Distribution Agreement

In presenting this thesis or dissertation as a partial fulfillment of the requirements for an advanced degree from Emory University, I hereby grant to Emory University and its agents the non-exclusive license to archive, make accessible, and display my thesis or dissertation in whole or in part in all forms of media, now or hereafter known, including display on the world wide web. I understand that I may select some access restrictions as part of the online submission of this thesis or dissertation. I retain all ownership rights to the copyright of the thesis or dissertation. I also retain the right to use in future works (such as articles or books) all or part of this thesis or dissertation.

Signature:

_____ _____

Kexin Qu Date

Efficient Classification for Ultra High Dimensional Variable Selection

By

Kexin Qu

Master of Science of Public Health

Emory University

Rollins School of Public Health

Department of Biostatistics and Bioinformatics

Jian Kang
Committee Chair

Tianwei Yu
Committee Member

Efficient Classification for Ultra High Dimensional Variable Selection

By

Kexin Qu

B.S., Emory University, 2014

MSPH, Emory University

Rollins School of Public Health

2015

Thesis Committee Chair: Jian Kang, PhD

An abstract of

A thesis submitted to the Faculty of the

Rollins School of Public Health of Emory University

in partial fulfillment of the requirements for the degree of

Master of Science in Public Health

in Department of Biostatistics and Bioinformatics

2015

Abstract

Efficient Classification for Ultra High Dimensional Variable Selection

By Kexin Qu

Rapid advances in technologies have demonstrated great needs for ultra-high dimensional data analysis in neuroimaging studies. Our work is motivated by the Autism Brain Imaging Data Exchange (ABIDE) study, where scientist are interested to identify important biomarkers for early detection of the autism spectrum disorder (ASD) using high resolution brain images that include hundreds of thousands voxels. However, most existing methods are not feasible to deal with such problems due to extensive computational cost coming as well model complexity. In our work, we propose a new spatial variable selection screening (SVSS) method which includes two components: 1) independent screening using each voxel as a predicator and 2) search for other predicators among neighbors based on spatial dependence. Our approach is computationally feasible and efficient; and it takes full advantage of using spatial configuration of the predicators without additional effort on building complex models. Applied to the resting state functional magnetic resonance imaging (R-fMRI) data in the ABIDE study, our methods identify voxel-level imaging biomarkers highly predictive of the ASD. Extensive simulations also show that our method achieve better performance in predication as well as variable selection compared to the widely used SIS method.

Efficient Classification for Ultra High Dimensional Variable Selection

By

Kexin Qu

B.S., Emory University, 2014

MSPH, Emory University

Rollins School of Public Health

2015

Thesis Committee Chair: Jian Kang, PhD

A thesis submitted to the Faculty of the

Rollins School of Public Health of Emory University

in partial fulfillment of the requirements for the degree of

Master of Science in Public Health

in Department of Biostatistics and Bioinformatics

2015

Acknowledgements

I want to thank the faculty, advisors, and staff of the Biostatistics Department at Rollins School of Public Health for the dynamic two years of learning that I have had. This thesis is only a sample of the vast knowledge that was attained and applied through my two years here at Rollins. I would especially like to thank Dr.Jian Kang for all of his advice and support to help me write this thesis. Also a special thanks to Dr. Tianwei Yu for taking the time to read my thesis. Lastly I want to give a special thanks to my parents for their loving support and encouragement that motivated me to pursue a degree at the graduate level.

Contents

1 Introduction

Rapid advances in technologies have demonstrated great needs for high dimensional data analysis in diverse fields, ranging from genomics, health sciences, to economics, and machine learning. Particularly in neuro-imaging study, the emergence of a large amount of high resolution imaging data with ultra high dimensionality requires efficient statistical methods to obtain accurate disease information.

In dealing with any statistical procedure, there are three major components: statistical accuracy, model interpretability and computational complexity (Fan and Lv, 2010). None of them need to be sacrificed if number of observations n is much larger than number of variables p. However, in ultra-high dimension problems where dimensionality is much larger than the number of observation size, significant challenges arise in terms of how to design a computationally efficient statistical procedure, build interpretable models, as well as obtain accurate inference.

A number of methods have been proposed. Regularization methods (Tibshirani 1996, Fan and Li 2001, Zou and Hastie 2005, Zou 2006, Yuan and Lin 2006) for variable selection can deal with high-dimensional feature space problems with structural information being incorporated. Bayesian methods such as Gibbs variable selection (Dellaportas et al. 2002) and stochastic search variable selection are also commonly used for variable selection and posterior simulation algorithms. The advantage of such methods is that, by specifying a positive prior probability for each parameter being zero, the posterior probability of each parameter being included in the model can be computed and used to quantify the uncertainty of variable selection. Based on the posterior inclusion probabilities, important variables can be obtained by specifying a threshold value. Structural information can be incorporated by using Ising or binary Markov random filed priors (Li and Zhang 2010, Stingo et

al. 2011, Smith et al. 2003, Smith and Fahrmeir 2007). Transdimensional sampling algorithms (Lamnisos et al. 2009) and adaptive Monte Carlo methods (Nott and Kohn 2005) are developed to improve efficiency of posterior simulations. Despite good performance in high dimensional space, the aforementioned methods are not feasible to handle problems involving hundreds of thousands or even millions of predictors. Therefore variable selection in ultra-high dimension feature space calls for extended statistical methods and theory. Sure independence screening (SIS), introduced by Fan and Lv (2008), is proposed to reduce computation in ultra-high dimensional variable selection and at the same time can achieve good theoretical properties ; however it can neither explicitly model the dependence among variables nor quantify the uncertainty of variable selection. Bottolo and Richardson (2010) proposed a sampling scheme in Bayesian modeling framework which allows 10,000 predicators, but it is infeasible when predicators are in number of 200,000 as in our motivating study. Johnson and Rossell (2012) proposed a Bayesian model selection method that can produce high selection accuracy in ultra high-dimensional problems but it fails to incorporate any structural information and thus can't be directly applied to the neruo-imaging problem.

Therefore there are needs to develop more efficient computation algorithms that can both utilize less complex models and incorporate structural information. We propose a new spatial variable selection screening (SVSS) which includes two components: 1) independent screening using each voxel as a predicator and 2) search for other predicators among neighbors based on spatial dependence. These are motivated by Sure Independent Screening (Fan and Lv, 2007) and EgoNet (Yang et al, 2014), respectively. Comparing to SIS which utilizes marginal sample correlation with the response variables, the initial screening in our method is guided by predication/classification accuracy coming from a simple generalized

linear model (GLM). This not only ensures goodness of model fit but also greatly reduces model complexity as well as computational cost. Due to the sparsity feature of ultra high dimensional problems (i.e., only a small subset of voxels are the true important predicators), only one voxel (or several if in ultra-high dimension space) with the best predication performance is selected and considered a starting point to search for other important variables; the other important ones are then recruited based on spatial dependence with the selected ones. Without introducing new parameters, we adopt in a similar fashion of the ego network model from Yang's study (2014): for each selected variable, the nearest 26 neighbors directly connected to it are altogether added to the model. The underlying assumption of our method is that if one voxel is selected, then its neighbors are likely to be included in the model. We fit GLM to assess the association between the voxels and the disease status; less significant voxels are indicated by lower coefficient values and are removed from the model to avoid overfitting.

Our work is primarily motivated by the Autism Brain Imaging Data Exchange (ABIDE) study (Di Martino et al. 2013). ABIDE study aims to find the association of brain activity with the autism spectrum disorder (ASD), a widely recognized disease due to its high prevalence and heterogeneity in Children (Rice 2009). The ABIDE study aggregated 20 resting-state functional magnetic resonance imaging (R-fMRI) data sets from 17 different sites including 539 ASDs and 573 age-matched typical controls. To characterize the local spontaneous brain activity, we plan to focus on the fractional amplitude of low-frequency fluctuations (fALFF)(Zou et al. 2008) based on the R-fMRI time series at each voxel for each subject. The fALFF is defined as the ratio of the power spectrum of low frequency (0.01-0.08 Hz) to the entire frequency range and has been widely used as a voxel-wise measure of the intrinsic functional brain architecture derived from the R-fMRI data (Zou

et al.2010). In our work, among 116 regions in the brain involving 185, 405 voxels, we focus on one particular region with 5104 voxels. We apply SVSS method to analyze the voxel-wise fALLF values and aim to identify imaging biomarkers in this particular region for ASD detection.

The remainder of the paper is organized as follows. In section 2, we present the algorithm for variable selection. In section 3, several simulation studies with different settings are presented to demonstrate the superiority of our proposed method. In section 4, we apply the proposed method to R-fMRI data in the ABIDE study to identify important voxel-level fALFF biomarkers that are predictive of the ASD risk. Finally we conclude with a discussion in Section 4.

2 Method

The underlying model adopted for linking disease status y, and predicators X_1,...,X_p, is logistic regression. Suppose there are n subjects in the study. Response variable $y_i \in \{0, 1\}$, is binary outcome indicating disease status of subject i (disease=1, control=0). The whole brain area contains a total number of V voxels. Let x_{iv} denote the imaging biomarker at voxel v for subject i. We consider a logistic regression model for variable selection $y_i = I\left[p_i \geq 0.5\right]$,

$$\log(\frac{p_i}{1-p_i}) = \sum_{v=1}^{V} c_v \beta_v x_{i,v} + \epsilon_i \quad \epsilon_i \sim N(0,1) \tag{1}$$

where indicator function $I(A) = 1$ if event A happens and 0 if not, β_v are coefficients of imaging biomarker $x_{i,v}$, $c_v \in \{0, 1\}$ is the selection indicator for voxel v. Therefore, a value of 0 indicates the corresponding voxel is not accosiated with the certain disease.

2.1 Step1: Screening

In the first step, only one voxel v at a time is introduced into the model with the corresponding data (x_{iv}, y_i),

$$y_i = I\left[p_i \geq 0.5\right], \quad \log(\frac{p_i}{1-p_i}) = \beta_v x_{i,v} + \epsilon_i \quad \epsilon_i \sim N(0,1) \tag{2}$$

We conduct one round of cross-validation by splitting the entire data into training and testing. By using each of the V fitted model obtained from the training data, we can make predications on disease status for each subject in the testing set and thus obtain the predication accuracy a_v given by each voxel,

$$a_v = \frac{\sum_{i=1}^{n} I(y_i = \widehat{y_{iv}})}{n} \tag{3}$$

where $\hat{y_{iv}}$ is the predicated disease status; I ($y_i = \widehat{y_{iv}}$) is 1 if $y_i = \widehat{y_{iv}}$, 0 otherwise. Based on ranking of a_v, select the voxel(s),v_{max} which generates the highest predication accuracy $a_{max} = \max\{a_1, ..., a_v\}$. We consider voxel v_{max} as the one bearing the most significant feature for disease predication and therefore make it the starting point to search for other important variables.

2.2 Step2: Variable Selection Incorporating Spatial Dependence

Denote the set of h selected voxel(s) as S_k where $k(k \geq 1)$ denotes the corresponding kth model, M_k, ($S_1 = \{v_{max}\}$) . The neighborhood of each voxel is defined as the set of adjacent voxels from six different directions (top, bottom, front, back, left, and right)(Figure 1). For each selected voxel in this model, recruit all of the 26 neighbors and fit into the logistic regression model, M_{k+1}. If the corresponding predication accuracy a_{k+1} decrease by more

than 0.01 from a_k, the search stops and we consider M_k as the optimal model where the set of voxels S_k are identified as important features associated with the disease; otherwise we continue to the next step to exclude variables with less significant association with the disease:

from the fitted model M_{k+1}, remove from the current model the voxel,$v_{k+1,r}$ ($r = 1, ..., h$), which has the least absolute value of the coefficient *beta* since smaller values indicate less association. Fit into the logistic regression model with the reduced set of variables, $S_{k+1,r+1} = S_{k+1,r} - \{v_{k+1,r}\}$. If the resulting accuracy $v_{k+1,r+1}$ decreases by more than 0.01 from $v_{k+1,r}$, then stop the remove step; if not, repeat until this criteria is satisfied or predication accuracy reaches 1. The resulting set of variables and the corresponding predication accuracy are the new S_{k+1} and a_{k+1}, respectively; these voxels are thus used as the root to recruit more neighbors in the next step. We choose " decrease by more than 0.01" instead of simply "decrease" to be signal indicating a less optimal model. As 0.01 being a small value, it won't significantly affect the model performance. One the other hand, it can help to exclude less important variables as many as possible and thus avoid overfitting and increase computation performance.

When there are multiple regions that contain true signals, some modifications are needed. In such cases, the entire space is partitioned into several sub-regions with equal number of voxels. Within each sub-region, the aforementioned screening step is implemented. Voxel(s) with highest accuracies from each sub-region are combined to one set to start the regular subsequent SVSS step.

3 Simulation Studies

We conduct simulation studies to evaluate the variable selection performance of the proposed method compared to SIS method.

To understand of the method performance at different scales of variable space, we focus on three cases: 1) a 25*25*25 cubic region with 15,625 voxels in total, 2) a 37*37*37 cubic region with 50,653 voxels, and 3) a 50*50*50 cubic region with 125,000 voxels. Sample size is also varied in order to find an acceptable one. Imaging biomarkers $\{x_{iv}\}_{v=1}^{V}$ are independently drawn from normal distribution $N(0, 2)$. We further set 125 voxels within a small cube region (Figure 2) to be true signals. The coefficients of true signals are set to be constant 1. Response variable is thus drawn from:

$$p(y_i = 1) = \frac{1}{1 + e^{-\beta x_{iv}}} \tag{4}$$

Table 1 presents the variable selection performance under different space dimension. Compared to SIS, our method has shown a obviously better performance with higher predication accuracy, sensitivity as well as specifity. Remarkably, in most cases our method has successfully covered all of 125 true voxels without including any of the non-important ones. When sample size is small ($N = 2000$), SIS has not given a good performance: predication accuracy is $0.65(V = 15,625)$ and $0.552(V = 50,653)$, respectively. Out of 125 true voxels, there are only $41(V = 15,625)$ and $11(V = 15,625)$ voxels successfully identified by SIS. Therefore our method has demonstrated a superiority over SIS particularly when sample size is small.

We further conduct a simulation study by setting two separate regions within the feature space to be true signals (Figure 3). Sample size N is 5000 and there are 15,625 voxels (64 true

voxels in each true region) in the space. The entire space is partitioned into 8 subregions involving 2000 voxels individually. The voxel with the highest predication accuracy in each subregion is selected at the screening step. Thus 8 voxels are involved as roots to recruit neighbors. SVSS has shown a predication accuracy as high as 0.984 (Table 2) and it successfully identified 128 true voxels without including any of the non-important ones. In comparison, SIS shows a predication accuracy of 0.799 and only 84 true voxels are covered by the total 93 voxels.

4 Application

We analyze the motivating ABIDE study introduced in Section 1 using the SVSS procedure. Our goal is to identify important voxel-wise image biomarkers that are predictive of ASD risk. Our analysis include 1071 subjects and for each subject fALFF values are computed for each of 5104 voxels . fALFF values are first standardized in order for GLM to work appropriately.

In the screening step, for each of 5104 voxels we implement GLM using the entire data as training set, and then randomly select 500 samples as testing set to obtain the predication accuracy. Top 10 voxels with the highest accuracies are candidates to incorporate spatial dependence in the next step. We conduct 10 rounds of testing and thus can correspondingly obtain 10 sets of best-performed voxels in total. By counting the number of occurrence across these 10 sets, we find the voxel with the highest occurrence frequency (Figure 4) which is selected in 5 rounds.

The selection results from SVSS procedure is presented in Table 3. Using a testing set with 500 randomly selected samples, SVSS achieves a considerably high predication accuracy (1), sensitivity (1) as well specificity(1), indicating a strong predictive power of

our method. 530 voxels, located at the upper right of the entire region, are selected in the final mode (Figure 5).

5 Discussion

In this work, we present a novel algorithm for variable selection in an ultra-high dimensional feature space. Our approach is computationally feasible and efficient; and it takes full advantage of using spatial configuration of the predicators without additional effort on building complex models; the variable selection is guided by the prediction/classification accuracy, which ensures goodness of model fit, reduces the model complexity and avoids the model overfitting. Furthermore, our method builds up a general framework for ultra-high dimensional variable selection; it can be readily extended to incorporate other existing statistical models/machine learning algorithms beyond the generalized linear model (GLM), such as the support vector machine (SVM), random forest and neural network.

The key components in our method are the set of selected voxels obtained from the screening step, which serve as the roots to find other adjacent voxels. It is extremely important to have the initially selected voxels be one of the true voxels; otherwise it will be time-consuming and computationally infeasible to find the true voxels. Since we only rely on spatial dependence and don't incorporate other parameters, the algorithm may fail to solve the problem if the initial voxel is far from the true region. This may happen when true signals are sparse relative to the voxel size or when there are multiple regions of interest. One solution, when sparsity occurs in ultra-high dimensional variable space, is to increase the number of initial candidates as to cover as least one of the true signals. In our simulation studies, when variable number is relatively lower ($V = 15625, 50653$), the first selected variable is always one of the true signals; when the variable number is in hundreds

of thousands ($V = 125,000$), however, it's essential to include more than one voxel (in our case, we selected 5 best-performed voxels obtained in the screening). It's a challenge to decide how many voxels are to be selected initially: on one hand a larger number indicates high possibility of covering at least one of the true signals while on the other hand a model involving too many voxels is not computationally feasible and induces overfitting. Another solution might be to partition the large space into subregions so that variable dimension is reduced. For example, in the last simulation study we partitioned the 15625 voxels into 8 subregions where each one contains approximately 2000 voxels. This method works well especially when there are multiple true regions in the entire space.

References

Bottolo, L. and Richardson, S. (2010), "Evolutionary stochastic search for Bayesian model exploration," *Bayesian Analysis*, **5**, 583–618.

Bowman, F. D., Zhang, L., Derado, G., and Chen, S. (2012), "Determining functional con- nectivity using fMRI data with diffusion-based anatomical weighting," *NeuroImage*, **62**, 1769–1779.

Dellaportas, P., Forster, J. J., and Ntzoufras, I. (2002), "On Bayesian model and variable selection using MCMC," *Statistics and Computing*,**12**, 27–36.

Di Martino, A., Yan, C., Li, Q., Denio, E., Castellanos, F., Alaerts, K., Anderson, J., Assaf, M., Bookheimer, S., Dapretto, M., et al. (2013), "The autism brain imaging data exchange: towards a large-scale evaluation of the intrinsic brain architecture in autism," *Molecular psychiatry*.

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

Fan, J. and Lv, J. (2008), "Sure independence screening for ultrahigh dimensional feature space," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*,**70**, 849–911.

Fan, J, and Lv. J (2010) "A selective overview of variable selection in high dimensional feature space." *Statistica Sinica 20.1*, **101**.

Fan, J. and Song, R. (2010), "Sure independence screening in generalized linear models with NP-dimensionality," *The Annals of Statistics*, **38**, 3567–3604.

Johnson, V. E. and Rossell, D. (2012), "Bayesian Model Selection in High-Dimensional Set- tings," *Journal of the American Statistical Association* , **107**, 649–660.

Lamnisos, D., Griffin, J. E., and Steel, M. F. (2009), "Transdimensional sampling algorithms for Bayesian variable selection in classification problems with many more variables than observations," *Journal of Computational and Graphical Statistics*, **18**, 592–612.

Li, F. and Zhang, N. (2010), "Bayesian variable selection in structured high-dimensional covariate spaces with applications in genomics," *Journal of the American Statistical Asso ciation*, **105**, 1202–1214.

Nott, D. J. and Kohn, R. (2005), "Adaptive sampling for Bayesian variable selection," *Biometrika*, **92**, 747–763.

Rice, C. (2009), "Prevalence of Autism Spectrum Disorders: Autism and Developmental Disabilities Monitoring Network, United States, 2006. Morbidity and Mortality Weekly Report. *Surveillance Summaries*. Volume 58, Number SS-10." Centers for Disease Control and Prevention.

Smith, M. and Fahrmeir, L. (2007), "Spatial Bayesian variable selection with application to functional magnetic resonance imaging," *Journal of the American Statistical Association*, **102**, 417–431.

Smith, M., Putz, B., Auer, D., and Fahrmeir, L. (2003), "Assessing brain activity through spatial Bayesian variable selection," *Neuroimage*, **20**, 802–815.

Stingo, F., Chen, Y., Tadesse, M., and Vannucci, M. (2011), "Incorporating biological infor- mation into linear models: a Bayesian approach to the selection of pathways and genes," *The Annals of Applied Statistics*, **5**, 1978–2002.

Tibshirani, R. (1996), "Regression shrinkage and selection via the lasso," *Journal of the Royal Statistical Society* . Series B (Methodological), 267–288.

Yang, Rendong, et al. "EgoNet: identification of human disease ego-network modules." *BMC genomics* 15.1 (2014): 314.

Yuan, M. and Lin, Y. (2006), "Model selection and estimation in regression with grouped variables," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **68**, 49–67.

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

Zou, H. and Hastie, T. (2005), "Regularization and variable selection via the elastic net," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **67**, 301–320.

Zou, Q.-H., Zhu, C.-Z., Yang, Y., Zuo, X.-N., Long, X.-Y., Cao, Q.-J., Wang, Y.-F., and Zang, Y.-F. (2008), "An improved approach to detection of amplitude of low-frequency fluctuation (ALFF) for resting-state fMRI: fractional ALFF," *Journal of neuroscience methods*, **172**, 137–141.

Zuo, X.-N., Di Martino, A., Kelly, C., Shehzad, Z. E., Gee, D. G., Klein, D. F., Castellanos, F. X., Biswal, B. B., and Milham, M. P. (2010), "The oscillating brain: complex and reliable," *Neuroimage*, **49**, 1432–1445.

Appendices

Tables

V	N	Method	Pedication Accuracy	Sensitivity	Specifity	NO_{voxel}(NO_{true})
15,625	2000	SVSS	0.942	0.935	0.950	123(123)
		SIS	0.65	0.616	0.685	54(41)
15,625	5000	SVSS	0.982	0.978	0.986	125(125)
		SIS	0.862	0.863	0.857	120(108)
50,653	2000	SVSS	0.891	0.873	0.904	176(124)
		SIS	0.552	0.593	0.514	36(11)
50,653	3000	SVSS	0.973	0.931	0.949	125(125)
		SIS	0.710	0.689	0.723	77(54)
50,653	5000	SVSS	0.981	0.974	0.988	125(125)
		SIS	0.856	0.855	0.857	120(110)
125,000	5000	SVSS	0.99	0.986	0.994	125(125)
		SIS	0.844	0.855	0.833	120(101)

Table 1: Variable selection performance focusing on one region. NO_{voxel} is the total number of selected voxels. NO_{true} is the number of true voxels covered by the selected ones in total. Sensitivity and specifity are obtained by using 80% of entire samples for training and the rest for testing.

V	N	Method	Pedication Accuracy	Sensitivity	Specifity	NO_{voxel}(NO_{true})
15625	5000	SVSS	0.984	0.983	0.984	128(128)
		SIS	0.799	0.799	0.798	93(84)

Table 2: Variable selection performance focusing on two regions. NO_{voxel} is the total number of selected voxels. NO_{true} is the number of true voxels covered by the selected ones in total. Sensitivity and specifity are obtained by using randomly selected 4000 samples for traning and 1000 for testing.

V	N	Method	Pedication Accuracy	Sensitivity	Specifity	NO_{voxel}
5104	1071	SVSS	1	1	1	530

Table 3: Selction results and predication accuracy for the ASD risk.NO_{voxel} is the total number of selected voxels. Sensitivity and specifity are obtained from 500 radomly selected voxles.

Figures

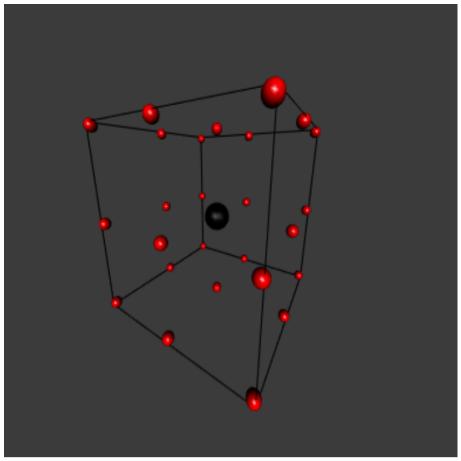

Figure 1: 26 adjacent neighbors around one voxel.

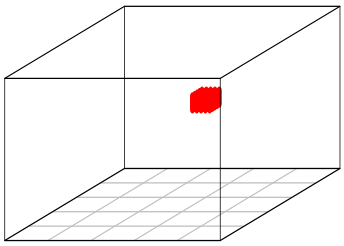

Figure 2: 125 true voxels (red) within one region are located in a general variale space

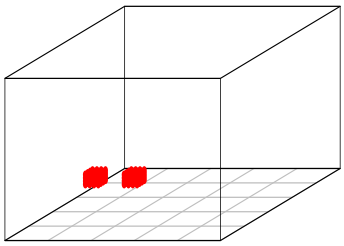

Figure 3: 128 true voxels are located within two separate regions in a general variable space

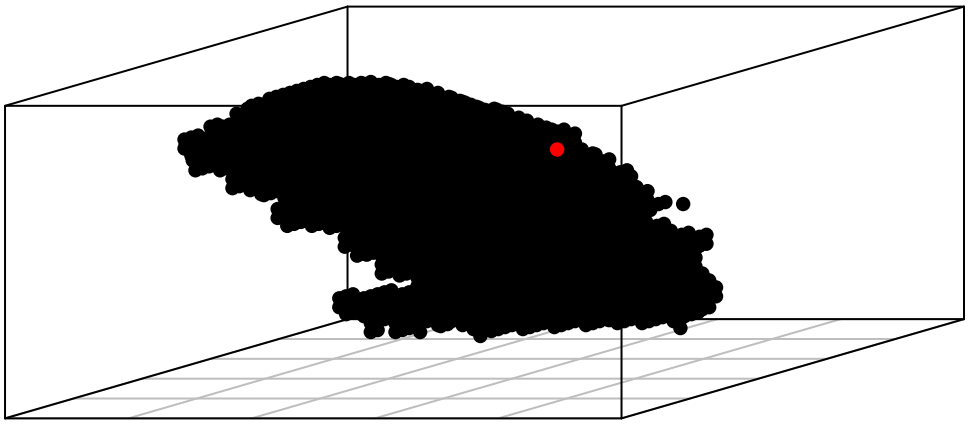

Figure 4: The selected voxel(red) is located on the right of the 5104 voxels (black).

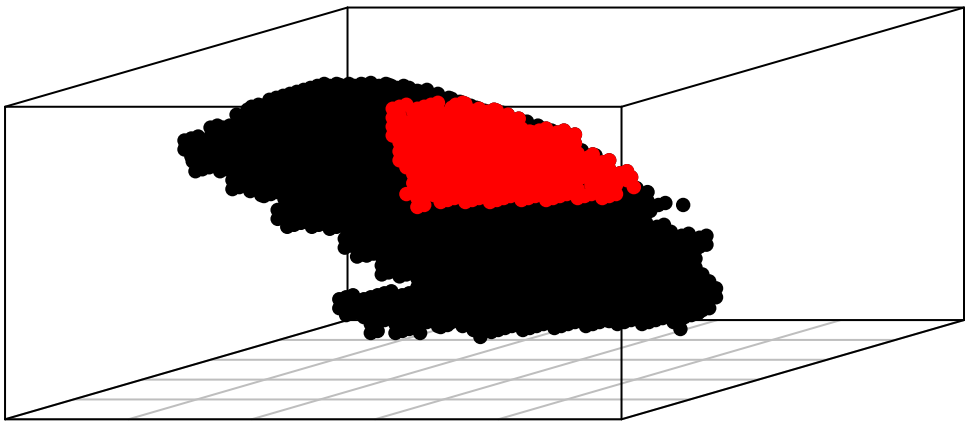

Figure 5: The selected 530 voxels (red) is located on the upper right of the 5104 voxels (black).