

# Supplementary Materials for “Incorporating Predictor Network in Penalized Regression with Application to Microarray Data”

Benhuai Xie, Wei Pan and Xiaotong Shen

## 1 Web Appendix A: Proof of Theorem 1

The penalty function has the form

$$p_\lambda(\beta) = \lambda 2^{1/\gamma'} \sum_{i \sim j} \|\beta_{(i,j)}\|_\gamma^{(w_i, w_j)}, \quad (1)$$

and the objective function is

$$L_P(\beta) = -Y'Y/2 + \beta'X'Y - \beta'X'X\beta/2 - p_\lambda(\beta) \quad (2)$$

The solution  $\beta_{(i,j)} = \mathbf{0}$  must satisfy the following condition

$$L_P(\mathbf{0}, \cdot) \geq L_P(\Delta\beta_{(i,j)}, \cdot) \text{ for all } \Delta\beta_{(i,j)} \text{ near } \mathbf{0}, \quad (3)$$

where  $\cdot$  in  $L_P(\mathbf{0}, \cdot)$  represents all parameters in  $L_P(\beta)$  except  $\beta_{(i,j)}$ .

Note that  $L_P(\beta) = \beta'_{(i,j)}(X'Y)_{(i,j)} - \beta'_{(i,j)}(X'X)_{(i,j;i,j)}\beta_{(i,j)}/2 - \beta'_{(i,j)}(X'X)_{(i,j;-(i,j))}\beta_{-(i,j)} - \lambda 2^{1/\gamma'} (\|\beta_{(i,j)}\|_\gamma^{(w_i, w_j)} + f_i(\beta_{(i)}) + f_j(\beta_{(j)})) + C$ , where  $C$  is a constant with respect to  $\beta'_{(i,j)}$ , and  $f_i$  and  $f_j$  are constants with respect to  $\beta_j$  and  $\beta_i$  respectively. For the above condition, we have

$$\begin{aligned} \text{LHS} &= -\lambda 2^{1/\gamma'} (f_i(0) + f_j(0)) + C \\ \text{RHS} &= \Delta\beta'_{(i,j)} [(X'Y)_{(i,j)} - (X'X)_{(i,j;-(i,j))}\beta_{-(i,j)}] - \Delta\beta'_{(i,j)} (X'X)_{(i,j;i,j)} \Delta\beta_{(i,j)}/2 \\ &\quad - \lambda 2^{1/\gamma'} (\|\Delta\beta_{(i,j)}\|_\gamma^{(w_i, w_j)} + f_i(\Delta\beta_{(i)}) + f_j(\Delta\beta_{(j)})) + C \end{aligned}$$

Thus,

$$\begin{aligned}
& \beta_{(i,j)} = \mathbf{0} \\
\iff & -\lambda 2^{1/\gamma'} (f_i(0) + f_j(0)) \\
& \geq \Delta\beta'_{(i,j)} [(X'Y)_{(i,j)} - (X'X)_{(i,j;- (i,j))} \beta_{-(i,j)}] - \Delta\beta'_{(i,j)} (X'X)_{(i,j;i,j)} \Delta\beta_{(i,j)} / 2 \\
& - \lambda 2^{1/\gamma'} (\|\Delta\beta_{(i,j)}\|_\gamma^{(w_i, w_j)} + f_i(\Delta\beta_{(i)}) + f_j(\Delta\beta_{(j)})) \\
\iff & \Delta\beta'_{(i,j)} [(X'Y)_{(i,j)} - (X'X)_{(i,j;- (i,j))} \beta_{-(i,j)}] - \Delta\beta'_{(i,j)} (X'X)_{(i,j;i,j)} \Delta\beta_{(i,j)} / 2 \\
& \leq \lambda 2^{1/\gamma'} (\|\Delta\beta_{(i,j)}\|_\gamma^{(w_i, w_j)} + [f_i(\Delta\beta_{(i)}) - f_i(0)] + [f_j(\Delta\beta_{(j)}) - f_j(0)]) \\
\iff & \Delta\beta'_{(i,j)} [(X'Y)_{(i,j)} - (X'X)_{(i,j;- (i,j))} \beta_{-(i,j)}] / \|\Delta\beta_{(i,j)}\|_\gamma^{(w_i, w_j)} \\
& - \Delta\beta'_{(i,j)} (X'X)_{(i,j;i,j)} \Delta\beta_{(i,j)} / (2 \|\Delta\beta_{(i,j)}\|_\gamma^{(w_i, w_j)}) \\
& \leq \lambda 2^{1/\gamma'} (1 + [f_i(\Delta\beta_{(i)}) - f_i(0)] / \|\Delta\beta_{(i,j)}\|_\gamma^{(w_i, w_j)} + [f_j(\Delta\beta_{(j)}) - f_j(0)] / \|\Delta\beta_{(i,j)}\|_\gamma^{(w_i, w_j)})
\end{aligned} \tag{4}$$

Note that  $\Delta\beta'_{(i,j)} (X'X)_{(i,j;i,j)} \Delta\beta_{(i,j)} / (2 \|\Delta\beta_{(i,j)}\|_\gamma^{(w_i, w_j)}) \rightarrow 0^+$  as  $\Delta\beta'_{(i,j)} \rightarrow \mathbf{0}$ . By the Hölder inequality, we have  $\Delta\beta'_{(i,j)} [(X'Y)_{(i,j)} - (X'X)_{(i,j;- (i,j))} \beta_{-(i,j)}] \leq \|\Delta\beta'_{(i,j)}\|_\gamma^{(w_i, w_j)} \|(X'Y)_{(i,j)} - (X'X)_{(i,j;- (i,j))} \beta_{-(i,j)}\|_\gamma^{(1/w_i, 1/w_j)}$ , where  $1/\gamma + 1/\gamma' = 1$ .  $f_j(\Delta\beta_{(j)}) - f_j(0)$  is always non-negative. Therefore, the last inequality above can be derived from (5) in the paper, a sufficient condition for  $\beta'_{(i,j)} = \mathbf{0}$ .

Further,  $[f_j(\Delta\beta_{(j)}) - f_j(0)] / \|\Delta\beta'_{(i,j)}\|_\gamma^{(w_i, w_j)} = ([f_j(\Delta\beta_{(j)}) - f_j(0)] / |\Delta\beta_j|) \cdot (|\beta_j| / \|\Delta\beta'_{(i,j)}\|_\gamma^{(w_i, w_j)}) \leq [f_j(\Delta\beta_{(j)}) - f_j(0)] / |\Delta\beta_j|$  and  $[f_j(\Delta\beta_{(j)}) - f_j(0)] / |\Delta\beta_j| \rightarrow \left. \frac{\partial f_j}{\partial \beta_j} \right|_{\beta_j=0} = d_j - 1$  as  $\Delta\beta_j \rightarrow 0$ . Thus, we obtain (6) in the paper as the necessary condition of  $\hat{\beta}_i = \hat{\beta}_j = 0$ .

## 2 Web Appendix B: Some Theory for Grouping Effects

We demonstrate the grouping effects of our proposed penalty: under some conditions, for two neighboring nodes, their non-zero regression coefficient estimates are shrunken

to be closer to each other as the tuning parameter or the correlation of their predictors increases. We assume that each predictor  $x_i$  for each gene has been standardized to have sample mean 0 and sample variance 1, and that  $\|Y\|_2$  is the  $L_2$  norm of the vector of response values. Similar to the proof of the grouping effect of Enet as in Zou and Hastie (2005), we can prove the next lemma.

**Lemma 1** *If  $p_\lambda(\beta)$  is differentiable at  $\hat{\beta}_i$  and  $\hat{\beta}_j$ , we have*

$$\left| \sum_{k:i \sim k} \frac{\partial p(\hat{\beta}_i, \hat{\beta}_k)}{\partial \hat{\beta}_i} - \sum_{k:j \sim k} \frac{\partial p(\hat{\beta}_j, \hat{\beta}_k)}{\partial \hat{\beta}_j} \right| \leq \frac{\|Y\|_2}{\lambda 2^{1/\gamma'}} \sqrt{2(1 - \rho_{i,j})}, \quad (5)$$

where  $i$ ghbors for gene  $i$  and gene  $j$  respectively,  $\rho_{i,j} = x'_i x_j$  is the sample correlation between  $x_i$  and  $x_j$ .

**Corollary 1** *Under the condition of Lemma 2, 1) if  $w_i = d_i^{(\gamma+1)/2}$ , then*

$$\left| \frac{1}{d_i} \sum_{k:i \sim k} \frac{\text{sgn}(\hat{\beta}_i) \left| \frac{\hat{\beta}_i}{\sqrt{d_i}} \right|^{\gamma-1}}{\left( \|(\hat{\beta}_i, \hat{\beta}_k)\|_\gamma^{(w_i, w_k)} \right)^{\gamma-1}} - \frac{1}{d_j} \sum_{k:j \sim k} \frac{\text{sgn}(\hat{\beta}_j) \left| \frac{\hat{\beta}_j}{\sqrt{d_j}} \right|^{\gamma-1}}{\left( \|(\hat{\beta}_j, \hat{\beta}_k)\|_\gamma^{(w_j, w_k)} \right)^{\gamma-1}} \right| \leq \frac{\|Y\|_2}{\lambda 2^{1/\gamma'}} \sqrt{2(1 - \rho_{i,j})}; \quad (6)$$

2) if  $w_i = d_i$ , then

$$\left| \frac{1}{d_i} \sum_{k:i \sim k} \frac{\text{sgn}(\hat{\beta}_i) |\hat{\beta}_i|^{\gamma-1}}{\left( \|(\hat{\beta}_i, \hat{\beta}_k)\|_\gamma^{(w_i, w_k)} \right)^{\gamma-1}} - \frac{1}{d_j} \sum_{k:j \sim k} \frac{\text{sgn}(\hat{\beta}_j) |\hat{\beta}_j|^{\gamma-1}}{\left( \|(\hat{\beta}_j, \hat{\beta}_k)\|_\gamma^{(w_j, w_k)} \right)^{\gamma-1}} \right| \leq \frac{\|Y\|_2}{\lambda 2^{1/\gamma'}} \sqrt{2(1 - \rho_{i,j})}; \quad (7)$$

3) if  $w_i = d_i'$ , then

$$\left| \frac{1}{d_i} \sum_{k:i \sim k} \frac{\text{sgn}(\hat{\beta}_i) \left| \frac{\hat{\beta}_i}{d_i} \right|^{\gamma-1}}{\left( \|(\hat{\beta}_i, \hat{\beta}_k)\|_\gamma^{(w_i, w_k)} \right)^{\gamma-1}} - \frac{1}{d_j} \sum_{k:j \sim k} \frac{\text{sgn}(\hat{\beta}_j) \left| \frac{\hat{\beta}_j}{d_j} \right|^{\gamma-1}}{\left( \|(\hat{\beta}_j, \hat{\beta}_k)\|_\gamma^{(w_j, w_k)} \right)^{\gamma-1}} \right| \leq \frac{\|Y\|_2}{\lambda 2^{1/\gamma'}} \sqrt{2(1 - \rho_{i,j})}. \quad (8)$$

Corollary 1 shows complicated shrinkage effects in a network.

### 3 Web Appendix C: Generalized Boosted Lasso

Suppose that a penalized loss function defined before can be expressed as

$$L_P(\beta) = L_P(\beta; \lambda) = L(\beta) + \lambda T(\beta).$$

where both the loss  $L(\beta)$  and penalty  $T(\beta)$  are convex in  $\beta$ . Denote  $\mathbf{1}_j$  as a vector of all 0 except that its  $j$ th element is 1. We fix a small step size  $e > 0$  and a small tolerance parameter  $\xi$ .

Step 1. Starting from  $\beta = 0$ , take a initial forward step

$$(\hat{j}, \hat{s}_j) = \arg \min_{j, s=\pm e} L(Z; s\mathbf{1}_j).$$

Then  $\hat{\beta}^{(0)} = \hat{s}_j \mathbf{1}_j$ , and

$$\lambda^{(0)} = \frac{L(0) - L(\hat{\beta}^{(0)})}{T(\hat{\beta}^{(0)}) - T(0)}.$$

Set  $t = 0$ .

Step 2. Find the steepest coordinate descent direction on the penalized loss

$$(\hat{j}, \hat{s}_j) = \arg \min_{j, s=\pm e} L_P(\hat{\beta}^{(t)} + s\mathbf{1}_j; \lambda^{(t)}).$$

If  $L_P(\hat{\beta}^{(t)} + \hat{s}_j \mathbf{1}_j; \lambda^{(t)}) - L_P(\hat{\beta}^{(t)}; \lambda^{(t)}) < -\xi$ , then

$$\hat{\beta}^{(t+1)} = \hat{\beta}^{(t)} + \hat{s}_j \mathbf{1}_j, \quad \lambda^{(t+1)} = \lambda^{(t)};$$

otherwise,

$$\begin{aligned} (\hat{j}, \hat{s}_j) &= \arg \min_{j, s=\pm\sqrt{2}e} L(\hat{\beta}^{(t)} + s\mathbf{1}_j), \\ \hat{\beta}^{(t+1)} &= \hat{\beta}^{(t)} + \hat{s}_j \mathbf{1}_j, \\ \lambda^{(t+1)} &= \min \left( \lambda^{(t)}, \frac{L(\hat{\beta}^{(t)}) - L(\hat{\beta}^{(t+1)})}{T(\hat{\beta}^{(t+1)}) - T(\hat{\beta}^{(t)})} \right). \end{aligned} \tag{9}$$

Step 3. Go back to Step 2 with  $t \leftarrow t + 1$  unless  $\lambda^{(t)} \leq 0$ .

The main modification to the original GBL algorithm (Zhao and Yu 2004) is the use of step size  $\sqrt{2}e$ , instead of  $e$  in expression (5); otherwise, there could be a dead loop in Step 2. This modification also illustrates that the GBL gives only an approximate solution. In our simulations, we used  $e = 0.1$  and  $x_i = 0.01$ ; for each simulated data with  $p = 110$  and  $n = 50$ , our implementation in R took about 30 seconds to obtain a final solution.

## 4 Web Appendix D: Penalized Cox Regression for the Example

One may be concerned that the linear model might not fit the data well. As an alternative, one can use the semi-parametric Cox proportional hazards model (PHM), perhaps the most popular model in survival analysis. Rather than modeling the mean of the response in a linear model, a PHM models hazard function at time  $t$  for a given predictor vector  $x_k$ :

$$h(t|x_k) = h_0(t) \exp\left(\sum_{i=1}^p x_{ki}\beta_i\right),$$

where  $h_0(t)$  is an unknown baseline hazard function. The coefficient vector  $\beta$  is usually estimated by the maximum partial likelihood estimator. In particular, the partial likelihood is robust in the sense that it only depends on the ranks of the observed response times, not their specific values. As in linear regression, we can incorporate a penalty, such as the Lasso penalty (Gui and Li 2005), into the partial likelihood, which is maximized to yield maximum penalized partial likelihood estimator (MPPLE)  $\hat{\beta}$ . It is known that fitting a PHM can be approximated by fitting a linear model: the null deviance residuals from a null PHM (i.e. with no predictor) are used as the response for a linear model; the linear model contains the same set of predictors as in the targeted PHM (Segal 2006). Hence, we can approximate an MPPLE by an MPLE.

We adopted this approximation strategy to fit a penalized PHM using the Lasso, Enet and our network-based penalties; they yielded results similar to their earlier ones (from the linear model) respectively, though our method seemed to be more stable in gene selection. Lasso selected 13 genes: SDC2 from the earlier list was missing while WNT2B, PRKCD and MPP5 were newly added, and the other ones remained the same. On the other hand, Enet selected 14 genes, the union of the two lists from the Lasso. In contrast, our method (with the same  $\gamma = 2$  and the same set of weights as before) included 15 genes, a subset of the previously identified ones with only FLNC and CD46 missing. For each method, the estimated coefficients, and even their solution paths, were similar to their counterparts from the linear model (except possibly flipped signs due to the use of deviance residuals here); estimated PMSEs by the tuning data from the Lasso and network-based method were also similar to the earlier ones (Fig 1). In particular, the minimum PMSEs from the Lasso and network-based method were 0.63 and 0.60 respectively.

## References

- Gui, J. and Li, H. (2005). Penalized Cox regression analysis in the high-dimensional and low-sample size settings, with applications to microarray gene expression data. *Bioinformatics*, **21**, 3001-3008.
- Segal M.R. (2006) Microarray gene expression data with linked survival phenotypes: diffuse large-B-cell lymphoma revisited. *Biostatistics*, **7**, 268-285.

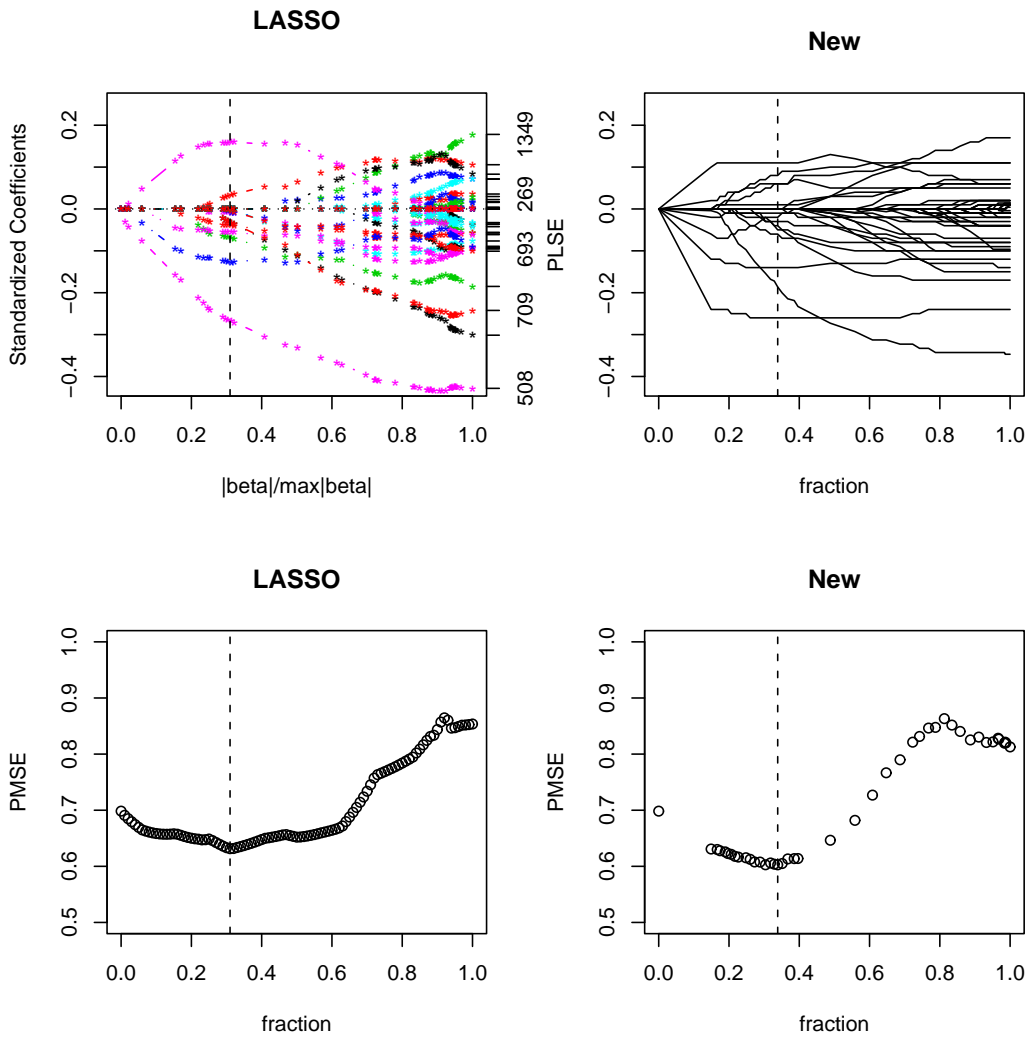


Figure 1: Solution paths and PMSE based on tuning data for Lasso and our proposed new method based on a Cox model for the first set of the glioblastoma data.