IMPERIAL

Predicting disease outcome using penalised regression

Fabio Feser[1]
Marina Evangelou[1]
[1] **Department of Mathematics, Imperial College London**

Regression

Logistic model

We have observed expression data for p genes in n individuals, forming a data matrix $\mathbf{X} \in \mathbb{R}^{n \times p}$, where $p >> n$. We are interested in discovering relationships between the genes and the disease outcome of a patient $y \in \{0, 1\}^n$, enabling prediction. The probability of having the disease is modelled as $\mathbb{P}(y = 1|\mathbf{X}) = 1/(1 + \exp(-\mathbf{X}\beta))$, where $\beta \in \mathbb{R}^p$ are the coefficients of the relationships.

Penalised regression

To solve the above problem, we fit a **penalised logistic model**, solving the optimisation problem

$$\hat{\beta} = \max_{b \in \mathbb{R}^p} \Big\{ \underbrace{\sum_{i=1}^{n} \left[y_i \left(b^\top x_i \right) - \log \left(1 + e^{b^\top x_i} \right) \right]}_{\text{loss function}} - \lambda \underbrace{f(b)}_{\text{penalty}} \Big\},$$

where $\lambda > 0$ is the penalisation parameter, and f is the penalty function. The most popular penalty function is the least absolute shrinkage and selection operator (**lasso**) [5]: $f_{\text{lasso}}(b) = \sum_{i=1}^{p} |b_i|$. The form of this penalty (Figure 1) allows variable selection to occur through shrinkage of the coefficients exactly to zero. That is, we obtain a set of associated genes $\{i \in \{1, \dots, p\} : \hat{\beta}_i \neq 0\}$.

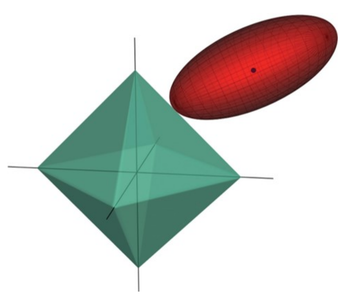

Figure 1: The lasso constraint region (green) in $\mathbb{R}^3$, shown with the loss function (red). [2]

Tuning

The parameter λ requires tuning, as it defines the level of sparsity in the model. We fit the model to an l-length path of parameters $\lambda_1 \geq \dots \geq \lambda_l \geq 0$, where λ_1 is the point at which the first coefficient becomes non-zero and λ_l is some fraction of λ_1. We apply k-fold cross-validation to these l models and pick the best predictive model.

Sparse-group models

Genes are naturally found in groups called pathways. We want to be able to exploit this grouping information in our prediction. To do this, assume that the genes sit in m non-overlapping groups $\mathcal{G}_1, \dots, \mathcal{G}_m$ of sizes $p_1, \dots, p_m$. An extension of the lasso to incorporate this grouping information is given by the group lasso (**gLasso**) [6]: $f_{\text{glasso}}(b) = \sum_{g=1}^{m} \sqrt{p_g} \|b^{(g)}\|_2$, where $b^{(g)} \in \mathbb{R}^{p_g}$ is a vector of coefficients. The gLasso shrinks whole groups to zero, performing group selection.

To avoid shrinking whole groups to zero, the sparse-group lasso (**SGL**) [4] combines the strengths of the lasso and gLasso convexly: $f_{\text{sgl}}(b; \alpha) = \alpha f_{\text{lasso}}(b) + (1 - \alpha) f_{\text{glasso}}(b)$, where $\alpha \in [0, 1]$, but is chosen subjectively at $\alpha = 0.99$.

SLOPE

The Sorted L-One Penalised Estimation (**SLOPE**) model [1] is an adaptive version of the lasso with the penalty: $f_{\text{slope}}(b) = \sum_{i=1}^{p} v_i |b|_{(i)}$, where $|b|_{(1)} \geq \dots \geq |b|_{(p)}$ are matched to $v_1 \geq \dots \geq v_p \geq 0$. SLOPE is motivated by genetics, as the weights, v, are set to the Benjamini-Hochberg critical values, to provide false-discovery rate (FDR) control.

Sparse-group SLOPE

We incorporate the SLOPE penalty into a sparse-group framework by introducing sparse-group SLOPE (**SGS**), defined as

$$f_{\text{sgs}}(b; \alpha) = \alpha \sum_{i=1}^{p} v_i |b|_{(i)} + (1 - \alpha) \sum_{g=1}^{m} \sqrt{p_g} w_g \|b^{(g)}\|_2,$$

where $|b|_{(i)}$ are matched as for SLOPE and additionally $\sqrt{p_1}\|b^{(1)}\| \geq \dots \geq \sqrt{p_g}\|b^{(g)}\|_2$ are matched to $w_1 \geq \dots \geq w_m \geq 0$. The weights, (v, w), were theoretically derived to provide bi-level (variable and group) FDR control. The SGS optimisation is solved using the Adaptive Three Operator Splitting (ATOS) algorithm [3].

(a) SLOPE **(b)** gSLOPE **(c)** SGS with $\alpha = 0.5$

Figure 2: Units balls in $\mathbb{R}^3$ for the penalty functions of SLOPE (a), gSLOPE (b), and SGS (c). SGS can be seen to be a convex combination of SLOPE and gSLOPE.

Predicting diseases

Using these models, we tackled the problem of predicting two diseases: ulcerative **colitis** and breast **cancer**. The colitis dataset contained blood cell gene expression data for $n = 127$ patients, of which 85 had colitis, with $p = 12031$ genes, grouped into 1408 pathways. The cancer dataset was of dimensions $n = 60$, $p = 12071$, and $m = 1132$, where the patients were treated with tamoxifen and classified into whether the cancer recurred.

Both datasets were split into train/validation sets to calculate the accuracy of the models by predicting the disease status of the patients in the validation sets. The models were fit along a path of 100 λ values. Figure 3 shows the peak classification scores of the SLOPE models and Table 1 for both SLOPE and lasso-based models.

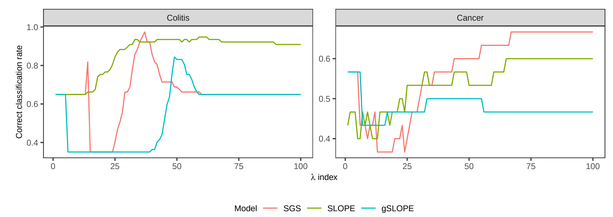

Figure 3: Validation classification scores (%) on the colitis and cancer datasets as a function of the penalisation parameter (λ) for the SLOPE-based models.

Table 1: Peak validation classification scores (%) achieved for each SLOPE and lasso-based model on the two genetics datasets.

	SLOPE-based models			Lasso-based model		
Dataset	SGS	SLOPE	gSLOPE	SGL	Lasso	gLasso
Colitis	**97.4**	94.8	84.4	92.2	93.5	89.6
Cancer	**66.7**	60.0	56.7	50.0	56.7	36.7

References

[1] Małgorzata Bogdan, Ewout van den Berg, Chiara Sabatti, Weijie Su, and Emmanuel J. Candès. Slope—adaptive variable selection via convex optimization. The Annals of Applied Statistics, 9, 9 2015.
[2] Trevor Hastie, Robert Tibshirani, and Martin Wainwright. Statistical Learning with Sparsity. Chapman and Hall/CRC, 5 2015.
[3] Fabian Pedregosa and Gauthier Gidel. Adaptive three operator splitting. In Jennifer Dy and Andreas Krause, editors, Proceedings of the 35th International Conference on Machine Learning, volume 80 of Proceedings of Machine Learning Research, pages 4085–4094. PMLR, 10–15 Jul 2018.
[4] Noah Simon, Jerome Friedman, Trevor Hastie, and Robert Tibshirani. A sparse-group lasso. Journal of Computational and Graphical Statistics, 22:231–245, 4 2013.
[5] Robert Tibshirani. Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society: Series B (Methodological), 58:267–288, 1 1996.
[6] Ming Yuan and Yi Lin. Model selection and estimation in regression with grouped variables. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 68:49–67, 2 2006.

Funders

UK Research and Innovation