Regularized MAVE through the elastic net with correlated predictors

A Alkenani¹ and E Rahman²
¹²Department of Statistics, College of Administration and Economics, University of Al-Qadisiyah, Al Diwaniyah, Iraq.
E-mail: ali.alkenani@qu.edu.iq

Abstract. In this article, we proposed a model-free variable selection method (SMAVE-EN). The concepts of sufficient dimension reduction (SDR) and regularization methods are combined to introduce SMAVE-EN. This method is proposed to produce a shrinkage estimation when the predictors are highly correlated under SDR settings. The advantage of SMAVE-EN is that SMAVE-EN extended Elastic net (EN) to nonlinear and multi-dimensional regression under SDR settings. From another side, the SMAVE-EN enables MAVE to work with problems were the predictors are highly correlated. In addition, SMAVE-EN can exhaustively estimate dimensions, while selecting informative covariates simultaneously under SDR framework. The effectiveness of SMAVE-EN is evaluated by both simulation and real data analysis.

1. Introduction

When the variables number 𝑝 is huge, the analysis of regression can be very difficult. A useful tool to tackle this problem is to reduce the 𝑝-dimensional predictors vector 𝑥 without the loss any information about the regression.

The sufficient dimension reduction (SDR) concept was introduced by [1] to achieve the mentioned goal. Let 𝑦 is an outcome variable and 𝑥 = (𝑥₁, ..., 𝑥_𝑝)ᵀ is a 𝑝 × 1 vector of predictors. The SDR explores a 𝑝 × 𝑑 matrix 𝐵, achieves 𝑦 ∥𝑥|𝐵, where ∥ is the independency. The intersection of all the dimension reduction subspace is called the central subspace (𝑆_𝑦|𝐱). The 𝑆_𝑦|𝐱 contains all the regression information of 𝑦|𝐱 [2]. Many methods were proposed for finding 𝑆_𝑦|𝐱. For example, SIR [3], SAVE [4] and PHD [5].

From other hand, [6] introduced the concept of the central mean subspace (𝑆_𝐸(𝑦|𝐱)). In order to estimate 𝑆_𝐸(𝑦|𝐱), a lot of methods were proposed, such as IHT [6] and MAVE [7].

SDR methods give us a good tool to obtain sufficient DR; however, they suffer from that each direction of DR contains all the original predictors. This makes it not ease to interpret the resulting estimates.
The role of predictors selection is a crucial in regression. The selection of significant predictors to be in the model can improve the model. Moreover, the interpretation of the results about the small subset of predictors is easier than huge subset. Many methods were employed for selection the predictors in regression. For example, Lasso [8], SCAD [9], Elastic Net (EN) [10], adaptive Lasso [11] and MCP [12].

Under the settings of SDR, the ideas of regularisation were combined with a number of SDR methods. See for example, [13], [14], [15], [16], [17]. [18]) incorporate Lasso with MAVE to result sparse MAVE (SMAVE). Penalised MAVE (P-MAVE) was suggested by [19] through emerging bridge penalty with MAVE. MAVE was combined with MCP, adaptive Lasso and SCAD by [20] to get MCP-MAVE, ALMAVE and SCAD-MAVE, respectively. Lasso was employed by [21] with group-wise MAVE [22].

In this article, SMAVE-EN is proposed. The SMAVE-EN is working under SDR settings. SMAVE-EN has advantages over the SMAVE [18], SPMAVE [20] and P-MAVE [19]. It benefits from EN. EN has high ability for selection groups of highly correlated predictors. The penalties which are used in the above SDR methods lack this ability.

The rest of this article is as follows. A review of SDR and MAVE is reported in Section 2. SMAVE-EN was proposed in Section 3. Examples are carried out in Section 4. In Section 5, the compared methods were applied to prostate cancer data. The conclusions are stated in Section 6.

2. SDR and MAVE

Assume the model:

\[ y = f(x_1, x_2, \ldots, x_p) + \epsilon, \]  

where \( y, x \) and \( \epsilon \) are the outcome variable, a \( p \times 1 \) vector of predictors and the error term, respectively. Let \( E(y|x) = f(x_1, x_2, \ldots, x_p) \) and \( E(\epsilon|x) = 0 \). The SDR is searching for a subspace \( S \) such that

\[ y \perp E(y|x)|P_S x, \]  

where \( P_S \) is a projection operator.

If \( d = \text{dim}(S) \) and \( B = (\beta_1, \beta_2, \ldots, \beta_d) \) are basis for \( S \), \( x \) can be replaced with \( x^T \beta_1, x^T \beta_2, \ldots, x^T \beta_d, d \leq p \). [6] showed that \( S_{E(y|x)} \) is the intersection of all subspace satisfying (2). MAVE is one of the more efficient methods to estimate \( S_{E(y|x)} \). It was proposed to obtain \( B \) as a solution of

\[ \min_B \{ E[y - E(y|x^T B)]^2 \}, \]  

where \( B^T B = I_d \). The conditional variance given \( x^T B \) is

\[ \sigma^2_B(x^T B) = E[(y - E(y|x^T B))^2 | x^T B]. \]  

Thus,

\[ \min_B E[y - E(y|x^T B)]^2 = \min_B \{ \sigma^2_B(x^T B) \}. \]  

For any given \( x_0, \sigma^2_B(x^T B) \) can be locally approximated as
\[ \sigma^2_B(x_i^T B) \approx \sum_{l=1}^{n} (y_l - E(y_l|x_l^T B))^2 \omega_{l0} \]
\[ \approx \sum_{l=1}^{n} \{y_l - \{a_0 + (x_l - x_0)^T B b_0\}\}^2 \omega_{l0}, \]
where \( \omega_{l0} \geq 0 \) are the kernel weights with \( \sum_{l=1}^{n} \omega_{l0} = 1 \). So, \( B \) can be found by solving
\[
\min_{B: B^TB = I_m} \left( \sum_{j=1}^{m} \left( \sum_{l=1}^{n} \left( y_l - \{a_l + (x_l - x_j)^T B b_l\}\right)^2 \omega_{lj} + \lambda_1|\beta_m|_1 + \lambda_2|\beta_m|_2 \right) \right). \quad (6)
\]

3. **The Sparse MAVE with Elastic Net penalty (SMAVE-EN)**

[18] incorporate \( l_1 \) penalty in (6) to get SMAVE. The SMAVE minimises:
\[
\sum_{j=1}^{m} \sum_{l=1}^{n} \left( y_l - \{a_j + (x_l - x_j)^T B b_j\}\right)^2 \omega_{lj} + \lambda_1|\beta_m|_1, \quad (7)
\]
for \( m = 1, ..., d \).

where, \( d \) is known. [20] combined the adaptive Lasso, SCAD and MCP penalties with the loss function of MAVE in (6).

[19] add the bridge penalty to least squares form of MAVE in (6). The mentioned methods employed penalties that fail to work with grouped variables situation.

In this article, SMAVE-EN is proposed to minimise
\[
\sum_{j=1}^{m} \sum_{l=1}^{n} \left( y_l - \{a_l + (x_l - x_j)^T B b_j\}\right)^2 \omega_{lj} + \lambda_1|\beta_m|_1^2 + \lambda_2|\beta_m|_1, \quad (8)
\]
where, \( |\beta_m|_2^2 \) is \( l_2 \) norm refers to ridge and \( |\beta_m|_1 \) is \( l_1 \) norm refers to Lasso. The first part in (8) is the loss function of MAVE. The second and third parts are the ridge and Lasso penalties, respectively. The collection of ridge and Lasso is the EN penalty.

Also, \( \lambda_1 \) and \( \lambda_2 \) are the tuning parameters of EN. The SMAVE-EN algorithm is as follows:

1. Let \( m = 1 \), and \( B = B_0 \), any arbitrary \( p \times 1 \) vector.
2. For known \( B \), get \( (a_j, b_j) \) where \( j = 1, ..., n \), from
\[
\min_{a_j,b_j, j=1,...,n} \left( \sum_j \sum_{l=1}^{n} \left( y_l - \{a_j + (x_l - x_j)^T B b_j\}\right)^2 \omega_{lj} \right). \quad (10)
\]
3. For a given \( (\tilde{a}_j, \tilde{b}_j), j = 1, ..., n \), solve \( \beta_{mSMAVE-EN} \) from
\[
\min_{B: B^TB = I_m} \left( \sum_{j=1}^{m} \sum_{l=1}^{n} \left( y_l - \{\tilde{a}_j + (x_l - x_j)^T (\tilde{\beta}_1, \tilde{\beta}_2, ..., \tilde{\beta}_{m-1}, \beta_m) \tilde{b}_j\}\right)^2 \omega_{lj} + \lambda_1|\beta_m|_1^2 + \lambda_2|\beta_m|_1 \right) \quad (11)
\]
4. Replace the \( m \)th column of \( B \) by \( \tilde{\beta}_{msMAVE-EN} \) and continue steps 2 and 3 to convergence.
5. Update \( B \) by \( (\tilde{\beta}_{1SMAVE-EN}, \tilde{\beta}_{2SMAVE-EN}, ..., \tilde{\beta}_{msMAVE-EN}, \beta_0) \), and set \( m \) to be \( m + 1 \).
6. If \( m < d \), continue steps 2 to 5 until \( m = d \), where \( \omega_{lj} \) are kernel weights.
\[
\omega_{ij} = K_h \left\{ (x_i - x_j)^T \bar{B} \right\} / \sum_{i=1}^{n} K_h \left\{ (x_i - x_j)^T \bar{B} \right\},
\]

\(K_h\) is the Gaussian kernel and \(h_{opt} = A(d)n^{-1/(4+d)}\) is the bandwidth, where \(A(d) = \left\{ \frac{4}{(d+2)} \right\}^{1/(4+d)}\), see [7].

4. Simulation study

The SMAVE-EN is compared with some methods to explain the behavior of SMAVE-EN. The performance of SMAVE-EN was explained via a number of examples. Within each example the data is divided to training set, a validation set and test set.

Model 1. The single-index model was considered, where \(d = 1\). The data were simulated from the following model

\[
y = 1 + 2(x^T \beta + 3)\log(3|x^T \beta| + 1) + \varepsilon,
\]

where \(\varepsilon \sim N(0,1)\).

Example 1.

100 data sets contain 20/20/200 observations and \(p = 8\) were simulated. We let \(\beta = (3,1,5,0,0,2,0,0,0)\). Let \(\rho_{ij} = 0.5|\!|i-j|\!|\) is the pairwise correlation between \(x_i\) and \(x_j\).

Example 2.

The settings for this example is the same as example 1, except that \(\beta_j = 0.85\) for all \(j\).

Example 3.

100 data sets contain 100/100/400 observations and \(p = 40\) were simulated. We set \(\beta = (0, ..., 0, 2, ..., 0, 2, ..., 2)\),

\[
\begin{bmatrix}
10 & 10 & 10 & 10 \\
\end{bmatrix}
\]

where \(\rho_{ij} = 0.5\) for all \(i\) and \(j\).

Example 4.

100 data sets contain 50/50/400 observations and \(p = 40\) were simulated. The \(\beta\) is selected as follows:

\(\beta = (3, ..., 3, 0, ..., 0)\)

\[
\begin{bmatrix}
15 & 15 \\
\end{bmatrix}
\]

and the predictors were

\(x_i = Z_1 + \varepsilon^*, Z_1 \sim N(0,1), i = 1, ..., 5\),
\(x_i = Z_2 + \varepsilon^*, Z_2 \sim N(0,1), i = 6, ..., 10\),
\(x_i = Z_3 + \varepsilon^*, Z_3 \sim N(0,1), i = 11, ..., 15\),
\(x_i\) is i.i.d from \(N(0,1), i = 16, ..., 40\).

where \(\varepsilon^* \sim N(0,0.01), i = 1, ..., 5\). There are three groups in this model, and each group contains 5 predictors. Also, there are 25 predictors with zero coefficients.

Model 2. (Multiple-index model)

Example 5.

The multiple-index model was considered, where \(d = 2\). The 100 data sets contain 20/20/200 observations and \(p = 8\) were simulated from

\[
y = \frac{x^T \beta_1}{0.5 + (1.5 + x^T \beta_2)} + \sigma \varepsilon,
\]
where $x_i$ is normally distributed and $\varepsilon \sim N(0,1)$. Also, $\beta_1 = (3,1.5,2,0,0,0,0,0)^T$, $\beta_2 = (0,0,0,0,0,3,1.5,2)^T$ and $\sigma = 3$. For $\beta_1$, the first three predictors were highly correlated with pairwise correlation $r = 0.7$, while the rest were uncorrelated. For $\beta_2$, the first five predictors were uncorrelated, while last three predictors were highly correlated with pairwise correlation $r = 0.7$.

Table 1: Median mean-squared errors (MMSE) for the five examples and six methods based on 100 replications. Standard errors are in parentheses.

| Methods      | Example1 | Example2 | Example3 | Example4 | Example5 |
|--------------|----------|----------|----------|----------|----------|
|              | $\beta_1$ | $\beta_2$ | $\beta_1$ | $\beta_2$ | $\beta_1$ | $\beta_2$ |
| SMAVE        | 3.18 (0.59) | 3.98 (0.70) | 67.79 (3.20) | 46.55 (4.27) | 5.18 (0.80) | 5.24 (0.81) |
| SCAD-MAVE    | 2.89 (0.56) | 3.72 (0.65) | 65.39 (3.03) | 44.39 (4.06) | 4.85 (0.76) | 4.83 (0.77) |
| MCP-MAVE     | 2.94 (0.57) | 3.79 (0.66) | 65.98 (3.10) | 45.19 (4.19) | 4.97 (0.77) | 5.02 (0.79) |
| ALMAVE       | 2.87 (0.53) | 3.65 (0.60) | 63.47 (2.96) | 42.55 (4.01) | 4.80 (0.68) | 4.81 (0.70) |
| P-MAVE       | 2.90 (0.56) | 3.75 (0.64) | 65.84 (3.07) | 44.89 (4.08) | 4.92 (0.76) | 4.90 (0.77) |
| SMAVE-EN     | 2.72 (0.49) | 3.34 (0.45) | 58.11 (1.91) | 38.67 (1.83) | 4.50 (0.53) | 4.53 (0.52) |

From Table 1, it is obvious that the worst performance is for SMAVE. The performance of SMAVE-EN is significantly better than the competitors. For all examples, the performance of ALMAVE was near of SMAVE-EN performance. In summary, SMAVE-EN dominates the competitors under the collinearity.

Table 2: The average of zero coefficients (Ave 0's).

| Methods      | Example1 | Example2 | Example3 | Example4 | Example5 |
|--------------|----------|----------|----------|----------|----------|
|              | $\beta_1$ | $\beta_2$ | $\beta_1$ | $\beta_2$ | $\beta_1$ | $\beta_2$ |
| SMAVE        | 3.01 0    | 10.13 12.03 | 2.88 2.73 |
| SCAD-MAVE    | 3.60 0    | 13.29 15.77 | 3.51 3.40 |
| MCP-MAVE     | 3.50 0    | 12.18 14.38 | 3.36 3.24 |
| ALMAVE       | 3.75 0    | 15.35 18.03 | 3.55 3.48 |
| P-MAVE       | 3.58 0    | 13.09 15.59 | 3.46 3.35 |
| SMAVE-EN     | 3.90 0    | 15.67 18.25 | 3.79 3.70 |

From Table 2, SMAVE-EN results sparse models. The true significant predictors were selected through SMAVE-EN more than the all competitors. In case of grouping, the performance of SMAVE-EN was excellent. The EN's ability to do ‘grouped selection’ makes the performance of SMAVE-EN better than the performance of the competitors. The behavior of ALMAVE was near from the behaviour of SMAVE-EN and better than
the behavior of the rest. In general, the worst performance was for SMAVE for all the examples.

5. Prostate cancer (P.C) data
In this section, the P.C data were analysed by the compared methods. The data related with prostate cancer study [24]. The data are public and available from "lasso2" package in R. The predictors are 8 of clinical indexes and \( n = 97 \). The predictors are: log (volume of cancer) (lcavol), log(weight of prostate) (lweight), log (prost. hyper.) (lbph), age, semi. ves. inv. (svi), log(caps. penetr.) (lcp), Gleas. score (gle.) and percent. gleas. 4 or 5 (pgg45). The \( y \) is log prost.- antig. (lpsa).

The P.C data were randomly split into a training and test sets with \( n = 67 \) and 30, respectively. Selection of penalty parameters by tenfold CV and model fitting were implemented on the training data. The considered methods performance was compared by their predicted MSE.

| Method       | Test MSE   | Variables selected |
|--------------|------------|--------------------|
| SMAVE        | 0.489 (0.137) | (1,2,4,5,8)        |
| SCAD-MAVE    | 0.460 (0.129) | (1,2,3,6)         |
| MCP-MAVE     | 0.477 (0.133) | (1,2,4,5,8)        |
| ALMAVE       | 0.456 (0.124) | (1,2,3,6)         |
| P-MAVE       | 0.463 (0.130) | (1,2,5,6,8)        |
| SMAVE-EN     | 0.442 (0.117) | (1,2,5,6,8)        |

From Table 3, it is clear that the SMAVE-EN is better than the competitors in terms of sparsity and prediction precision. SMAVE is the worst method among the competitors. The performance of ALMAVE is near from the performance of SMAVE-EN and better than the performance of the rest methods. The SMAVE-EN selects lcavol, lweight, svi, lcp and pgg45 as the most important predictors. Also, the prediction error of the SMAVE-EN is lower than that of all the competitors.

6. Conclusion
SMAVE-EN is proposed. It incorporates EN in MAVE. MAVE estimates \( S_{E|Y|X} \) while EN does continuous selection and it helps groups selection. The SMAVE-EN has the strength of MAVE and EN. SMAVE-EN helps EN to work under MAVE settings. SMAVE-EN is carried out numerically with an efficient algorithm. From the results, it is obvious that SMAVE-EN gives accurate prediction and encourages groups selection under SDR settings.

We can extend this idea to other SDR methods, such as SIR [3], SAVE [4] and PHD [5]. Also, we can improve SMAVE-EN to work with binary response models.

References
[1] Cook, R. (1998). Regression graphics: ideas for studying the regression through graphics. New York, Wily.

[2] Yu, Z. and Zhu, L. (2013). Dimension reduction and predictor selection in semiparametric models. Biometrika, 100, 641-654.

[3] Li, K. (1991). Sliced inverse regression for dimension reduction (with discussion). Journal of the American Statistical Association 86, 316–342.

[4] Cook, R. D. and Weisberg, S. (1991). Discussion of Li (1991). Journal of the American Statistical Association 86, 328–332.

[5] Li, K. C. (1992). On principal Hessian directions for data visualization and dimension reduction: Another application of Stein’s lemma. Journal of the American Statistical Association 87, 1025–1039.

[6] Cook, R. D. and Li, B. (2002). Dimension reduction for the conditional mean in regression. The Annals of Statistics 30, 455–474.

[7] Xia, Y., Tong, H., Li, W. and Zhu, L. (2002). An adaptive estimation of dimension reduction space. Journal of the Royal Statistical Society Series B 64, 363–410.

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

[9] Fan, J. and Li, R. Z. (2001). Variable selection via non-concave penalized likelihood and its oracle properties. Journal of the American Statistical Association 96, 1348–1360.

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

[11] Zou, H. (2006). The adaptive Lasso and its oracle properties. Journal of the American Statistical Association 101, 1418–142.

[12] Zhang, C. H. (2010). Nearly unbiased variable selection under minimax concave penalty. Annals of Statistics 38, 894–942.

[13] Li, L., Cook, R. D. and Nachtsheim, C. J. (2005). Model-free variable selection. Journal of the Royal Statistical Society Series B, 67, 285–299.

[14] Ni, L., Cook, R. D. and Tsai, C. L. (2005). A note on shrinkage sliced inverse regression. Biometrika 92, 242–247.

[15] Li, L. and Nachtsheim, C. J. (2006). Sparse sliced inverse regression. Technometrics 48, 503–510.

[16] Li, L. (2007). Sparse sufficient dimension reduction. Biometrika 94, 603–613.

[17] Li, L. and Yin, X. (2008). Sliced Inverse Regression with regularizations. Biometrics 64, 124–131.

[18] Wang, Q. and Yin, X. (2008). A Nonlinear Multi-Dimensional Variable Selection Method for High Dimensional Data: Sparse MAVE. Computational Statistics and Data Analysis 52, 4512–4520.

[19] Wang, T., Xu, P. and Zhu, L. (2013). Penalized minimum average variance estimation. Statist. Sinica 23 543–569.

[20] Alkenani, A. and Yu, K. (2013). Sparse MAVE with oracle penalties. Advances and Applications in Statistics 34, 85–105.

[21] Wang, T., Xu, P., Zhu, L. (2015). Variable selection and estimation for semiparametric multiple-index models. Bernoulli 21 (1), 242–275.
[22] Li, L., Li, B. and Zhu, L.-X. (2010). Groupwise dimension reduction. Journal of the American Statistical Association. 105, 1188–1201.
[23] Efron, B., Hastie, T., Johnstone, I. and Tibshirani, R. (2004). Least angle regression. The Annals of Statistics 32, 407–499.
[24] Stamey, T., Kabalin, J., McNeal, J., Johnstone, I., Freiha, F., Redwine, E. and Yang, N. (1989) Prostate specific antigen in the diagnosis and treatment of adenocarcinoma of the prostate ii: radical prostatectomy treated patients. J. Urol., 16, 1076–1083.