Informative Measures of Significance for Constructing Intelligent Feature Weights

Samuel Müller[1,4], Tanya P. Garcia[2], Raymond J. Carroll[3]

[2]School of Mathematics and Statistics, University of Sydney, NSW 2006 Australia

[2]Department of Epidemiology and Biostatistics, Texas A&M University, College Station, TX 77843-3143

[3]Department of Statistics, Texas A&M University, College Station, TX 77843-3143

[4]Corresponding author: Samuel Mueller, e-mail:
samuel.mueller@sydney.edu.au

Abstract

When part of the regressors can act on both the response and some of the other explanatory variables, the already challenging problem of selecting variables in a $p > n$ context becomes more difficult. A recent methodology for variable selection in this context links the concept of q-values from multiple testing to the weighted Lasso. In this talk, we show that different informative measures of significance to q-values, such as partial correlation coefficients or Benjamini-Hochberg adjusted p-values, give similarly promising performance as when using q-values.

Keywords: adjusted p-values, q-values, partial correlation coefficients, variable selection

1. Introduction

A common issue in modeling biological and social phenomena is variable selection, where data in such fields often involve more predictor variables than samples. A new difficulty to this already challenging task is selecting predictor variables in a *structured* way so that an existing hierarchy among the model variables is obeyed. In some instances, for example, some predictors are known to act on both the response and other candidate predictors; thus, one must select which *candidate* variables affect the response after accounting for those predictors *known* to affect the response. Recently, Garcia et al. (2013) proposed a novel method for handling such structured variable selection problems; their method involves extracting q-values in multiple hypothesis testing (Storey, 2003) and using them as weights in the weighted Lasso (Zou, 2006) to appropriately direct the selection procedure. In this paper, we take a closer look at their proposed method and through various simulation studies, we determine if weights other than the q-values could improve the procedure.

Performing structured variable selection is a challenge and extends beyond what earlier selection procedures can handle. Earlier methods include the Lasso (Tibshirani, 1996) and its extensions (Yuan and Lin, 2006; Zou 2006; Meinshausen and Bühlmann, 2010) least angle regression (Efron et al, 2004); and selection via controlling false discovery rates (Benjamini and Hochberg, 1995; Storey, 2003). To remedy this gap in the literature, the method of Garcia et al. (2013) was developed. The method modifies the weights in the weighted Lasso in such a way that certain variables are ensured to be in the final model, and that important candidate variables are selected over less important ones. The method provides a proper multivariate analysis by collectively considering all relevant information in the model variables, and ultimately results in selections with acceptable false positive rates and low false discovery rates. This contrasts from individually assessing which predictors are related to the response through simple measures of correlations or partial correlations.

Our aim in this paper is to further explore the weight selection to perhaps further improve the method of Garcia et al. (2013). The rest of the paper is organized as follows. Section

2 provides a brief overview of the modified weighted Lasso proposed by Garcia et al. (2013). We discuss additional weights that could be considered. Section 3 describes various simulation studies to assess the use of different weights. Section 4 concludes the paper.

2. Main Results

Let the sample size be n, $y = (y_1, \ldots, y_n)^T$ be the response variable, v_j, $j = 1, \ldots, m$ denote the $n \times 1$ covariates that we suppose are linearly related to y. The covariates are divided into two groups: those that need to be included in the model (i.e., fixed covariates), and those that are subject to selection. Let m_0 denote the number of fixed covariates, which are denoted as $v_1 := z_1, \ldots, v_{m_0} := z_{m_0}$. Let m_1 denote the number of covariates subject to selection, which are denoted as $v_{m_0+1} := x_1, \ldots, v_{m_0+m_1} := x_{m_1}$. We have that $m = m_0 + m_1$. Without loss of generality, we assume all variables are standardized to have mean zero and sample variance one, so that the intercept is excluded from the regression model. We also suppose that $m_0 + 1 \leq n$, but we allow for $m_0 + m_1 = m > n$. For ease in presentation, we refer collectively to all covariates as v's, whereas we may refer to the fixed covariates as z's and covariates subject to selection as x's.
For the $m > n$ variable selection problem, a commonly used method is the weighted Lasso (Zou, 2006) which minimizes

$$Q(\beta) = \frac{1}{2} \left\| y - \sum_{k=1}^{m} v_k \beta_k \right\|^2 + \lambda \sum_{k=1}^{m} w_k |\beta_k|,$$

with respect to $\beta = (\beta_1, \ldots, \beta_m)^T$. Here, $\lambda > 0$ is a regularization parameter, $w_k > 0$, $k = 1, \ldots, m$, are weights, and $\|\cdot\|$ denotes the L_2-norm. We denote the minimizer as $\widehat{\beta}$, which will be a function of λ and the weights $w = (w_1, \ldots, w_m)^T$.
To gain insight into the minimizer $\widehat{\beta}$, let $r_{(-k)} = y - \sum_{j \neq k} v_j \beta_j$, $k = 1, \ldots, m$, denote the partial residual after removing the kth covariate. Through a careful derivation involving subgradients (see Garcia and MÏller (2013) for a full derivation), we have that if

$$\left| v_k^T r_{(-k)} \right| \leq \lambda w_k, \tag{1}$$

then $\widehat{\beta}_k = 0$; otherwise,

$$\widehat{\beta}_k = \text{sign}(v_k^T r_{(-k)})(|v_k^T r_{(-k)}| - \lambda w_k)/v_k^T v_k,$$

for $k = 1, \ldots, m$.
From (1), it is apparent that for a fixed λ, variables v_k with large weights w_k will generally not be included in the model (i.e., $\widehat{\beta}_k = 0$); whereas, those variables with small weights w_k will generally be included in the model (i.e., $\widehat{\beta}_k \neq 0$). Using this key property, Garcia et al. (2013) extended the work of Charbonnier et al. (2010) and Bergersen et al. (2011) by formulating a method so that important variables were included in the model before less important ones.
First, choosing the weights so that $\max(w_1, \ldots, w_{m_0}) / \min(w_{m_0+1}, \ldots, w_{m_0+m_1})$ is arbitrarily close to zero will ensure that the z's are selected before the x's. For example, setting $w_1 = \ldots = w_{m_0} = 0$ guarantees that the z's are always selected. In our simulations we will set $w_j = 10^{-5}$ on z_j, $j = 1, \ldots, m_0$, to explore to what extent such small weight values lead to the exclusion of the z's.
Second, the x's are weighted according to some measure of significance of their relationship to y after accounting for the z's. For example, this measure of significance can be based on the partial correlation between x and y after accounting for z's; i.e.,

$\rho_{x_k,y|z_1,\ldots,z_{m_0}}$. In this case, we could weigh each x_k with $w_{m_0+k} = 1/|\rho_{x_k,y|z_1,\ldots,z_{m_0}}|$, so that x's most correlated with y are included first in the model selection before those x's least correlated with y.

Similarly, we could also use results from m_1 separate linear regressions: for $k = 1, ..., m_1$, run a linear regression of y on $(z_1, ..., z_{m_0}, x_k)$, and compute the p-value p_k for the effect of x_k in this regression. Assuming $m_0 + 1 \leq n$, these regressions are valid. Then, a measure of the significance of x's could be the computed p-values or adjusted p-values to account for the multiplicity of tests. For example, we could adjust the p-values using Benjamini-Hochberg methods (Benjamini and Hochberg, 1995) or using q-values (Storey and Tibshirani, 2003). The p-values, or their adjusted versions, can then be used as weights in the weighted Lasso. Statistically significant x's tend to have small (adjusted) p-values and non-significant x's have large (adjusted) p-values. Thus, weighing each x_k with its corresponding (adjusted) p-value will generally lead to including statistically significant x's in the final model.

The proposed method of Garcia et al. (2013) focused on q-values as weights for the x's. In this paper, we also consider other weights; specifically,

1. $w_{m_0+k} = 1/|\rho_{x_k,y|z_1,\ldots,z_{m_0}}|$ on x_k, $k = 1, \ldots, m_1$, where $\rho_{x_k,y|z_1,\ldots,z_{m_0}}$ is the partial correlation between x_k and y after controlling for $z_1, \ldots, z_{m_0}$;

2. $w_{m_0+k} = 1/|t_k|$ on x_k, $k = 1, \ldots, m_1$, where $t_k = \widehat{\beta}_k^* / se(\widehat{\beta}_k^*)$ is the t-statistic obtained from the individual linear regressions and $\widehat{\beta}_k^*$ is the estimated coefficient associated with x_k in the individual linear regressions;

3. $w_{m_0+k} = p_k$ on x_k, $k = 1, \ldots, m_1$, where p_k are the p-values obtained from the individual linear regressions;

4. $w_{m_0+k} = p_k^{\mathrm{BH}}$ on x_k, $k = 1, \ldots, m_1$, where p_k^{BH} are the Benjamini-Hochberg (Benjamini and Hochberg, 1995) adjusted p-values obtained from the individual linear regressions;

5. $w_{m_0+k} = q_k$ on x_k, $k = 1, \ldots, m_1$, where q_k are the q-values obtained from the individual linear regressions.

In Section 3, we explore the influence these different weights have on the weighted Lasso.

In practice, the weighted Lasso is solved using a least angle regression (LARS) algorithm (Efron et al, 2004) which provides the entire sequence of model fits in the Lasso path, along with estimated parameter coefficients. The best descriptive model among all those in the Lasso path is the one that minimizes the penalized loss function

$$M_n(\delta, p) = \mathrm{SSE}_{p^*} / \widehat{\sigma}^2 - n + \delta p^*. \tag{2}$$

Here, $\delta > 0$, p^* is the number of predictors in the selected model, SSE_{p^*} is the residual sum of squares for the selected model, and $\widehat{\sigma}^2$ is an appropriate estimator of the model error variance. For example, when $n > p^*$, $\widehat{\sigma}^2$ can be the residual mean square when using all available variables, and when $n < p^*$, $\widehat{\sigma}^2$ can be the variance of the response vector y (Hirose et al, 2011).

An important detail of (2) is the choice δ as different δ values yield different model fits and observed FDR. Garcia et al. (2013) proposed a modified cross-validation procedure to appropriately select δ; in this paper, however, we consider δ fixed at $\delta = 1$, and focus on the choice of weights.

3. Simulation

We evaluated the performance of the different weighting schemes on simulated data similar to the simulation study in Garcia et al. (2013) that mimics the real microbiata data in Garcia et al. (2013) and Garcia and Müller (2013), where a diet variable is known to act on both, the response (weight related phenotype) and possibly on some of the other regressors (microbes). We supposed there were two diet groups with 20 subjects in each, and generated m_1+1 explanatory variables as follows. First, we generated a binary diet indicator z where for each subject $i = 1,\ldots,40$, $z_i = I(i > 20) - I(i \leq 20)$. Then we generated $x_k = (x_{1,k},\ldots,x_{40,k})^{\mathrm{T}}$, $k = 1,\ldots,m_1$, such that $x_{ik} = u_{ik} + z_i s_k$, where u_{ik} were independent uniform (0,1) random variables, $s_1,\ldots,s_{0.75m_1}$ were independent uniform (0.25,0.5) and $s_{0.75m_1+1},\ldots s_{m_1}$ were identically zero. Thus, we created m_1 variables, $x_1,\ldots,x_{m_1}$ where the first 75% of the x's depend on z. Finally, we generated the response vector as

$$y = \beta_1 z + \beta_2 x_1 + \beta_3 x_2 + \beta_4 x_3 + \sum_{k=5}^{m_1} \beta_k x_k + \beta_{m_1+1} x_{m_1} + \varepsilon, \tag{3}$$

where ε is normally distributed with mean 0 and covariance $\sigma^2 I$. In summary, $x_1,\ldots,x_{m_1}$ were generated according to four distinct categories:

Group 1. x_1, x_2, x_3 depend on diet and act on y even after taking into account diet;

Group 2. $x_4,\ldots,x_{0.75m_1}$ depend on diet and do not act on y;

Group 3. $x_{0.75m_1+1},\ldots,x_{m_1-1}$ neither depend on diet, nor act on y;

Group 4. x_{m_1} does not depend on diet, but acts on y.

To evaluate the performance of the different weights in the weighted Lasso, we considered two different parameter settings. We set $m_1 = 40$, $\sigma^2 = 0.5$ with

$$\beta = (4.5, 3, -3, -3, 0^{\mathrm{T}}, 3)^{\mathrm{T}} \quad \text{and} \quad \beta = (2.5, 1.5, -1.5, -1.5, 0^{\mathrm{T}}, 1.5)^{\mathrm{T}}$$

where 0^{T} is an (m_1-4)-dimensional vector of zeros. Under each parameter setting, we generated 1000 independent data sets, and applied the weighted Lasso with the proposed weights in Section 2. We report the averaged percentages of time variables in each group were selected, the observed false discovery rate (FDR), and the average weight placed on variables in each group. We refer to Garcia and Müller (2013) for more simulation results, including investigating effect

From Table 1, we observe that the weighted Lasso has a high rate of true positives and an acceptable false positive rate. Most interestingly, we observed that weights based on the inverted absolute partial correlations and inverted absolute t-statistics equally selected the same percentage of variables in each group. Likewise, weights based on any of the p-values (either with or without adjustment) led to similar results in the variable selection. In fact, both the BH-adjusted p-values and q-values led to exactly the same percentages of selection for each group. This performance is not surprising given that, for each group, the average BH-adjusted p-value was nearly the same as the average q-value (see last two columns of Table 2). Interestingly, weights based on p-values without any adjustment led to the most reasonable results with variables in Groups 1 and 4 selected most frequently, and variables in Groups 2 and 3 least frequently. This suggests that when using the weighted Lasso for variable selection, no transformation of the p-value based weights is necessary. This contrasts from using adjusted p-values to select significant variables where transformation is needed to account for the compounded error in multiple hypothesis testing.

Weights	Average Variable Selection $\lvert\rho_{x,y\lvert z}\rvert^{-1}$	$\lvert t\rvert^{-1}$	p	p^{BH}	q
		$\beta = (4.5, 3, -3, -3, 0^{\mathrm{T}}, 3)^{\mathrm{T}}$			
Diet	100.00	100.00	100.00	100.00	100.00
Group 1	75.53	75.10	73.00	73.03	73.03
Group 2	0.39	0.36	0.27	0.29	0.29
Group 3	1.88	1.60	0.44	0.50	0.50
Group 4	85.70	84.80	75.70	77.80	77.80
FDR	0.08	0.07	0.04	0.04	0.04
		$\beta = (2.5, 1.5, -1.5, -1.5, 0^{\mathrm{T}}, 1.5)^{\mathrm{T}}$			
Diet	100.00	100.00	100.00	100.00	100.00
Group 1	40.57	41.33	46.80	43.53	43.53
Group 2	0.51	0.50	0.57	0.47	0.47
Group 3	2.19	1.96	0.78	0.97	0.97
Group 4	60.20	59.40	52.00	51.50	51.50
FDR	0.16	0.14	0.10	0.10	0.10

Table 1: Simulation results from 1000 simulations. Averaged percentages of time variables in each group were selected and observed false discovery rate (FDR). Ideal weighted Lasso will largely select variables in Groups 1 and 4, and not select variables in Groups 2 and 3.

Weights	Average Weights $\lvert\rho_{x,y\lvert z}\rvert^{-1}$	$\lvert t\rvert^{-1}$	p	p^{BH}	q
		$\beta = (4.5, 3, -3, -3, 0^{\mathrm{T}}, 3)^{\mathrm{T}}$			
Group 1	0.38	2.55	0.03	0.14	0.11
Group 2	7.22	43.97	0.50	0.78	0.63
Group 3	8.11	49.39	0.50	0.78	0.64
Group 4	0.37	2.51	0.03	0.15	0.12
		$\beta = (2.5, 1.5, -1.5, -1.5, 0^{\mathrm{T}}, 1.5)^{\mathrm{T}}$			
Group 1	0.75	4.75	0.07	0.28	0.23
Group 2	24.01	146.11	0.50	0.79	0.65
Group 3	11.55	70.32	0.50	0.79	0.65
Group 4	0.53	3.44	0.07	0.29	0.24

Table 2: Simulation results from 1000 simulations. Average weights in each group. Ideally the weights are small in Groups 1 and 4 and (relatively) large in Groups 2 and 3.

4. Conclusion

We conclude that in the $p > n$ context, when part of the regressors can act on both the response and some of the other explanatory variables, using structural information to construct feature weights in the weighted Lasso greatly aids the variable selection. We

have shown that the results from Garcia et al. (2013) extend from using q-values to any other informative measure of significance, such as p-values and adjusted p-values, or partial correlation coefficients and test statistic values.

Acknowledgements

Samuel Müller was supported by a grant from the Australian Research Council (DP130100488).

References

Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society, Series B* **57**, 289–300.

Bergersen, L. C., Glad, I. K. and Lyng H. (2011). Weighted Lasso with data integration. *Statistical Applications in Genetics and Molecular Biology*, **10**, 1–29.

Charbonnier, C., Chiquet, J. and Ambroise C. (2010). Weighted-Lasso for structured network inference from time course data. *Statistical Applications in Genetics and Molecular Biology*, **9**, Article 15.

Efron, B., Hastie, T., Johnstone, I. and Tibshirani, R. (2004). Least angle regression. *Annals of Statistics*, **32**, 407-499.

Garcia, T. P. and Müller, S. (2013). Influence of informative measures of significance based weights in the weighted Lasso. *Preprint*.

Garcia, T. P., Müller, S., Carroll, R.J., Dunn, T.N., Thomas, A.P., Adams, S.H., Pillai, S.D., and Walzem, R.L. (2013). Structured variable selection with *q*-values. *Biostatistics*, in press.

Hirose, K., Tateishi, S. and Konishi, S. (2011). Efficient algorithm to select tuning parameters in sparse regression modeling with regularization. *Science*, 2011, 1-25.

Meinshausen, N. and Bühlmann, P. (2010) Stability selection (with discussion). *Journal of the Royal Statistical Society, Series B* **72**, 417–473.

Storey, J. D. (2003). The positive false discovery rate: A Bayesian interpretation and the *q*-value. *Annals of Statistics* **31**, 2013–2035.

STOREY, J. D. AND TIBSHIRANI, R. (2003). Statistical significance for genomewide studies. *Proc. Natl. Acad. Sci.* **100**, 9440–9445.

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

Yuan, M. and Lin, Y. (2006). Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society, Series B* **68**, 49–67.

Zou, H. (2006). The adaptive Lasso and its oracle properties. *Journal of the American Statistical Association*, **101**, 1418-1429.