Simulated Annealing and Particle Swarm Optimization Approach to Variable Selection Under Misspecified Sparse Bridge Regression Using Information Complexity

Yaojin Sun[1], Hamparsum Bozdogan[1,*]

[a]*Department of Business Analytics and Statistics*
The University of Tennessee, Knoxville, TN 37996 U.S.A.

Abstract

This paper proposes a novel Sparse Bridge Regression model for subset selection of variables under model misspecification when the number of variables is much larger than the number of observations. The misspecification resistant information complexity criterion is developed as the fitness function for subset selection that allows both the robust statistical inference and proper interpretation of the results. To estimate the shrinkage factor q, tuning parameter λ, and model coefficients jointly, a hybridized Simulated Annealing and Particle Swarm Optimization algorithm, abbreviated as SA-PSO, is proposed. We introduce a smoothed covariance estimator to resolve the problem of ill-conditioned Gram matrix. The proposed new approach enables simultaneous model selection in Sparse Bridge Regression. It eliminates the costly and time-consuming cross-validation in high dimensions when the number of observations is limited. We show numerical examples using Monte Carlo simulation and a high-dimensional data to illustrate the model's sparsity and the efficiency of the proposed methodological and computational approach.

Keywords: Sparse Bridge Regression; High Dimensions; Misspecification; Information Complexity; SA-PSO Algorithm; Smoothed Covariance Estimators

1. Introduction

In high-dimensional data modeling, a statistical technique's success depends on identifying and selecting the most informative predictor variables. High dimensional data often have many redundant variables (or features) and a small number of relevant variables. It is crucial to identify and choose the relevant variables. Therefore, in high dimensions, the principle of parsimony or the Occam's razor, that is, *"entities should not be multiplied without necessity,"* attributed to the English Franciscan monk Occam (1287 - 1347), is necessary to data modeling and machine learning to understand the underlying data generation process.

In the literature, many penalized regression models have been proposed and used for variable selection. It has been shown that penalized regression methods yield consistent variable selection and oracle parameter estimation under correct model specification. However, the robustness of the penalized regression methods has not been convincingly studied. Many of the penalized regression methods still depend on the Gram matrix $\hat{\Sigma} := (X'X)$ to estimate the model coefficients. Given that the dataset is high dimensional ($p \gg n$), the

*Please address correspondence to Professor Bozdogan.
Email addresses: ysun52@vols.utk.edu (Yaojin Sun), bozdogan@utk.edu (Hamparsum Bozdogan)

Gram matrix is singular, and we cannot calculate its inverse. Because of the singularity of the Gram matrix, the regression model cannot provide reliable predictive inference.

Under singularity, one solution is to use a regularized covariance matrix instead of a raw Gram matrix to compute the coefficients of the variables. Therefore, further data-adaptive regularization is required on covariance matrix of the estimated coefficients to allow the selection of the best subset of predictors in the model. This approach achieves consistent variable selection, and allows dimension reduction further and eliminates irrelevant variables.

To address and resolve the existing problems in currently practiced penalized-regression modeling, we propose a new Sparse Bridge Regression (SBR) modeling approach under model misspecification based on the generalization of the work of Bozdogan (2004), Bozdogan and Pamukcu (2016),Mohebbi et al. (2019), Kawano (2012, 2014). The primary assumption in the literature is that the regression model is almost always correctly specified. For high-dimensional data, even if we assume Gaussianity on the model, this assumption may not be accurate in practice.

The sparsity of statistical models is important when the original dataset includes large variables, and statisticians cannot interpret the relationship between predictors and the response variable. To find the best subset of variables, we develop the misspecification resistant information complexity (ICOMP) criterion of Bozdogan (2000); Bozdogan and Howe (2012) for variable selection in SBR modeling. To derive the ICOMP criterion, we utilize the celebrated Fisher information matrix $\mathcal{F}$ in the Hessian form and the outer-product form $\mathcal{R}$ to obtain the robust covariance matrix. The robust covariance matrix is also known as a sandwich covariance matrix. We can use the sandwich covariance matrix in SBR model to guard against distributional misspecification and to obtain standard error estimates of the coefficients for further inference such as constructing confidence intervals.

Cross-validation (CV) and Grid-search (GS) type of methods are well-known in the literature and often utilized to estimate the tuning parameter λ and the shrinkage parameter q. It has been noted that these methods are too time consuming, especially, when we have under sample wide data sets. As an alternative to these methods, to estimate the shrinkage parameter q, the tuning parameter λ, and the coefficients of the SBR model jointly, here we propose and use a hybridized Simulated Annealing and Particle Swarm Optimization algorithm, abbreviated as SA-PSO.

To further resolve the ill-conditioned Gram matrix $\hat{\Sigma} := (X'X)$ in high dimensional datasets, we introduce the idea of Smoothed Covariance Estimators (SCEs) of the predictor variables. By regularizing the commonly used Local Quadratic Approximation (LQA) algorithm, we provide a Regularized Local Quadratic Approximation (RLQA) algorithm using one of the SCE's to make the Gram matrix $\hat{\Sigma} := (X'X)$ well-behaved to estimate the coefficients of the SBR model. We then apply variable selection via the ICOMP criterion as our fitness function.

There are several forms of SCEs available in the literature, as discussed in Mohebbi et al. (2019) and **?**. Here, we only use one type of SCEs for space consideration. However, our proposed approach is flexible to use other forms of SCEs.

In summary, our goals and contributions in this paper are multifaceted.

- We focus on Sparse Bridge Regression (SBR) when $0 < q \leq 1$ and $p \gg n$, the undersized sample problem.
- We develop the misspecification-resistant information complexity ICOMP criterion for the selection of best variables in the SBR models.
- We propose a new hybridized Simulated Annealing-based Particle Swarm Optimization (SA-PSO) algorithm to jointly estimate the SBR model parameters.

- To resolve the ill-conditioning of the Gram matrix $\hat{\Sigma} := (X'X)$, we introduce Smoothed Covariance Estimators (SCEs) of the predictor variables. We use SCE in the Local Quadratic Approximation (LQA) algorithm and provide a Regularized Local Quadratic Approximation (RLQA) algorithm in the SBR model.

We organize the rest of the paper as follows. In Section 2, we present the statistical background of the SBR model. In particular, Section 2.1 has a brief introduction of the least-squares (LS) regression model under misspecification. As a continuation, we discuss the use of the SBR model under misspecification setting and provide the derived form of the estimated sandwich covariance matrix (Section 2.2). In Section 3, we discuss the general form of the misspecification resistant information complexity ICOMP criterion. Section 4 presents our newly proposed estimation of SBR model parameters simultaneously and provides computational algorithms. In Subsection 4.1, we present the hybridized SA-PSO algorithm to obtain the λ's optimal values, the tuning parameter, and q, the shrinkage factor using the ICOMP criterion. In Subsection 4.2, we present the idea of Smoothed Covariance Estimators (SCEs), which has been largely used to resolve the "large p, small n" problem in the SBR model. Subsection 4.3 presents the Regularized Local Quadratic Approximation (RLQA) algorithm and provides the necessary steps to obtain the estimated coefficients $\hat{\beta}_{Bridge}$ of the SBR model. By combining the results from Subsections 4.1, 4.2, and 4.3, we score the misspecification resistant information complexity ICOMP criterion for the SBR model. In Section 5, we provide our numerical examples. In Subsection 5.1, we provide the structure of the Monte Carlo Simulation protocol. In Subsection 5.2, we conduct several Monte Carlo simulations to illustrate the performance of our proposed approach using the OAS smoothed covariance estimator. In Subsection 5.3, we compare the SBR model with that of LASSO using different simulated data sets as we vary the number of predictor variables from small 32 to large 10,000 variables. We provide the results for variable performance metrics such as Precision, Accuracy, Recall, and F-1 Score, respectively.

Further, in Section 6, a real-world benchmark dataset, Riboflavin (Vitamin B_2), is analyzed. This benchmark dataset has $n = 71$ observations and has $p = 4,088$ dimensional normalized gene-expressions (predictor variables). As can be seen, this benchmark data set is extremely undersized. Our results from both simulated numerical examples and real Vitamin B_2 data show the efficiency and accuracy of the proposed SBR model using the misspecification resistant information complexity criterion. Section 7 concludes the paper with a discussion for future research.

2. Sparse Bridge Regression Model

2.1. A Brief Background: The Least Squares Regression Model

We first consider the usual linear regression model

$$y = X\beta + \varepsilon, \tag{1}$$

where $y = (y_1, y_2, \ldots, y_n)$ is a response variable vector, $X_{n\times p} = [\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n]'$ represents non-stochastic predetermined variables. Further, we assume that ε is a $(n\times 1)$ vector of random noise term that is distributed as Gaussian. That is,

$$\varepsilon \sim N(0, \sigma^2 I). \tag{2}$$

The log likelihood of $(y_1, y_2, \ldots, y_n)$, conditional on $(\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n)$, under Guassian assumption is given by

$$\log L(\beta, \sigma^2) = -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log(\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \mathbf{x}_i'\beta). \tag{3}$$

The least-squares method is used to estimate the unknown β by minimizing the residual sum of squares

$$\hat{\beta}_{OLS} = \arg\min_{\beta} \|y - X\beta\|_2^2 . \tag{4}$$

Under the Gauss-Markov assumptions, specifically, X is of full rank, the least-squares estimator of β is

$$\hat{\beta}_{OLS} = (X'X)^{-1}X'y \tag{5}$$

with $\mathbb{E}\left(\hat{\beta}_{OLS}\right) = \beta$ and the covariance matrix $Cov\left(\hat{\beta}_{OLS}\right) = \sigma^2 (X'X)^{-1}$.

If the entries of $\varepsilon_1 \ldots \varepsilon_n$ of the noise vector ε are uncorrelated and have common variance σ^2 , then

$$\mathbb{E}\left\|X\left(\hat{\beta}_{OLS} - \beta\right)\right\|_2^2 = \sigma^2 p, \tag{6}$$

where $\| X\left(\hat{\beta}_{OLS} - \beta\right) \|_2^2 / n$ is referred to as the *prediction error*.

Because of the simplicity, unbiasedness, and having a minimum variance, the least square estimator is widely applied to estimate coefficients and the covariance. However, Godfrey (1991) and White (1994) pointed out that least squares estimator can be affected when the linear model is misspecified. Misspecification occurs when there is high collinearity, heteroskedasticity, and autocorrelation. Following the research of Huber (1967), White (1982), and Kauermann and Carroll (2001), we use the "robust variance matrix" estimation to deal with misspecification.

We define the two forms of the Fisher information matrix to check the misspecification of classical regression models. From Bozdogan (2004), these are given as follows.

The Hessian form of the estimated Fisher information matrix (in inner-product form) is

$$\widehat{\mathcal{F}} = \begin{bmatrix} \frac{1}{\widehat{\sigma}^2}(X'X) & 0 \\ 0' & \frac{n}{2\widehat{\sigma}^4} \end{bmatrix}, \tag{7}$$

and the estimated outer-product form of the Fisher information matrix is

$$\widehat{\mathcal{R}} = \begin{bmatrix} \frac{1}{\widehat{\sigma}^4}X'D^2X & X'\mathbf{1}\frac{Sk}{2\widehat{\sigma}^3} \\ (X'\mathbf{1}\frac{Sk}{2\widehat{\sigma}^3})' & \frac{(n-q)(Kt-1)}{4\widehat{\sigma}^4} \end{bmatrix}. \tag{8}$$

where $D = diag(\hat{\varepsilon}_1, ..., \hat{\varepsilon}_n)$, and Sk denotes the estimated residual skewness, and Kt is the estimated kurtosis. The symbol $\mathbf{1}$ is a $(n \times 1)$ vector of ones. Sk and Kt are given by

$$\begin{aligned} Sk &= \frac{(\frac{1}{n}\sum_{i=1}^{n}\widehat{\varepsilon}_i^3)}{\widehat{\sigma}^3}, \text{and} \\ Kt &= \frac{(\frac{1}{n}\sum_{i=1}^{n}\widehat{\varepsilon}_i^4)}{\widehat{\sigma}^4}. \end{aligned} \tag{9}$$

Under model misspecification, we denote the sandwich covariance matrix as

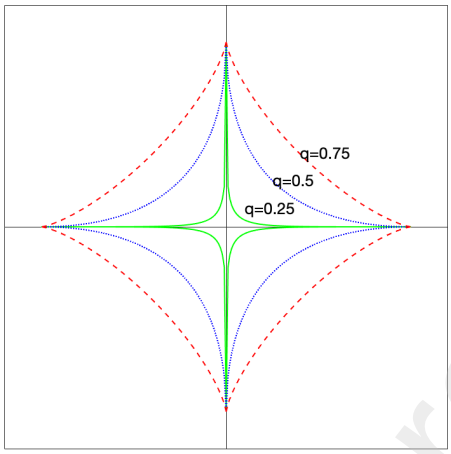


Figure 1: Plot of the penalty $\sum_{j=1}^{p} |\beta_j|^q$ for different values of $0 < q < 1$.

$$Cov(\beta, \sigma^{2*}_k) = \mathcal{F}^{-1}\mathcal{R}\mathcal{F}^{-1}. \tag{10}$$

The consistent covariance estimator of the sandwich matrix is given by

$$\widehat{Cov}(\hat{\beta}, \hat{\sigma}^2)_{\texttt{Miss}} = \hat{\mathcal{F}}^{-1}\hat{\mathcal{R}}\hat{\mathcal{F}}^{-1}. \tag{11}$$

2.2. *Sparse Bridge Regression Under Misspecification*

One of the most popular classes of estimators is the L_q-regularized least squares (LQLS), known as the Bridge regression (BR). Frank and Friedman (1993) proposed the Bridge regression (BR) with no solution.

The optimization of Bridge Regression is through minimizing both the residual sum of squares and the penalty term given by

$$\hat{\beta}_{Bridge} = arg\,min \left\{ \| y - X\beta \|^2 + \lambda \sum_{j=1}^{p} \left|\beta_j\right|^q \right\}, \tag{12}$$

where $\| \cdot \|$ is the L_2 norm of the vector, $\lambda > 0$ is the tuning parameter, $q > 0$ is the shrinkage or bridge parameter, and $\hat{\beta}_{Bridge}$ is called the bridge estimator.

When $0 < q \leq 1$, the Bridge regression (BR) tends to shrinkage the variable coefficients leading to variable selection and estimation simultaneously. It is therefore called the Sparse Bridge Regression (SBR), in which the feature of sparsity is the focus of the paper. In Figure 1 we show the structure of the penalty $\sum |\beta_j|^q$ for different values of $0 < q < 1$.

We note that any choice of q in the interval $(0, 1]$ can generate sparse coefficients, whereas the model is non-sparse if $q > 1$. When $q = 0$, the penalty term $\| \beta \|^0$ gives many non-zero coefficients in β, which

corresponds to L_0 norm, known as the "pseudo-norm" that enforces model sparsity. As the regularization parameter λ increases, $\| \beta \|^0$ decreases, and the solution is more sparse, which can be observed from Figure 1. However, the penalty function is not continuous, and the optimization problem is challenging to solve and generally leads to combinatorial and intractable solutions by Natarajan Natarajan (1995). In practice, the L_0 pseudo-norm is relaxed to relieve the computational difficulties while preserving sparsity.

When $q = 1$ in bridge regression (BR), the SBR model is equivalent to the so-called Least Absolute Shrinkage and Selection Operator (LASSO) model. LASSO penalty has been an important research area over the past twenty years by Tibshirani (1996). Assuming a squared error loss function, $\hat{\beta}_{LASSO}$ minimizes

$$\hat{\beta}_{LASSO} = \arg\min_{\beta} \|y - X\beta\|^2 + \lambda\|\beta\|_1. \tag{13}$$

LASSO automatically identifies an essential subset of predictors that constitute a simpler predictive model for the response y. However, it sacrifices some estimation bias for a reduction in variance and better interpretability. The LASSO penalty is not strictly convex. Hence LASSO estimates generally do not assign identical coefficients to predictors that are perfectly correlated Zou and Hastie Zou and Hastie (2005). Instead, LASSO tends to identify one variable from a category of highly correlated predictors that lead to inaccurate model interpretation.

When $q = 2$, we obtain the Ridge Regression (RR) introduced by Hoerl and Kennard (1970b,a). Assuming a squared error loss function, $\hat{\beta}_{Ridge}$ minimizes:

$$\hat{\beta}_{Ridge} = \arg\min_{\beta} \|y - X\beta\|^2 + \lambda \|\beta\|_2^2 . \tag{14}$$

Ridge Regression tends to retain all the predictors in the model. Thus the model cannot maintain the sparsity.

When $q > 2$, the true coefficient vector would be scattered in the directions oblique to the variable axis. However, when we set $0 < q < 2$, the true coefficients are aligned with the axis. In general, one can set $q \in (0, \infty]$. However, when $q > 2$, bridge regression does not provide new insights for sparsity or variable selection.

As previously pointed out, the bridge regression belongs to a broad class of the penalized regression model that deserves close attention and further study for improvements.

During the past two decades, in the literature, we see that there has been active research and studies toward understanding the mechanism of the bridge regression since it was initially introduced by Frank and Friedman (1993) without a provided solution. For example, Fu (1998) was one of the first who studied the bridge estimator and developed an algorithm to solve the bridge estimator for any fixed $q \geq 1$. Fu (1998) derived the covariance of the bridge estimator using the delta method. Knight and Fu (2000) later studied the asymptotic properties of the bridge estimator when the number of predictors, p, is finite or fixed. These authors showed that, for $0 < q \leq 1$, under certain regularity conditions, when the actual value of the parameter is zero, the limiting distributions of bridge estimators can have positive probability mass at zero. Huang et al. (2008) extended the work of Knight and Fu (2000) to study the asymptotic properties of bridge estimators for $0 < q < 1$, given the situation that the number of variables p might be larger than the sample size n. Huang et al. (2008) showed that for $0 < q < 1$, the bridge regression estimators can correctly select variables with nonzero coefficients. The conclusion is derived under the assumption that all the nonzero coefficients $\left(\hat{\beta}_j \neq 0\right)$ share the same asymptotic distribution and all the predictors with zero coefficients were known in advance. Thus these studies showed that the bridge estimator for $0 < q \leq 1$ provides a way to combine variable selection and parameter estimation at the same time.

There are many other studies published on bridge regression. For space considerations, we will not cite all these research papers but do acknowledge their contributions.

Coupled with the misspecification problem in the SBR model, when we have wide data sets, that is, when $p \gg n$ ("large p, small n"), to our knowledge, their is no research work carried out for the SBR model.

Under the misspecified SBR model, following the work of Bozdogan (2000, 2004) and generalizing the results of Konishi et al. (2004), and Kawano (2014), we provide the analytical form of the *sandwich covariance matrix* for the SBR model to guard against misspecification. We are only interested in those cases where the probabilistic assumption is away from the Gaussian distribution under the SBR framework.

To this end, let $\mathcal{A} = \{j : \hat{\beta}_j \neq 0\}$ be an active set of the coefficients β, and $\mathcal{A}_C = \{j : \hat{\beta}_j = 0\}$ be the non-active set or number of coefficients equal to zero.

From Kawano (2014), the estimated Fisher information $\hat{\mathcal{F}}$, in Hessian form, is a $(|\mathcal{A}| + 1) \times (|\mathcal{A}| + 1)$ matrix given by

$$\hat{\mathcal{F}} = \frac{1}{n\hat{\sigma}^2}\begin{bmatrix} X'_{\mathcal{A}}X_{\mathcal{A}} + n\lambda\hat{\sigma}^2 q(q-1)K_1 & \frac{1}{\hat{\sigma}^2}X'_{\mathcal{A}}D\mathbf{1}_n \\ \frac{1}{\hat{\sigma}^2}\mathbf{1}'_n DX_{\mathcal{A}} & \frac{n}{2\hat{\sigma}^2} \end{bmatrix}, \tag{15}$$

where $D = diag(\hat{\varepsilon}_1, ..., \hat{\varepsilon}_n)$ is a $n \times n$, and $K_1 = diag(|\hat{\beta}_{k_1}|^{q-1}/2, ..., |\hat{\beta}_{k_r}|^{q-r}/2)$ is a $|\mathcal{A}| \times |\mathcal{A}|$ diagonal matrix.

The estimated Fisher information matrix in outer-product form is given by

$$\hat{\mathcal{R}} = \frac{1}{n\hat{\sigma}^2}\begin{bmatrix} \frac{1}{\hat{\sigma}^2}X'_{\mathcal{A}}D^2X_{\mathcal{A}} - \lambda K_2\mathbf{1}_n^T DX_{\mathcal{A}} & \frac{1}{2\hat{\sigma}^4}X'_{\mathcal{A}}D^3\mathbf{1}_n - \frac{1}{2\hat{\sigma}^2}X'_{\mathcal{A}}D\mathbf{1}_n \\ \frac{1}{2\hat{\sigma}^4}\mathbf{1}'_n D^3X_{\mathcal{A}} - \frac{1}{2\hat{\sigma}^2}\mathbf{1}_n DX_{\mathcal{A}} & \frac{1}{4\hat{\sigma}^6}\mathbf{1}'_n D^4\mathbf{1}_n - \frac{n}{4\hat{\sigma}^2} \end{bmatrix}, \tag{16}$$

where, K_2 is also a $|\mathcal{A}| \times |\mathcal{A}|$ diagonal matrix given by

$$K_2 = diag(\frac{|\hat{\beta}_{k_1}|^{q-1} sgn(\hat{\beta}_{k_1})}{2}, ..., \frac{|\hat{\beta}_{k_r}|^{q-1} sgn(\hat{\beta}_{k_r})}{2}), \tag{17}$$

and where *sgn* denotes the sign function.

Following the work of Huber (1967) and White (1982), under misspecification, we construct the *sandwich covariance matrix* for the SBR model given by

$$\widehat{Cov}(\hat{\beta}, \hat{\sigma}^2)_{\texttt{Miss}} = \hat{\mathcal{F}}^{-1}\hat{\mathcal{R}}\hat{\mathcal{F}}^{-1}, \tag{18}$$

which is the correct form of the covariance matrix whether the model is correct or not. This is form of the covariance matrix we should be using in general, and not just $\hat{\mathcal{F}}^{-1}$. We note that using the sandwich covariance matrix in equation (11) for the usual regression model and the equation in (18) for the SBR model, we penalize the presence of skewness and kurtosis in the data set. This is another advantage of the sandwich covariance matrix in model selection.

3. Misspecification Resistant Information Complexity

The issue of model misspecification has been a challenge for statisticians during the model fitting and selection process. In order to drive valid variable selection, we generalize the information complexity criterion given that the Gaussian assumption is not valid for the original input data. We formalize the work of Bozdogan (2000, 2004); Koc and Bozdogan (2015) to the case of SBR model under misspecification.

$$\begin{aligned} \mathcal{ICOMP}(\text{SBR Model})_{\texttt{Miss}} &= -2logL(\widehat{\beta}, \widehat{\sigma}^2|y) + 2C_1(\widehat{Cov}(\widehat{\beta}, \widehat{\sigma}^2)_{\texttt{Miss}}) \\ &= -2logL(\widehat{\beta}, \widehat{\sigma}^2|y) + 2C_1(\hat{\mathcal{F}}^{-1}\hat{\mathcal{R}}\hat{\mathcal{F}}^{-1}), \end{aligned} \tag{19}$$

where $L(\widehat{\beta}, \widehat{\sigma}^2|y)$ is the maximized likelihood function of the probability density function $f\left(x \mid \beta, \sigma^2\right)$. The second term $C_1(\widehat{Cov}(\widehat{\beta}, \widehat{\sigma}^2)_{\texttt{Miss}})$ denotes the maximal information-theoretic complexity of the sandwich covariance matrix $\widehat{Cov}(\widehat{\beta}, \widehat{\sigma}^2)_{\texttt{Miss}} = \hat{\mathcal{F}}^{-1}\hat{\mathcal{R}}\hat{\mathcal{F}}^{-1}$ given by

$$C_1(\widehat{Cov}(\widehat{\theta})_{\texttt{Miss}}) := \frac{s}{2}\log\left(\frac{tr\left((\widehat{Cov}(\widehat{\beta}, \widehat{\sigma}^2)_{\texttt{Miss}})\right)}{s}\right) - \frac{1}{2}\log \mid (\widehat{Cov}(\widehat{\beta}, \widehat{\sigma}^2)_{\texttt{Miss}}) \mid, \tag{20}$$

where $s := rank\left(\widehat{Cov}(\widehat{\beta}, \widehat{\sigma}^2)_{\texttt{Miss}}\right)$. Then we have

$$\mathcal{ICOMP}(\text{SBR Model})_{\texttt{Correct}} = -2logL(\widehat{\beta}, \widehat{\sigma}^2|y) + 2C_1(\hat{\mathcal{F}}^{-1}). \tag{21}$$

where $\mathcal{F}^{-1}$ is from equation (15).

Note that when

$$\mathcal{ICOMP}(\text{SBR Model})_{\texttt{Miss}} \neq \mathcal{ICOMP}(\text{SBR Model})_{\texttt{Correct}}, \tag{22}$$

or, equivalently, if

$$C_1(\widehat{Cov}(\widehat{\theta})_{\texttt{Miss}}) \neq C_1(\hat{\mathcal{F}}^{-1}), \tag{23}$$

then we say that the two models are misspecified.

There are other forms of $\mathcal{ICOMP}$. These different forms of information complexity are derived from both the likelihood function and the covariance matrix of the estimated parameters of the proposed model. The various derivation of $\mathcal{ICOMP}$ also takes consideration of datasets from finite sampling distributions and Bayesian justification, widely used in linear and nonlinear models. We refer the readers more on this and other applications of $\mathcal{ICOMP}$ to **?**.

Throughout the following sections of the paper, we use and score
$\mathcal{ICOMP}(Model)_{\texttt{Miss}}$ as our fitness function for variable selection in the SBR model under misspecification.

4. New Proposed Estimation of the Sparse Bridge Regression Model Parameters

In this section, we present our new proposed estimation of the SBR model parameters to answer the following questions.

- How do we efficiently estimate the regularization (or tuning) parameter λ?
- What is the optimal value of bridge (or shrinkage) parameter $q \in (0, 1]$?
- How do we simultaneously estimate λ , q, and the SBR coefficients β?

Researchers typically utilize the grid search method or cross-validation method to search for hyperparameters. Here, we present an alternative search method that combines simulated annealing and particle swarm optimization methods to quickly and efficiently estimate the values of the hyperparameters in the SBR model.

Table 1: Parameters of SA-based Particle Swarm Optimization Algorithm.

Parameters	**Values**
Fitness Function	$\mathcal{ICOMP}_{\texttt{Miss}}$
Size of Population (N)	50
c_1	2.05
c_2	2.05
Number of Generations	50
Dimension of Inputs	2

4.1. Simulated Annealing and Particle Swarm Optimization: SA-PSO

In this subsection, we propose an alternative approach to select the optimal values of the regularization (or tuning) parameter λ, and bridge (or shrinkage) parameter q, simultaneously in the penalty function of the SBR model by hybridizing Simulated Annealing (SA) and Particle Swarm Optimization (PSO), abbreviated as SA-PSO. Simulated Annealing (SA) is a method for solving unconstrained and bound-constrained optimization problems. Particle Swarm Optimization (PSO), on the other hand, is a new method of evolutionary computation. Since both SA and PSO are highly efficient optimization algorithms, we borrow the strengths from these two algorithms and, by hybridizing them as SA-PSO, one can create a new family of optimization algorithms for function optimization in high dimensions. This proposed approach enables us to objectively identify appropriate values of the adjusted hyper-parameters of the SBR model.

PSO can find globally optimal results, and SA can find local search optimization. By hybridization of the two, the SA-PSO algorithm achieves a balance between global and local optimization. This idea of a hybridized algorithm was introduced by Wang and Li (2004) and further developed by Zhan et al. (2009). Hybridization provides a better global search efficiency that narrows down the search space and speeds up the convergence rate.

The full implementation of SA-PSO is provided in Algorithms 2, 3 and 4. Algorithm 2 shows the initialization of SA-PSO analogs to the genetic algorithm (GA) and sets up the parameters in Table 1. Algorithms 3 and 4 are the main bodies of SA-PSO. For each generation of particles, we select individuals with the best performance and store the coordinates (estimates of each variable coefficient). By comparing the best candidate with other particles, we calculate the velocity to update each particle's position in the next generation. The position of each candidate would be updated randomly to ensure the searching result is a globally optimal rather than locally optimal result.

Throughout the numerical examples in this paper, we set the parameter values of SA-PSO algorithm given in Table 1 to optimally learn the parameter values of the SBR model by minimizing $\mathcal{ICOMP}_{\texttt{Miss}}$. In Table 1, the user determines the size of the population and the number of generations. However, c_1 and c_2 are recommended by Clerc and Kennedy (2002) based on repeated experimentation. For further details, we refer the readers to their paper.

4.2. Smoothed Covariance Estimators (SCEs)

The last two decades have witnessed an increase in research on penalized regression methods, including the bridge regression, due to the emergence of high-dimensional data with different structures and increasing complexity with a small number of observations and with high-dimensions. Such situations created new challenges to existing statistical models and covariance matrix estimation when $p \gg n$. In these scenarios, obtaining the inverse covariance matrix under singularity conditions became a severe challenge to statistical modelers.

Many smoothed covariance estimators have been proposed, including Ledoit-Wolf (LW) estimator by Ledoit and Wolf (2004) and the optimal shrinkage covariance estimator by Ollila (2017). Chen et al. (2010) improved the LW estimator by conditioning on a sufficient statistic, and proposed the oracle approximating shrinkage (OAS) estimator. Bozdogan and Pamukcu (2016) and Mohebbi et al. (2019) gave a list of Smoothed Covariance Estimators. In this paper, we will use only the OAS estimator to make the Gram matrix well-conditioned for under-sample data. Certainly, the performance of other Smoothed Covariance Estimators needs a separate study. For space considerations, we did not pursue this here.

Let $\widehat{\mathbf{S}} = \frac{1}{n}\sum_{i=1}^{n}\mathbf{x}_i\mathbf{x}_i^T$ the gram matrix, then

$$\hat{\rho}_{OAS} = min\left(1, \frac{(1-2/p)tr(\widehat{\mathbf{S}}^2) + tr^2(\widehat{\mathbf{S}})}{(n+1-\frac{2}{p})(tr(\widehat{\mathbf{S}}^2) - tr^2(\widehat{\mathbf{S}})/p)}\right). \tag{24}$$

Chen et al. (2010) proved that $\hat{\rho}_{OAS}$ is bounded within $(0, 1]$. Then, the final form of the *SCE* is given by

$$\widehat{\mathbf{\Sigma}}_{SCE} = (1-\hat{\rho}_{OAS})\widehat{\mathbf{S}} + \hat{\rho}_{OAS}\frac{tr(\widehat{\mathbf{S}})}{p}I_p. \tag{25}$$

We use $\widehat{\mathbf{\Sigma}}_{SCE}$ in our computational results.

4.3. Regularized Local Quadratic Approximation (RLQA)

When we select values of λ and q, we must also find an efficient algorithm to calculate regression coefficients. When $0 < q \leq 1$, the sparse bridge penalty is non-concave, which means traditional convex optimization could not be applied. Fan and Li (2001) proposed the Local Quadratic Approximation (LQA) method to estimate the model coefficients. The method can determine the estimated coefficients within a limited number of steps by repeatedly using the Gram matrix. However, the Gram matrix is not invertible under singularity situations. As a result, the LQA might not be reliable when $p \gg n$.

We propose replacing the Gram matrix with one of the smoothed covariance estimators to regularize the Local Quadratic Approximation (LQA) method. We call this the Regularized Local Quadratic Approximation (RLQA) method. Here, we only considered using the oracle approximating shrinkage (OAS) estimator. Furthermore, our computational method is flexible enough to study the results of the other SCEs, which can be of further study in the SBR model.

In what follows, we show the steps of RLQA by modifying the steps of LQA in Fan and Li (2001) and Kawano (2014) to obtain the estimated values of the SBR model coefficients.

The main purpose of the algorithm is to transform a non-concave optimization problem to a concave one. Under some mild conditions, we initialize the estimated values of coefficients first, which are denoted as $\beta_{\texttt{OLD}} = (\beta_{1,\texttt{OLD}}, \beta_{2,\texttt{OLD}}, ..., \beta_{p,\texttt{OLD}})$. Then we update the values of $\beta_{\texttt{NEW}}$ by using $\beta_{\texttt{OLD}}$ and partial derivatives.

We can write the parameter update function as

$$\widehat{\beta}_{\texttt{NEW}} = (\widehat{\mathbf{\Sigma}}_{SCE} + \mathbf{\Sigma}_{\lambda,q}(\widehat{\beta}_{\texttt{OLD}}))^{-1}\mathbf{X}'\mathbf{y}. \tag{26}$$

In equation (26), once a coefficient is set to zero, then in all the subsequent steps, the estimated values would be zero. Following Kawano (2014), our proposed Regularized Local Quadratic Approximation (RLQA) steps are outlined in Algorithm 1.

Algorithm 1 Regularized Local Quadratic Approximation (RLQA).

Input: $\lambda > 0$, $0 < q < 1$, X, y, $\delta = 10^{-4}$, $\gamma = 10^{-5}$, $\texttt{maxSteps} = 1000$
$\hat{\beta}_{\texttt{OLD}} = \hat{\beta}_{\texttt{NEW}} = (\widehat{\mathbf{\Sigma}}_{SCE} + n\gamma I_p)^{-1}\mathbf{X}'\mathbf{y}$
▷ To improve the calculation efficiency, the initialized estimated coefficients should be close to the MLE estimation by Fan and Li (2001).
$\hat{\sigma}^2_{\texttt{OLD}} = \hat{\sigma}^2_{\texttt{NEW}} = \frac{1}{n}(\mathbf{y} - \mathbf{X}\hat{\beta}_{\texttt{NEW}})'(\mathbf{y} - \mathbf{X}\hat{\beta}_{\texttt{NEW}})$
for $i = 1 : \texttt{maxSteps}$ **do**
  $\mathbf{\Sigma}_{\lambda,q}(\widehat{\beta}_{\texttt{OLD}}) = diag(n\lambda\hat{\sigma}^2_{\texttt{OLD}}q|\hat{\beta}_{1,\texttt{OLD}}|^{q-2}/4, ..., n\lambda\hat{\sigma}^2_{\texttt{OLD}}q|\hat{\beta}_{p,\texttt{OLD}}|^{q-2}/4)$
  $\widehat{\beta}_{\texttt{NEW}} = (\widehat{\mathbf{\Sigma}}_{SCE} + \mathbf{\Sigma}_{\lambda,q}(\widehat{\beta}_{\texttt{OLD}}))^{-1}\mathbf{X}'\mathbf{y}$
  ▷ $\widehat{\mathbf{\Sigma}}_{SCE}$ represents the smoothed covariance estimators
  $\sigma^2_{\texttt{NEW}} = \frac{1}{n}(\mathbf{y} - \mathbf{X}\hat{\beta}_{\texttt{NEW}})'(\mathbf{y} - \mathbf{X}\hat{\beta}_{\texttt{NEW}})$
  if $max(|\widehat{\beta}_{\texttt{NEW}} - \widehat{\beta}_{\texttt{OLD}}|) < \delta$ **then**
    Break
  end if
end for
return $\widehat{\beta}_{\texttt{NEW}}$

We define $\mathcal{ICOMP}_{\texttt{Miss}}$ for the SBR model given by

$$\mathcal{ICOMP}_{\texttt{Miss}}(SBR) = -2log\widehat{L(\theta|y)} + 2C_1(\widehat{Cov}(\hat{\sigma}^2, \hat{\beta})_{\texttt{Miss}}), \tag{27}$$

where both the likelihood function and sandwich covariance matrix in closed-form expression are given in Section 2.

To summarize, we use the SA-PSO algorithm and the $\mathcal{ICOMP}$ criteria to find hyperparameters. Then, with the λ and q in hand, we can use the RLQA algorithm to estimate the parameter. As a sparse model, SBR selects variables by considering only those parameters with non-zero coefficients.

In our investigation, we cannot find any existing criteria in the literature that handles misspecification, except perhaps the Takeuchi (1976) information criterion (TIC), AIC_T, also known as $GAIC$ discussed in Kawano (2012, 2014), where the $2tr(\hat{\mathcal{F}}^{-1}\hat{\mathcal{R}})$ is used in the penalty term of $GAIC$.

Next, we show our numerical examples.

5. Monte Carlo Simulation Results

In this section, we compare the performance of *the SA-PSO* based SBR model to the LASSO regression using simulated datasets. For LASSO, there are multiple methods for choosing λ adaptively from the data. These methods include, but are not limited to: 1) performing grid-search, and 2) using $(min(eigenvalues(X'X)))^{-1}$ values and other parameter tuning techniques. In our computations, we use the OAS estimator in equation (25) for tuning λ values in LASSO. All our computations are carried out using a newly developed MATLAB® computational routines.

5.1. Simulation Protocol

In our simulations, we focus on comparing performances on variable selection under different scenarios of complex data structures. In order to create different levels of collinearity among the variables, we use the following simulation protocol. This simulation protocol considers both pairwise correlations among

actual variables and autoregressive (AR) structure among irrelevant variables. Roozbeh (2018) and Mohebbi et al. (2019) suggested and used this simulation protocol to obtain X_{ij}, reflecting the convoluted collinear relationships among variables for $j = 1, 2, ..., p$ and $i = 1, 2, ..., n$ observations. We denote independent explanatory variables by Z_{ij} as follows.

$$X_{ij} = (1 - \gamma^2)^{\frac{1}{2}} Z_{ij} + \gamma Z_{ip}, \tag{28}$$

where Z_{ij} has zero mean and unit variance. To answer the question of whether SBR outperforms the LASSO in variable selection, given the complicated data structures, we fix a set of variables ($X^{(1)}$) as the ground truth and $X^{(2)}$ as redundant set of variables in

$$y = X^{(1)}\beta_1 + X^{(2)}\beta_2 + \epsilon,$$

where

$$\epsilon \sim N_n(0, \sigma^2\Sigma),$$

and

$$\Sigma_{i,j} = exp(-7|i - j|).$$

The true coefficient values is set as $\beta_1 = (-1.25, 1, 2.5, 4, -3, -5)^T$, where $p_1 = 6$ and $n = 100$ observations. To create the sparse model with redundant variables, $p_2 = p - p_1, \beta_2$ is sampled from multivariate Gaussian distribution $N_{p_2}(0, 0.01I_p)$. In our simulation, we vary the total number of predictor variable sets $p = 32$ to $10,000$.

5.2. Effects of Smoothed Covariance Estimator

Next, we analyze the effects of oracle approximating shrinkage (OAS) estimator by answering two questions,

1. How does OAS help adjust the eigenvalues of the Gram matrix when $p \gg n$?
2. Can OAS stabilize the calculation of SBR coefficients when $p \gg n$?

We first illustrate the eigenvalues of the Gram matrix with and without utilizing the OAS smoothed covariance estimator. In Figure 2, we see that the OAS method smooths the eigenvalue and regularizes them. In other words, OAS is a data-adaptive method for dealing with ill-conditioned Gram matrix $\hat{\Sigma} := (X'X)$. One can use other SCEs as well.

Figure 3 shows the visualization of the SBR coefficients given different values of λ. The first panel displays the results using the usual LQA algorithm. The second panel shows how the RLQA algorithm performs with the usage of the OAS smoothed covariance estimator. We can see that the change of coefficient values become smoother when using the OAS, and estimated results are stable when $log(\lambda) \leq -4$.

In Figure 4, we show the results of our approach and demonstrate why one needs to regularize further the LQA algorithm employing SCE, using the data generated from our simulation protocol. Looking at Figure 4, we can see that the utilization of the usual Gram matrix in the SBR model, adopted by many authors, exhibits an erratic behavior and is not smooth because there are singularities in the Gram matrix. However,

when we use one of the smoothed covariance estimators and its shrinkage intensity, $\hat{\rho}$, for example, using the $\hat{\rho}$ of OAS, we see much improved regularized results. Therefore, we propose and recommend the usage of the SCEs within the SBR model for the best subset selection of the predictors.

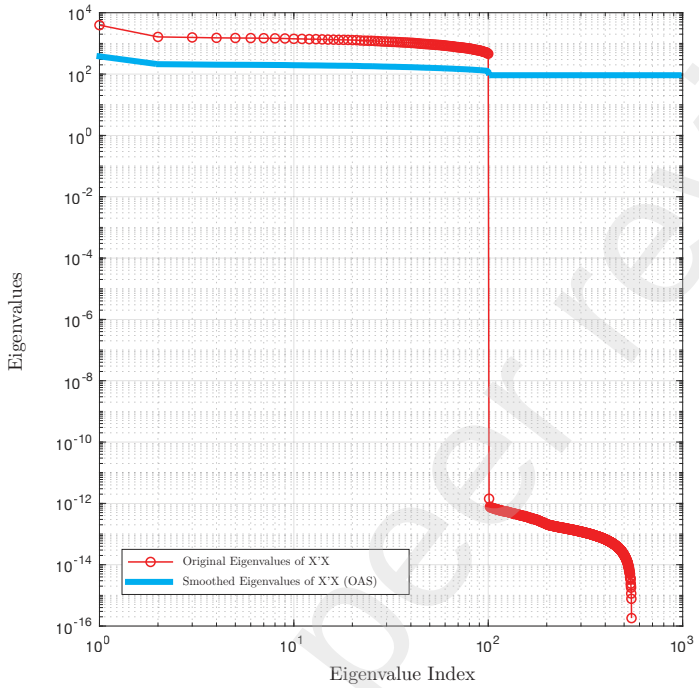
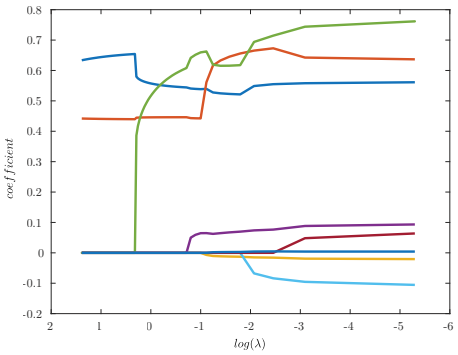
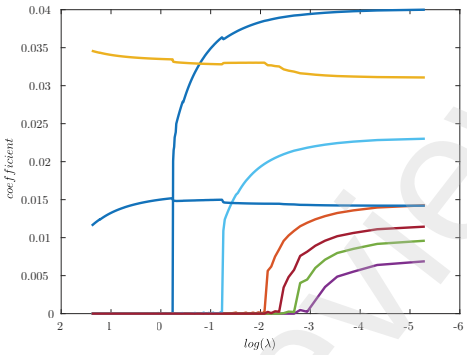


Figure 2: Comparison of eigenvalues w/o smoothed covariance estimator (SCE) of the Gram matrix ($X'X$) for the simulated dataset with $\gamma = 0.3$, $n = 100$ and $p = 1000$. Note that the eigenvalues of SCE is smoother after index = 100.



(a) Coefficients of variables from original ($X'X$).



(b) Coefficients of variables with SCE estimator of ($X'X$).

Figure 3: Estimated coefficients w/o smoothed covariance estimators (SCE) in SBR using LQA algorithm.

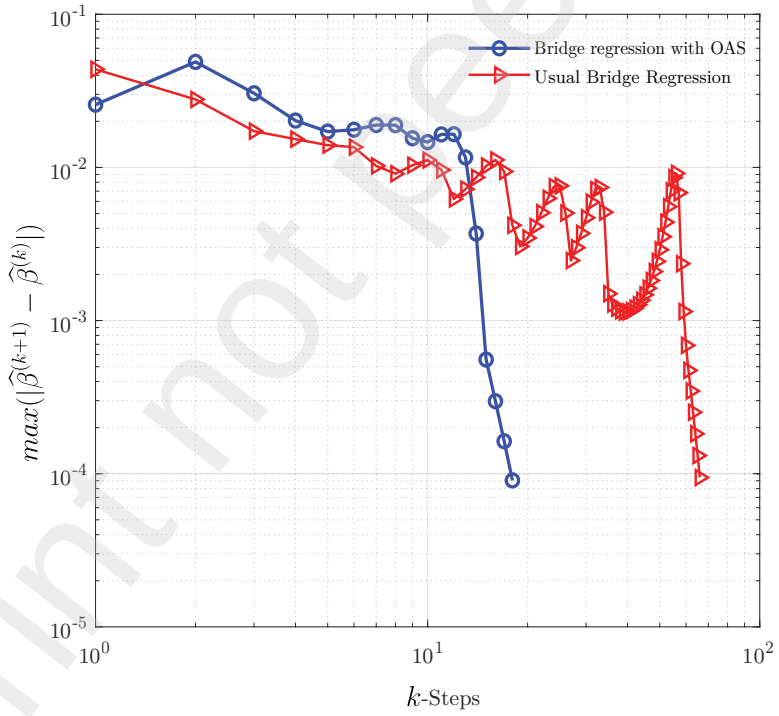


Figure 4: The behavior of $max(\widehat{\beta}^{(k+1)} - \widehat{\beta}^{(k)}) < \delta = 10^{-4}$.

5.3. Comparative Performance of Sparse Bridge Regression with Lasso

In this section, we compare and study the performance metrics of variable selection between the SBR model and the LASSO. To this end, we generate multiple datasets using the simulation protocol in subsection 5.1. We fix the ground truth variables in each experiment and change the number of irrelevant variables $X^{(2)}$ through different simulated datasets.

The LASSO estimators are calculated using $\hat{\lambda} = \widehat{\rho}_{OAS}$. For SBR, we use the SA-PSO algorithm to find the optimal solution of the tuning parameters (λ, q). RLQA is adopted to calculate the SBR coefficient values. Table 1 lists the parameters of the SA-PSO algorithm.

We use four classification performance metrics, 1) Precision, 2) Accuracy, 3) Recall, and 4) F-1 Score to evaluate the classification performance of the proposed SBR model and LASSO. The accuracy measures the proportion of correctly selected variables among all the selected factors. Recall stands for choosing all the relevant variables within the dataset. To seek a balance between precision and recall, we include the F-1 score in our comparison.

$$\text{Precision} = \frac{\text{Number of correctly selected variables}}{\text{Number of selected variables}},$$

$$\text{Accuracy} = \frac{\text{Number of correctly selected variables + Number of corrected ommitted variables}}{\text{Total Number of variables}},$$

$$\text{Recall} = \frac{\text{Number of correctly selected variables}}{\text{Number of correctly selected variables + Number of erroneously omitted variables}},$$

$$\text{F-1 Score} = \frac{2 \cdot \text{Precision} \cdot \text{Recall}}{\text{Precision + Recall}}.$$

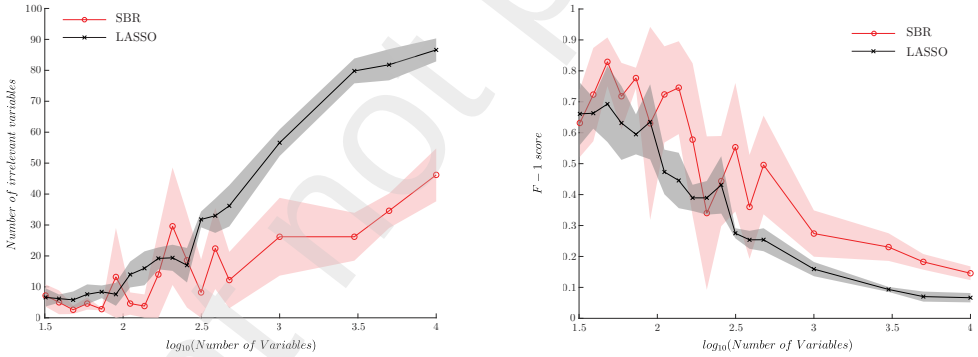
Results of the number of incorrectly selected variables (false positive) under SBR and LASSO are presented in Figure 5a. The gray color band indicates a 1SE (one standard error) region of LASSO, and the red color band indicates a 1SE (one standard error) region of the SBR model. A striking feature of Figure 5a is that, compared with LASSO, SBR selects fewer irrelevant variables as the number of irrelevant variable sets grows. When the number of variables increases to more than the number of observations ($n = 100$), LASSO selects two times the irrelevant variables than does the SBR model. In other words, SBR's performance is two times better than LASSO. The comparison clearly shows that despite its popularity, when $p \gg n$, the LASSO tends to over-shrink significant coefficients and include unimportant variables in order to compensate for this over-shrinkage Fan and Li (2001).

In terms of F-1 scores shown in Figure 5b, both LASSO and SBR decrease as the number of variables increases. When it comes to 10,000 variables with 100 observations in the dataset, the covariance matrix becomes highly ill-conditioned. Even though the penalized regression models (LASSO and SBR model) can still select the real coefficients, both models select some unnecessary variables in the final results. Since SBR has a more robust performance in false-positive detection (FPD), the F-1 scores of SBR also outperform the LASSO results.

In summary, from the results in Table 2, we see that the SA-PSO based SBR outperforms the LASSO in terms of false-positive rates and F-1 scores for all eighteen simulations.

Table 2: Comparison of SBR with LASSO as the number of variable sets increases.

	Precision		Accuracy		Recall		$F-1$ score	
Number of variables	Lasso	SBR	Lasso	SBR	Lasso	SBR	Lasso	SBR
32	0.502	0.485	0.794	0.769	1.000	0.967	0.661	**0.632**
39	0.498	0.603	0.841	0.867	1.000	0.967	0.663	**0.723**
48	0.545	0.716	0.879	0.946	1.000	1.000	0.693	**0.829**
59	0.474	0.576	0.871	0.919	1.000	0.967	0.631	**0.718**
73	0.426	0.668	0.885	0.956	1.000	0.933	0.595	**0.777**
90	0.477	0.625	0.916	0.844	1.000	0.867	0.635	**0.630**
111	0.313	0.618	0.874	0.955	1.000	0.933	0.473	**0.724**
137	0.291	0.666	0.883	0.968	1.000	0.900	0.446	**0.746**
168	0.243	0.456	0.886	0.917	1.000	1.000	0.389	**0.578**
207	0.243	0.268	0.906	0.852	1.000	0.833	0.389	**0.340**
256	0.280	0.316	0.934	0.926	1.000	0.933	0.431	**0.444**
315	0.159	0.553	0.899	0.968	1.000	0.667	0.275	**0.553**
388	0.147	0.250	0.914	0.940	0.933	0.833	0.254	**0.360**
477	0.146	0.405	0.924	0.972	1.000	0.833	0.254	**0.496**
1000	0.087	0.172	0.943	0.972	0.900	0.767	0.159	**0.274**
3000	0.050	0.141	0.973	0.991	0.700	0.667	0.093	**0.230**
5000	0.038	0.106	0.983	0.993	0.533	0.667	0.070	**0.183**
10000	0.036	0.082	0.991	0.995	0.533	0.667	0.067	**0.146**



(a) Number of incorrectly selected variables (false positive) under SBR and LASSO.

(b) Comparison of F-1 scores of SBR model with LASSO.

Figure 5: Comparison of Sparse Bridge Regression (SBR) with LASSO.

Table 3: Variable selection for riboflavin production data.

Model	Covariance Estimator	No. of selected variables
EN, Mohebbi et al.Mohebbi et al. (2019)	MLE	68
AEN, Mohebbi et al.Mohebbi et al. (2019)	OAS	74
AEN, Mohebbi et al.Mohebbi et al. (2019)	BCSE	62
Sparse Bridge Regression (SBR)	OAS	54

6. Real Data Example: Riboflavin (Vitamin B2) Dataset

In order to demonstrate the performance of SA-PSO based SBR model and to show the importance of tuning parameters on model selection, the Riboflavin data is analyzed in detail to identify the relationship between production rate and thousands of gene expression measures. Tännler et al. (2008) originally collected the Riboflavin dataset. Furthermore, Bühlmann et al. (2014) made it publicly available for research purposes.

Many researchers analyzed the Riboflavin production data as a benchmark dataset over the past several years. This dataset has $n = 71$ observations, and each observation includes $p = 4,088$ gene expressions (predictor variables). The riboflavin production rate in the logarithm scale is the response variable. Our primary purpose is to identify a subset of genes that is relevant to predict the Riboflavin production rate. We also want to maintain the sparsity of the model, ensuring the interpretation of gene expression data is clean and bright.

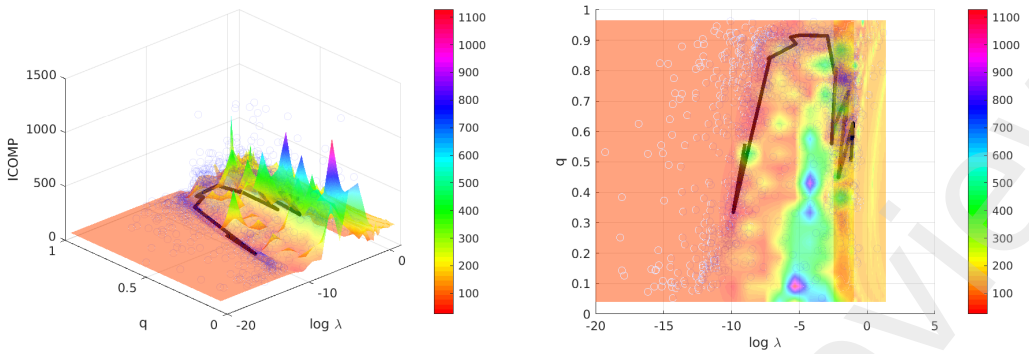
We now elaborate on the selection performance of SA-PSO based SBR for the Riboflavin production data.

Our analysis shows that the SA-PSO algorithm can search the model space and identifies the optimal values of (λ, q) automatically and efficiently. Figure 6 illustrates the model space of SBR for the Riboflavin production data. The first panel is the surface plot of $\mathcal{ICOMP}_{\texttt{Miss}}$ values over $(log(\lambda), q)$. Specifically, a ridge-like area is formed when $log(\lambda) \in [-7, -3]$. SA-PSO algorithm generates particles for each generation (marked as blue circles in Figure 6). It develops that particle swarms left behind ridges and spread with fluidity, as illustrated in the black line in both panels. The second panel shows the top-down view of the surface plot.

SA-PSO based SBR represents the right balance between precision and recall and identifies 54 relevant genes. Compared with other penalized regression models, such as the adaptive elastic net (AEN) based regression model reported by Mohebbi et al. (2019). In Table 3, we select fewer variables compare to EN, AEN models.

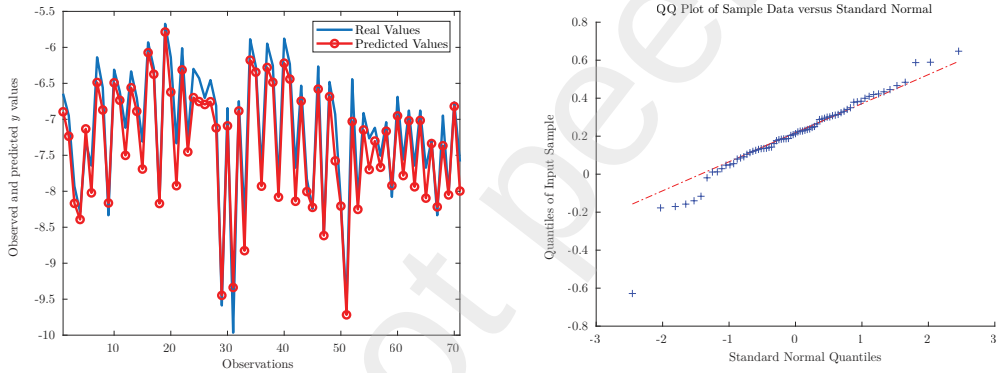
In Figure 7, we show our predicted y values against the real values in panel (a) and the QQ-plot of the residuals in panel (b). We can observe the shrinkage effects because some predicted values get closer to zero compared with real values. However, despite this shrinkage effect, the predicted values are still close to the original values. The QQ-plot of residuals also confirmed this observation.

Figure 8 shows the efficiency of the SA-PSO algorithm based on the $\mathcal{ICOMP}_{\texttt{Miss}}$ results. Compared with the grid-search method, SA-PSO is data-adaptive in searching for the optimal value of λ and q. As a by-product of SBR model, we present Figure 9 illustrating the hierarchical structure of selected genes based on the correlation matrix of the 54 relevant genes. This result can be used in building classification models.



(a) Surface Plot (b) Top-down view

Figure 6: ICOMP values of the entire model space from SBR of Riboflavin data using SA-PSO algorithm. The model space is illustrated over $(log(\lambda), q)$. The black solid line represents the path of the best particle over each generation of the SA-PSO algorithm.



(a) (b)

Figure 7: Plot of predicted y values against real values and QQ-plot of the residuals.

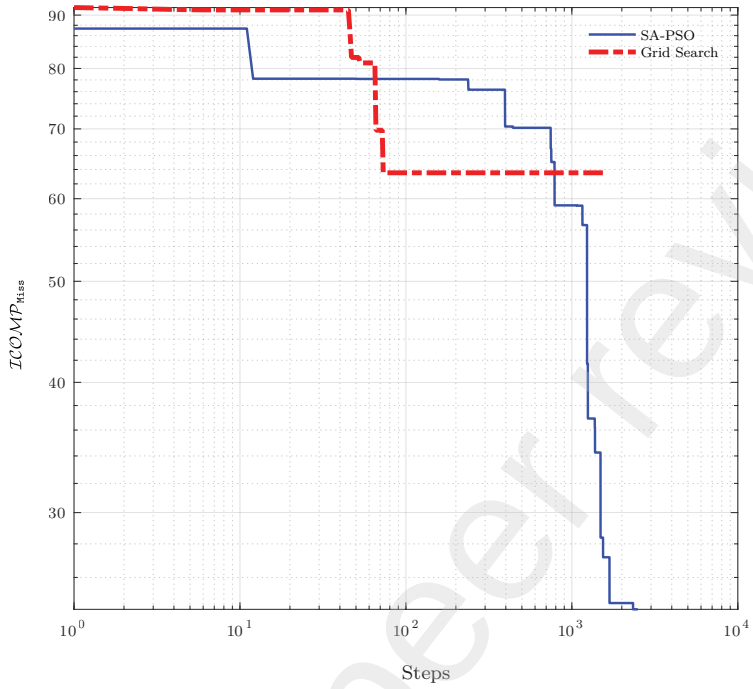


Figure 8: Comparison of SA-PSO and grid-search results for Riboflavin data for $0 < q \leq 1$, $0 < \lambda \leq 4$. This comparison is labeled on a 40×40 grid search and SA-PSO with population size 50 and 50 generations.

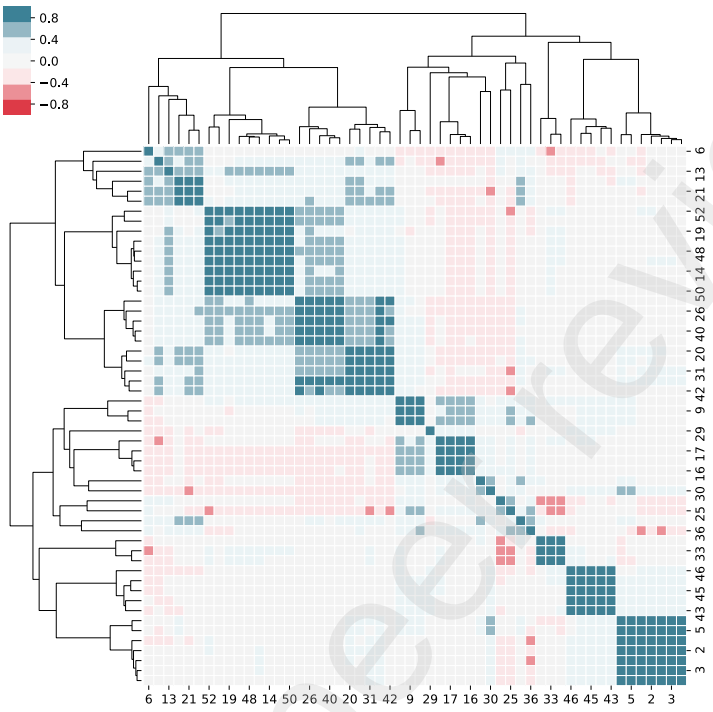


Figure 9: Hierarchically-clustered heatmap of selected genes. The heatmap plot identifies the correlational relationship between the selected 54 genes. The hierarchical tree shows the similarities and differences arising between selected genes of SBR.

7. Conclusions and Discussion

In this work, we introduced and proposed a novel Sparse Bridge Regression (SBR) model for subset selection of best predictors when the number of observations is less than the number of predictors. We developed the misspecification resistant information complexity ($\mathcal{ICOMP}_{\texttt{Miss}}$) criterion as a fitness function for subset selection to allow robust statistical inference and their proper interpretations for the Sparse Bridge Regression (SBR) model.

In the literature, researchers mostly ignore the model misspecification issue. To resolve this issue, we provided the closed-form expression of the robust covariance matrix by using the derived form of the celebrated Fisher information matrix $\mathcal{F}$ in the Hessian form, and the outer-product form $\mathcal{R}$. This robust covariance matrix is useful to calculate the standard errors of the model parameters for further frequentist inference.

We also proposed and used a novel hybridized Simulated Annealing (SA) and Particle Swarm Optimization (PSO) algorithm, SA-PSO. We showed that this new approach enables us to determine appropriate values of the estimated tuning parameter λ and the shrinkage parameter q jointly by eliminating the costly and time-consuming cross-validation (CV) in such problems.

We introduced one of the Smoothed Covariance Estimators (SCEs), namely the OAS covariance estimator, to regularize the commonly used Local Quadratic Approximation (LQA) algorithm. The regularized LQA algorithm (RLQA) and SCE can resolve the ill-conditioned Gram matrix.

Finally, the SBR model can select the best subset of predictors via the $\mathcal{ICOMP}_{\texttt{Miss}}$ criterion as our fitness function.

Under a large scale Monte Carlo misspecified simulation setting, we studied and compared the performance metrics of our proposed SBR approach to the LASSO. Through its application to a high-dimensional real dataset, we also presented some evidence that the SBR model is a worthy competitor to LASSO, to Elastic Net (EN) regression, and other penalized regression methods.

There are still unresolved problems that remain in SBR research deserving of further study. Areas, including the SBR model from a Bayesian perspective under misspecification, kernel SBR, and robust SBR, lack the rigorous solutions. We intend to conduct future research in some of these areas and will report our results under separate work. In the Appendix, we provide our proposed SA-PSO Algorithm, SA-PSO Main, and SA-PSO Update as our computational routines.

Appendix A. SA-PSO Algorithm

Algorithm 2 SA-PSO Initialization .

Input: N: size of population; M: number of generations; $\mathcal{ICOMP}_{\texttt{Miss}}$: fitness function; X, Y, the dataset
Initialize values of $v_0(\lambda)$, $v_0(q)$ and T.
$\Lambda_0 = \{\lambda_{i,0} | \lambda_{i,0} \in (0, 4), i = 1, 2, ..., N\}$
$Q_0 = \{q_{i,0} | q_{i,0} \in (0, 1), i = 1, 2, ..., N\}$
$F_0 = \{f_{i,0} | f_{i,0} = \mathcal{ICOMP}_{\texttt{Miss}}(\lambda_{i,0}, q_{i,0}, X, Y), i = 1, 2, ..., N\}$
$(\lambda^*, q^*) = \text{argmin}_{\Lambda_0, Q_0} \mathcal{ICOMP}_{\texttt{Miss}}(X, Y)$
$f^* = \mathcal{ICOMP}_{\texttt{Miss}}(\lambda^*, q^*, X, Y)$
$Y = (\Lambda_0, Q_0)$ ▷ $Y_i(\lambda)$ means λ value of i^{th} particle.
$P = F_0$ ▷ P_i records $\mathcal{ICOMP}_{\texttt{Miss}}$ value of the i^{th} particle.

Algorithm 3 SA-PSO Main.

After initializing the parameters using Algorithm SA-PSO (Initialization).
for $t = 1 : M$ **do**
    $K = \{k_i | k_i = \frac{Exp(-\frac{f_{i,t-1} - f^*}{T})}{\sum_{j=1}^{N} Exp(-\frac{f_{j,t-1} - f^*}{T})}\}$
    $threshold = rand()$ ▷ $rand() \in [0, 1]$ is a (pseudo)random number generator.
    for $e = 1 : N$ **do**
        if $threshold \leq sum(K(1 : e))$ **then** ▷ $sum(K(1 : e))$ represents summation from the first value to e^{th} value .
            $\lambda'_i = \lambda_{e,t-1}$
            $q'_i = q_{e,t-1}$
            break
        end if
        Run algorithm SA-PSO (Update)
    end for
end for

Algorithm 4 SA-PSO Update.

$C = c_1 + c_2$
$\chi = \frac{2}{|2-C-\sqrt{C^2-4C}|}$
for $i = 1 : N$ **do**
  $r_1 = rand()$, $r_2 = rand()$
  $v_{i,t}(\lambda) = \chi\{v_i(\lambda) + c_1 r_1(Y_i(\lambda) - \lambda_{i,t-1})\} + c_2 r_2(\lambda_i' - \lambda_{i,t-1})$
  $v_{i,t}(q) = \chi\{v_i(q) + c_1 r_1(Y_i(q) - q_{i,t-1})\} + c_2 r_2(q_i' - q_{i,t-1})$
  $\lambda_{i,t} = \lambda_{i,t} + v_{i,t}(\lambda)$, $q_{i,t} = q_{i,t} + v_{i,t}(q)$
  $f_{i,t} = \mathcal{ICOMP}_{\texttt{Miss}}(\lambda_{i,t}, q_{i,t}, X, Y)$
  if $f_{i,t} < f^*$ **then**
    $\lambda^* = \lambda_{i,t}$, $q^* = q_{i,t}$
  end if
  if $f_{i,t} < P_i$ **then**
    $P_i = f_{i,t}$, $Y_i(\lambda) = \lambda_i$, $Y_i(q) = q_i$
  end if
end for

References

Bozdogan, H. (2000). Akaike's information criterion and recent developments in information complexity. *Journal of Mathematical Psychology 44*, 62–91.

Bozdogan, H. (Ed.) (2004). *Statistical Data Mining and Knowledge Discovery*. Chapman and Hall/CRC.

Bozdogan, H. and J. A. Howe (2012). Misspecified multivariate regression models using the genetic algorithm and information complexity as the fitness function. *European Journal of Pure and Applied Mathematics 5*(2), 211–249.

Bozdogan, H. and E. Pamukcu (2016). Novel dimension reduction techniques for high dimensional data using information complexity. In A. Gupta and A. Capponi (Eds.), *Optimization Challenges in Complex, Networked, and Risky Systems*, pp. 140–170. INFORMS.

Bühlmann, P., M. Kalisch, and L. Meier (2014). High-dimensional statistics with a view toward applications in biology. *Annual Review of Statistics and Its Application 1*, 255–278.

Chen, Y., A. Wiesel, Y. C. Eldar, and A. O. Hero (2010). Shrinkage algorithms for mmse covariance estimation. *IEEE Transactions on Signal Processing 58*(10), 5016–5029.

Clerc, M. and J. Kennedy (2002). The particle swarm-explosion, stability, and convergence in a multidimensional complex space. *IEEE transactions on Evolutionary Computation 6*(1), 58–73.

Fan, J. and R. Li (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American statistical Association 96*(456), 1348–1360.

Frank, L. E. and J. H. Friedman (1993). A statistical view of some chemometrics regression tools. *Technometrics 35*(2), 109–135.

Fu, W. J. (1998). Penalized regressions: the bridge versus the lasso. *Journal of computational and graphical statistics 7*(3), 397–416.

Godfrey, L. G. (1991). *Misspecification tests in econometrics: the Lagrange multiplier principle and other approaches*. Number 16. Cambridge University Press.

Hoerl, A. E. and R. W. Kennard (1970a). Ridge regression: applications to nonorthogonal problems. *Technometrics 12*(1), 69–82.

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

Huang, J., J. L. Horowitz, and S. Ma (2008). Asymptotic properties of bridge estimators in sparse high-dimensional regression models. *The Annals of Statistics 36*(2), 587–613.

Huber, P. J. (1967). The behavior of maximum likelihood estimates under nonstandard conditions. In *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, Volume 1, pp. 221–233. University of California Press.

Kauermann, G. and R. J. Carroll (2001). A note on the efficiency of sandwich covariance matrix estimation. *Journal of the American Statistical Association 96*(456), 1387–1396.

Kawano, S. (2012). Adaptive bridge regression modeling and selection of the tuning parameters.

Kawano, S. (2014). Selection of tuning parameters in bridge regression models via bayesian information criterion. *Statistical Papers 55*(4), 1207–1223.

Knight, K. and W. Fu (2000). Asymptotics for lasso-type estimators. *The Annals of statistics 28*(5), 1356–1378.

Koc, E. K. and H. Bozdogan (2015). Model selection in multivariate adaptive regression splines (mars) using information complexity as the fitness function. *Machine Learning 101*(1-3), 35–58.

Konishi, S., T. Ando, and S. Imoto (2004). Bayesian information criteria and smoothing parameter selection in radial basis function networks. *Biometrika 91*(1), 27–43.

Ledoit, O. and M. Wolf (2004). A well-conditioned estimator for large-dimensional covariance matrices. *Journal of multivariate analysis 88*(2), 365–411.

Mohebbi, S., E. Pamukcu, and H. Bozdogan (2019). A new data adaptive elastic net predictive model using hybridized smoothed covariance estimators with information complexity. *Journal of Statistical Computation and Simulation*, 1–30.

Natarajan, B. K. (1995). Sparse approximate solutions to linear systems. *SIAM journal on computing 24*(2), 227–234.

Ollila, E. (2017). Optimal high-dimensional shrinkage covariance estimation for elliptical distributions. In *2017 25th European Signal Processing Conference (EUSIPCO)*, pp. 1639–1643. IEEE.

Roozbeh, M. (2018). Generalized ridge regression estimator in high dimensional sparse regression models. *Statistics, Optimization & Information Computing 6*(3), 415–426.

Takeuchi, K. (1976). Distribution of information statistics and a criterion of model fitting. *Suri-Kagaku, Mathematical Sciences 153*, 12–18.

Tännler, S., N. Zamboni, C. Kiraly, S. Aymerich, and U. Sauer (2008). Screening of bacillus subtilis transposon mutants with altered riboflavin production. *Metabolic Engineering 10*(5), 216–226.

Tibshirani, R. (1996). Regression shrinkage and selection via lasso. *Journal of the Royal Statistical Society Series B 58*, 267–288.

Wang, X.-H. and J.-J. Li (2004). Hybrid particle swarm optimization with simulated annealing. In *Proceedings of 2004 International Conference on Machine Learning and Cybernetics (IEEE Cat. No. 04EX826)*, Volume 4, pp. 2402–2405. IEEE.

White, H. (1982). Maximum likelihood estimation of misspecified models. *Econometrica*, 1–25.

White, H. (1994). *Estimation, Inference and Specification Analysis*. Number 22. Cambridge University Press.

Zhan, Z.-H., J. Zhang, Y. Li, and H. S.-H. Chung (2009). Adaptive particle swarm optimization. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics) 39*(6), 1362–1381.

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