sparse multiresponse regression estimation to select the variables affecting thyroid disorders

Saja Mohammad Hussein
Abdulqader Ahmed Jasim
Department of Statistics, Faculty of Administration and Economics/University of Baghdad, Iraq
Corresponding Author : Abdulqader Ahmed Jasim1 *Email: abdulqaderahmed79@gmail.com*

***Abstract* :** With the advance of technology, the collection and storage of data have become routine. Huge amounts of data are increasingly produced from biology, meteorology, psychology, chemistry, and economics experiments. As technology progresses, these high-dimension problems are becoming more and more common. The "large p, small n" problem, in which there are more variables than samples, is currently a challenge that many statisticians face especially when it becomes multiresponse. Many researchers have resorted to using the sparse regression of the response variable in the multivariate case. A sparse matrix is defined as a matrix with the majority of its members equal to zero. A sparse matrix's zero entries reduced the number of parameters that may be interpreted. In this paper, we focus on the comparison between the four method: SiER, SRRR, remMap, SPLS in the selection of variables that affect the hormones that cause for the data of a group of patients with thyroid disorders. Software implementing the method is publicly available in the R package sparse-reg.

. Keywords - High dimension, multiresponse, sparse, Dimension reduction, SiER, SRRR, SPLS, remap

INTRODUCTION: As the capacity of data storage evolves, high-dimensional data are ubiquitous in many facets of research. People are able to mine and store thousands or millions of features of a given object while the number of objects is limited due to nature.

This phenomenon displays widely in various areas: in biological and medical research, the microarray data that contains tens of thousands gene expressions of the patients is recorded yet the number of patients is usually a few hundred, large volatility matrices with dimension thousands by thousands need to be estimated while the history of high-quality data is no longer than 20 years. Thus, high-dimension statistics has become a focus of the statisticians in the last two decades, and it becomes a critical problem to detect meaningful statistical relationships when the useful information is immersed in the peripheral noises.

Estimation of coefficient matrix, particularly in situations where the data dimension p is comparable to or larger than the sample size n.

So in this paper we try to overcome the problem of p > n with multiresponse by using estimation methods sparse regression as an estimation method for the regression model. sparse regression has gone through a process of improvement in statistics over a long period, and is often used in many fields. According to its meaning, "sparse" estimation only selects meaningful explanatory variables from a large number of candidates, and estimates the parameters of other variables to be exactly zero. Popular choices of regularizations methods including ridge[1], least absolute shrinkage and selection operator (lasso)[10], principal components regression[5], and partial least squares regression[7], can produce biased regression estimates but they may improve prediction performance because the regression coefficients could have smaller variances compared to coefficients from OLS..

affecting thyroid disorders is one of the most common diseases found in human being, it is not a deadly disease, but it is chronic disease which can give rise to other diseases, several sparse estimation regression methods for select variables affecting thyroid disorders.

Sparse Penalized Regression [4]

In a typical linear regression model, the unknown vector is estimated using the ordinary least squares (OLS) approach, which is a type of estimation method. The residual sum of squares (RSS) criteria function is minimized using the OLS estimator.

$$\min_{\boldsymbol{\beta}\in\boldsymbol{R^p}} \|y - X\beta\|^2 \qquad (1)$$

When n > p and X is of full rank, the preceding minimization yields the well-known closed-form solution $\widehat{\boldsymbol{\beta}} = (\boldsymbol{X^T X})^{-1}\boldsymbol{X^T Y}(i.e.,$ rank $p)$.

All components of the OLS estimate, on the other hand, will be typically non-zero. Furthermore, in the scenario when p > n, the issue is ill-posed and does not have a single solution that can be identified. In order to find a sparse solution, The technique of variable selection is frequently used to select the best subset of predictors that produce a good fit while using the fewest number of variables possible.

A subset selection strategy was traditionally used to choose the variables. For example, in the first technique, the best subset of *K* variables is chosen that fits the data the best amongst all possible, $p!/k!\,(p-k)!$, which are combinations of K predictors.

However, when K increases, the number of possible combinations rapidly grows to an unmanageably enormous amount. The second approach is to employ stage wise subset selection procedures: The process of backward elimination consists of starting with the full model (i.e., the model that includes all predictors) and sequentially removing the least impactful predictor from the model at each step until only K predictors are left; the process of forwarding selection involves starting with a null model (that does not contain any predictors), then adding the best predictor one by one, until only K predictors are left. Iterative forward stage-wise selection of predictive variables is a less greedy and computationally expensive method for subset selection of predictive variables than iterative backward stage-wise selection because it requires multiple short stages to get the final model.

Sparse Partial Least Squares[2]

Sparse partial least squares (SPLS) regression, introduced by Chun and Keles (2010) [2], is an alternative technique for variable selection. SPLS was developed based on partial least squares (PLS) regression, which was introduced by Wold (1966) [11]. PLS, which is a dimension reduction technique coupled with a regression model, has been widely used for handling the collinearity among variables. PLS attempts to obtain a smaller number of latent components that are linear combinations of the original variables. It has been shown to achieve good predictive performance when the number of variables (p) is not too large, typically, when the sample size (n) is larger than p.

Suppose in a study that p variables are under consideration. Let $\boldsymbol{W}$ be a loading matrix consisting of $\boldsymbol{K}$ vectors $\boldsymbol{(w_i, ..., w_k)}$ of length $\boldsymbol{p}$ such that $\boldsymbol{XW}$ gives the linear combination of the original design matrix. It follows that $\boldsymbol{K}$ is the number of latent components. Similar to the idea of the LASSO, SPLS imposes an $\boldsymbol{l_1}$ constraint on the $\boldsymbol{w}$ in order to obtain a sparse solution. The solution of $\boldsymbol{w_k}$ is

$$\boldsymbol{w_k} = \arg\max \{ w^T X^T Y Y^T X w \} \qquad (2)$$

such that $w^T w = 1$ and $|w| < \lambda$, for $k = 1, \ldots, K$. Here, λ is the tuning parameter that determines the amount of sparsity, **X** is the design matrix, and **Y** is the response vector or response matrix for multivariate responses. However, the solutions are found to be insufficiently sparse and convexity is not achieved. SPLS reformulates the objective function by imposing the $\boldsymbol{l_1}$ penalty onto the surrogate direction vector **c** instead of the original direction vector **w**. The reformulated objective function is given by

$$\min - \kappa \boldsymbol{w^T M w} + (1-\kappa)(c-w)^T \boldsymbol{M} (c-w) + \lambda_1 \|c\|_1 + \lambda_2 \|c\|_2 \qquad (3)$$

such that $\boldsymbol{w^T w} = 1$, where $M = X^T Y Y^T X$. The first $\boldsymbol{l_1}$ penalty encourages sparsity and the second $\boldsymbol{l_2}$ penalty avoids potential singularity in **M**. The objective function has four tuning parameters: κ , λ_1, λ_2 , and K. So, here, only $\lambda_1$and K need to be determined. In the SPLS algorithm, $\lambda_1$is not determined directly. Instead, a soft thresholded estimate of w, given by

$$\widetilde{w} = (\|\widehat{w}\| - \eta \max \|\widehat{w}_i\|) I_{(\|\widehat{w}\| \geq \eta \max_{1 \leq i \leq p} \|\widehat{w}_i\|)} \operatorname{sign}(\widehat{w}) \qquad (4)$$

where $0 \leq \eta \leq 1$, plays the role of λ_1 (Chun and Keles, 2010) [2]. Thus, we need to find optimal values for η and K. We will do this by cross-validation.

REgularized Multivariate regression for identifying MAster Predictors (remMap)

Peng, 2010[8] introduced this method. It was developed for data sets with a higher number of explanatory and response variables than its number of samples.

RemMap, Regularized Multivariate regression for identifying MAster Predictors, was proposed to fit multivariate response regression models under the high dimension-low-sample-size setting. The motivation of remMap is to explore the regulatory relationships among different biological molecules from multiple types of high dimensional genomic data, especially the modulation effect of copy number variation on gene expression. RemMap can build a multivariate linear regression model with an $\boldsymbol{l_1}$ norm penalty to control the overall sparsity of the coefficient matrix and a group lasso penalty [3] to control the total number of predictors entering the model. Consequently, the detection of master predictors can be facilitated.

Peng et al. applied remMap method to gene expression and copy number variation data to investigate the influences of DNA copy number alterations on RNA transcript levels based on 172 breast cancer tumor samples. RemMap can also be utilized to study the relationships between other types of biological molecules [8]. It can be applied to other models as well to select a group of variables in a multiple regression model. In our study, remMap was utilized to select genes whose expression could affect single drug sensitivity.

This method aims to induce overall sparsity so that only significant genetic relationships are included, and it aims to induce group sparsity so that explanatory variables that affect a high number of response variables are prioritized. To do this, this technique includes an

$$\ell_{(\lambda_1,\lambda_2)}(Y,B) = \frac{1}{2}\|Y - XB\|_F^2 + \lambda_1 \sum_{j=1}^{p}\left\|C_j B_j\right\|_1 + \lambda_2 \sum_{j=1}^{p}\left\|C_j B_j\right\|_2 \qquad (5)$$

Where $C = (C_{pq})$ is a p × q 0-1 matrix indicating the coefficients of B on which penalization is imposed. In the above equation C_j and B_j are the ***jth*** rows of C and B respectively, while$\|\cdot\|_F$ denotes the Frobenius norm (Euclidean norm) of matrices, $\|\cdot\|_1$ and $\|\cdot\|_2$ are respectively the ℓ_1 and $\ell_2$norms of vectors and " · " stands for the Hadamard product (entry-wise multiplication).

The selection matrix C is pre-specified based on prior knowledge: if we know in advance that predictor x_i affects response y_j, then the corresponding regression coefficient B_{ij}will not be penalized and we set C_{ij}= 0. When there is no such prior information, C can be simply set to a constant matrix $C_{ij} \equiv 1$. Finally, an estimate of the coefficient matrix B :

$$\widehat{\boldsymbol{B}}_{(\lambda_1,\lambda_2)} := \mathbf{argmin}_{\boldsymbol{B}}\boldsymbol{\ell}_{(\lambda_1,\lambda_2)}(\boldsymbol{Y},\boldsymbol{B}) \qquad (6)$$

In the above criterion function where λ_1 and λ_2 are two nonnegative tuning parameters the combined penalty in equation (5) is referred to as the MAP (MAster Predictor) penalty. The first penalty term is ℓ_1 penalty induces the overall sparsity of the coefficient matrix B. while the second penalty is the ℓ_2 penalty on the row vectors $C_j \cdot B_j$ induces row sparsity of the product matrix $\boldsymbol{C}\cdot\boldsymbol{B}$. As a result, some rows are shrunken to be entirely zero. predictors which affect relatively more response variables are more likely to be selected into the model.

Sparse Reduced-Rank Regression

Alternating minimization was proposed by Chen and Huang [6] to solve this non-convex optimization problem, where two algorithms — subgradient descent and a variational method were considered. The subgradient method was shown to be faster when $p \gg n$ and the variational method faster when $p \ll n$. However, in each iteration, the computational complexity of either method is at least quadratic in the number of variables p. It makes the problem almost intractable in ultrahigh-dimension problems, which is common, for example, in modern genetic studies.

Given a rank r, we are going to solve the following penalized bounded-rank optimization problem:

$$\text{Minimize} \quad \frac{1}{2}\|\mathrm{Y} - \mathrm{X}\,\mathrm{B}\|_F^2 + \lambda \sum_{j=1}^{p}\left\|\mathrm{B}_{j\cdot}\right\|_2 \qquad (7)$$

$$\text{s.t} \quad \text{rank}(\mathrm{B}) \leq \mathrm{r}.$$

where the superscript denotes a row of the matrix so that $\mathbf{B}^i$ is a *row* vector, the constraint $\mathbf{A}^T\mathbf{A} = \mathbf{I}$ is introduced for identifiability purposes, and $\boldsymbol{\lambda_i} > \mathbf{0}$ are penalty parameters.

In (7), each row of B is treated as a group and $\|\boldsymbol{B^i}\| = 0$ is equivalent to setting the *ith* row of **B** as zeros. Thus the group lasso penalty encourages row-wise sparsity on the **B** matrix. An alternative way of introducing sparsity is to directly use the lasso penalty (Tibshirani 1996) on the entire matrix B that encourages element-wise sparsity.

Signal extraction approach to multivariate regression

Proposed by X. Qi, R. Luo(2017)[9], for dimension reduction and regression in multiple response linear model with high-dimension predictor variables. This approach considered the decomposition of the coefficient matrix B, let K denote the rank of B, each coefficient vector βj can be expressed as a linear combination of $w_1, \ldots, w_K$. So we have the decomposition

$$\boldsymbol{B} = \mathrm{A}\,\mathrm{W}^{\mathrm{T}} = \boldsymbol{\alpha}_{\mathbf{k}}\mathbf{w}_{\mathbf{1}}^{\mathbf{T}} + \cdots + \boldsymbol{\alpha}_{\mathbf{k}}\mathbf{w}_{\boldsymbol{K}}^{\mathbf{T}} \qquad (8)$$

Where A= [$\alpha_1, \ldots, \alpha_K$] and W= [$w_1, \ldots, w_K$] are $p \times K$ and $q \times K$ matrices, respectively. There are infinitely many choices of $w_1, \ldots, w_K$ hence the decomposition (8) is not unique, we will consider a different decomposition which leads to the best lower rank approximation to $\mathbf{X}\boldsymbol{B}$, We call $\mathbf{X}\mathfrak{B}$ the signal matrix in the response matrix Y.

We make the following sparsity assumption only a small number of the coefficient vectors $\beta_1,\ldots,\beta_p$, are nonzero vectors. Since these vectors are the row vectors of $\boldsymbol{B}$, this assumption is the row-wise sparsity of $\boldsymbol{B}$. implies that $\boldsymbol{\alpha_k}$ is a sparse vector and the number of its nonzero coordinates is less than or equal to the number of nonzero vectors among $\beta_1,\ldots,\beta_p$. Motivated by the sparsity of $\boldsymbol{\alpha_k}$, we propose the following penalized optimization problem whose solution is the sparse estimate$\widehat{\boldsymbol{\alpha}}_{\mathbf{k}}$ of $\boldsymbol{\alpha_k}$

$$\text{Max}\ \frac{\alpha^{\mathrm{T}}\widehat{\mathrm{B}}\alpha}{\alpha^{\mathrm{T}}\mathrm{S}\alpha + \boldsymbol{\tau}\|\boldsymbol{\alpha}\|_{\boldsymbol{\lambda}}^2} \quad \text{subject } \alpha^{\mathrm{T}}\mathrm{S}\alpha = 0 \ , 1 \leq l \leq \mathrm{k}-1 \qquad (9)$$

Where $\|\alpha\|_{\lambda}^{2}$ = (1- λ) $\|\alpha\|_{2}^{2}$ + λ $\|\alpha\|_{1}^{2}$, is a mixture of the squared l_2 and squared l_1 norms. and both $\tau \geq 0$ and $0 \leq \lambda < 1$are tuning parameters. In the penalty $\tau \|\alpha\|_{\lambda}^{2}$ the l_2 term is used to overcome the singularity problem of **S** and the l_1 term encourages the *sparsity of* $\hat{\boldsymbol{\alpha}}_{\boldsymbol{k}}$.

Real data application

Thyroid disorders are one of the most important and common diseases in the world because they affect almost all tissue cells of the body, causing changes in vital activities, as they play an important role in maintaining the body's metabolic rate, affecting the nervous system, pituitary glands, general circulation, and plasma proteins, as well as various metabolic mechanisms. Such as its effect on increasing the expenditure of the basic energy of the body used for the metabolism of proteins, carbohydrates, and fats. The thyroid gland is necessary for life, but its absence causes a mental and physical slowdown and poor resistance to cold. In children, it causes short stature and mental retardation. On the contrary, hyperthyroidism leads to weight loss, nervousness, increased heart rate, tremors, and an increase in temperature. The study community includes data for people suffering from a thyroid disorder, in which the sample size (n = 39) is less than the number of variables (p = 52), collected from the laboratories of Azad Hospital - Kirkuk Governorate - Iraq country, where both the concentration of the hormone, hormone, thyroid hormone were studied, this is done by performing the examinations indicated Thyroid stimulating hormone ,Thyroid hormone ,Thyroid hormone ,free triiodothyronine and its relationship and the effect of a set of examinations on it : age, gender, weight, height, bodymassindex, RBS, HbA1c, Total cholesterol, LDL (cholesterol), HDL(Cholesterol),Triglycerides,SystolicBP,Diastolic(BP),pulse,Creatinine,Urea,Haemoglobin,RDW,Sodium,Calcium,Chlorid(CL),osmolality,Magnisium,potassium,Phosphorus,Uricacid,Albumin,Iron,S.Ferritin,TSB,DirectSB,IndirictB,S.GOT,S.GPT,S.ALKPh,PCV,ALP,Zinc,CO2 , alkaline phosphatase ,platelets per microliter, folic acid, Vitamin B12 test ,Vitamin D ,C-reactive protein (CRP) , Glucose urine, Total Urine Protein , Ketones in Urine ,Phosphate in Urine. The Mean square errors (MSE) and Parameter estimates for each of the methods are shown in Table (1) and (2).

Table 1: The Parameter estimates by methods
(SRRR, SiER, SPLS, remMap)

method		Y4	Y3	Y2	Y1
SRRR	β_6	0.1463	0.1496	0.1904	0.1678
	β_7	0.1421	0.1442	0.1888	0.1519
	β_8	0.1394	0.1400	0.1808	0.1515
	β_9	0.1382	0.1331	0.1601	0.1425
	β_{10}	0.1372	0.1013	0.1191	0.1421
	β_{11}	0.1288	0.0915	0.1168	0.1408
	β_{15}	0.1125	0.0821	0.0907	0.0966
	β_{16}	0.0904	0.0635	0.0559	0.0646
	β_{17}	0.0890	0.0538	0.0444	0.0502
SiER	β_7	0.0170	0.0637	0.0562	0.0629
	β_8	0.0128	0.0480	0.0424	0.0475
	β_9	0.0127	0.0476	0.0421	0.0471
	β_{10}	0.0026	0.0099	0.0087	0.0098
	β_{11}	0.0018	0.0067	0.0059	0.0066
	β_{15}	0.0212	0.0300	0.0140	0.0000
	β_{16}	0.0110	0.0188	0.0145	0.0222
	β_{17}	0.0237	0.0310	0.0275	0.0166
	β_{18}	0.0205	0.0766	0.0676	0.0757
	β_{41}	0.0232	0.0662	0.0762	0.0558
	β_{42}	0.0170	0.0637	0.0562	0.0629
SPLS	β_6	0.1534	0.1289	0.1790	0.1512
	β_7	0.1423	0.1234	0.1717	0.1266
	β_8	0.1417	0.1198	0.1529	0.1219
	β_9	0.1342	0.1094	0.1483	0.1028
	β_{10}	0.1266	0.1023	0.1383	0.0991
	β_{11}	0.1169	0.0891	0.1242	0.0846
	β_{17}	0.0661	0.0339	0.0589	0.0342
	β_{27}	0.0185	0.0416	0.0109	0.0463
	β_{28}	0.0136	0.0434	0.0292	0.0505
	β_{41}	0.0959	0.1582	0.1299	0.1442
remMap	β_7	0.0796	0.0473	0.0523	0.0424
	β_8	0.0774	0.0353	0.0507	0.0259
	β_9	0.0733	0.0131	0.0359	0.0155
	β_{10}	0.0582	0.0018	0.0111	0.0110
	β_{11}	0.0525	0.0000	0.0106	0.0095
	β_{15}	0.0450	0.0018	0.0214	0.0000
	β_{16}	0.0020	0.0135	0.0164	0.0192
	β_{17}	0.0606	0.0137	0.0170	0.0058

	β_{18}	0.0318	0.0195	0.0000	0.0000
	β_{28}	0.0199	0.0160	0.0154	0.0160
	β_{36}	0.0000	0.0334	0.0102	0.0102
	β_{41}	0.0062	0.0699	0.0506	0.0433
	β_{42}	0.0065	0.0.845	0.0653	0.0514

According to Table (1), when we applied the SiER method, we obtained estimated the parameters of the variables, HbA1c, Total cholesterol, LDL cholesterol, HDL Cholesterol, Triglycerides, Urea, Haemoglobin, RDW, platelets per microliter, folic acid. get estimated the parameters of the variables, RDW, platelets per microliter, folic acid.RBS, HbA1c, Total cholesterol, LDL cholesterol, HDL cholesterol, Triglycerides, Creatinine, Urea, and Haemoglobin, when applied the SRRR method. Get estimated the parameters of the variables, HbA1c, Total cholesterol, LDL cholesterol, HDL Cholesterol, Triglycerides, Creatinine, Urea, Haemoglobin, RDW, Iron, PCV, platelets, folic acid, when applied the remMap method, and get estimated the parameters of the variables, RBS, HbA1c, Total cholesterol, LDL cholesterol, HDL Cholesterol, Triglycerides, Haemoglobin, Albumin, Iron, when applied the SPLS method.

Table 2: The Mean Square Error (MSE) for methods

method	MSE
SRRR	4.4456
SiER	5.7676
SPLS	5.7877
reMmap	5.9507

From the table (2), it is clear that a method SRRR is the best performance, followed by method, SiER method, SPLS, and then remMap method

Conclusions

The above results show that SRRR method is better than other methods where we comparison the MSE four methods, followed, followed by SiER method, SPLS, and then remMap method .When we applied the SRRR method, we obtained the estimated parameters of (9) of the variables, obtained the estimated parameters of (10) of the variables When applied the SiER method, obtained the estimated parameters of (9) of the variables when applied the SPLS method, obtained the estimated parameters of (13) of the variables when applied the remMap method,and there are several common variables chosen by the four methods: HbA1c, Total cholesterol, LDL cholesterol, HDL Cholesterol, Triglycerides**.**

References

1. A. E. Hoerl and R. W. Kennard.Ridge regression: biased estimation for nonorthogonal problems. *Technometrics, vol(12), no. 1, pp. 55–67*, 1970.

2. Chun, S. Keles, Sparse partial least squares regression for simulatenous dimension reduction and variable selection, *J. R. Stat. Soc. Ser. B Stat.Methodol.vol(72*), pp.3–25. 2010.

3. Friedman J, Hastie T, Tibshirani R. A note on the group lasso and a sparse group lasso. *arXiv preprint arXiv:1001.0736*, 2010.

4. Guan Yu, Liang Yin, Shu Lu & Yufeng Liu. Confidence Intervals for Sparse Penalized Regression with Random Designs. *Journal of the American Statistical Association. pp. 794-809*. 2020.

5. Jolliffe, I.T.,. A note on the use of principal components in regression. Appl. *Stat. 31, p.p.300–303*. 1982.

6. Lisha Chen and Jianhua Z Huang. Sparse reduced-rank regression for simultaneous dimension reduction and variable selection. *Journal of the American Statistical Association, Vol(107),p.p.1533-1545*,2012.

7. P. Geladi and B. R. Kowalski. A Partial least-squares regression: a tutorial. *Anal. Chim. Acta. Vol(185), pp. 1–17*. 1986.

8. Peng, J. Zhu, A. Bergamaschi, W. Han, D.Y. Noh, J.R. Pollack, P. Wang, Regularized multivariate regression for identifying master predictors with application to integrative genomics study of breast cancer, A*nn. Appl. Stat. 4 53*. 2010.

9. R. Luo X. Qi. Signal extraction approach for sparse multivariate response regression. *J Multivar Anal.pp. 97-83*. 2017.

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

11. Wold, H. Estimation of Principal Components and Related Models by Iterative Least Squares. New York: Academic Press.1966.