$ and $\mathbf{Y} = [y_1, \cdots, y_n]^T \in \mathbb{R}^n$. Therefore a matrix form alternative of equation (3), namely Regular Linear Kernel Regression, is given by,

$$\hat{\theta} \leftarrow \arg\min \frac{1}{2n} \|\mathbf{Y} - \mathbf{X}\alpha^T \theta\|_2^2 + \lambda_n \|\theta\|_1. \tag{4}$$

For the not differentiable of l_1-norm, due to its sharp point at the origin, there's no hope for a close form solution. Since this is a convex optimization problem, many optimization methods can be used to solve the problem, e.g. Gradient descent or Least Angle Regression(LARS)[25].

So the method of estimating $\hat{\theta}$ is not the focus of this paper. The explanation of the estimation result of $\hat{\theta}$ and the condition of consistence between the estimation and the truth θ is our focus. Here we give the explanation of $\hat{\theta}$. From the model (2), for a zero $\theta_j = 0$, which means y is irrelevant with α_j, the location of α_j shall not be the emission source. For a non-zero θ_j, which means that y is relevant with α_j, the location of α_j shall be the emission source. For a positive θ_j, the corresponding location of α_j shall be see as emission source in the emission detection case. And a negative θ_j, the corresponding location of α_j shall be see as absorption source.

3 Condition Of Consistence

Consider an S-sparse linear kernel regression model and the true support of $\theta^\star$ is S. Given an estimation $\hat{\theta}$ with support $\hat{S}$. The question is by which condition the estimation $\hat{S}$ is consistent with S? We begin by assuming $\mathbf{X}$ be deterministic design matrices. Let S stands for the supports of truth $\theta^\star$ and α_S stands for the subset of $\alpha = [\alpha_1, \cdots, \alpha_m]^T$ with indices of S. First the truth emission source α_S cannot be linearly correlated, in order to ensure that the model is identifiable, even if the support set S were known a priori. Second the fake source α_{S^C} shall be irrelevant with the truth emission source α_S, which makes it possible to exclude irrelevant areas. In general, we cannot expect this orthogonality to hold, but a type of approximate orthogonality. Especially, The conditions are as follows:

(A1) *Full rank* : The smallest eigenvalue of the hat matrix formed by the true kernels indexed by S is greater than zero:

$$\gamma_{min}\left(\frac{\alpha_S \mathbf{X}^T \mathbf{X} \alpha_S^T}{n}\right) \geq c_{min} > 0. \tag{5}$$

(A2) *Irrelevance* : The correlation between irrelevant kernels and true kernels is low enough, that's for some $p \in [0, 1)$

$$\max_{j \in S^C} \|(\alpha_S \mathbf{X}^T \mathbf{X} \alpha_S^T)^{-1} \alpha_S \mathbf{X}^T \mathbf{X} \alpha_j\|_1 < p. \tag{6}$$

With this set-up, the following **Theorem 1** applied with the regular least square optimization (4) shows that the true support of parameter $\theta^\star$ is consistent with support of the result estimation $\hat{\theta}$, with a proper regularization parameter λ_n. In order to state the result, we introduce the convenient shorthand $\Pi_{S^\perp}(\mathbf{X}\alpha^T) = \mathbf{I}_n - \mathbf{X}\alpha_S^T(\alpha_S \mathbf{X}^T \mathbf{X} \alpha_S^T)^{-1} \alpha_S \mathbf{X}^T$, a type of orthogonal projection matrix.

Theorem 1 *Let the linear kernel regression has the form (2) and assume that the kernel matrix fulfills the full rank condition (5) and irrelevance condition (6). Try to solve the optimization problem defined by (4), with super-parameter λ_n satisfying*

$$\lambda_n \geq \frac{2}{1-p} \left\| \alpha_{S^C} X^T \Pi_{S^\perp(\boldsymbol{X}\alpha^T)} \frac{w}{n} \right\|_\infty, \tag{7}$$

the following conclusions about the solution of the optimization (4) hold:
(a) Existence: The optimal solution $\hat{\theta}$ is unique.
(b) No false inclusion: The support set $\hat{S}$ of optimal solution is contained within the true support set S.
(c) l_∞-bounds: The difference between the solution $\hat{\theta}$ with the truth $\theta^\star$ is upper

bounded

$$\|\hat{\theta}_S - \theta_S^\star\|_\infty \leq \underbrace{\left\| \left(\frac{\alpha_S \boldsymbol{X}^T \boldsymbol{X} \alpha_S^T}{n} \right)^{-1} \alpha_S \boldsymbol{X}^T \frac{w}{n} \right\|_\infty + \left\| \left(\frac{\alpha_S \boldsymbol{X}^T \boldsymbol{X} \alpha_S^T}{n} \right)^{-1} \right\|_\infty}_{B(\lambda_n;\ \boldsymbol{X}\alpha^T)} \lambda_n \,. \quad (8)$$

(d) No mission: The support set $\hat{S}$ of solution contains all indices $i \in S$ which satisfies $\mid \theta_i^\star \mid > B(\lambda_n;\ \boldsymbol{X}\alpha^T)$, and hence the consistence of kernel selection holds when $\min_{i \in S} \mid \theta_i^\star \mid > B(\lambda_n;\ \boldsymbol{X}\alpha^T)$.

The prove of this result utilizes a technique called a primal-dual witness method, which constructs a solution and then the solution was shown be optimal and unique, see Appendix. Before that, let us explain its main results. Result (a) is the basis for subsequent results. Although the objective function is convex, it is not always strictly convex if $m > \min\{d, n\}$. Based on this existence and uniqueness result, we can further discuss the consistency of kernel selection and the inclusion of false kernels. Result (b) states that for sufficiently large λ_n, the result of kernel selection will exclude all false kernels, that's $\hat{\theta}_{S^C} = 0$. Result (d) is a natural corollary of result (c): the coefficient which is large enough thus non-negligible, then it can be successfully identified by the solution support set. And the consistence condition of kernel selection follows from result (d).

Further more, if we make specific assumptions about the noise vector w, more concrete results can be obtained. Before giving the results, we give the definition of the Sub-Gaussian variable[26].

Definition 1 Define a variable w as sub-gaussian variable, when the moment generation function satisfies the following inequality: for some positive σ,

$$\mathbb{E}[e^{\lambda(w-\mu)}] \leq e^{\sigma^2 \lambda/2} \qquad for \quad all \quad \lambda \in \mathbb{R},$$

where $\mu = \mathbb{E}[w]$ stands for the mean of w.

With this in mind, we given a third assumption, that's

(A3) The observation noise $w_i, i \in \{1, \cdots, n\}$ is i.i.d. σ-sub-Gaussian variables.

Before given the concrete corollary about Linear Kernel Regression, a Lemma about lower bounds for sub-Gaussian variables is needed.

Lemma 2 Let $\{Z_i\}_{i=1}^n$ be a sequence of sub-gaussian variables, with the same parameter σ. For simplicity, take the mean of all these variables be zero. A upper bound of the maximum of all these variables, $Z = \max_{i=1,\cdots,n} \mid Z_i \mid$ holds as follows,

$$\mathbb{P}[Z \geq t] \leq 2ne^{-\frac{t^2}{2\sigma^2}}, \quad (9)$$

valid for all $n \geq 2$. Notice that there is no independence required.

Corollary 3 Let the linear kernel regression has the form (2) and assume that the kernel matrix fulfills the full rank condition (5) and irrelevance condition (6). And the l_2-norm of all kernels is upper bounded: $\max_{k=1,\cdots,m} \|\mathbf{X}\alpha_k\|_2/\sqrt{n} \leq C$(**kernel normalization condition**). Try to solve the optimization problem defined by (4), with super-parameter λ_n defined by

$$\lambda_n = \frac{2C\sigma}{1-p}\left\{\sqrt{\frac{2\log(m-s)}{n}} + \delta\right\} \tag{10}$$

for some $\delta > 0$. Then with high probability, there is an unique optimal solution $\hat{\theta}$ with no false kernels contained with the truth support set S. And there is a l_∞-error bound for difference between the estimation of coefficients with the truth coefficients,

$$\|\hat{\theta}_S - \theta^\star_S\|_\infty \leq \frac{\sigma}{\sqrt{c_{min}}}\left\{\sqrt{\frac{2\log s}{n}} + \delta\right\} + \left\|\left(\frac{\alpha_S \mathbf{X}^T \mathbf{X}\alpha_S^T}{n}\right)^{-1}\right\|_\infty \lambda_n, \tag{11}$$

with probability at least $1 - 4e^{-\frac{n\delta^2}{2}}$.

Proof Firstly, we show that the lower bound of super-parameter λ_n, (7), holds with high probability, given the assumption (10). Consider the following random variables,

$$Z_k := \alpha_k^T \mathbf{X}^T \underbrace{[\mathbf{I}_n - \mathbf{X}\alpha_S^T(\alpha_S \mathbf{X}^T \mathbf{X}\alpha_S^T)^{-1}\alpha_S \mathbf{X}^T]}_{\Pi_{S^\perp}(\mathbf{X})}\left(\frac{w}{n}\right) \qquad for \quad k \in S^c.$$

From the definition of $\Pi_{S^c}(\mathbf{X})$, it follows that

$$\|\Pi_{S^\perp}\mathbf{X}\alpha_k\|_2 \leq \|\mathbf{X}\alpha_k\| \leq C\sqrt{n},$$

where the last inequality comes from the kernel normalization condition. Hence, all Z_j is proved to be sub-Gaussian with parameters smaller than $C^2\sigma^2/n$. By Lemma 2, it holds that,

$$\mathbb{P}\left[\max_{j\in S^c} \mid Z_j \mid \geq t\right] \leq 2(m-s)e^{-\frac{nt^2}{2C^2\sigma^2}}.$$

Transforming this to the l_∞-norm form, we have

$$\mathbb{P}\left[\frac{2}{1-p}\left\|\alpha_{S^C} X^T \Pi_{S^\perp(\mathbf{X}\alpha^T)}\frac{w}{n}\right\|_\infty \geq \lambda_n\right] \leq 2(m-s)e^{-\frac{n(1-p)^2\lambda_n^2}{8C^2\sigma^2}}.$$

Substitute the choice of λ_n (10), we have

$$\lambda_n \geq \frac{2}{1-p}\left\|\alpha_{S^C} X^T \Pi_{S^\perp(\mathbf{X}\alpha^T)}\frac{w}{n}\right\|_\infty,$$

with probability at least $1 - 2e^{-\frac{n\delta^2}{2}}$.

It remains to deduce the error bound (11) from equation (8). Define variables $\tilde{Z}_i := e_i^T\left(\frac{1}{n}\alpha_S \mathbf{X}^T \mathbf{X}\alpha_S^T\right)^{-1}\mathbf{X}\alpha_S^T w/n$. From assumption, we know that elements of the observation noise w are i.i.d. σ-sub-Gaussian. Hence $\tilde{Z}_i$ is zero-mean and sub-Gaussian variable with parameter satisfying

$$\frac{\sigma^2}{n}\left\|\left(\frac{1}{n}\alpha_S \mathbf{X}^T \mathbf{X}\alpha_S^T\right)^{-1}\right\|_2 \leq \frac{\sigma^2}{c_{\min} n}.$$

Here the **Full rank** condition (5) is needed. From **Lemma 2**, it holds that, for any $\delta > 0$,

$$\mathbb{P}\left[\max_{i=1,\cdots,s} \mid \tilde{Z}_i \mid > \frac{\sigma}{\sqrt{c_{\min}}}\left\{\sqrt{\frac{2\log s}{n}} + \delta\right\}\right] \leq 2e^{-\frac{n\delta^2}{2}}.$$

we complete the proof by the probability simplification, $(1-2e^{-\frac{n\delta^2}{2}})^2 \geq 1-4e^{-\frac{n\delta^2}{2}}$. □

Upper to here, we have given the Kernel selection consistence conditions. By careful analysis of the original optimization problem (3), we find that the values of $\{\hat{\theta}_k\}_{k=1}^m$ gradually diminish, as the penalize parameter λ_n is large enough. From condition (7), if the given parameter λ_n is too small, (7) condition is violated, we have no hope for excluding all fake Kernels. From part (d) of Theorem 1, the $\hat{\theta}_i, i \in S$, which cannot be zero, may be miscalculated as zero. In real world data detection, the choice of λ shall be carefully checked. Besides, Cross-validation methods[27] are usually utilized for the decision of λ_n, or other improved methods [28].

4 Prediction Error

Another interesting twist is to do a prediction of response on all candidate kernels. Consider the Linear Kernel Regression Model (2). Given an estimation of $\hat{\theta}$, the prediction of response $\hat{y}$ is the linear combination of the Kernels at the position of x, that's

$$\hat{y} = \sum_{k\in[m]} \hat{\theta}_k < \alpha_k, x > .$$

The prediction $\hat{y}$ is the linear combination of the Kernel functions, and can also be seen as the result of the superposition of the radiation domains of all emitting sources, at the point x. We see $\hat{y}$ as the irradiance assessment of the point x. Given a S-sparse Linear Kernel Regression model, we devoted ourselves to bound on the mean-squared prediction error

$$\frac{\|\mathbf{X}\alpha^T(\hat{\theta}-\theta^\star)\|_2^2}{n} = \frac{1}{n}\sum_{i=1}^n \left[\sum_{k=1}^m < \alpha_k, x_i > (\hat{\theta}_k - \theta_k^\star)\right]^2. \tag{12}$$

Before giving the results, some more constraints about the design matrix, namely restricted eigenvalue (RE) condition shall be advertised here, which is first proposed by [29].

Definition 2 The matrix $\mathbf{X}$ satisfies the restricted eigenvalue (RE) condition over S with parameters (κ, γ) if

$$\frac{1}{n}\|\mathbf{X}\Delta\|_2^2 \geq \kappa\|\Delta\|_2^2 \qquad for \quad all \quad \Delta \in \mathbb{C}_\gamma(S).$$

Here the subset $\mathbb{C}_\gamma(S) = \{\Delta \in \mathbb{R}^d \mid \|\Delta_{S^c}\|_1 \leq \gamma\|\Delta_S\|_1\}$, is a collection of vectors whose l_1-norm off the support is dominated by the l_1-norm on the support.

The restricted eigenvalue (RE) condition constructs a link between the recovering error $\|\hat{\theta} - \theta^\star\|_2$ and the prediction error $\|\mathbf{X}\alpha^T(\hat{\theta} - \theta^\star)\|_2$. With this set-ups, we give the bounds on prediction errors as follows.

Theorem 4 *(Prediction error bounds) Consider the Regular Linear Kernel Regression Optimization Problem (3) with a positive regularization parameter* $\lambda_n \geq 2\|\frac{\alpha \boldsymbol{X}^T w}{n}\|_\infty$.
(a) Given any optimal estimation $\hat{\theta}$*, the prediction error follows*

$$\frac{\|\boldsymbol{X}\alpha^T(\hat{\theta} - \theta^\star)\|_2^2}{n} \leq 12\|\theta^\star\|_1\lambda_n.$$

(b) If $\theta^\star$ *is supported on a subset* S *of cardinality* s*, and the design matrix* $\boldsymbol{X}\alpha^T$ *satisfies the* $(\kappa;\ 3)$*-RE condition over* S*, then the optimal solution satisfies the bound*

$$\frac{\|\boldsymbol{X}\alpha^T(\hat{\theta} - \theta^\star)\|_2^2}{n} \leq \frac{9}{\kappa}s\lambda_n^2.$$

The proof of this prediction bounds can be seen in Appendix. Before that, let's see what we are informed from the bounds. The two upper bounds are all related to λ_n. In order to reduce the forecast errors, a critical step is to reduce the regularization parameter λ_n.

5 Simulation Studies

Consider a S-sparse Linear Kernel Regression model (2). Suppose each row x_i of the design matrix $\mathbf{X} \in \mathbb{R}^{n\times d}$ is independently draw from a Gaussian distribution $N(0, \mathbf{I}_d)$. And each α_j of the candidates kernels $\alpha \in \mathbb{R}^{m\times d}$ is also i.i.d of Gaussian distribution $N(0, \mathbf{I}_d)$. The true support S is a subset of $\{1, \cdots, m\}$ uniformly distributed in $[m]$. The real parameter of $\theta_j^\star, j \in S$ is uniformly distributed on $[-1, -0.1] \cup [0.1, 1]$, and $\theta_j^\star = 0$ for $j \in S^c$. Let the entries of the noise vector w follow i.i.d of a Gaussian distribution of $N(0, 0.01)$. Then the response variable Y is calculated by

$$Y = \mathbf{X}\alpha^T\theta + w.$$

Applied by the regular method of l_1-norm, we are interested in the false-inclusion rate and consistency. False-inclusion rate is defined as,

$$False - inclusion = \frac{\hat{S}\backslash S}{\hat{S}}, \tag{13}$$

Table 1 The False-inclusion rate and consistency of given levels of λ_n, where $n = m = 1000$, $d = 500$.

s	λ_n	F	Con	s	λ_n	F	Con
$s = 5$	$0 \sim 0.03$	0.992	0.007	$s = 15$	$0 \sim 0.05$	0.97	0.029
	$0.03 \sim 55.5$	0	1		$0.05 \sim 68.7$	0	1
	$55.5 \sim 95.9$	0	0.8		$68.7 \sim 72.7$	0	0.933
	$95.9 \sim 242$	0	0.6		$72.7 \sim 103$	0	0.867
$s = 10$	$0 \sim 0.10$	0.988	0.011	$s = 20$	$0 \sim 0.05$	0.964	0.035
	$0.10 \sim 103$	0	1		$0.05 \sim 82.8$	0	1
	$103 \sim 254$	0	0.9		$82.8 \sim 105.5$	0	0.95
	$254 \sim 263.6$	0	0.8		$105.5 \sim 128.3$	0	0.9

where $\hat{S}$ stands for the support of estimation $\hat{\theta}$. And the consistency between $\hat{S}$ and S, is given by

$$Consistency = \frac{\hat{S} \cap S}{\hat{S} \cup S}. \tag{14}$$

Since the choice of λ_n is critical for consistence. We checked the false-inclusion (F) and consistency (Con) under given levels of λ_n in Table 1. From Table 1, we known that the best result is $F = 0$, which means no false inclusion and $Con = 1$, which means the estimation support $\hat{S}$ is identical with the true kernel support S. For λ_n that's not too small, the no false inclusion holds. And the consistence grows up and down, as λ_n increases. The consistence holds for a long time as we increase the value of λ_n, which means that for a long period, the support of $\hat{\theta}$ makes no changes. And this information will be useful for empirical selection of the regularization parameter λ_n.

Further, we find that the sample dimension has an important relationship with the consistency. We calculated the average consistency under different dimensions of $d \in [40, 100, 200, 300, 500]$. The average is the result of 100 replicates, where the λ_n is chosen according to maximizing consistence. From the results in Table 2, it's easy to find that as the dimension increases, the average consistency increases steadily. As the number of candidate Kernels increase, picking out exactly the true supports becomes more difficult. From here, we get the revelation that increasing the dimension of the data is of great significance to correctly pick out the Kernels. And reducing the number of candidate kernels is also useful for reducing $false-inclusion$ rates and improving $consistency$.

Finally, We care about the changing law of the prediction error, by adding tests check. We draw the test data-set by choosing a test design matrix $\mathbf{X}^{test} \in \mathbb{R}^{T \times d}$, with each row x_i^{test} follows a Gaussian distribution $N(0, \mathbf{I}_d)$. And the test response variable Y^{test} is calculated by

$$Y^{test} = \mathbf{X}^{test} \alpha^T \theta + w.$$

The prediction error is calculated by equation (12). The curves of prediction error along λ_n were plotted in Figure 1 under given levels of $s \in [5, 10, 15, 20]$, which stands for the number of true Kernels. Here $n = 1000, m = 1000, d = 500, T = 500$, where T stands for the size of test data-set. From the results of

Table 2 The average False-inclusion rate and consistency of dimensions of $d \in [40, 100, 200, 300, 500]$. Here $n = 1000$, $s = 20$ and the average is calculated over 100 replicates.

m	d	$\bar{F}$	$\bar{Con}$	m	d	$\bar{F}$	$\bar{Con}$
$m = 40$	$d = 40$	0.205	0.772	$m = 300$	$d = 40$	0.603	0.234
	$d = 100$	0.043	0.956		$d = 100$	0.330	0.593
	$d = 200$	0.004	0.995		$d = 200$	0.074	0.923
	$d = 300$	0.001	0.998		$d = 300$	0.014	0.984
	$d = 500$	0	1		$d = 500$	0.001	0.999
$m = 100$	$d = 40$	0.462	0.430	$m = 1000$	$d = 40$	0.731	0.118
	$d = 100$	0.158	0.830		$d = 100$	0.475	0.366
	$d = 200$	0.019	0.980		$d = 200$	0.169	0.800
	$d = 300$	0.002	0.997		$d = 300$	0.049	0.948
	$d = 500$	0	1		$d = 500$	0.002	0.997

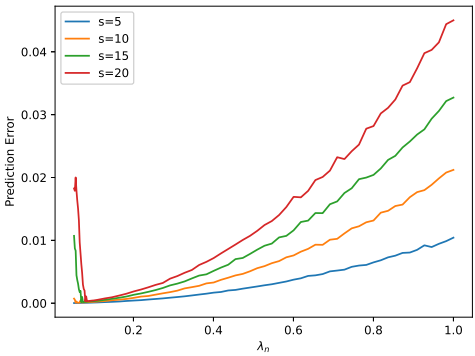

Fig. 1 The Prediction error of Regular Linear Kernel Regression.

the Figure 1, the prediction error grows as λ_n goes beyond some fixed point. And for the different levels of s, the prediction error grows proportional to s, which is coincident with the result from part (b) of Theorem 4.

6 Propagation Sources Selection and Risk Assessment of Covid-19

In this section, we try to apply regular linear kernel regression method to exploring the propagation and purification sources of Covid-19 among all Unites States in situations where Covid-19 transmission and treatment are balanced. The data of confirmed cases of each states in U.S. comes from https://raw.githubusercontent.com/nytimes/covid-19-data/master/us-counties.csv provided by the US Centers for Disease Control and Prevention(CDC). The dataset contains all confirmed cases in the state level from January 21, 2020 to March 26, 2022. With the daily data in hand, we count the number of confirmed cases each week for each state. There

Table 3 The propagation and purification states selected by Regular Linear Kernel Regression.

λ_n	*propagation states*	*purification states*
$\lambda_n = 0.2$	California, New York, Florida	Washington, Texas,
$\lambda_n = 0.1$	California, New York, Florida,Ohio	Washington, Texas, Illinois
$\lambda_n = 0.07$	California, New York, Florida,Ohio	Washington, Texas, Illinois,Nebraska

are 113 weeks of data, which form the features of each state. Take number of cumulative confirmed cases as the response variable. Denote $x = [w_1, w_2, \cdots, w_{113}]$ with its element w_i, the number of weekly confirmed cases, as the representation of a state. Take all states be candidate kernels $K_x(\cdot) =< \cdot, x >$.

Before the regular Linear Kernel Regression model being implemented, a normalize scale of data set is employed, as follows

$$x_{scale} = \frac{x - \bar{x}}{\sqrt{S^2(x)}}.$$

Here, $\bar{x}$ stands for sample mean of x and $S^2(x)$ stands for sample covariates, $S^2(x) = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$. Given several levels of λ_n, the Kernel areas selected were given in Table 3. In the table, propagation states are kernels, whose coefficients are positive. Purification states are kernels, whose coefficients are negative.

Another meaningful task is to carry out a risk assessment of all candidate states. Given $\lambda_n = 0.07$, we made a risk assessment for all 56 states. The risk is calculated by

$$Y_{risk} = \sum_{k=1}^{m} \hat{\theta}_k < x, \alpha_k > .$$

As all variables of covariates were scaled with zero-mean and standard deviation $S^2(x) = 1$, including response variable, the number of cumulative confirmed cases. So the risk Y_{risk} of each area takes values around 0. For high-risk areas, the response Y_{risk} is also large. In particularly, the assessment results are presented in Table 4. From the risk assessment table, we see that California, Texas, Florida, Illinois and New York are of high risk in risk assessment program. While in purification areas selection, Texas and Illinois are viewed as purification areas. The reason lies in the those areas are close in "distance" with high-risk areas and actually they are victims. They are surrounded by the true Covid-19 emission sources. The relation between kernels selection with risk assessment is that kernel selection is finding the causes or emission sources and risk assessment is the presentation of spreading of Covid-19. For some areas that are judged to be of great risk assessment, they are likely to be innocent and may be purification areas. With Kernel selection of

Table 4 The risk assessment of all 189 countries or areas.

index	state	*risk*	index	state	*risk*
1	California	4.41	29	Arkansas	-0.33
2	Texas	3.19	30	Nevada	-0.35
3	Florida	2.51	31	Iowa	-0.35
4	New York	2.04	32	Mississippi	-0.38
5	Illinois	1.06	33	Oregon	-0.43
6	Pennsylvania	0.76	34	Kansas	-0.43
7	North Carolina	0.68	35	Connecticut	-0.43
8	Georgia	0.62	36	Nebraska	-0.5
9	Ohio	0.6	37	Idaho	-0.53
10	Michigan	0.6	38	New Mexico	-0.55
11	New Jersey	0.41	39	West Virginia	-0.57
12	Arizona	0.39	40	Puerto Rico	-0.6
13	Tennessee	0.26	41	Rhode Island	-0.65
14	Washington	0.22	42	Montana	-0.67
15	Virginia	0.16	43	Maine	-0.68
16	Massachusetts	0.13	44	New Hampshire	-0.68
17	Indiana	0.12	45	Delaware	-0.7
18	Missouri	0.07	46	North Dakota	-0.7
19	Wisconsin	0.06	47	Alaska	-0.7
20	Minnesota	0.04	48	South Dakota	-0.7
21	Colorado	0.02	49	Hawaii	-0.74
22	South Carolina	-0.02	50	Wyoming	-0.75
23	Alabama	-0.06	51	District of Columbia	-0.77
24	Kentucky	-0.09	52	Vermont	-0.79
25	Louisiana	-0.1	53	Guam	-0.82
26	Maryland	-0.2	54	Virgin Islands	-0.84
27	Oklahoma	-0.27	55	Northern Mariana Islands	-0.84
28	Utah	-0.28	56	American Samoa	-0.85

propagation areas, only then can we decide whether an area is a true source of transmission, not just by risk assessment.

7 Conclusion

In this paper, we considered the condition of Kernel selection consistence, in Regular Linear Kernel Regression Model with l_1-norm. And bounds on the prediction error of RLKR are also given with mild conditions. Then we applied the RLKR to propagation sources selection and risk assessment for Covid-19, in a view point of area-to-area spreading mode. In the simulation of random design matrix, we explored the $False - inclusion$ and $Consistency$ under different levels of λ_n and the number m of candidate kernels. The prediction error bounds were also checked with a no-wise result. In real world data of Covid-19, the risk assessment of 56 states were given. We believe this will provide valid advices on traveling.

Acknowledgments. Our thanks go to Professor Bo Li for stimulating this research. Our research is supported by NSF of China (Grant No. 61877023) and the Fundamental Research Funds for the Central Universities (Grant No. CCNU19TD009).

Appendix A

Proof **(Theorem 1)** As the non-differentiable of l_1-norm, dual to the sharp point at 0. A tool named subgradient is needed. Consider a convex function $f : \mathbb{R}^m \to \mathbb{R}$, $z \in \mathbb{R}^m$ is called the subgradient of f, simplified as $z \in \partial f(\theta)$, when z satisfies

$$f(\theta + \Delta) \geq f(\theta) + < z, \Delta > \qquad for \quad all \quad \Delta \in \mathbb{R}^m.$$

For $f(\theta) = \|\theta\|_1$, it holds that $z \in \partial\|\theta\|_1$ if and only if $z_j = sign(\theta_j)$, for $j = 1, \cdots, m$. Here $sign(0)$ takes value in $[-1, 1]$. For Regular Linear Kernel Regression program (3), a tuple $(\hat{\theta}, \hat{z}) \in \mathbb{R}^m \times \mathbb{R}^m$ is optimal dual when $\hat{\theta}$ is a minimizer and $z \in \partial\|\theta\|_1$. For regular linear kernel regression, the zero-subgradient must have the form,

$$\frac{1}{n}\alpha\mathbf{X}^T(\mathbf{X}\alpha^T\hat{\theta} - \mathbf{Y}) + \lambda_n\hat{z} = 0, \tag{15}$$

which generalizes the zero-gradient condition to non-differentiable situation. For the existence and uniqueness, an optimal solution $(\hat{\theta}, \hat{z})$ must satisfy the zero-subgradient (15), where $\hat{\theta}$ has right support. The constructive program namely Primal Dual Witness (PDW) is as follows

1. Given $\hat{z}_{S^c} = 0$.
2. Estimate $(\hat{\theta}_S, \hat{z}_S) \in \mathbb{R}^s \times \mathbb{R}^s$ by

$$\hat{\theta}_S \leftarrow \arg\min_{\theta_S \in \mathbb{R}^s} \left\{ \underbrace{\frac{1}{2n}\|\mathbf{Y} - \mathbf{X}\alpha_S^T\theta_S\|_2^2}_{f(\theta_S)} + \lambda_n\|\theta_S\|_1 \right\}. \tag{16}$$

   Then compute $\hat{z}_S \in \partial\|\theta_S\|_1$ from equation $\nabla f(\theta_S)\,|_{\theta_S = \hat{\theta}_S} + \lambda_n\hat{z}_S = 0$.
3. Compute $\hat{z}_{S^c} \in \mathbb{R}^{m-s}$ by substitute $\hat{z}_{S^c} = 0$ and $(\hat{\theta}_S, \hat{z}_S)$ into equation (15). If $\|\hat{z}_{S^c}\|_\infty < 1$ follows, then we say this PDW program succeed.

We announce the PDW program succeed when $\|\hat{z}_{S^c}\|_\infty < 1$ in the last step. From Lemma 5, we have the unique solution $(\hat{\theta}_S, 0)$ for the program (3). It remains to show the PDW program succeed, that's $\hat{z}_{S^c} < 1$. Let's give a block matrix form of zero-subgradient equation (15) first, that's

$$\frac{1}{n}\begin{bmatrix} \alpha_S\mathbf{X}^T\mathbf{X}\alpha_S^T & \alpha_S\mathbf{X}^T\mathbf{X}\alpha_{S^c}^T \\ \alpha_{S^c}\mathbf{X}^T\mathbf{X}\alpha_S^T & \alpha_{S^c}\mathbf{X}^T\mathbf{X}\alpha_{S^c}^T \end{bmatrix}\begin{bmatrix} \hat{\theta}_S - \theta_S^\star \\ 0 \end{bmatrix} - \frac{1}{n}\begin{bmatrix} \alpha_S\mathbf{X}^Tw \\ \alpha_{S^c}\mathbf{X}^Tw \end{bmatrix} + \lambda_n\begin{bmatrix} \hat{z}_S \\ \hat{z}_{S^c} \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \tag{17}$$

Solving this condition, we have

$$\hat{z}_{S^c} = -\frac{1}{\lambda_n n}\alpha_{S^c}\mathbf{X}^T\mathbf{X}\alpha_S^T(\hat{\theta}_S - \theta_S^\star) + \alpha_{S^c}\mathbf{X}^T\left(\frac{w}{\lambda_n n}\right).$$

Similarly, use the invertibility of $\alpha_S\mathbf{X}^T\mathbf{X}\alpha_S^T$, we have

$$\hat{\theta}_S - \theta_S^\star = (\alpha_S\mathbf{X}^T\mathbf{X}\alpha_S^T)^{-1}\alpha_S\mathbf{X}^Tw - \lambda_n n(\alpha_S\mathbf{X}^T\mathbf{X}\alpha_S^T)^{-1}\hat{z}_S. \tag{18}$$

Substituting this solution into the calculation of $\hat{z}_{S^c}$, we must have

$$\hat{z}_{S^c} = \underbrace{\alpha_{S^c}\mathbf{X}^T\mathbf{X}\alpha_S^T(\alpha_S\mathbf{X}^T\mathbf{X}\alpha_S^T)^{-1}\hat{z}_S}_{u} + \underbrace{\alpha_{S^c}\mathbf{X}^T[\mathbf{I} - \alpha_S\mathbf{X}^T(\alpha_S\mathbf{X}^T\mathbf{X}\alpha_S^T)^{-1}\mathbf{X}\alpha_S^T]\left(\frac{w}{\lambda_n n}\right)}_{v}.$$

With the triangle inequality, it follows that $\|\hat{z}_{S^c}\|_\infty \leq \|u\|_\infty + \|v\|_\infty$. With *Irrelevance* condition (A2), we have $\|u\|_\infty \leq p$. From the definition of λ_n (7), it follows that $\|v\|_\infty \leq \frac{1}{2}(1-p)$. Combining these results, it holds that $\|\hat{z}_{S^c}\|_\infty < 1$.

Further, we establish an upper bound on the error $\|\hat{\theta}_S - \theta^\star_S\|_\infty$. According to equation (18), it follows that

$$\|\hat{\theta}_S - \theta^\star_S\|_\infty = \left\| \left(\frac{\alpha_S \mathbf{X}^T \mathbf{X} \alpha_S^T}{n} \right)^{-1} \alpha_S \mathbf{X}^T \frac{w}{n} \right\|_\infty + \lambda_n \left\| \left(\frac{\alpha_S \mathbf{X}^T \mathbf{X} \alpha_S^T}{n} \right)^{-1} \right\|_\infty . \quad (19)$$

From the construction of the unique optimal solution $\hat{\theta}$, the result (b) holds naturally. And result (d) is a general corollary of result (c). □

Lemma 5 With the *Full Rank* condition (A1) at hand, if the PDW program succeed, then the constructed solution $(\hat{\theta}_S, 0) \in \mathbb{R}^m$ is the unique optimal solution of the RLKR program.

Proof Let the PDW program succeeds, hence $(\hat{\theta}_S, 0)$ is a minimizer and there is a subgradient vector $\hat{z} \in \mathbb{R}^m$ with its subindex child suffices $\|\hat{z}_{S^c}\|_\infty < 1$, and $< \hat{z}, \hat{\theta} >= \|\hat{\theta}\|_1$. Suppose we have another minimizer $\tilde{\theta}$. If we make a simple notation $F(\theta) = \frac{1}{2n}\|\mathbf{Y} - \mathbf{X}\alpha^T\theta\|_2^2$, it must be follows that $F(\hat{\theta}) + \lambda_n < \hat{z}, \hat{\theta} >= F(\tilde{\theta}) + \lambda_n \|\tilde{\theta}\|_1$, and thus

$$F(\hat{\theta}) - \lambda_n < \hat{z}, \tilde{\theta} - \hat{\theta} >= F(\tilde{\theta}) + \lambda_n(\|\tilde{\theta}\|_1 - < \hat{z}, \tilde{\theta} >).$$

From the zero-subgradient equation (15), it holds that $\lambda_n \hat{z} = -\nabla F(\hat{\theta})$, and hence

$$F(\hat{\theta}) + < \nabla F(\hat{\theta}), \tilde{\theta} - \hat{\theta} > - F(\tilde{\theta}) = \lambda_n(\|\tilde{\theta}\|_1 - < \hat{z}, \tilde{\theta} >).$$

Given $F(\theta)$ a convex function, the left expression is always negative, and hence $\|\tilde{\theta}\|_1 \leq < \hat{z}, \tilde{\theta} >$. Combined with $< \hat{z}, \tilde{\theta} >\leq \|\hat{z}\|_\infty \|\tilde{\theta}\|_1$, it holds that $\|\tilde{\theta}\|_1 =< \hat{z}, \tilde{\theta} >$. As $\hat{z}_{S^c} < 1$, it follows that $\tilde{\theta}_{S^c} = 0$.

Thus, all minimizers shall can only be supported on S , and the non-zero coeffiences can be estimated by equation (16) with no exceptions. With the assumption of *Full Rank* (A1), the strictly convexity of optimizaton (16) can be infered, and hence only one unique solution can be obtained. □

Proof **(Theorem 4)** We denote $\hat{\Delta} = \hat{\theta} - \theta^\star$.

(a) As $\hat{\theta}$ is the minimizer of the Regular Linear Kernel Regression program, we have

$$\frac{1}{2n}\|\mathbf{Y} - \mathbf{X}\alpha^T\hat{\theta}\|_2^2 + \lambda_n\|\hat{\theta}\|_1 \leq \frac{1}{2n}\|\mathbf{Y} - \mathbf{X}\alpha^T\theta^\star\|_2^2 + \lambda_n\|\theta^\star\|_1.$$

By some calculation, it follows that

$$0 \leq \frac{1}{2n}\|\mathbf{X}\hat{\Delta}\|_2^2 \leq \frac{w^T\mathbf{X}\alpha^T\hat{\Delta}}{n} + \lambda_n\{\|\theta^\star\|_1 - \|\hat{\theta}\|_1\}. \quad (20)$$

By Hölder inequality and the definition of λ_n, it holds that

$$| \frac{w^T\mathbf{X}\alpha^T}{n} | \leq \left\| \frac{\alpha\mathbf{X}^T w}{n} \right\|_\infty \|\hat{\Delta}\|_1 \leq \frac{\lambda_n}{2}\{\|\theta^\star\|_1 + \|\hat{\theta}\|_1\}.$$

Hence

$$0 \leq \frac{\lambda_n}{2}\{\|\theta^\star\|_1 + \|\hat{\theta}\|_1\} + \lambda_n\{\|\theta^\star\|_1 - \|\hat{\theta}\|_1\}.$$

It can be derived that $\|\hat{\theta}\|_1 \leq 3\|\theta^\star\|_1$. Consequently, by triangle inequality, we have $\|\hat{\Delta}\|_1 \leq \|\theta^\star\|_1 + \|\hat{\theta}\|_1 \leq 4\|\theta^\star\|_1$.

Returning to the inequality (20), it holds that

$$\frac{1}{2n}\|\mathbf{X}\hat{\Delta}\|_2^2 \leq \frac{\lambda_n}{2}\|\hat{\Delta}\|_1 + \lambda_n\{\|\theta^\star\|_1 - \|\theta^\star - \hat{\Delta}\|_1\} \leq \frac{3\lambda_n}{2}\|\hat{\Delta}\|_1.$$

Substitute $\|\hat{\Delta}\|_1 \leq 4\|\theta^\star\|_1$ into this inequality, the claim fulfilled.

(b) With the same argument with the proof of (a), we have

$$\begin{aligned} 0 \leq \frac{1}{2n}\|\mathbf{X}\hat{\Delta}\|_2^2 &\leq \frac{\lambda_n}{2}\|\hat{\Delta}\|_1 + \lambda_n\{\|\theta^\star\|_1 - \|\hat{\theta}\|_1\} \\ &\overset{i}{\leq} \frac{\lambda_n}{2}\|\hat{\Delta}\|_1 + \lambda_n\{\|\hat{\Delta}_S\|_1 - \|\hat{\Delta}_{S^c}\|_1\} \\ &\leq \frac{\lambda_n}{2}\{3\|\hat{\Delta}_S\|_1 - \|\hat{\Delta}_{S^c}\|_1\}. \end{aligned} \tag{21}$$

The inequality (i) comes from the S-sparse of $\theta^\star$. Consequently, $\hat{\Delta} \in \mathbb{C}_3(S)$, by the the $(\kappa;\, 3)$-RE condition of $\mathbf{X}\alpha^T$, it holds that

$$\kappa\|\hat{\Delta}\|_2^2 \leq 3\lambda_n\|\hat{\Delta}_S\|_1 \leq 3\lambda_n\sqrt{s}\|\hat{\Delta}_S\|_2.$$

From the inequality (21), we have

$$\frac{1}{n}\|\mathbf{X}\hat{\Delta}\|_2^2 \leq 3\lambda_n\|\hat{\Delta}_S\|_1 \leq 3\lambda_n\sqrt{s}\|\hat{\Delta}_S\|_2.$$

Putting together this pieces, we have

$$\frac{1}{n}\|\mathbf{X}\hat{\Delta}\|_2^2 \leq \frac{9}{\kappa}s\lambda_n^2.$$

□

Proof **(Lemma 2)** Consider the moment generation function of zero-mean sub-Gaussian variable X,

$$\mathbb{E}\left[e^{\lambda X}\right] \leq e^{\frac{\sigma^2\lambda^2}{2}}.$$

By Hoeffding bound, we have

$$P\left[X \geq t\right] \leq e^{-\frac{t^2}{2\sigma^2}}.$$

From the definition of Z, it holds that

$$\begin{aligned} P[Z \geq t] &\leq \sum_{i=1}^{n}\{P[X_i \geq t] + P[-X_i \geq t]\} \\ &\leq 2ne^{-\frac{t^2}{2\sigma^2}}. \end{aligned}$$

□

References

[1] Radenović, F., Tolias, G., Chum, O.: Fine-tuning cnn image retrieval with no human annotation. IEEE transactions on pattern analysis and machine intelligence **41**(7), 1655–1668 (2018)

[2] Minaee, S., Kalchbrenner, N., Cambria, E., Nikzad, N., Chenaghlu, M., Gao, J.: Deep learning–based text classification: a comprehensive review. ACM Computing Surveys (CSUR) **54**(3), 1–40 (2021)

[3] Heinze, G., Wallisch, C., Dunkler, D.: Variable selection–a review and recommendations for the practicing statistician. Biometrical journal **60**(3), 431–449 (2018)

[4] Zhao, P., Yu, B.: On model selection consistency of lasso. The Journal of Machine Learning Research **7**, 2541–2563 (2006)

[5] Zou, H.: The adaptive lasso and its oracle properties. Journal of the American statistical association **101**(476), 1418–1429 (2006)

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

[7] Zou, H., Hastie, T.: Regularization and variable selection via the elastic net. Journal of the royal statistical society: series B (statistical methodology) **67**(2), 301–320 (2005)

[8] Vert, J.-P., Tsuda, K., Schölkopf, B.: A primer on kernel methods. Kernel methods in computational biology **47**, 35–70 (2004)

[9] Hejazi, M., Al-Haddad, S.A.R., Singh, Y.P., Hashim, S.J., Aziz, A.F.A.: Ecg biometric authentication based on non-fiducial approach using kernel methods. Digital Signal Processing **52**, 72–86 (2016)

[10] Wang, T., Zhang, L., Hu, W.: Bridging deep and multiple kernel learning: A review. Information Fusion **67**, 3–13 (2021)

[11] Crawford, L., Wood, K.C., Zhou, X., Mukherjee, S.: Bayesian approximate kernel regression with variable selection. Journal of the American Statistical Association **113**(524), 1710–1721 (2018)

[12] Jang, B.-W., Kim, C.-G.: Acoustic emission source localization in composite stiffened plate using triangulation method with signal magnitudes and arrival times. Advanced Composite Materials **30**(2), 149–163 (2021)

[13] Al-Jumaili, S.K., Pearson, M.R., Holford, K.M., Eaton, M.J., Pullin, R.: Acoustic emission source location in complex structures using full automatic delta t mapping technique. Mechanical Systems and Signal Processing **72**, 513–524 (2016)

[14] Lowell, J.L., Wagner, D.M., Atshabar, B., Antolin, M.F., Vogler, A.J., Keim, P., Chu, M.C., Gage, K.L.: Identifying sources of human exposure to plague. Journal of Clinical Microbiology **43**(2), 650–656 (2005)

[15] Zouhal, L.M., Denoeux, T.: An evidence-theoretic k-nn rule with parameter optimization. IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews) **28**(2), 263–271 (1998)

[16] Indyk, P., Motwani, R.: Approximate nearest neighbors: towards removing the curse of dimensionality. In: Proceedings of the Thirtieth Annual ACM Symposium on Theory of Computing, pp. 604–613 (1998)

[17] Keerthi, S.S., Lin, C.-J.: Asymptotic behaviors of support vector machines with gaussian kernel. Neural computation **15**(7), 1667–1689 (2003)

[18] Wang, J., Lu, H., Plataniotis, K.N., Lu, J.: Gaussian kernel optimization for pattern classification. Pattern recognition **42**(7), 1237–1247 (2009)

[19] Takeda, H., Farsiu, S., Milanfar, P.: Kernel regression for image processing and reconstruction. IEEE Transactions on image processing **16**(2), 349–366 (2007)

[20] Fan, J., Heckman, N.E., Wand, M.P.: Local polynomial kernel regression for generalized linear models and quasi-likelihood functions. Journal of the American Statistical Association **90**(429), 141–150 (1995)

[21] Köhler, M., Schindler, A., Sperlich, S.: A review and comparison of bandwidth selection methods for kernel regression. International Statistical Review **82**(2), 243–274 (2014)

[22] Velavan, T.P., Meyer, C.G.: The covid-19 epidemic. Tropical medicine & international health **25**(3), 278 (2020)

[23] Feder, M., Weinstein, E.: Parameter estimation of superimposed signals using the em algorithm. IEEE Transactions on acoustics, speech, and signal processing **36**(4), 477–489 (1988)

[24] Upadhya, K., Vorobyov, S.A., Vehkapera, M.: Superimposed pilots are superior for mitigating pilot contamination in massive mimo. IEEE Transactions on Signal Processing **65**(11), 2917–2932 (2017)

[25] Efron, B., Hastie, T., Johnstone, I., Tibshirani, R.: Least angle regression. The Annals of statistics **32**(2), 407–499 (2004)

[26] Lugosi, G., Mendelson, S.: Sub-gaussian estimators of the mean of a random vector. The annals of statistics **47**(2), 783–794 (2019)

[27] Platt, J., *et al.*: Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. Advances in large margin classifiers **10**(3), 61–74 (1999)

[28] Braga, I., Monard, M.C.: Improving the kernel regularized least squares

method for small-sample regression. Neurocomputing **163**, 106–114 (2015)

[29] Wainwright, M.J.: High-dimensional Statistics: A Non-asymptotic Viewpoint vol. 48, pp. 208–209