Published in final edited form as:

Stat Med. 2018 November 20; 37(26): 3764–3775. doi:10.1002/sim.7832.

Genetic risk prediction using a spatial autoregressive model with adaptive lasso

Yalu Wen¹, Xiaoxi Shen², and Qing Lu²

¹Department of Statistics, University of Auckland, Auckland, New Zealand ²Department of Epidemiology and Biostatistics, Michigan State University, East Lansing, Michigan, USA

Abstract

With rapidly evolving high-throughput technologies, studies are being initiated to accelerate the process toward precision medicine. The collection of the vast amounts of sequencing data provides us with great opportunities to systematically study the role of a deep catalog of sequencing variants in risk prediction. Nevertheless, the massive amount of noise signals and low frequencies of rare variants in sequencing data pose great analytical challenges on risk prediction modeling. Motivated by the development in spatial statistics, we propose a spatial autoregressive model with adaptive lasso (SARAL) for risk prediction modeling using high-dimensional sequencing data. The SARAL is a set-based approach, and thus, it reduces the data dimension and accumulates genetic effects within a single-nucleotide variant (SNV) set. Moreover, it allows different SNV sets having various magnitudes and directions of effect sizes, which reflects the nature of complex diseases. With the adaptive lasso implemented, SARAL can shrink the effects of noise SNV sets to be zero and, thus, further improve prediction accuracy. Through simulation studies, we demonstrate that, overall, SARAL is comparable to, if not better than, the genomic best linear unbiased prediction method. The method is further illustrated by an application to the sequencing data from the Alzheimer's Disease Neuroimaging Initiative.

Keywords

adaptive lasso; high-dimensional sequencing data; risk prediction; spatial autoregressive model

1 | INTRODUCTION

With rapidly evolving high-throughput technologies and ever-decreasing cost, studies have been initiated to accelerate the process toward precision medicine.¹ A large number of genetic variants have been revealed to be associated with complex diseases through large-scale genomic studies, but they only explain a small proportion of heritability.² Complex

Correspondence: Yalu Wen, Department of Statistics, University of Auckland, Auckland, New Zealand. y.wen@auckland.ac.nz, Qing Lu, Department of Epidemiology and Biostatistics, Michigan State University, East Lansing, Michigan, USA. qlu@epi.msu.edu.

CONFLICT OF INTEREST

The authors declare no potential conflict of interests.

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section at the end of the article.

diseases are very likely to be caused by the joint effect of genetic variants through complex biological pathways. Therefore, the joint analysis of sets of genetic variants can aggregate signals from multiple variants, including rare variants, for risk prediction modeling.^{3,4} The joint analysis of single-nucleotide variant (SNV) sets can further enhance the detection of disease-related SNV sets and, thus, improve prediction accuracy. While the conventional methods, such as logistic and simple linear regression models, can be used to simultaneously estimate the effects of each individual predictor given a moderate number of predictors, it is impractical to model high-dimensional genomic data partially due to the exponentially increased number of parameters and the potential multicollinearity problem.

Several statistical methods, such as the kernel machine-based (known as SKAT) and random field-based methods, have been proposed for the joint association analysis of SNVs within a set and have achieved various levels of successes.^{5–7} These methods jointly estimate the effects of multiple genetic variants within an SNV set and then test for their statistical significance. The basic idea of these methods is that if an SNV set is associated with the disease, then the genetic similarity will lead to phenotypic similarity.^{5–7} This is also the intuition of our proposed method. As the SNV set that includes multiple SNVs is the evaluation unit, the number of tests can be remarkably reduced. Using appropriate kernel functions, these methods can also account for linkage disequilibrium and potential high-order interactions, which may further improve their performance.⁷ Despite the advantages, these methods are designed for association studies, and they can only consider one SNV set at a time, making it not applicable to prediction studies. Moreover, the goal of a prediction study is, in general, different from association analyses. Validity and robustness are the two key concerns for a prediction model, and a causal interpretation is not absolutely necessary for a good predictive model.⁸ Therefore, a model that is capable of jointly estimating the predictive effects of all SNV sets and achieves high and robust prediction accuracy is critically important for risk prediction research.

Best linear unbiased prediction (BLUP)-based methods have been proposed for prediction analyses, and the basic premise for these methods is that instead of estimating the effects of each individual genetic variant, they attempt to estimate their cumulative effects.^{9–13} The correlation between genomic risk scores would be higher for those that are more genetically similar. The genetic similarity of individuals can be estimated based on either known pedigree information^{9,10} or genomic data.^{11,12} One of the implicit assumptions for most of these methods is that all genetic variants, regardless of their functions and locations, have a similar effect-size distribution that is too simple and unrealistic.¹⁴ As noted by Byrnes et al, variable selection algorithms can substantially improve the performance of a risk prediction model in the absence of good biological annotations. Allowing different SNV sets to contribute differently to the disease risk can also better reflect the underlying disease mechanisms and potentially enhance the model's prediction accuracy.¹⁵

In this paper, we propose a simultaneous autoregressive model with adaptive lasso (SARAL) for risk prediction modeling based on high-dimensional genomic data. The SARAL is motivated by the idea that genetic similarities will lead to phenotypic similarity if the SNV sets are predictive. Our perspective is very distinct from the BLUP-based methods in the sense that SARAL regresses the outcome of each individual on all the others and makes

predictions at unobserved values in the same genetic space based on the values from nearby locations. The proximity is determined by the disease-relevant SNV sets in the genetic space, and SARAL attenuates the value with decreasing proximity from the observed location. With the adaptive lasso implemented,¹⁶ SARAL can enjoy many appealing features of adaptive lasso and achieve comparable or better performance as existing methods, as demonstrated through simulation studies and a real data application.

Simultaneous autoregressive (SAR) models have long been used for mining geospatial and economic data, where prediction is one of major research interests in many applications.^{17–19} Many statistical properties of SAR models have been extensively studied.^{20–26} For example, many authors addressed the prediction problem in the context of spatial lag models.^{22,23,25} Recently, Thomas-Agnan et al have systematically investigated the prediction problem using spatial autoregressive models.²⁶ Kato explored the best linear prediction problem in the framework of spatial error models.²⁴ Kelejian and Prucha noticed that the efficiency of predictors is determined by the information set on which the predictors are based and that the predictors are efficient if the information set is large.²¹ The information set quantified by the relationship between spatial errors can improve the prediction accuracy, and it contains information on the relationship between the spatial errors for both unobserved values and values in the estimation sample. The successful use of SAR models in geospatial/economic data can be naturally extended to genomic data, where we can construct a genomic space based on high-dimensional genetic data and then apply the SAR model to form a risk prediction model. Different from the existing standard SAR models, we incorporate the adaptive lasso into the SAR model for the selection of autoregressive coefficients, which is important for genomic studies with a large number of nonpredictive genetic variants. In the following sections, we first lay out the details of the proposed methods and then compare its accuracy with genomic BLUP.^{10,27} We further illustrate the SARAL method through an application to a whole-genome sequencing data set from the Alzheimer's Disease Neuroimaging Initiative (ADNI).²⁸

2 | METHODS

SAR models have long been used to model geospatial and ecological data. Spatial autocorrelation is a common phenomenon in these data because observations from nearby locations tend to be more similar than those would be expected on a random basis. Our SARAL model is motivated by a similar idea that adjacent points in the genetic space share similar outcomes, where adjacencies between subjects in the genetic space are gauged by predictive genetic variants. Quite similar to the SAR model, we propose to directly model the outcome of an individual as a linear function of all the others' outcomes and the relative contribution of each outcome is determined by genetic similarities of the relevant genetic variants. By adopting a spatial autoregressive model framework, we transform millions of SNVs into genetic similarities of predictive variants on the genetic space, and thus reduce the dimension of the original data substantially.

2.1 | The spatial autoregressive model

Consider a study of n subjects and a total of L SNVs located on K SNV sets (eg, genes), where each set has m_k SNVs ($\sum_k m_k = L$). For simplicity, we treat the SNV sets as genes, but it can be other functional units (eg, regulatory DNA regions). Throughout the manuscript, we use genes and SNV sets interchangeably. For subject i , let $G_i^k = (g_{i,1}^k, g_{i,2}^k, \dots, g_{i,m_k}^k)$ denote the genotype of m_k variants located on the k th gene, where $1 \leq k \leq K$ and $g_{i,j}^k \in \{0, 1, 2\}$. We use Y_i to denote the outcome and X_i to denote the other covariates, such as demographic and environmental factors.

If the k th SNV set is predictive, then the similarity in the k th SNV set will lead to the similarity in the outcomes. Based on this idea, a subject's outcome is modeled as a linear function of the other subjects' outcomes weighted by the genetic similarities of each predictive SNV set. The model can be written as

$$Y_i = X_i^T \beta + \sum_k \gamma_k \sum_{j \neq i} s_{i,j}^k (Y_j - X_j^T \beta) + \epsilon_i, \quad (1)$$

where β is the regression coefficients for covariates, γ_k is the predictive effect of the k th set on the outcome, and $s_{i,j}^k$ measures the genetic similarity of set k between subjects i and j . $s_{i,j}^k$ can be defined by a prespecified function, such as a weighted Euclidean distance function (ie, $s_{i,j}^k = \sqrt{\sum_{l=1}^{m_k} \omega_l^k (g_{i,l}^k - g_{j,l}^k)^2}$), where ω_l^k is a weight defined based on the minor allele frequency of variant l on SNV set k to capture the effects of rare variants.

For simplicity, we further denote $\mathbf{Y} = (Y_1, Y_2, \dots, Y_n)^T$, $\mathbf{X} = (X_1, X_2, \dots, X_n)^T$, and $\boldsymbol{\epsilon} = (\epsilon_1, \epsilon_2, \dots, \epsilon_n)^T$, and Equation (1) can be written in a matrix format, as follows:

$$\mathbf{Y} = \sum_k \gamma_k \mathbf{S}_k (\mathbf{Y} - \mathbf{X}\beta) + \mathbf{X}\beta + \boldsymbol{\epsilon}, \quad (2)$$

where $\mathbf{S}_k$ is an $n \times n$ similarity matrix with zeros on the diagonal and $s_{i,j}^k$ on the i th row and j th column. If we assume $\boldsymbol{\epsilon} \sim N(\mathbf{0}, \sigma^2 \mathbf{I}_n)$ and $\mathbf{V}^T \mathbf{V}$ with $\mathbf{V} = (\mathbf{I}_n - \sum_k \gamma_k \mathbf{S}_k)$ is invertible, then the marginal distribution of $\mathbf{Y}$ is $\mathbf{Y} \sim N(\mathbf{X}\beta, (\mathbf{V}^T \mathbf{V})^{-1} \sigma^2)$, and the corresponding parameters can be estimated by maximizing the log-likelihood function (denoted by $l(\boldsymbol{\eta})$ with $\boldsymbol{\eta} = (\beta^T, \boldsymbol{\gamma}^T)^T$ and $\boldsymbol{\gamma} = (\gamma_1, \gamma_2, \dots, \gamma_K)^T$).

Suppose we want to predict outcomes for n' individuals. Let $\mathbf{V}_{n+n'} = (\mathbf{I}_{n+n'} - \sum_k \gamma_k \mathbf{S}_k^{n+n'})$ and $\Sigma_{n+n'} = (\mathbf{V}_{n+n'}^T \mathbf{V}_{n+n'})^{-1}$. We can partition the similarity matrix $\mathbf{S}_k^{n+n'}$ as follows:

$$S_k^{n+n'} = \begin{pmatrix} S_k^{nn} & S_k^{nn'} \\ S_k^{n'n} & S_k^{n'n'} \end{pmatrix},$$

where S_k^{nn} is the $n \times n$ similarity of set k in all the training samples, $S_k^{nn'} = (S_k^{n'n})^T$ is the $n \times n'$ similarity of set k between the training samples and the testing samples, and $S_k^{n'n'}$ is the $n' \times n'$ similarity of set k between testing samples. Similarly, we can also partition $\Sigma_{n+n'}$ as follows:

$$\Sigma_{n+n'} = \begin{pmatrix} \Sigma^{nn} & \Sigma^{nn'} \\ \Sigma^{n'n} & \Sigma^{n'n'} \end{pmatrix}.$$

Given the estimates of β and γ , the predicted values of new subjects' outcomes can be estimated using the Goldberger formula, ie, $E(\hat{Y}_{n'} | Y_n) = X_n \hat{\beta} + \hat{\Sigma}^{n'n} (\hat{\Sigma}^{nn})^{-1} (Y_n - X_n \hat{\beta})$.

2.2 | Selection with adaptive lasso

2.2.1 | Penalized likelihood function—For most complex diseases, the genetic causes are not known in advance. Therefore, it is quite likely that a significant proportion of SNV sets included in the analyses are not disease related. This is especially true for whole-genome sequencing studies. Variable selection and feature extraction are crucial for predictive modeling with high dimensionality. A natural choice for the selection of SNV sets (ie, γ_k) from model (2) is to utilize the idea of adaptive lasso.¹⁶ The corresponding penalized log-likelihood function is $l_p(\eta) = l(\eta) - \sum_k^K \lambda_k |\gamma_k|$, where $\lambda_k = \frac{\lambda \log(n)}{|\gamma_k^0|}$ is the

regularization parameter and γ_k^0 is the maximum likelihood estimator for $\ell(\eta)$. Maximizing $l_p(\eta)$ will enable variable selection and parameter estimation simultaneously, as the effects of less important factors will be shrunk to zeros under the L_1 -penalty. The regularization parameters (λ_k) are allowed to vary with the genetic effects, which is similar to the idea of adaptive lasso. It is worth noting that the penalties are only put on the autoregressive parameters (ie, γ_k) as the focus is on the selection of predictive SNV sets.

Although adaptive lasso has been used widely for variable selection in generalized linear regression models, it has been rarely used in spatial models for the selection of autoregressive coefficients partially due to its nonconcave and nondifferentiable nature. We propose to adopt a one-step approximation, where a local approximation algorithm is used.^{29,30} For a given initial value $\eta^{(0)}$, the log-likelihood function can be locally approximated, and thus, the corresponding penalized log-likelihood function can be approximated by

$$l_p(\boldsymbol{\eta}) = l(\boldsymbol{\eta}) - \sum_k^K \lambda_k |\gamma_k| \quad (3)$$

$$\approx l(\boldsymbol{\eta}^{(0)}) + l'(\boldsymbol{\eta}^{(0)})^T (\boldsymbol{\eta} - \boldsymbol{\eta}^{(0)}) + \frac{1}{2} (\boldsymbol{\eta} - \boldsymbol{\eta}^{(0)})^T l''(\boldsymbol{\eta}^{(0)}) (\boldsymbol{\eta} - \boldsymbol{\eta}^{(0)}) - \sum_k^K \lambda_k |\gamma_k|.$$

If we set the initial value to be the maximum likelihood estimator of $\mathcal{L}(\boldsymbol{\eta})$ ($\boldsymbol{\eta}^{(0)} = \hat{\boldsymbol{\eta}}_{\text{mle}}$), then $l'(\boldsymbol{\eta}^{(0)}) = \mathbf{0}$. It can be shown that the maximizers for Equation (3) can be attained equivalently by

$$\hat{\boldsymbol{\eta}} = \arg \min_{\boldsymbol{\eta}} \left\{ \frac{1}{2} (\boldsymbol{\eta} - \boldsymbol{\eta}^{(0)})^T (-l''(\boldsymbol{\eta}^{(0)})) (\boldsymbol{\eta} - \boldsymbol{\eta}^{(0)}) + \sum_k^K \lambda_k |\gamma_k| \right\}. \quad (4)$$

Clearly, Equation (4) can be solved by a least-angle regression algorithm that allows computing of the entire regularization path very efficiently.³¹ The regularization parameter λ is determined according to the Bayesian information criteria (BIC), ie,

$$\text{BIC}(\lambda) = -2l(\hat{\boldsymbol{\eta}}; \lambda) + \sum_k^K l(\gamma_k \neq 0).$$

2.2.2 | Asymptotic properties—In this subsection, we focus on the asymptotic properties of our estimators. Let $\boldsymbol{\theta}^0 = ((\boldsymbol{\eta}^0)^T, (\boldsymbol{\sigma}^2)^0)^T = ((\boldsymbol{\beta}^0)^T, (\gamma_1^0)^T, (\gamma_2^0)^T, (\boldsymbol{\sigma}^2)^0)^T$ denote true parameter values, where, without loss of generality, we assume that $\gamma_1^0 = (\gamma_1^0, \gamma_2^0, \dots, \gamma_p^0)$ is a p -dimensional vector of nonzero coefficients (ie, predictive SNV sets) and $\gamma_2^0 = (\gamma_{p+1}^0, \gamma_{p+2}^0, \dots, \gamma_K^0)$ is a $(K-p)$ -dimensional vector of zero coefficients (ie, noise SNV sets). Let $\hat{\boldsymbol{\theta}} = ((\hat{\boldsymbol{\eta}})^T, (\hat{\boldsymbol{\sigma}}^2)^0)^T = ((\hat{\boldsymbol{\beta}})^T, (\hat{\gamma}_1)^T, (\hat{\gamma}_2)^T, (\hat{\boldsymbol{\sigma}}^2)^0)^T$, where $\hat{\boldsymbol{\beta}}$, $\hat{\gamma}_1$, $\hat{\gamma}_2$, and $\hat{\boldsymbol{\sigma}}^2$ are the parameter estimates for $\boldsymbol{\beta}^0$, γ_1^0 , γ_2^0 , and $(\boldsymbol{\sigma}^2)^0$, respectively. Let $c_n = \max(\lambda_k/n : k = 1, 2, \dots, p)$ and $d_n = \min(\lambda_k/n : k = p+1, p+2, \dots, K)$. We further use $\mathcal{I}(\boldsymbol{\eta}^0)$ to denote the Fisher information matrix and $I_1(\boldsymbol{\eta}^0)$ to denote the Fisher information matrix knowing $\gamma_2^0 = \mathbf{0}$. Note that $I_1(\boldsymbol{\eta}^0)$ is a submatrix of $\mathcal{I}(\boldsymbol{\eta}^0)$. Under some regularity conditions, it can be shown that $n^{-1}l''(\hat{\boldsymbol{\eta}}_{\text{mle}}) \xrightarrow{P} -\mathcal{I}(\boldsymbol{\eta}^0)$ and $\sqrt{n}(\hat{\boldsymbol{\eta}}_{\text{mle}} - \boldsymbol{\eta}^0) \xrightarrow{D} N(\mathbf{0}, \mathcal{I}^{-1}(\boldsymbol{\eta}^0))$. On the basis of these facts, we can establish the oracle and sparsity properties of our one-step estimators.

Theorem 1. *If $\sqrt{n}d_n \rightarrow \infty$ and $c_n = o(n^{-1/2})$ as $n \rightarrow \infty$, then the one-step estimates $\hat{\boldsymbol{\eta}}$ must satisfy the following properties.*

- a. Sparsity: with probability tending to 1, $\hat{\gamma}_2 = \mathbf{0}$

b. Asymptotic normality: $n^{1/2}((\hat{\beta}^T, \hat{\gamma}_1^T)^T - (\beta_0^T, (\gamma_1^0)^T)^T) \xrightarrow{D} N(\mathbf{0}, I_1^{-1}(\eta^0))$

The details of the proof are shown in Appendix A in the online Supplementary Material. From part (a) of the theorem, the effect estimates for nonpredictive SNV sets are zero with probability tending to 1 when the sample size is large enough. This property ensures the sparsity of our predictive model and allows us to select predictive SNV sets. Part (b) of the theorem shows the asymptotic normality property of the estimators that are estimated efficiently through the least-angle regression algorithm. Therefore, our proposed model can select predictive SNV sets and estimate their effects at the same time, which is of great importance for genetic risk prediction research.

3 | SIMULATION STUDIES

Simulation studies were conducted to evaluate the performance of the proposed method. We further compared the proposed method with the genomic BLUP method (gBLUP).^{10,27} gBLUP has been examined in both animal breeding and human genetics and has been shown to have comparable or better accuracy as against the pedigree-based BLUP.³² gBLUP fits all SNVs in a prediction model, and thus can be applied to studies with millions of SNVs. Yang et al showed that in the content of the genome-wide association study, all SNVs can explain a larger proportion of heritability of human height than the significant SNVs.¹¹

In all simulation studies, Pearson correlations between the predicted and observed values in the testing samples are compared between SARAL and gBLUP based on 500 replicates under various numbers of noncausal SNV sets and causal/noncausal SNV ratios. For the SARAL method, we also calculate the probabilities of correctly identifying the predictive/nonpredictive SNV sets. To mimic the distribution of minor allele frequencies and linkage disequilibrium in the real genome, in all simulations described below, the genomic data are drawn from the 1000 Genomes Project. In particular, we divided the genome into SNV sets, each having 500 SNVs, and then randomly selected these sets for each replicate. We set sample sizes to be 500 and 1000, with 75% of the samples serving as training samples and the remaining serving as testing samples.

3.1 | Simulation 1: the impact of the number of noncausal SNV sets

For sequencing studies, usually, many non-disease-related SNV sets are measured. In this set of simulations, we evaluated our method by gradually increasing the number of noncausal SNV sets, with the total sample size being 500 and 1000. We first randomly chose one set of genome from the 1000 Genomes Project and simulated the phenotype based on causal variants in this set under an additive model, ie,

$$Y_i = \beta \sum_{l=1}^{l=L} G_{il} + \epsilon_i,$$

where $\epsilon_i \sim N(0, \sigma^2)$. We then gradually increased the number of noncausal SNV sets from 9 to 19, to 29, to 39, and to 59, and thus, the total number of SNVs included in the analyses changed from 5000 to 10 000, to 15 000, to 20 000, and to 30 000 (ie, each SNV set has 500

SNVs). For each given number of noise SNV sets, we further varied the effect sizes (ie, β changed from 0.2 to 1.5). We generated 500 replicates for each setting. For each replicate, we used 75% of the samples to train the model and used the remaining testing data to calculate Pearson correlations and mean square error (MSE) ratios (defined as $\frac{\text{MSE of SARAL}}{\text{MSE of gBLUP}}$) for both SARAL and gBLUP. We further reported the results of variable selection for SARAL, in terms of the probabilities of correctly identifying causal and noncausal SNV sets.

Pearson correlations between the predicted and observed values in the testing samples are reported in Figure 1 (β ranges from 0.2 to 0.6) and Figure S1 of Appendix B (β ranges from 0.6 to 1.5). As the number of noise SNV sets increases, the prediction accuracies remain robust for the proposed SARAL method, whereas they decrease for gBLUP. This trend is preserved regardless of sample sizes. We reported the MSE ratios in Table S1 of Appendix B. As expected, as the number of noise SNV sets increases, the MSE ratios decrease. The proposed method has comparable, if not better, performance with gBLUP. Regarding the selection performance of the proposed method, as shown in Table 1, our method has a relatively high probability of correctly selecting the predictive SNV set, and it is robust against the increasing number of noise SNV sets. When the sample size increases, this probability increases as well. With regard to the probability of correctly shrinking the effects of noise SNV sets to zero, it increases with effect sizes and remains relatively robust against the number of noise SNV sets. With a high probability of correctly identifying the causal SNV sets and teasing out the effects of noise SNV sets, our predictive model is most likely to be built based solely on causal SNV sets, and thus it tends to have higher accuracy than the commonly used gBLUP-based methods, and its prediction accuracy remains stable with the increasing number of noise SNV sets. We consider this important, as the majority of measured SNV sets or genes is not predictive for high-dimensional data. The risk prediction models built with a substantial amount of noise signals are subject to low prediction accuracy.

As various criteria can be used to determine the regularization parameter λ , we further compared the results with the optimal model selected by Mallows' C_p , bias-corrected Akaike information criterion (AIC), generalized cross validation, BIC, and information-theoretic minimum message length.³³ As shown in Table 2, the model selected by BIC has the best prediction accuracy, and thus we recommend using the BIC in determining the regularization parameter.

3.2 | Simulation 2: the impact of signal/noise SNV ratios

In practice, it is quite likely that there are some noncausal variants located on the causal SNV sets (ie, not all variants within a set are causal). In this simulation, we evaluated the performance of our method given different signal/noise SNV ratios within a causal SNV set. We randomly chose one SNV set and set it as causal, and then, we varied the proportion of the predictive markers within the set. We simulated the phenotype based on the causal variants in this set under an additive model, ie,

$$Y_i = \beta \sum_{l=1}^{l=L} G_{il}g_l + \epsilon_i,$$

where g_l is an indicator with $g_l = 1$ if the l th marker is causal and $g_l = 0$ if the l th marker is noncausal, and $P(g_l = 1)$ represents the proportion of causal markers on the causal SNV set. We changed the proportion of causal variants from 100% to 10% (ie, causal/noise SNV ratios changed from ∞ to 1/9) to evaluate the effects of signal/noise SNV ratios on the methods' performance. For simplicity, we set the number of noise SNV sets to be 29 (ie, the total number of SNV sets was 30, and the total number of SNVs was 15 000) for this set of simulations.

Pearson correlations and MSE ratios in the testing samples are reported in Table 3 and Table S2 of Appendix B, respectively. As expected, when the signal/noise ratio increases, the Pearson correlations increase for both methods regardless of sample sizes. Moreover, as there are 29 noise SNV sets in the simulated data, the proposed model tends to have better performance than the commonly used gBLUP method. Regarding the variable selection, it is expected that the selection tends to be more accurate with the increasing proportion of causal variants. As shown in Table 4, the selection accuracy is, in general, robust against the signal/noise ratio. For example, the probability of correctly selecting the causal SNV set increases from 0.886 to 0.930, and the probability of correctly identifying noncausal SNV sets increases from 0.972 to 0.996 for $\beta = 1$ and $N = 500$. When the sample size increases, the method tends to be more accurate, but the general trend is preserved. As shown through simulations, with only 10% of variants being causal, our method has more than 74% chance of correctly identifying the causal SNV set and shrinking the effects of noncausal SNV sets to zero, and this could explain why our method has better performance than gBLUP as our method tends to use the disease-relevant sets for risk prediction analysis.

4 | REAL DATA APPLICATION

We analyzed the whole-genome sequencing data from the ADNI using both our proposed SARAL and gBLUP methods. The ADNI is a multisite longitudinal study that assesses clinical, imaging, genetic, and biospecimen biomarkers through the process of normal aging to early mild cognitive impairment, to late mild cognitive impairment, and to dementia or Alzheimer's disease (AD). It has three phases: ADNI1, ADNI GO, and ADNI2. Participants were followed and reassessed over time to track the pathology of the disease. DNA samples were obtained from ADNI participants and were sent to Illumina where non-Clinical Laboratory Improvements Amendments (CLIA) whole genome sequencing (WGS) were performed on each sample. Magnetic resonance image data (ie, functional magnetic resonance imaging and diffusion tensor imaging), Positron emission tomography (PET) image data (ie, florbetapir (AV-45) PET Analysis, fluorodeoxyglucose (FDG), and Pittsburgh Compound-B(PIB)), and clinical data (eg, biospecimen and cognitive tests) were also collected, both at baseline and follow-up visits.

4.1 | PET imaging outcomes

In this analysis, we are interested in predicting the PET imaging outcomes, including the FDG and Florbetapir (AV-45). AV-45 PET and FDG PET imaging were performed on all newly enrolled participants, and scans were completed within 2 weeks before or 2 weeks after the baseline in-clinic assessments. As there were only 20 patients from the ADNI1 cohort that received PIB scans due to the change in the study protocol (ie, use the Florbetapir scans over PIB scans due to processing time limits), we did not include PIB as our final outcomes for PET imaging prediction analyses.

We annotated the genetic variants based on GRch37 assembly and included a total of 96 genes that have been previously reported to be associated with AD. The complete list of genes used in our analyses is listed in Appendix C. In total, 151159 SNVs were included in the final analyses. To avoid the over-fitting problem, we randomly chose 80 subjects to serve as the testing data and used the remaining data to train the model. The sample sizes for all PET imaging outcomes are listed in Appendix D. We calculated the Pearson correlations based on the testing data. To avoid the chance findings, we repeated this process 500 times. The results from both methods are presented in Figure 2. As shown in the histogram, our method outperforms the gBLUP method for both outcomes, which suggests that excluding noise genes from prediction analyses can help improve prediction accuracy. Our method is also relatively robust in selecting predictive genes. More than 88% of the genes are selected less than 15% of times, and three genes were selected more than 90% of times. The details of the probability of each gene being selected by our method for FDG and AV-45 are summarized in Appendix C. The *APOE* gene on chromosome 19, a well-known risk factor for AD, has been selected 100% for both outcomes. The fibroblast growth factor (*FGF-1*) located on chromosome 5 has been selected about 98% times for FDG, and the *MCM3AP* gene family located on chromosome 21 has been selected more than 92% times for AV45. *FGF-1* promotes the survival of neurons, and it has been found that serum *FGF-1* in patients with AD was higher than that in patients without AD.³⁴ It has also been found that variants within *FGF-1* are associated with AD in the Han Chinese population.³⁵ Genetic variants on *MCM3AP* have been reported to be associated with AD.³⁶ It has been found that abnormal gene expressions of *MCM3AP* have initiated the dysfunction of the biological process in the brains of senescence-accelerated mouse prone 8 (*SAMP8*) mice, which are considered a robust animal model for AD investigation.³⁷

4.2 | Rey auditory verbal learning test outcomes

We also performed similar analyses on the outcomes from the Rey Auditory Verbal Learning Test (RAVLT), which is a widely used cognitive assessment tool for AD. The RAVLT is a test of episodic memory that assesses the ability for acquisition of 15 words from five learning trails, to recall the words immediately after an intervening interference list and to recall the words after a 30-minute delay interval. The sample sizes for all RAVLT outcomes are listed in Appendix D. The accuracies for learning, immediate recalling, and 30-minute delay recalling are presented in Figure 3. Similar to the PET imaging outcomes, our method performs comparable with or better than the commonly used gBLUP method. The probabilities for each gene being selected are summarized in Appendix C. For learning, genes *HSPA5* and *APOE* were selected more than 85% times. For immediate recalling,

genes *APOE* and *ATF7* were selected more than 85% times. For 30-minute delay interval recalling, genes *APOE*, *COX10*, and *SERPINA2* were selected more than 85% times.

5 | DISCUSSION

We have proposed a novel framework for modeling disease outcomes using high-dimensional sequencing data from the perspective of spatial statistics. We utilized the idea that adjacent points share similar outcomes and predicted the disease outcome based on their genetic similarities using a spatial autoregressive model. We further developed a variable selection technique using the idea of adaptive lasso and proposed to use the one-step estimator to get the approximate parameter estimates with the oracle and sparsity properties. We consider the use of the adaptive lasso to select autoregressive coefficients in the SAR model innovative, as few existing methods are capable of selecting neighborhood structures from high-dimensional data. While many criteria (eg, Mallows' C_p , bias-corrected AIC, generalized cross validation, and information-theoretic minimum message length) can be used to determine the regularization parameter λ , we recommend using the BIC as it has better prediction performance than many existing model selection criteria (Table 2). Through simulations, we showed that SARAL achieves better prediction performance than the commonly used gBLUP method when a substantial amount of noise SNV sets was included in the analysis. We further demonstrated our method through the application to the ADNI data set, where a number of brain imaging and cognitive test outcomes were predicted using genomic data. Among all the outcomes considered, our proposed method achieves comparable, and sometimes even much better, performance relative to the commonly used gBLUP method. Regarding the variable selection, our method, on average, has a relatively high probability of correctly selecting disease-associated SNV sets and shrinking the effects of noise SNV sets to zero, as demonstrated through simulation studies. The increase in the number of noise SNV sets has little effect on the method, and it is relatively robust against the signal/noise ratios. Through the real data application, we further demonstrated that the selection is relatively consistent. About 84% of the genes have been consistently excluded from the prediction model (more than 85% chance), and about 2% of the genes have been consistently selected as predictive genes (more than 85% of chance). We also demonstrated that teasing out the noise effects can improve prediction accuracy, as shown in Figures 2 and 3.

Although the adaptive lasso has been widely used in generalized linear models, it is rarely used to select autoregressive coefficients for spatial models partially due to computational complexity. We adopted the local approximation method to estimate parameters for high-dimensional genetic data sets, which substantially reduces the computational cost and makes it applicable to sequencing data.^{29,30} It has been shown that under a partial orthogonality condition where covariates with zero coefficients are weakly correlated with nonzero covariates, the adaptive lasso has the oracle properties with initial estimators determined from marginal regression even when $p > n$. However, this condition is not readily extendable to autoregressive models. In practice, researchers can adopt several preprocessing procedures when they analyze whole-genome data. For example, researchers can only focus on the genes that may potentially be related to the disease under study. Alternatively, researchers can group genes based on pathway information or other prior knowledge.

Although it is a general practice in high-dimensional genetic studies to use a linear model to predict disease outcomes regardless of their original distributions and we can adopt a similar strategy for our model, it would be interesting to study, within the framework of the spatial autoregressive model, other link functions (eg, logit and log) for the prediction of outcomes with various distributions (eg, binary and Poisson) in the future.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

ACKNOWLEDGEMENTS

The Project was supported by the National Natural Science Foundation of China (Award No. 81502887), the Faculty Research Development Funds from the University of Auckland, and the National Institute on Drug Abuse (Award No. R01DA043501). The authors wish to acknowledge the contribution of the New Zealand eScience Infrastructure (NeSI) High Performance Computing facilities to the results of this research. New Zealand's national facilities are provided by the NeSI and funded jointly by NeSI's collaborator institutions and through the Ministry of Business, Innovation & Employment's Research Infrastructure program.

Funding information

National Natural Science Foundation of China, Grant/Award Number: 81502887; Faculty Research Development Funds from the University of Auckland; National Institute on Drug Abuse, Grant/Award Number: R01DA043501

REFERENCES

1. Collins FS, Varmus H. A new initiative on precision medicine. *N Engl J Med*. 2015;372:793–795. [PubMed: 25635347]
2. Chatterjee N, Wheeler B, Sampson J, Hartge P, Chanock SJ, Park JH. Projecting the performance of risk prediction based on polygenic analyses of genome-wide association studies. *Nat Genet*. 2013;45(4):400–405. [PubMed: 23455638]
3. Neale BM, Sham PC. The future of association studies: gene-based analysis and replication. *Am J Hum Genet*. 2004;75(3):353–362. [PubMed: 15272419]
4. Chatterjee N, Kalaylioglu Z, Moslehi R, Peters U, Wacholder S. Powerful multilocus tests of genetic association in the presence of gene-gene and gene-environment interactions. *Am J Hum Genet*. 2006;79(6):1002–1016. [PubMed: 17186459]
5. Wu MC, Kraft P, Epstein MP, et al. Powerful SNP-set analysis for case-control genome-wide association studies. *Am J Hum Genet*. 2010;86(6):929–942. [PubMed: 20560208]
6. Wu MC, Lee S, Cai T, Li Y, Boehnke M, Lin X. Rare-variant association testing for sequencing data with the sequence kernel association test. *Am J Hum Genet*. 2011;89(1):82–93. [PubMed: 21737059]
7. He Z, Zhang M, Zhan X, Lu Q. Modeling and testing for joint association using a genetic random field model. *Biometrics*. 2014;70(3):471–479. [PubMed: 24628067]
8. Abraham G, Inouye M. Genomic risk prediction of complex human disease and its clinical application. *Curr Opin Genet Dev*. 2015;33:10–16. [PubMed: 26210231]
9. Meuwissen THE, Hayes BJ, Goddard ME. Prediction of total genetic value using genome-wide dense marker maps. *Genetics*. 2001;157(4):1819–1829. [PubMed: 11290733]
10. VanRaden PM. Efficient methods to compute genomic predictions. *J Dairy Sci*. 2008;91(11):4414–4423. [PubMed: 18946147]
11. Yang JA, Lee SH, Goddard ME, Visscher PM. GCTA: a tool for genome-wide complex trait analysis. *Am J Hum Genet*. 2011;88(1):76–82. [PubMed: 21167468]
12. de los Campos G, Vazquez AI, Fernando R, Klimentidis YC, Sorensen D. Prediction of complex human traits using the genomic best linear unbiased predictor. *PLoS Genet*. 2013;9(7).

13. Zhou X, Stephens M. Genome-wide efficient mixed-model analysis for association studies. *Nat Genet.* 2012;44(7):821–824. [PubMed: 22706312]
14. Speed D, Balding DJ. MultiBLUP: improved SNP-based prediction for complex traits. *Genome Res.* 2014;24(9):1550–1557. [PubMed: 24963154]
15. Byrnes AE, Wu MC, Wright FA, Li M, Li Y. The value of statistical or bioinformatics annotation for rare variant association with quantitative trait. *Genet Epidemiol.* 2013;37(7):666–674. [PubMed: 23836599]
16. Zou H The adaptive lasso and its oracle properties. *J Am Stat Assoc.* 2006;101(476):1418–1429.
17. Cressie N The origins of kriging. *Math Geol.* 1990;22(3):239–252.
18. Bivand R Spatial econometrics functions in R: classes and methods. *J Geogr Syst.* 2002;4(4):405–421.
19. Wimberly MC, Baer AD, Yabsley MJ. Enhanced spatial models for predicting the geographic distributions of tick-borne pathogens. *Int J Health Geogr.* 2008;7(1):15. [PubMed: 18412972]
20. Dubin R Robustness of spatial autocorrelation specifications: some Monte Carlo evidence. *J Reg Sci.* 2003;43(2):221–248.
21. Kelejian HH, Prucha IR. Prediction efficiencies in spatial models with spatial lags. College Park, MD: University of Maryland, Economics Department; 2004 Working paper.
22. LeSage JP, Pace RK. Models for spatially dependent missing data. *J Real Estate Finance Econ.* 2004;29(2):233–254.
23. Kelejian HH, Prucha IR. The relative efficiencies of various predictors in spatial econometric models containing spatial lags. *Reg Sci Urban Econ.* 2007;37(3):363–374.
24. Kato T A further exploration into the robustness of spatial autocorrelation specifications. *J Reg Sci.* 2008;48(3):615–639.
25. LeSage JP, Pace RK. Spatial econometric models, prediction In: Shekhar S, Xiong H, eds. *Encyclopedia of Geographical Information Science.* New York, NY: Springer; 2008.
26. Thomas-Agnan C, Thibault L, Goulard M. About predictions in spatial autoregressive models: optimal and almost optimal strategies. Toulouse, France: Toulouse 1 University Capitole, Toulouse School of Economics 2014:13–452. Working paper.
27. Habier D, Fernando RL, Dekkers J. The impact of genetic relationship information on genome-assisted breeding values. *Genetics.* 2007;177(4):2389–2397. [PubMed: 18073436]
28. Saykin AJ, Shen L, Foroud TM, et al. Alzheimer's Disease Neuroimaging Initiative biomarkers as quantitative phenotypes: genetics core aims, progress, and plans. *Alzheimer's Dement J Alzheimer's Assoc.* 2010;6(3):265–273.
29. Zou H, Li RZ. One-step sparse estimates in nonconcave penalized likelihood models. *Ann Stat.* 2008;36(4):1509–1533. [PubMed: 19823597]
30. Zhu J, Huang HC, Reyes P. On selection of spatial linear models for lattice data. *J Royal Stat Soc Ser B Stat Methodol.* 2010;72(3):389–402.
31. Efron B, Hastie T, Johnstone I, Tibshirani R. Least angle regression. *Ann Stat.* 2004;32(2):407–451.
32. VanRaden PM, Van Tassell CP, Wiggans GR, et al. Invited review: reliability of genomic predictions for North American Holstein bulls. *J Dairy Sci.* 2009;92(1):16–24. [PubMed: 19109259]
33. Xu Z, Schmidt DF, Makalic E, Qian G, Hopper JL. Bayesian sparse global-local shrinkage regression for selection of grouped variables. *arXiv preprint arXiv:1709.04333.* <https://arxiv.org/abs/1709.04333>
34. Mashayekhi F, Hadavi M, Vaziri HR, Naji M. Increased acidic fibroblast growth factor concentrations in the serum and cerebrospinal fluid of patients with Alzheimer's disease. *J Clin Neurosci.* 2010;17(3):357–359. [PubMed: 20079650]
35. Tao QQ, Sun YM, Liu ZJ, et al. A variant within FGF 1 is associated with Alzheimer's disease in the Han Chinese population. *Am J Med Genet B Neuropsychiatr Genet.* 2014;165(2):131–136.
36. Taguchi K, Yamagata HD, Zhong W, et al. Identification of hippocampus-related candidate genes for Alzheimer's disease. *Ann Neurol.* 2005;57(4):585–588. [PubMed: 15786443]

37. Cheng XR, Cui XL, Zheng Y, et al. Nodes and biological processes identified on the basis of network analysis in the brain of the senescence accelerated mice as an Alzheimer's disease animal model. *Front Aging Neurosci.* 2013;5:65. [PubMed: 24194717]

FIGURE 1.

Pearson correlations between the predicted and observed values in the testing samples with various numbers of noise single-nucleotide variant sets. A, Sample size $N=500$; B, Sample size $N=1000$

FIGURE 2.

Histogram of Pearson correlations between the predicted and observed values in the testing samples for FDG and AV45

FIGURE 3.
Histogram of Pearson correlations between the predicted and observed values in the testing samples for the Rey Auditory Verbal Learning Test (RAVLT) outcomes

TABLE 1

Variable selection for the simultaneous autoregressive model with adaptive lasso with sample sizes $N = 500$ and $N = 1000$ given different numbers of noise single-nucleotide variant sets

Effects	No. Genes=10		No. Genes=20		No. Genes=30		No. Genes=40		No. Genes=60	
	Zero	Nonzero	Zero	Nonzero	Zero	Nonzero	Zero	Nonzero	Zero	Nonzero
<i>Sample size: $N = 500$</i>										
$\beta = 0.2$	0.826	0.994	0.862	1.000	0.874	0.998	0.876	0.994	0.892	0.974
$\beta = 0.4$	0.854	0.998	0.896	1.000	0.914	1.000	0.912	0.990	0.914	0.954
$\beta = 0.6$	0.900	1.000	0.922	0.998	0.914	0.994	0.920	0.990	0.922	0.938
$\beta = 0.8$	0.888	1.000	0.926	1.000	0.920	0.994	0.932	0.990	0.926	0.950
$\beta = 1.0$	0.918	1.000	0.960	1.000	0.926	0.998	0.932	0.996	0.942	0.964
$\beta = 1.5$	0.932	1.000	0.946	1.000	0.920	0.994	0.934	0.992	0.950	0.962
<i>Sample size: $N = 1000$</i>										
$\beta = 0.2$	0.852	1.000	0.878	1.000	0.882	1.000	0.886	0.996	0.904	0.982
$\beta = 0.4$	0.912	1.000	0.918	1.000	0.920	0.996	0.932	0.994	0.944	0.994
$\beta = 0.6$	0.922	1.000	0.952	1.000	0.940	1.000	0.948	0.994	0.960	0.978
$\beta = 0.8$	0.942	1.000	0.950	1.000	0.952	0.996	0.948	0.990	0.970	0.994
$\beta = 1.0$	0.946	1.000	0.962	1.000	0.960	1.000	0.960	0.996	0.982	0.994
$\beta = 1.5$	0.958	1.000	0.968	1.000	0.954	0.996	0.968	0.994	0.976	0.988

TABLE 2

Comparisons between different model selection criteria with $N = 500$

Criteria	No. Genes=10		No. Genes=20		No. Genes=30		No. Genes=40		No. Genes=60	
	CORR	MSE	CORR	MSE	CORR	MSE	CORR	MSE	CORR	MSE
<u>Effect size: $\beta = 0.2$</u>										
BIC	0.590	0.869	0.607	0.740	0.610	0.711	0.616	0.691	0.615	0.679
Cp	0.583	0.868	0.596	0.747	0.603	0.715	0.609	0.691	0.597	0.694
AICc	0.554	0.874	0.542	0.810	0.475	0.867	0.434	0.908	0.369	1.009
GCV	0.541	0.890	0.509	0.854	0.431	0.925	0.381	0.997	0.276	1.201
MML	0.541	0.890	0.509	0.854	0.431	0.925	0.381	0.997	0.276	1.201
<u>Effect size: $\beta = 0.4$</u>										
BIC	0.728	0.902	0.768	0.700	0.775	0.604	0.786	0.526	0.776	0.502
Cp	0.724	0.895	0.751	0.717	0.772	0.593	0.784	0.512	0.769	0.501
AICc	0.714	0.841	0.730	0.726	0.670	0.779	0.714	0.630	0.652	0.689
GCV	0.709	0.850	0.675	0.850	0.598	0.912	0.675	0.700	0.586	0.804
MML	0.709	0.850	0.675	0.850	0.598	0.912	0.675	0.700	0.586	0.804
<u>Effect size: $\beta = 0.6$</u>										
BIC	0.785	0.892	0.811	0.687	0.819	0.598	0.819	0.522	0.823	0.458
Cp	0.771	0.894	0.792	0.701	0.815	0.573	0.820	0.492	0.826	0.431
AICc	0.752	0.833	0.787	0.672	0.767	0.697	0.803	0.540	0.781	0.515
GCV	0.759	0.807	0.727	0.814	0.725	0.793	0.777	0.590	0.719	0.610
MML	0.759	0.807	0.727	0.814	0.725	0.793	0.777	0.590	0.719	0.610
<u>Effect size: $\beta = 0.8$</u>										
BIC	0.792	0.936	0.827	0.703	0.825	0.630	0.830	0.506	0.836	0.439
Cp	0.786	0.938	0.811	0.709	0.825	0.572	0.846	0.467	0.843	0.402
AICc	0.764	0.835	0.807	0.665	0.760	0.734	0.811	0.557	0.821	0.465
GCV	0.765	0.827	0.750	0.822	0.698	0.878	0.790	0.602	0.766	0.549
MML	0.765	0.827	0.750	0.822	0.698	0.878	0.790	0.602	0.766	0.549
<u>Effect size: $\beta = 1.0$</u>										
BIC	0.786	0.945	0.842	0.687	0.832	0.629	0.847	0.503	0.843	0.429
Cp	0.778	0.935	0.818	0.709	0.826	0.590	0.847	0.474	0.846	0.395
AICc	0.762	0.854	0.825	0.625	0.788	0.703	0.836	0.520	0.833	0.446
GCV	0.775	0.822	0.775	0.780	0.738	0.824	0.806	0.587	0.784	0.530
MML	0.775	0.822	0.775	0.780	0.738	0.824	0.806	0.587	0.784	0.530
<u>Effect size: $\beta = 1.5$</u>										
BIC	0.811	0.964	0.845	0.701	0.835	0.619	0.852	0.512	0.859	0.417
Cp	0.799	0.938	0.821	0.717	0.826	0.595	0.849	0.494	0.864	0.371
AICc	0.770	0.878	0.837	0.620	0.802	0.677	0.844	0.525	0.866	0.398
GCV	0.774	0.870	0.768	0.819	0.745	0.814	0.820	0.580	0.820	0.478
MML	0.774	0.870	0.768	0.819	0.745	0.814	0.820	0.580	0.820	0.478

Abbreviations: AICc, bias-corrected Akaike information criterion; BIC, Bayesian information criteria; CORR, Pearson correlation; Cp, Mallows' C_p ; GCV, generalized cross validation; MML, information-theoretic minimum message length; MSE, mean square error ratio = $\frac{\text{MSE of SARAL}}{\text{MSE of gBLUP}}$.

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

TABLE 3

Pearson correlation comparisons as the signal/noise ratio increases

Effects	% causal=10%		% causal=25%		% causal=50%		% causal=75%		% causal=100%	
	SARAL	gBLUP	SARAL	gBLUP	SARAL	gBLUP	SARAL	gBLUP	SARAL	gBLUP
<i>Sample size: N = 500</i>										
$\beta = 0.2$	0.048	0.018	0.189	0.060	0.362	0.127	0.500	0.200	0.609	0.279
$\beta = 0.4$	0.181	0.061	0.406	0.138	0.593	0.274	0.706	0.413	0.764	0.506
$\beta = 0.6$	0.308	0.099	0.517	0.210	0.688	0.381	0.764	0.513	0.799	0.584
$\beta = 0.8$	0.399	0.132	0.590	0.263	0.729	0.462	0.790	0.564	0.809	0.612
$\beta = 1.0$	0.467	0.173	0.644	0.319	0.740	0.489	0.791	0.574	0.826	0.625
$\beta = 1.5$	0.555	0.253	0.706	0.420	0.771	0.538	0.803	0.610	0.833	0.637
<i>Sample size: N = 1000</i>										
$\beta = 0.2$	0.071	0.026	0.203	0.069	0.380	0.167	0.514	0.308	0.601	0.425
$\beta = 0.4$	0.222	0.082	0.409	0.196	0.588	0.407	0.681	0.545	0.733	0.612
$\beta = 0.6$	0.318	0.128	0.532	0.324	0.674	0.542	0.737	0.621	0.769	0.664
$\beta = 0.8$	0.397	0.182	0.602	0.428	0.707	0.594	0.764	0.653	0.773	0.684
$\beta = 1.0$	0.470	0.257	0.641	0.481	0.733	0.627	0.782	0.672	0.794	0.693
$\beta = 1.5$	0.575	0.381	0.690	0.566	0.756	0.655	0.787	0.686	0.806	0.704

Abbreviations: gBLUP, genomic best linear unbiased prediction; SARAL, simultaneous autoregressive model with adaptive lasso.

TABLE 4

Variable selection for the simultaneous autoregressive model with adaptive lasso with sample sizes $N = 500$ and $N = 1000$ given different signal/noise ratios

Effects	% causal=10%		% causal=25%		% causal=50%		% causal=75%		% causal=100%	
	Zero	Nonzero	Zero	Nonzero	Zero	Nonzero	Zero	Nonzero	Zero	Nonzero
<i>Sample size: N = 500</i>										
$\beta = 0.2$	0.818	0.736	0.848	0.946	0.876	0.994	0.884	1.000	0.914	1.000
$\beta = 0.4$	0.846	0.860	0.880	1.000	0.898	0.994	0.922	0.996	0.920	0.994
$\beta = 0.6$	0.870	0.948	0.882	0.996	0.898	1.000	0.908	1.000	0.926	0.998
$\beta = 0.8$	0.882	0.968	0.886	1.000	0.898	0.996	0.930	1.000	0.920	0.994
$\beta = 1.0$	0.886	0.972	0.896	0.998	0.918	0.998	0.916	0.998	0.930	0.996
$\beta = 1.5$	0.898	0.984	0.898	1.000	0.914	0.996	0.918	0.994	0.928	0.996
<i>Sample size: N = 1000</i>										
$\beta = 0.2$	0.820	0.752	0.862	0.970	0.880	0.994	0.912	0.996	0.920	0.996
$\beta = 0.4$	0.870	0.936	0.890	0.996	0.912	0.998	0.944	0.996	0.952	0.996
$\beta = 0.6$	0.882	0.952	0.902	0.998	0.932	1.000	0.952	0.998	0.960	1.000
$\beta = 0.8$	0.884	0.974	0.906	0.996	0.942	0.998	0.956	1.000	0.954	0.996
$\beta = 1.0$	0.906	0.986	0.912	1.000	0.946	0.998	0.960	1.000	0.966	0.998
$\beta = 1.5$	0.922	0.994	0.924	1.000	0.950	0.996	0.956	0.998	0.968	1.000