entropy

MDPI

Article

Feature Selection in High-Dimensional Models via EBIC with Energy Distance Correlation

Isaac Xoese Ocloo [1,*] and Hanfeng Chen [2]

[1] Department of Statistics, University of Georgia, Athens, GA 30602, USA
[2] Department of Mathematics and Statistics, Bowling Green State University, Bowling Green, OH 43403, USA
* Correspondence: isaac.ocloo@uga.edu

Abstract: In this paper, the LASSO method with extended Bayesian information criteria (EBIC) for feature selection in high-dimensional models is studied. We propose the use of the energy distance correlation in place of the ordinary correlation coefficient to measure the dependence of two variables. The energy distance correlation detects linear and non-linear association between two variables, unlike the ordinary correlation coefficient, which detects only linear association. EBIC is adopted as the stopping criterion. It is shown that the new method is more powerful than Luo and Chen's method for feature selection. This is demonstrated by simulation studies and illustrated by a real-life example. It is also proved that the new algorithm is selection-consistent.

Keywords: energy distance; extended Bayesian information criteria; feature variable selection

Citation: Ocloo, I.X.; Chen, H. Feature Selection in High-Dimensional Models via EBIC with Energy Distance Correlation. *Entropy* **2023**, *25*, 14. https://doi.org/10.3390/e25010014

Academic Editor: Yuehua Wu

Received: 20 November 2022
Revised: 16 December 2022
Accepted: 18 December 2022
Published: 21 December 2022

1. Introduction

Advancements in technology have led to the production of sophisticated machines which are able to measure many details about every observational or experimental unit in a system. This results in data with more features (p or predictors) than the number of observational or experimental units (sample size n), referred to as high-dimensional data. Most of these data come from genetic research, e-commerce, biomedical imaging, and functional magnetic resonance imaging, among many others.

Since there are many more features (p) than the sample size (n), they are analyzed using a sparse high-dimensional regression (SHR) model:

$$y_i = \beta_0 + \sum_{j=1}^{p} x_{ij}\beta_j + \epsilon_i, \text{ for } i = 1, \ldots, n. \tag{1}$$

It is assumed that there is only a relatively small number of the nonzero β_j's. The main goal in their analysis is feature selection. As stated by [1], feature selection typically has two goals. The first is for model building using desirable prediction properties. The second is for identifying the features with nonzero coefficients. For convenience, such features are referred to as relevant features in this paper.

One approach to the SHR model is to estimate the β_j's by a regularization method, which is done by simultaneously minimizing the penalized least squares below:

$$\sum_{i=1}^{n}\left(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j x_{ij}\right)^2 + \sum_{j=1}^{p} p_\lambda(|\beta_j|),$$

where λ is the regulating parameter and p_λ is a penalty function. When p_λ is based on the L_1 norm, thus $\|\beta\|_1 = \sum_{j=1}^{p}|\beta_j|$, it is referred to as the LASSO [2]. The L_1 norm penalty is able to shrink the coefficients of redundant predictors to zero. Thus, the LASSO usually results in sparse models that are easier to interpret. Other penalty functions such as the

SCAD [3] and adaptive LASSO (ALasso) [4] have also been reported in the literature. SCAD smoothly clips the L_1 penalty (for small $|\beta_j|$), and assigns a constant penalty (for large $\mid \beta_j \mid$s). On the other hand, adaptive LASSO utilizes the minimax concave penalty [5], $p_\lambda(\mid \beta_j \mid) = \lambda w_j \mid \beta_j \mid$, where w_j represents the weights. The regulating parameter, λ, is usually chosen using cross-validation (CV).

Another approach to analyzing the SHR model, large-p-small-n problem, is sequential variable selection, which is designed to reduce the dimension of the data such that $d < p$. There are two forms of sequential variable selection. The first uses the sure screening property, which selects from the many features a subset which contains the relevant features (predictors). This is usually followed by a regularization method such as SCAD or ALasso to identify and estimate the relevant predictors from the reduced feature space. The other form is to sequentially select the relevant features through a repetitive process which terminates when a stopping criterion is met.

A recent addition to sequential feature selection is the sequential LASSO cum EBIC in ultra high-dimensional feature space (SLasso) by [1], which sequentially solves partially penalized least squares problems and uses EBIC as the stopping criteria. The EBIC proposed by [6] are suitable for model selection in large model spaces. It has the ordinary BIC as a special case. For large model spaces, the ordinary BIC tends to select a model with many spurious variables. Let k and $k+1$ be the number of predictors in two models, respectively. Using EBIC as the selection or stopping criteria, the model with k predictors is selected if the EBIC(k + 1) $>$ EBIC(k).

In SLasso, sequentially solving the partially penalized least squares reduces to selecting the feature(s) which maximize the ordinary correlation coefficient between the features and the response variable at each step. It is well-known that the Pearson correlation coefficient is used for measuring the strength of linear associations. Thus, maximizing the Pearson correlation coefficient might not work well for data structures where the relationship between at least one feature and the response variable is nonlinear.

In this article, we propose the use of the energy distance correlation instead of the ordinary correlation coefficient to identify and maximize both the linear and nonlinear relationships that might exist between each feature and the response. Energy distance is a metric that measures the distance between the distributions of random vectors. The name 'energy' is motivated by analogy to the potential energy between objects in a gravitational space. The potential energy is zero if and only if the locations (the gravitational centers) of the two objects coincide, and increases as their distance in space increases [7]. The energy distance correlation has an explicit relationship with the product-moment correlation, but unlike the classical definition of correlation, energy distance correlation is zero only if the random vectors are independent. The empirical energy distance correlation is based on Euclidean distances between sample elements rather than sample moments.

The remainder of the article is arranged as follows: in Section 2, we discuss the derivation of the energy distance correlation, extended Bayesian information criteria and our proposed method (energy distance correlation with EBIC (Edc + EBIC)). In Section 3, we report simulation studies comparing Edc + EBIC with various other methods and provide an analysis of real data. In Section 4, we conclude the article with a discussion of the results.

2. EBIC with Energy Distance Correlation

2.1. Energy Distance Correlation

The authors in [8] proposed the energy distance correlation between two random variables. Suppose that $W \in \mathbb{R}^p$ and $Z \in \mathbb{R}^q$ are two random vectors with $\mathbb{E}\|W\| < \infty$, and $\mathbb{E}\|Z\| < \infty$, where $\|.\|$ is the euclidean norm and $\mathbb{E}$ is the expected value. Let F and G be the cumulative distribution function (CDF) of W and Z, respectively. Further, let $W^{'}$ denote an independent and identically distributed (iid) copy of W; that is, W and $W^{'}$ are iid. Similarly, Z and $Z^{'}$ are iid.

The squared energy distance can be defined in terms of expected distances between the random vectors

$$D^2(F,G) := 2\mathbb{E}\|W-Z\| - \mathbb{E}\|W-W'\| - \mathbb{E}\|Z-Z'\| \geq 0,$$

and the energy distance between distributions F and G is defined as the square root of $D^2(F,G)$.

The energy distance correlation between random vectors W and Z with finite first moments is the nonnegative number $\mathcal{R}(W,Z)$ defined by

$$\mathcal{R}(W,Z) = \begin{cases} \frac{\nu^2(W,Z)}{\sqrt{\nu^2(W)\nu^2(Z)}}, & \nu^2(W)\nu^2(Z) > 0 \\ 0, & \nu^2(W)\nu^2(Z) = 0 \end{cases}$$

where $\nu^2(W,Z)$ is the energy distance covariance between W and Z, $\nu^2(W)$ and $\nu^2(Z)$ are the energy distance variance of W and Z respectively.

For a statistical sample $(w,z) = \{(w_k, z_k), k = 1,2,\ldots,n\}$ from a pair of real-valued or vector-valued random variables (W,Z), the sample energy distance correlation, $\mathcal{R}_n(W,Z)$, is calculated by first computing the n by n distance matrices $(a_{j,k})$ and $(b_{j,k})$ containing all pairwise distances $(a_{j,k}) = \|W_j - W_k\|$, $j,k = 1,2,\ldots,n$ and $(b_{j,k}) = \|Z_j - Z_k\|$, $j,k = 1,2,\ldots,n$ where $\|.\|$ denotes euclidean norm. Secondly, calculate all doubly centered distances $A_{j,k} = a_{j,k} - \bar{a}_{j.} - \bar{a}_{.k} + \bar{a}_{..}$, $B_{j,k} = b_{j,k} - \bar{b}_{j.} - \bar{b}_{.k} + \bar{b}_{..}$ where $\bar{a}_{j.}$ is the j^{th} row mean, $\bar{a}_{.k}$ is the k^{th} column mean, and $\bar{a}_{..}$ is the grand mean of the distance matrix of the w sample. The notation is similar for the b values.

The squared sample distance covariance (a scalar) is the arithmetic average of the products $A_{j,k}B_{j,k}$.

$$\nu_n^2(w,z) = \frac{1}{n^2}\sum_{j=1}^{n}\sum_{k=1}^{n} A_{j,k}B_{j,k}.$$

The sample energy distance variance for sample w

$$\nu_n^2(w) = \frac{1}{n^2}\sum_{j,k=1}^{n} A_{j,k}^2.$$

The sample energy distance variance for sample z

$$\nu_n^2(z) = \frac{1}{n^2}\sum_{j,k=1}^{n} B_{j,k}^2.$$

The sample energy distance correlation is

$$\mathcal{R}_n(W,Z) = \begin{cases} \frac{\nu_n^2(W,Z)}{\sqrt{\nu_n^2(W)\nu_n^2(Z)}}, & \nu_n^2(W)\nu_n^2(Z) > 0 \\ 0, & \nu_n^2(W)\nu_n^2(Z) = 0. \end{cases}$$

Some basic properties of the distance correlation are as follows:

(i) $0 \leq R_n(W,Z) \leq 1$;
(ii) If $E(|W|_p + |Z|_q) < \infty$, then $R_n(W,Z) = 0$ if and only if W and Z are independent;
(iii) Suppose that $R_n(W,Z) = 1$. Then, there exist a vector a, a nonzero real number b and an orthogonal matrix C such that $Z = a + bWC$.

For further details on the energy distance correlation, see [7].

2.2. EBIC

Ref. [6] derived the EBIC which have special cases as AIC and BIC. Let $\{(y_i, x_i) : i = 1, 2, \ldots, n\}$ be independent observations. Suppose that the conditional density function of y_i given x_i is $f(y_i|x_i, \beta)$, where $\beta \in \Theta \subset R^{p_n}$, p_n being a positive integer. The likelihood function of β is given by

$$L_n(\beta) = f(x; \beta) = \prod_{i=1}^{n} f(y_i|x_i, \beta).$$

Denote $\mathbf{y} = (y_1, y_2, \ldots, y_n)^\tau$. Let $s \subset \{1, 2, \ldots, p_n\}$ and $\beta(s)$ be the parameter vector β with those components outside s set to 0. Let S be the underlying model space, i.e., $S = \{s : s \subseteq \{1, 2, \ldots, p_n\}\}$, let $p(s)$ be a prior for model s. Assume that, given s, the prior density of $\beta(s)$ is $\pi(\beta(s))$. The posterior is

$$p(s|y) = \frac{m(y|s)p(s)}{\sum_{s \in S} m(y|s)p(s)},$$

where $m(y|s)$ is the likelihood in model s, i.e.,

$$m(Y|s) = \int f(y; \beta(s))\pi(\beta(s))d\beta(s).$$

Suppose S is partitioned into $\cup_{j=1}^{p} S_j$, such that models within each S_j have an equal dimension. Let $\tau(S_j)$ be the size of S_j. Assign the prior distribution $P(S_j)$ proportional to $\tau^\eta(S_j)$ for some η between 0 and 1. For each $s \in S_j$, assign equal probability, $p(s|S_j) = 1/\tau(S_j)$; this is equivalent to $P(s)$ for $s \in S_J$ proportional to $\tau^{-\gamma}(S_j)$, where $\gamma = 1 - \eta$. Then, the extended BIC family is given by

$$EBIC_\gamma(s) = -2 \log L_n\{\hat{\beta}(s)\} + |s| \log(n) + 2\gamma \ln(\tau(S_{|s|})), 0 \leq \gamma \leq 1,$$

where $\hat{\beta}(s)$ is the maximum likelihood estimator of $\beta(s)$ and $|s|$ is the number of components in s.

2.3. Energy Distance Correlation with EBIC (Edc + EBIC) Algorithm

We propose a sequential model selection method which we call energy distance correlation with EBIC, and for convenience abbreviate it as Edc + EBIC. Let $y_i, i = 1, \ldots, n$ be a continuous response variable and $x_j, j = 1, \ldots, p$ be an $n \times p$ data matrix. Let S be the index set of all predictors. Let $s_0 = \{j : \beta_j \neq 0, j = 1, \ldots, p\}$. For $s \subset S$, let $s^- = s^c \cap s_0$. If $s \subset s_0$, then s^- is the complement of s in s_0. Let $p_0 = |s_0|$ be the number of elements in the set s_0.

At the initial stage we standardize all the variables. Next, we find the energy distance correlation between the response variable and each of the predictor variables—$\{\mathcal{R}(x_j, y)\ j = 1, \ldots, p.\}$. We then select the predictor (feature) which has the highest distance correlation with the response and store it in the active set s_{*1}.

Let $\mathcal{L}(s)$ be the linear space spanned by the columns of $X(s)$ and $H(s)$ its corresponding projection matrix, i.e., $H(s) = X(s)[X^\tau(s)X(s)]^{-1}X^\tau(s)$. Next, we compute $I - H(s_{*1})$, EBIC(s_{*1}), $\tilde{y} = [I - H(s_{*k})]y$ and $\tilde{x}_j = [I - H(s_{*k})]x_j$. The variable $\tilde{y}$ is the unexplained part of y by $X(s_{*1})$. This gives $X(s_{*1})$ close to a zero chance of being selected in the subsequent steps.

For the general step where $k > 1$, we calculate $\{\mathcal{R}(\tilde{x}_j, \tilde{y})\ j = 1, \ldots, p.\}$ and update the active set to s_{*k+1}, which is the union of all the previous selected variables and the current one. We then compute EBIC(s_{*k+1}) and compare it with EBIC(s_{*k}). The procedure stops if EBIC(s_{*k+1}) > EBIC(s_{*k}). The selected variables which we call the relevant variables will be $X(s_{*k})$. We can then fit a linear regression model between the response y and the relevant variables.

We wish to note that care must be taken in fitting this model because some of the predictors might be non-linearly related to y, and thus some of the predictors may have to enter into the model in their quadratic or cubic form, etc. Alternatively, a Box–Cox transformation can be performed on the data before fitting the model.

The algorithm details are given in the following.

- *Initial Step:* With $y, x_j, j = 1, \ldots, p$ satisfying $y^\tau 1 = 0, x_j^\tau 1 = 0$ and $y^\tau y = n, x_j^\tau x_j = n$, compute $\mathcal{R}(x_j, y)$ for $j \in S$. Let

$$s_{TEMP} = \{j : \mathcal{R}(x_j, y) = \max_{j' \in S} \mathcal{R}(x_{j'}, y)\}.$$

  Let $s_{*1} = s_{TEMP}$ be the active set. Compute $I - H(s_{*1})$ and EBIC(s_{*1}), where $H(s) = X(s)[X^\tau(s)X(s)]^{-1}X^\tau(s)$.
- *General Step:* In the selection step k, compute $\mathcal{R}(\tilde{x}_j, \tilde{y})$ for $j \in s_{*k}^c$, where $\tilde{y} = [I - H(s_{*k})]y, \tilde{x}_j = [I - H(s_{*k})]x_j$. Let

$$s_{TEMP} = \{j : \mathcal{R}(\tilde{x}_j, \tilde{y}) = \max_{j' \in S_{*k}^c} \mathcal{R}(\tilde{x}_j, \tilde{y})\}.$$

  Let $s_{*k+1} = s_{*k} \cup s_{TEMP}$. Compute EBIC$(s_{*k+1})$. If EBIC$(s_{*k+1}) >$ EBIC(s_{*k}), stop; otherwise, continue computing $I - H(s_{*k+1})$.
- When the process terminates, return the least-squares estimates for parameters in the selected model.

2.4. Selection Consistency of Edc + EBIC

We attempt to establish the large sample property for the Edc + EBIC. We will show that under regular conditions, the Edc + EBIC is selection-consistent. The proof essentially follows the approach in [9]. We proceed with the regularity conditions.

Assumption 1. *Random vectors X and Y possess the subexponential tail probabilities, uniformly in p, specified as follows. There is a constant $a_0 > 0$, such that for any $0 < a \leq 2a_0$, $\sup_p \max_{1 \leq k \leq p} E\{\exp(a\|X_k\|_1^2)\} < \infty$ and $E\{\exp(a\|Y\|_q^2)\} < \infty$.*

Assumption 2. *The minimum distance correlation of predictors on which y functionally depends satisfies $\min_{j \in s_0} \mathcal{R}(\tilde{X}_j, \tilde{Y}) \geq 2cn^{-d}$, for some constants $0 < c < 1$ and $0 \leq d < 1/2$.*

Assumption 3. *For the index set S of all predictors, let $s_0 = \{j : \beta_j \neq 0, j = 1, \ldots, p\}$ and $p_0 = |s_0|$ (p_0 is the number of elements in the set s_0). For $s \subset S$ let $s^- = s^c \cap s_0$. If $s \subset s_0$ then s^- is the complement of s in s_0. For $s \subset s_0$, $\max_{j \in s_0^c} \mathcal{R}(\tilde{X}_j, \tilde{Y}) < q \max_{j \in s^-} \mathcal{R}(\tilde{X}_j, \tilde{Y})$ for some $0 < q < 1$, where $\tilde{Y} = [I - H(s_{*k})]Y$, $\tilde{X}_j = [I - H(s_{*k})]X_j$. For $k = 0$, s_{*0} is defined as the empty set $\emptyset$.*

Details for requiring Assumption 1 and 2 are stated in [9]. Intuitively, Assumption 1 is required to make it easy to establish a relationship between the energy distance correlation and the squared Pearson correlation to aid with the derivations in the proof. Assumption 2 requires that the energy distance correlation for the relevant predictors cannot be too small. Assumption 3 requires that the maximum energy distance correlation between the selected features and the residual response $\tilde{Y}$ is smaller than the maximum energy distance correlation between the remaining features and the residual response in the sequential step of the algorithm.

Theorem 1. *Suppose that Assumptions 1–3 hold. The proposed Edc + EBIC with the energy distance correlation is consistent, i.e.,*

$$\lim_{n \to \infty} P(s_{*k^*} = s_{0n}) = 1,$$

where s_{*k^*} is the set of features selected at the k^{*th} step of Edc + EBIC such that $|s_{*k^*}| = p_{0n}$, s_{0n} is the set of relevant features and $p_{0n} = |s_{0n}|$.

Proof. Suppose that $X \in \mathbb{R}^p$ and $Y \in \mathbb{R}^q$ with cumulative distribution function (CDF) F and G, respectively, where $\mathbb{E}\|X\| < \infty$, and $\mathbb{E}\|Y\| < \infty$. The energy distance correlation $\mathcal{R}(X, Y)$ is the square root of the standardized coefficient:

$$\mathcal{R}(X,Y) = \begin{cases} \frac{\nu^2(X,Y)}{\sqrt{\nu^2(X)\nu^2(Y)}}, & \nu^2(X)\nu^2(Y) > 0 \\ 0, & \nu^2(X)\nu^2(Y) = 0 \end{cases}$$

where $0 \leq \mathcal{R}(X, Y) \leq 1$. In the numerator is the distance covariance defined by [8], as

$$dcov^2(x,y) = S_1 + S_2 - 2S_3,$$

where $S_j, j = 1, 2,$ and 3 are defined as:

$$\begin{aligned} S_1 =& \mathbb{E}\|X - X'\|\|Y - Y'\| \\ S_2 =& \mathbb{E}\|X - X'\|\mathbb{E}\|Y - Y'\| \\ S_3 =& \mathbb{E}\|X - X'\|\|Y - Y^{''}\| \end{aligned} \tag{2}$$

where $(X, Y), (X', Y')$, and $(X^{''}, Y^{''})$ are independently and identically distributed.

For a random sample $\{(x_i, y_i), i = 1, \ldots, n\}$ from (x, y), [8] estimated S_1, S_2, S_3 as:

$$\begin{aligned} \hat{S_1} =& \frac{1}{n^2}\sum_{k,l=1}^{n} |x_k - x_l|_p |y_k - y_l|_q \\ \hat{S_2} =& \frac{1}{n^2}\sum_{k,l=1}^{n} |x_k - x_l|_p \frac{1}{n^2}\sum_{k,l=1}^{n} |y_k - y_l|_q \\ \hat{S_3} =& \frac{1}{n^3}\sum_{k=1}^{n}\sum_{l,m=1}^{n} |x_k - x_l|_p |y_k - y_l|_q \end{aligned}$$

so the sample distance covariance is $\widehat{dcov}^2 = \hat{S_1} + \hat{S_2} - 2\hat{S_3}$.

The remaining part of the proof is to show that the energy distance correlation is uniformly consistent and has the sure screening property. The numerator and denominator of the energy distance correlation are similar, so to show the uniform consistency of the energy distance correlation it suffices to show that both the numerator and the denominator are uniformly consistent.

The uniform consistency of the numerator, $\widehat{dcov}^2 = \hat{S_1} + \hat{S_2} - 2\hat{S_3}$, of the energy distance correlation between the random vectors (x, y) is shown by [9]. However, in the general step of the sequential algorithm for Edc + EBIC, the energy distance correlation is calculated between the residuals $\tilde{y} = [I - H(s_{*k})]y$, and $\tilde{x_j} = [I - H(s_{*k})]x_j$ at each step of the algorithm. Thus, to show the uniform consistency of Edc+EBIC it is equivalent to follow the proof by [9].

Additionally, in [9] they showed that the energy distance correlation has the sure screening property. They showed that the energy distance is able to select a subset of the features which contains the relevant features. Their argument applies here because we used the energy distance correlation as well, thus the Edc + EBIC has the sure screening property.

Therefore, the Edc + EBIC is selection-consistent since it is uniformly consistent and has the sure screening property. The proof is complete. □

3. Simulation Studies and Data Analysis

3.1. Sure Independence Screening Using Energy Distance Correlation

We establish the need for Ebc + EBIC by firstly examining the performance of a sure independence screening method introduced by [9] called Distance Correlation Sure Independence Screening (DC-SIS). This is similar to the Sure Independence Screening (SIS) introduced by [10].

In SIS, they perform a componentwise regression between each predictor and the response and select the first $n-1$ or $[n/log(n)]$ predictors with the largest estimates. Performing a componentwise regression is equivalent to finding the ordinary correlation between the response and each predictor when the two variables are standardized. Hence, in DC-SIS, they replaced the ordinary correlation with the energy distance correlation.

We examine the performance of DC-SIS through a simulation study. We are interested in observing, on average, the model size selected by SCAD or ALasso if we screened the data first using DC-SIS. We present two simulation set-ups. For each simulation we generated two hundred datasets, and for each dataset we ran SCAD, ALasso, DC-SIS + SCAD, DC-SIS + ALasso and found the average model size and the standard deviation.

In [10], details of two simulation setups we adapted for this subsection are discussed, namely independent features setup and dependent features setup. In Tables 1 and 2 we present results under the independent features setup and in Tables 3–5 we present results under the dependent features setup. In each simulation, n is the sample size, p is the number of features and s is the true model size. For the screening using the energy distance correlation we chose $d = [n/\log n]$ features and applied SCAD or ALasso.

In Tables 1 and 2, we report the average selected model size and their standard deviations. We observe that applying the sure screening by distance correlation before either SCAD or ALasso in all cases did not lead to significant differences in the average model size when SCAD and ALasso were applied directly to the data. This suggests that either applying distance correlation before SCAD or ALasso did not yield the intended result, and thus needs some improvement.

Table 1. Comparing model size selected with or without screening for $n = 200, s = 8, p = 1000$.

Methods	MSize (SD)
SCAD	12.87 (7.292)
DC-SIS + SCAD	10.7 (3.1575)
ALasso	25.24 (9.0365)
DC-SIS + ALasso	11.74 (4.419)

Table 2. Comparing model size selected with or without screening for $n = 800, s = 14, p = 3000$.

Methods	MSize (SD)
SCAD	16.62 (2.78807)
DC-SIS + SCAD	16.69 (3.5525)
ALasso	14.78 (0.7860)
DC-SIS + ALasso	14.78 (3.8522)

Table 3. Comparing model size selected with or without screening for $n = 200, p = 1000, s = 5$.

Methods	MSize (SD)
SCAD	12.215 (12.2560)
DC-SIS + SCAD	7.335 (2.6109)
ALasso	44.485 (13.8041)
DC-SIS + ALasso	8.21 (2.5844)

Table 4. Comparing model size selected with or without screening for $n = 200, p = 1000, s = 8$.

Methods	MSize (SD)
SCAD	14.625 (10.3324)
DC-SIS + SCAD	10.905 (2.3158)
ALasso	19.95 (3.2789)
DC-SIS + ALasso	12.36 (3.9875)

Table 5. Comparing model size selected with or without screening for $n = 800, p = 3000, s = 14$.

Methods	MSize (SD)
SCAD	19.185 (5.0146)
DC-SIS + SCAD	17.675 (4.6038)
ALasso	38.125 (5.6372)
DC-SIS + ALasso	31.845 (13.1011)

In Tables 3–5 we report the selected model size and the standard deviation. We observe that applying DC-SIS followed by either SCAD or ALasso did not yield any significant difference in the average model size, as was also observed in the independent features setup.

3.2. Simulation Studies to Compare Edc + EBIC with Other Feature Selection Methods

In this simulation study we adopted two simulation setups from [1], which they call group A and group B, respectively. Under their group A we considered four settings of the covariance structure for the design matrix X, namely GA1, GA2, GA3, and GA5. In their group B setup we considered all three settings of the covariance structure for the design matrix X, namely GB1, GB2, and GB3. We compared the performance of adaptive LASSO (ALasso) [11], SCAD [12], SIS+SCAD [10], SLasso [1], and the energy distance correlation with EBIC (Edc + EBIC) based on the model size (MSize), positive discovery rate (PDR), PDR $= \frac{|s_{*k^*} \cap s_o|}{|s_o|}$, and false discovery rate, FDR $= \frac{|s_{*k^*} \cap s_0^c|}{|s_{*k^*}|}$ averaged over 200 and 500 simulations, respectively.

We considered the diverging pattern $(n, p, p_o) = (n, [5e^{n^{0.3}}], [4n^{0.16}])$, meaning that as the sample size increased, the number of predictors increased and the number of relevant predictors also increased. The coefficients were generated as independent random variables distributed as $(-1)^u(4n^{-0.15} + |z|)$, where $u \sim PBernoulli(0.4)$ and z is a normal random variable with mean 0 and satisfies $P(|z| \geq 0.1) = 0.25$. The variance of the error term in the linear model was determined by

$$h = \frac{\beta^\tau \Sigma \beta}{\beta^\tau \Sigma \beta + \sigma^2} = 0.8$$

where Σ is the variance-covariance matrix of relevant features. The response variable is simulated from the sparse high-dimensional regression (SHR) model

$$y_i = \beta_0 + \sum_{j=1}^{p} \beta_j x_{ij} + \epsilon_i, i = 1, ..., n$$

3.2.1. Group A Simulations and Results

We used two sample sizes, $n = 100$ and $n = 200$. By the diverging pattern considered for the simulation, for a sample size of 100 we have $(n, p, p_0) = (100, 268, 8)$ and for a sample size of 200 we have $(n, p, p_0) = (200, 672, 9)$. Under the two sample sizes we had eight and nine relevant features, respectively, and we expected a well-performed selection model to select the right number of relevant predictors (model size) on average. The details about the covariance structure for GA1, GA2, GA3, and GA5 are in [1].

In Tables 6 and 7, we report the simulation results under the conditions for GA1, GA2, GA3, and GA5 using sample sizes 100 and 200, respectively. We observed that under all setups, Edc + EBIC improved in terms of the average model size, PDR, and FDR, as we increased the sample size. This demonstrates the selection consistency of Edc + EBIC.

Table 6. We compared the methods using PDR, FDR, and model size (MSize) averaged over 200 simulation replications. The relevant predictors were 8 and the sample size was 100. The standard deviations are in parentheses.

Setting	Methods	MSize	PDR	FDR
GA1	ALasso	34.105 (13.95)	1.000 (0.000)	0.721 (0.120)
	SCAD	25.735 (5.020)	1.000 (0.000)	0.676 (0.065)
	SIS + SCAD	8.100 (1.790)	0.866 (0.239)	0.157 (0.167)
	SLasso	8.565 (0.848)	1.000 (0.000)	0.058 (0.081)
	Edc + EBIC	8.365 (1.375)	0.978 (0.125)	0.056 (0.085)
GA2	ALasso	34.455 (11.095)	1.000 (0.000)	0.754 (0.108)
	SCAD	25.650 (6.720)	0.876 (0.141)	0.709 (0.075)
	SIS + SCAD	7.335 (1.740)	0.813 (0.182)	0.103 (0.107)
	SLasso	6.055 (1.725)	0.688 (0.185)	0.080 (0.104)
	Edc + EBIC	6.075 (1.713)	0.717 (0.195)	0.050 (0.082)
GA3	ALasso	14.710 (3.847)	1.000 (0.000)	0.423 (0.131)
	SCAD	26.27 (5.244)	1.000 (0.000)	0.680 (0.070)
	SIS + SCAD	8.165 (1.160)	0.951 (0.113)	0.062 (0.078)
	SLasso	8.625 (1.005)	1.000 (0.000)	0.062 (0.089)
	Edc + EBIC	8.265 (1.373)	0.976 (0.132)	0.048 (0.074)
GA5	ALasso	23.845 (7.005)	0.964 (0.057)	0.652 (0.092)
	SCAD	24.070 (6.147)	0.997 (0.020)	0.642 (0.102)
	SIS + SCAD	7.605 (2.020)	0.832 (0.245)	0.127 (0.141)
	SLasso	7.650 (2.182)	0.856 (0.217)	0.089 (0.113)
	Edc + EBIC	7.180 (2.453)	0.842 (0.270)	0.050 (0.081)

Table 7. We compared the methods using PDR, FDR, and model size (MSize) averaged over 200 simulation replications. The relevant predictors were 9 and the sample size was 200. The standard deviations are in parentheses.

Setting	Methods	MSize	PDR	FDR
GA1	ALasso	27.670 (12.996)	1.000 (0.000)	0.638 (0.180)
	SCAD	17.035 (7.746)	1.000 (0.000)	0.454 (0.168)
	SIS + SCAD	9.215 (1.507)	1.000 (0.000)	0.112 (0.123)
	SLasso	8.710 (0.944)	1.000 (0.000)	0.072 (0.088)
	Edc + EBIC	8.42 (0.712)	1.000 (0.000)	0.045 (0.071)
GA2	ALasso	27.92 (9.686)	1.000 (0.000)	0.675 (0.124)
	SCAD	15.11 (6.241)	1.000 (0.000)	0.397 (0.171)
	SIS + SCAD	9.16 (1.509)	1.000 (0.000)	0.108 (0.171)
	SLasso	8.72 (0.947)	1.000 (0.000)	0.073 (0.089)
	Edc + EBIC	8.49 (0.763)	1.000 (0.000)	0.051 (0.076)
GA3	ALasso	27.115 (12.867)	1.000 (0.000)	0.632 (0.177)
	SCAD	16.245 (7.770)	1.000 (0.000)	0.434 (0.162)
	SIS + SCAD	9.22 (1.617)	1.000 (0.000)	0.110 (0.128)
	SLasso	8.70 (0.857)	1.000 (0.000)	0.072 (0.084)
	Edc + EBIC	8.47 (0.694)	1.000 (0.000)	0.050 (0.071)
GA5	ALasso	38.95 (8.308)	0.939 (0.075)	0.797 (0.054)
	SCAD	19.075 (7.427)	1.000 (0.000)	0.520 (0.159)
	SIS + SCAD	9.975 (1.858)	1.000 (0.000)	0.174 (0.132)
	SLasso	8.765 (1.125)	0.989 (0.061)	0.087 (0.094)
	Edc + EBIC	8.44 (0.768)	0.998 (0.025)	0.048 (0.073)

3.2.2. Group B Simulations and Results

We considered three different covariance structures named GB1, GB2, and GB3 for the features (predictors), as used in [1]. We also increased the signal-to-noise ratio by increasing the value of the expected predictors.

GB1. All the features had constant pairwise correlation $p_{ij} = 0.5.(n, p, p_0) = (100, 200, 15)$. $\sigma = 1.5$. The coefficients of the relevant features were specified as $|\beta_j| = 2.5$ for $1 \leq j \leq 5$, 1.5 for $6 \leq j \leq 10, 0.5$ for $11 \leq j \leq 15$. The signs of the coefficients were determined as $(-1)^{u_i}$, where the u_is were iid Bernoulli random variables with probability of success $p = 0.5$.

GB2. This structure was the same as in GB1, that is, $(n, p, p_0) = (100, 200, 15)$ and $\sigma = 1.5$. The covariance structure of the features was specified such that the partially orthogonality condition [11] was satisfied. Specifically, while s_0 was taken as $\{1, \ldots 5, 11, \ldots, 15, 21, \ldots, 25\}$, the correlations were specified as $\rho_{ij} = 0.5^{|i-j|}$ for $1 \leq i \leq 215$ and $1 \leq j \leq 215$. The coefficients were specified as $|\beta| = 2.5$ for $1 \leq j \leq 5, 1.5$ for $10 \leq j \leq 15, 0.5$ for $21 \leq j \leq 25$. The signs of the coefficients were determined in the same way as in GB1.

GB3. $(n, p, p_0) = (100, 1000, 10)$ and $\sigma = 1$. The relevant features were generated as iid standard normal variables with coefficients (3, 3.75, 4.5, 5.25, 6, 6.75, 7.5, 8.25, 9, 9.75). The irrelevant features were generated as

$$x_j = 0.25Z_j + \sqrt{0.75} \sum_{k \in s_0} X_k, j \notin s_0,$$

where $Z_j s$ are iid standard normal and independent from the relevant features.

In Table 8, we report the simulation results under the conditions for GB1, GB2, and GB3. We observed that SLasso and Edc + EBIC performed better. SLasso had the highest PDR while Edc + EBIC had the lowest FDR. Thus, when there was some correlation among the features, Edc + EBIC still performed well.

Table 8. We compared the methods using PDR, FDR, and model size (MSize) averaged over 500 simulation replications. The standard deviations are in parentheses.

Setting	Methods	MSize	PDR	FDR
GB1	ALasso	23.32 (3.018)	0.766 (0.062)	0.501 (0.066)
	SCAD	14.08 (1.644)	0.853 (0.065)	0.085 (0.068)
	SIS + SCAD	10.656 (1.688)	0.694 (0.112)	0.025 (0.067)
	SLasso	14.916 (2.194)	0.893 (0.081)	0.092 (0.089)
	Edc + EBIC	14.094 (2.035)	0.869 (0.088)	0.067 (0.076)
GB2	ALasso	40.474 (11.7331)	0.447 (0.0858)	0.710 (0.0605)
	SCAD	20.966 (7.6121)	0.517 (0.0614)	0.315 (0.1896)
	SIS + SCAD	10.314 (1.0797)	0.403 (0.0427)	0.042 (0.0712)
	SLasso	13.65 (2.038)	0.499 (0.052)	0.077 (0.081)
	Edc + EBIC	14.006 (1.657)	0.67 (0.014)	0.0273 (0.0785)
GB3	ALasso	22.464 (2.4414)	1.000 (0.000)	0.5495 (0.0498)
	SCAD	11.000 (0.000)	1.000 (0.000)	0.091 (0.000)
	SIS + SCAD	9.964 (0.6897)	0.992 (0.0764)	0.107 (0.050)
	SLasso	10.182 (0.475)	0.667 (0.006)	0.015 (0.039)
	Edc + EBIC	10.158 (0.440)	1.000 (0.000)	0.0139 (0.038)

3.2.3. Real Data Example

The new method was applied to the gene expression data used in [1]. For the data and details of data collection and variable definitions, see [1].

This study aimed to find the probes among the remaining 18,975 probes most closely related to TRIM32. The response variable was the expression level of probe 1389163_at. The features were the expression levels of the remaining 18,975 probes. Of the 18,975 probes,

the top 3000 probes with the largest variances were considered. The expression levels were standardized to have mean 0 and standard deviation 1.

In our analysis of the data, for each of the 100 replications we selected a random sample of size 100 from 120 rats and applied Edc + EBIC to the sample. From these 100 replications, Edc + EBIC selected the distinct probes 1367705_*at* and 1367728_*at*. In [1], the probes selected by five (5) variable selection methods are reported. The probes selected by Edc + EBIC did not intersect with any of the probes these methods selected. Among the probes selected by these five (5) methods, some intersected, but this is not a surprise because these methods essentially maximized only the linear relationship between TRIM32 and each of the probes. Since Edc + EBIC is capable of detecting and maximizing both the linear and nonlinear relationships that might exist between TRIM32 and each of the probes, as evidenced in the simulation studies by the high PDR and low FDR, we are convinced that the two probes selected by our method are the most associated with TRIM32.

4. Conclusions and Discussion

From the simulation results in Tables 6 and 7 we observed that, as the sample size (n) increased, Edc + EBIC selected on average the expected number of predictors and did so with decreasing standard deviations, meaning that through the simulation runs more and more of the selected predictors were close to the expected number of relevant predictors. We also observed the positive discovery rate of 100%, indicating that on average for each simulation run, out of the selected features all of the relevant features were selected. Of greater importance was the small false discovery rates recorded as the sample size increased.

We found through simulation studies that when we applied the Energy Distance Correlation Sure Independence Screening proposed by [9] for variable screening followed by a regularization method such as SCAD and ALasso, the average model size selected was higher than expected and with high standard deviations.

Author Contributions: Conceptualization, I.X.O. and H.C.; Methodology, I.X.O. and H.C.; Supervision, H.C. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data presented in this study is openly available and reported in Luo and Chen (2014) [1] at https://doi.org/10.1080/01621459.2013.877275, accessed on 19 November 2022.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Luo, S.; Chen, Z. Sequential Lasso cum EBIC for feature selection with ultra-high dimensional feature space. *J. Am. Stat. Assoc.* **2014**, *109*, 1229–1240. [CrossRef]
2. Tibshirani, R. Regression shrinkage and selection via the lasso. *J. R. Stat. Soc. Ser. B (Methodol.)* **1996**, *58*, 267–288. [CrossRef]
3. Fan, J.; Li, R. Variable selection via nonconcave penalized likelihood and its oracle properties. *J. Am. Stat. Assoc.* **2001**, *96*, 1348–1360. [CrossRef]
4. Zou, H. The adaptive lasso and its oracle properties. *J. Am. Stat. Assoc.* **2006**, *101*, 1418–1429. [CrossRef]
5. Zhang, C.H. Nearly unbiased variable selection under minimax concave penalty. *Ann. Stat.* **2010**, *38*, 894–942. [CrossRef] [PubMed]
6. Chen, J.; Chen, Z. Extended Bayesian information criteria for model selection with large model spaces. *Biometrika* **2008**, *95*, 759–771. [CrossRef]
7. Rizzo, M.L.; Székely, G.J. Energy distance. *Wiley Interdiscip. Rev. Comput. Stat.* **2016**, *8*, 27–38. [CrossRef]
8. Székely, G.J.; Rizzo, M.L.; Bakirov, N.K. Measuring and testing dependence by correlation of distances. *Ann. Stat.* **2007**, *35*, 2769–2794. [CrossRef]
9. Li, R.; Zhong, W.; Zhu, L. Feature screening via distance correlation learning. *J. Am. Stat. Assoc.* **2012**, *107*, 1129–1139. [CrossRef] [PubMed]

10. Fan, J.; Lv, J. Sure independence screening for ultrahigh dimensional feature space. *J. R. Stat. Soc. Ser. B (Stat. Methodol.)* **2008**, *70*, 849–911. [CrossRef] [PubMed]
11. Huang, J.; Ma, S.; Zhang, C.H. Adaptive Lasso for sparse high-dimensional regression models. *Stat. Sin.* **2008**, *18*, 1603–1618.
12. Kim, Y.; Choi, H.; Oh, H.S. Smoothly clipped absolute deviation on high dimensions. *J. Am. Stat. Assoc.* **2008**, *103*, 1665–1673. [CrossRef]