Optimum Tuning Parameter Selection in Generalized lasso for Clustering with Spatially Varying Coefficient Models

S Rahardiantoro\(^12\), Wataru Sakamoto\(^1\)

\(^1\)Department of Human Ecology, Graduate School of Environmental and Life Science, Okayama University, Okayama, 700-8530, Japan
\(^2\)Department of Statistics, Faculty of Mathematics and Natural Science, IPB University, Bogor, 16680, Indonesia

E-mail: septianrahardiantoro@apps.ipb.ac.id

Abstract. Spatial clustering with spatially varying coefficient models is useful for determining the region with common effects of variables in spatial data. This study focuses on selecting the optimum tuning parameter of the generalized lasso for clustering with the spatially varying coefficient model. The \(k\)-fold cross-validation (CV) may fail to split spatial data into a training set and a testing set, if a region contains only a few observations. Moreover, the \(k\)-fold CV is known to give a biased estimate of the out-of-sample prediction error. Therefore, we investigated the performance of approximate leave-one-out cross-validation (ALOCV) in comparison with \(k\)-fold CV for selecting the tuning parameter in a simulation study on 2-dimensional grid. The ALOCV yielded smaller error than \(k\)-fold CV and could detect edges with differences shrunk by generalized lasso appropriately. Then, the ALOCV for selecting the optimum tuning parameter of the generalized lasso in fitting the spatially varying coefficient model is applied to the Chicago crime data. The result of selection by ALOCV was in accordance with the conclusion suggested in the preceding literature. Clustering into regions in advance for making \(k\)-fold CV feasible may lead to a wrong result of clustering with a spatially varying coefficient model.

1. Introduction

Spatial cluster analysis is one of the essential methods to identify areas with high and low prevalence. During the last few decades, identification of spatial clusters on environmental problems has been widely carried out. Spatial clustering can be applied on the chronic diseases clustering, especially on physical activity and obesity, in U.S. states [1]. Moreover, a spatial cluster analysis can be implemented for detecting the air pollution exposure inequities in the United States [2]. We can also apply it on the social aspect, as in [3], who considered a spatial cluster analysis of social and environmental inequalities of infant mortality.

Spatial clustering with spatially varying coefficient models [4] is useful for determining the region with common effects of variables in spatial data. This study considers applying the generalized lasso for clustering with the spatially varying coefficient models. The generalized lasso is a generalized form of lasso by expressing constraints on the coefficients as the \(\ell_1\) penalty of linear forms [5]. Consider a multiple linear regression model \(\mathbf{y} = \mathbf{X}\beta + \mathbf{e}\), where \(\mathbf{y} \in \mathbb{R}^n\) is a response vector, \(\mathbf{X} \in \mathbb{R}^{n \times p}\) is a predictor matrix, \(\beta \in \mathbb{R}^n\) is a parameter vector, and \(\mathbf{e} \in \mathbb{R}^n\) is a random vector. In the lasso [6], the estimator of \(\beta\) can be expressed as
\begin{equation}
\hat{\beta} = \arg \min_\beta \left\{ \frac{1}{2} \|y - X\beta\|_2^2 + \lambda \|\beta\|_1 \right\}.
\end{equation}

where \(\|a\|_1 = \sum_i |a_i|\) and \(\|a\|_2 = \sqrt{\sum_i |a_i|^2}\); with \(a\) an arbitrary vector and \(\lambda \in [0, \infty)\) is a tuning parameter. If we select \(\lambda\) as zero, then equation (1) is equivalent to ordinary least square (OLS), while if we select a positive \(\lambda\), each component of \(\hat{\beta}\) is penalized and shrunk toward zero, and some components are estimated as zero. In addition, we define \(D \in \mathbb{R}^{m \times p}\) as a penalty matrix, where each row of \(D\) indicates the structure or geometry on the components of \(\beta\). For example, each row of \(D\) contains \(-1\) and \(1\) corresponding to adjacent \(\beta_i\) for the fused lasso problem [7]. Then, the generalized lasso estimator of \(\beta\) can be expressed as

\begin{equation}
\hat{\beta} = \arg \min_\beta \left\{ \frac{1}{2} \|y - X\beta\|_2^2 + \lambda \|D\beta\|_1 \right\}.
\end{equation}

If we have \(D = I\), then the generalized lasso (2) becomes the lasso (1). The concept of applying the generalized lasso to spatially varying coefficient models is motivated by the study of [8], which can be helpful in clustering regions based on spatially varying coefficients [9].

Our crucial theme in this research is to select the optimum tuning parameter \(\lambda\) in applying this generalized lasso for clustering with spatially varying coefficient models. The choice of inappropriate \(\lambda\) may lead to wrong interpretation of the clustering results. Figure 1 illustrates a case study of Chicago crime data [10][11], to which we applied the generalized lasso (1) with four different \(\lambda\) values, \(\lambda = \{0.45, 0.25, 0.08, 0.04\}\). It shows that, as \(\lambda\) value gets smaller, the number of regional clusters increases. Therefore, the selection of the \(\lambda\) value is essential in determining the number of clusters. We consider this case study in more detail in Section 4.

![Illustration of clustering regions with generalized lasso for Chicago crime data with different \(\lambda\) values. (a) \(\lambda = 0.45\); (b) \(\lambda = 0.25\); (c) \(\lambda = 0.08\); (d) \(\lambda = 0.04\).](image)

Figure 1. Illustration of clustering regions with generalized lasso for Chicago crime data with (a) \(\lambda = 0.45\); (b) \(\lambda = 0.25\); (c) \(\lambda = 0.08\); (d) \(\lambda = 0.04\).

A common method that has been often used in selecting \(\lambda\) values is the \(k\)-fold cross-validation (CV). However, in the context of spatial clustering, the \(k\)-fold CV may fail to split the data into training and testing data due to the limited number of observations in each region. Moreover, it is known that \(k\)-fold cross-validation produces a biased estimate of the out-of-sample prediction error, especially when both \(n\) and \(p\) are large and \(k\) is small [12][13]. We may reduce the bias issue by applying a large \(k\) value. We may also consider using \(k = n\) as the special case, which is equal to leave-one-out cross-validation (LOOCV) [14]. The LOOCV minimizes the total prediction error when each one observation is predicted as the testing set using the remaining observations as the training set. LOOCV shows better performance, both in numerical and theoretical aspects [13]. On the other hand, for general regularization problems, LOOCV requires direct computation of the predicted value for each leave-one-out data set, which is computationally expensive. Therefore, the approximate leave-one-out cross-validation (ALOCV) was proposed to reduce computation time [12][15]. The ALOCV gives an approximation of the leave-one-out predicted values based on the primal and dual formulations of the general regularization problems, which is explained in the next session.
In this paper, we investigate performance of ALOCV in comparison with k-fold CV to select the optimum $\lambda$ of the generalized lasso (2) for clustering with the spatially varying coefficient model to spatially connected data. We conduct a simulation study to compare the performance of these CV methods for the generalized lasso in fitting the spatially varying coefficient model to 2-dimensional grid data. Our comparison of the performance is based on cross-validation error between predicted values and the actual values, the distribution of selected $\lambda$, and the number of edges with zero differences of $\hat{\beta}$'s between adjacent nodes. Moreover, we illustrate a case study to show usefulness of ALOCV in applying the generalized lasso to a spatial data.

In Section 2, we explain the generalized lasso for clustering with spatially varying coefficient model, and selection of $\lambda$ by k-fold CV and ALOCV. Furthermore, in Section 3, we describe the results of our simulation study. Finally, in Section 4, we describe the case study of the Chicago crime data.

2. Methodology

First, we explain the concept of applying the generalized lasso for clustering with spatially varying coefficient models. Next, in subsection 2.2, we discuss how to use k-fold CV and ALOCV for the generalized lasso in spatially varying coefficient models.

2.1. The generalized lasso in the spatially varying coefficient models

Suppose that we have spatial data with $L$ regions, where the $l$-th region consists of $q_l$ observations, so that $\sum_{l=1}^{L} q_l = n$. For clustering purposes, let $X$ be the $L$-block diagonal matrix $\tilde{I} = \text{diag}\{1_{q_1}, \ldots, 1_{q_L}\}$, where $1_{q_l} \in \mathbb{R}^{q_l}, l = 1, \ldots, L$ is a vector of all ones. It means that $(i, l)$-th element of $X$ is 1 if $y_l$ belongs to the $l$-th region, and 0 otherwise. Therefore, the generalized lasso in this case can be expressed as

$$\hat{\beta} = \arg\min_\beta \left\{ \frac{1}{2} \|y - \beta\|_2^2 + \lambda \|D\beta\|_1 \right\}$$

(3)

where $\beta \in \mathbb{R}^l$ is a vector of coefficients. The matrix $D \in \mathbb{R}^{m \times L}$ is constructed based on the adjacencies between regions, that is, letting $m$ be the number of edges, the $s$-th row of $D$ can be expressed as

$$D_s = (0, \ldots, -1, \ldots, 1, \ldots, 0), \quad s = 1, \ldots, m,$$

with the $l$-th element $-1$ and $l'$-th element $1$ indicating that the $l$-th and $l'$-th nodes are connected.

2.2. Implementation of k-fold CV and ALOCV in spatially varying coefficient models

Implementing the k-fold CV method in spatially varying coefficient models may fail to split spatial data into a training set and a testing set. If we split spatial data into a training set and a testing set arbitrarily, then each fold allows regions with no observation in the training set. The design matrix $\tilde{I}$ for the training set may have columns of all zeros, and the corresponding elements of $\hat{\beta}$ may not be identified. Therefore, we need to choose sufficiently large $k$ so that every region contains at least one observation in the training set.

The $k$-fold CV error for a specified $\lambda$ can be stated as

$$CV_k(\lambda) = \frac{1}{k} \sum_{c=1}^{k} E_c(\lambda)$$

(4)

where $E_c(\lambda) = \frac{1}{n_c} \sum_{i \in K_c} (y_i - \tilde{x}_i^T \hat{\beta}_{c/}^c)^2$, and

$$\hat{\beta}_{c/}^c = \arg\min_\beta \left\{ \frac{1}{2} \sum_{i \in K_c} (y_i - \tilde{x}_i^T \beta)^2 + \lambda \|D\beta\|_1 \right\}$$

(5)

where $c$ indicates the index of a fold ($c = 1, 2, \ldots, k$), and $K_c$ is the set of indices of observations in the $c$-th fold. $K_c$ should contain at least one observation for each region.

The $k$-fold CV method requires fitting by (5) and computing the prediction error for each division into training and testing sets. The leave-one-out CV, the special case when $k = n$, can solve the issue in
splitting the spatial data into training and data sets described above, but requires heavy computation for large data set.

We follow the ALOCV algorithm by [15] applied to the generalized lasso problem. For each given $\lambda$, the step of ALOCV can be expressed as follows

1. Solve both the primal and dual problems of (3). The generalized lasso estimate $\hat{\beta}$ can be obtained through the primal problem. Then, we estimate $\hat{u}$ as a solution of the dual formulation of (3) which can be represented as:

$$\arg\min_{\theta, u} \frac{1}{2} \| \theta - y \|_2^2 \text{ s.t. } \| u \|_\infty \leq \lambda \text{ and } \tilde{I}^T \theta = D^T u$$  \hspace{1cm} (6)

2. Obtain $E = \{ s : |\hat{u}_s| = \lambda \}$.

3. Define $D_{-E}$. The subscript $-E$ denotes to index over all rows of $D$ except those in $E$.

4. Construct the matrix $B$ whose columns span the null space of $D_{-E}$, or in other words $D_{-E} B = 0$.

5. Compute $H = A A^+$, where $A = X B$ and $A^+$ indicates the Moore-Penrose pseudoinverse of $A$.

6. Obtain the approximate leave-one out predicted value of $y_i$ as $\hat{y}_i = x_i^T \tilde{\beta} + \frac{H_{ii}}{1 - H_{ii}} (x_i^T \hat{\beta} - y_i)$, where $H_{ii}$ is the i-th diagonal component of $H$.

7. Calculate ALOCV error:

$$\frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$ \hspace{1cm} (7)

We select $\lambda$ that minimizes the ALOCV error (7).

3. Simulation Study

We describe our simulation study to compare the performance of k-fold CV and ALOCV in the selection of tuning parameter $\lambda$ of the generalized lasso in fitting the spatially varying coefficient model. This section consists of two parts; the first part explains the simulation scheme and criteria for performance assessment, and the second part discusses the simulation results.

![Figure 2. 2-dimensional grid simulation data pattern.](image)

3.1. Simulation scheme and criteria for checking performance

Here we define our simulation design on a 2-dimensional grid through generating artificial data using the pseudo random number generator of R software. We build a $10 \times 10$ grid, and we consider three sample sizes: 5, 10 and 20 observations in each cell. The response vector $y$ is generated by the formula $y = \tilde{I} \beta + \varepsilon$, where each component of true $\beta$ is given by, according to Figure 2,

$$\beta_i = \begin{cases} 
5 & \text{for lower – right brown cells,} \\
2.5 & \text{for upper – left red cells,} \\
1 & \text{otherwise} 
\end{cases}$$ \hspace{1cm} (8)
and each component of $\varepsilon$ is drawn independently from $N(0, \sigma)$, where standard deviation $\sigma = \{0.5, 0.7, 1\}$. In our simulation, we have $D \in \mathbb{R}^{180 \times 100}$, with $n = 500$, $n = 1000$, and $n = 2000$.

We applied the generalized lasso for 100 replicated datasets, and selected $\lambda$ from a sequence between 0 and 3 by the four methods: 3-fold CV, 5-fold CV, 10-fold CV, and ALOCV. We compared the performance of the four methods based on the following three criteria: the minimum CV error, selected $\lambda$, and the number of edges between adjacent cells $(l, l')$ with zero differences, that is, $|\hat{\beta}_l - \hat{\beta}_{l'}| = 0$.

The algorithm of our simulation is described below.

```
Compute $\hat{\beta}$ according to (8);  
For each replication 1 to 100, do;  
    Draw $\varepsilon$ to generate $y = \hat{\beta} + \varepsilon$;  
    Run the generalized lasso based on formula (3) for each $\lambda$;  
    Apply 3-fold CV, 5-fold CV, 10-fold CV and ALOCV, then do;  
        Save the minimum CV errors and $\lambda$ selected by each method;  
        Save $\hat{\beta}$ based on the $\lambda$ selected by each method, then do;  
            Count the number of edges with $|\hat{\beta}_l - \hat{\beta}_{l'}| < 10^{-5}$;  
        End;  
    End;  
End;
```

Figure 3. Boxplots of simulation results when $\varepsilon \sim N(0, \sigma = 0.5)$: (a) min CV error; (b) selected $\lambda$; (c) the number of edges with zero differences of $\hat{\beta}$'s.
3.2. Simulation results and interpretations

In this part, we show the simulation results in Figure 3 – 5. Figure 3 shows boxplots of the simulation results of the three performance criteria for the three sample sizes with generated error $\varepsilon \sim N(0, \sigma = 0.5)$. We can see that for all sample sizes $n$, ALOCV yields slightly smaller error compared to $k$-fold CV. The range of the CV error for all methods becomes narrower as $n$ increases. ALOCV tends to select higher $\lambda$ value than $k$-fold CV, and the range of selected $\lambda$ become wider as $n$ increases. The $k$-fold CV with $k = 5$ and 10 produces a fairly wide range of selected $\lambda$ when $n = 1000$ and 2000.

Furthermore, on the number of edges between adjacent cells with zero differences of $\hat{\beta}$’s, ALOCV tends to give the largest number, and its range becomes the widest among all methods when $n = 500$, while it gets narrower as $n$ increases. On the other hand, the range of the number of edges with zero differences for $k$-fold CV becomes wider as $k$ and $n$ increase. In summary, ALOCV has a good consistency in selecting edges with shrunk difference.

![Boxplots](image1)

(a) $n = 500$
(b) $n = 500$
(c) $n = 500$

![Boxplots](image2)

(a) $n = 1000$
(b) $n = 1000$
(c) $n = 1000$

![Boxplots](image3)

(a) $n = 2000$
(b) $n = 2000$
(c) $n = 2000$

**Figure 4.** Boxplots of simulation results when $\varepsilon \sim N(0, \sigma = 0.7)$: (a) min CV error; (b) selected $\lambda$; (c) the number of edges with zero differences of $\hat{\beta}$’s.

Figure 4 and Figure 5 show boxplots of the simulation results with generated error $\varepsilon \sim N(0, \sigma = 0.7)$ and $\varepsilon \sim N(0, \sigma = 1)$, respectively. These results are similar with the simulation results when $\sigma = 0.5$. In summary, ALOCV tends to produce a slightly smaller CV error than $k$-fold CV. In addition, ALOCV tends to select higher $\lambda$ value than $k$-fold CV, although the range of selected $\lambda$ for all methods becomes wider as $n$ increases. Moreover, ALOCV tends to give the largest number of edges with zero differences of $\hat{\beta}$’s, with its range getting narrower as $n$ increases.
Figure 5. Boxplots of simulation results when $\varepsilon \sim N(0, \sigma = 1)$: (a) min CV error; (b) selected $\lambda$; (c) the number of edges with zero differences of $\beta$’s.

4. Case Study: Chicago Crime Data

We apply the generalized lasso for clustering with a spatially varying coefficient problem to Chicago crime data [10], which have been also discussed in [11] as an application of fused lasso. The dataset provides the number of burglaries per household over the year 2005 to 2009, spatially aggregated within 2010 census block groups. Although there are totally 2167 blocks in this dataset, we have observations at 2162 blocks (nodes) because 5 blocks with only 1 household have been deleted, and 6995 connections (edges) between neighboring blocks.

First, we applied the generalized lasso (3) to the original dataset directly. We can use ALOCV for selecting optimum $\lambda$ to the dataset with one observation for each node, because it uses only the result of the original fit. On the other hand, $k$-fold CV is not feasible in applying generalized lasso to the original dataset. Therefore, we applied the generalized lasso (3) to the dataset clustered into regions in advance, with $\lambda$ selected by 3-fold CV, 5-fold CV, 10-fold CV and ALOCV. In this setting, we constructed $K = 50$ regions using the $K$-means method, based on the Euclidian distance between the coordinates of each block census. Then, we used the Voronoi tessellation based on the centroids of these regions to identify the adjacency structure between constructed regions.
Figure 6. Fitting the generalized lasso to Chicago crime data (non-clustered).
(a) The ALOCV error for the specified $\lambda$ sequences;
(b) Plot of the estimated coefficient at each block, with $\lambda$ selected by ALOCV.

Figure 6 shows the result of using ALOCV to select the optimum $\lambda$ in fitting the generalized lasso to non-clustered data. In this setting, we have $y \in \mathbb{R}^{2162 \times 1}, X = I \in \mathbb{R}^{2162 \times 2162}, D \in \mathbb{R}^{6995 \times 2162}$. We consider a sequence of 50 tuning parameters $\lambda$ from 0.01 to 0.7. As shown in Figure 6(a), the optimum $\lambda$ was selected at 0.0385, and ALOCV error was $2.681 \times 10^{-4}$. We estimated the coefficient at each block as shown in Figure 6(b). This result is quite similar with the result described in [11], in which they provided a solution for a particular value $\lambda = 0.037$.

Figure 7. The adjacency of regions constructed by Voronoi tessellation.

Figure 7 shows adjacency of regions constructed by Voronoi tessellation. Then, we applied the generalized lasso with formula (3) to the data clustered into the regions, with using 3-fold CV, 5-fold CV, 10-fold CV and ALOCV to select $\lambda$ along to a sequence from 0.01 to 0.7. In this setting, we have $y \in \mathbb{R}^{2162 \times 1}, X = I \in \mathbb{R}^{2162 \times 50}, D \in \mathbb{R}^{133 \times 50}$, since there are 50 regions (nodes) and 133 connections (edges) between regions. Table 1 shows the selected $\lambda$ and CV errors by each method, and Figure 8 shows the plots of the estimated coefficient for $\lambda$ selected by each method.

Table 1. Selected $\lambda$ and CV error by each method in fitting the generalized lasso to Chicago crime data clustered into 50 regions

| Criteria | Methods |
|----------|---------|
|          | 3-fold CV | 5-fold CV | 10-fold CV | ALOCV |
| $\lambda$ selected | 0.0082 | 0.0011 | 0.0011 | 0.0035 |
| CV error | 0.000216 | 0.000215 | 0.000213 | 0.000271 |
Figure 8. Fitting the generalized lasso to Chicago crime data clustered into 50 regions: plots of the estimated coefficient at each region, with $\lambda$ selected by: (a) 3-fold CV; (b) 5-fold CV; (c) 10-fold CV; (d) ALOCV.

As shown in Table 1, the CV errors obtained by ALOCV is slightly greater than the CV error obtained by $k$-fold CV. Then, as shown in Figure 8, the plots of the estimated coefficient at each region for $\lambda$ selected by the four methods are quite similar. However, the estimated coefficients around the southern-east corner for the pre-clustered data are different from those by ALOCV for the original non-clustered data shown in Figure 6 (b). We suspect that clustering into 50 regions in advance might affect the results of estimation and clustering with generalized lasso. Thus, it would be more appropriate to apply the generalized lasso to the original non-clustered Chicago crime data, with using the ALOCV method for selecting $\lambda$. In conclusion, we can determine the burglaries risk clusters in Chicago based on the above result, that is, the high-risk cases occurred relatively in the middle to southern side of the city.

5. Conclusion
Based on this study, we can conclude that ALOCV can be a useful method for selecting the tuning parameter $\lambda$ in the generalized lasso for clustering with spatially varying coefficients models. Through our simulation study, ALOCV provides stable performance in selecting $\lambda$ for various conditions of sample sizes and error variances. Moreover, ALOCV provides relatively smaller CV error, and can detect edges with differences shrunk by generalized lasso appropriately. ALOCV in generalized lasso can be also applied to dataset that contains one or few observations in each node, while $k$-fold CV is not feasible. On the other hand, $k$-fold CV provides larger CV error and under-estimated number of shrunk edges for smaller $k$, while it tends to select extremely small $\lambda$ when the sample size is large for larger $k$, which leads to high variability of the estimated number of shrunk edges.

In the case study of Chicago crime data, the selection of $\lambda$ by ALOCV for original non-clustered data was reasonable, and the result was in accordance with the conclusion suggested in the preceding literature. On the other hand, clustering into regions in advance for making $k$-fold CV feasible may lead to a wrong result of clustering with a spatially varying coefficient model. We should further investigate the effect of clustering in advance in our future study.

Acknowledgement
This research is supported by JICA (Japan International Cooperation Agency).

References
[1] Tamura K, Puett R C, Hart J E, Starnes H A, Laden F, and Troped P J 2014 Spatial clustering of physical activity and obesity in relation to built environment factors among older women in three U.S. states *BMC Public Health*. 14:1322
[2] Zou B, Peng F, Wan N, Mamady K, and Wilson G J 2014 Spatial Cluster Detection of Air
Pollution Exposure Inequities across the United States. *PLoS ONE*. 9(3):1-14

[3] Padilla C M, Deguen S, Lalloue B, Blanchard O, Beaugard C, Troude F, Navier D Z, and Vieira V M 2013 Cluster analysis of social and environment inequalities of infant mortality. A spatial study in small areas revealed by local disease mapping in France. *Sci Total Environ.* 0: 433–441

[4] Gelfand A E, Kim H-J, Sirmans C, Banerjee S 2003 Spatial modeling with spatially varying coefficient processes. *J. Amer. Statist. Assoc.* 98(462), 387–396

[5] Tibshirani R J and Taylor J 2011 The solution path of the generalized lasso. *Ann. Statist*. 39(3): 1335–1371

[6] Tibshirani R 1996 Regression shrinkage and selection via the lasso. *J. R. Stat. Soc. Ser. B Stat. Methodol.*. 267–288.

[7] Tibshirani R J and Saunders M 2005 Sparsity and smoothness via the fused lasso. *J. R. Statist. Soc. B*. 67(1): 91-108

[8] Zhao Y and Bondell H 2020 Solution paths for the generalized lasso with applications to spatially varying coefficients regression. *Computational Statistics and Data Analysis* 142(106821): 1-14

[9] Rahardiantoro S and Sakamoto W 2021 Clustering Regions Based on Socio-Economic Factors Which Affected the Number of COVID-19 Cases in Java Island. *J. Phys.: Conf. Ser.* 1863 012014

[10] Chicago Police Department 2014 *City of Chicago Data Portal, Crimes—2001 to Present*, Chicago, IL: Chicago Police Department

[11] Arnold T and Tibshirani R J 2016 Efficient implementations of the generalized lasso dual path algorithm. *Journal of Computational and Graphical Statistics*. 25(1): 1–27

[12] Rad K R and Maleki A 2020 A scalable estimate of the out-of-sample prediction error via approximate leave-one-out cross-validation. *Journal of the Royal Statistical Society Series B*, 82(4), 965-996.

[13] Rad K R, Zhou W, and Maleki A 2020 Error bounds in estimating the out-of-sample prediction error using leave-one-out cross validation in high-dimensions. *Proceedings of the 23rd International Conference on Artificial Intelligence and Statistics (AISTATS)* 108

[14] Stone M 1974 Cross-validatory choice and assessment of statistical predictions. *Journal of the royal statistical society. Series B (Methodological)* 111-147

[15] Wang S, Zhou W, Maleki A, Lu H, and Mirrokni V 2018 Approximate Leave-One-Out for High-Dimensional Non-Differentiable Learning Problems. *arXiv:1810.02716v1 [cs.LG]*