Statistical nugget

LASSO regression

J Ranstam, JA Cook

Regression models are commonly used in statistical analyses (1-2). A popular usage is to model the predicted risk of a future event or likely outcome (e.g. survival or recurrence of disease). Unfortunately, applying standard regression methods to a set of candidate variables to generate a model tends to lead to overfitting in terms of the number of variables ultimately included in the model, and also overestimation of how well the model performs in terms of using the included variables to explain the observed variability ("optimism bias"). The model tends to perform particularly poorly with predicting observations more "extreme" (very high or very low) risk. Various (penalised or regularisation) regression techniques, can be used to address these problems. LASSO regression, a shrinkage and variable selection method for regression models, is an attractive option as it address both problems (3). Gains in computational power and incorporation into statistical software also mean that its computer intensive nature is no longer off-putting. One area where it has been used is for handing genetic data as the number of potential predictors is often large relative to the number of observations, and there is often little or no a-priori knowledge to inform variable selection.

LASSO regression aims to identify the variables and corresponding regression coefficients that leads to a model that minimises the prediction error. This is achieved by imposing a constraint on the model parameters, which "shrinks" the regression coefficients towards zero, i.e. by forcing the sum of the absolute value of the regression coefficients to be less than a fixed value (λ). It a practical sense this constraints the complexity of the model. Variables having a regression coefficient of zero after shrinkage are excluded from the model. The choice of λ is often made by using an automated k-fold cross-validation approach. Under this approach, the dataset is randomly partitioned into *k* sub-samples of equal size. While the *k*-1 sub-samples are used for developing a prediction model, the remaining sub-sample is used for validating this model. The procedure is then repeated *k* times, with each one of the *k* sub-samples in turn being used for validation and the other ones for model development. An overall result is produced by combining the *k* separate results and choosing the preferred λ which is then used to determine the final model. A particular advantage with this technique is

that it reduces overfitting without restricting a subset of the dataset to sole use for internal validation.

LASSO approaches have been shown to outperform standard approaches in some settings. However, they are not a panacea to the problems of overfitting and optimism bias, and do not remove the need of validation of a model in an external dataset. Additionally, the LASSO approach trades-off potential bias in estimating individual parameters for a better expected overall prediction. A corresponding important disadvantage of the LASSO approach is that the regression coefficients may not be reliably interpretable in terms of independent risk factors as the focus is on the best combined prediction, not on the accuracy of the estimation and interpretation of the contribution of individual variables. Variants on the general LASSO approach exist, such as ridge regression and Elastic Net (4), and their relative merits of penalisation regression techniques is an area of ongoing research.

References

1. Ranstam J, Cook JA. Statistical models: an overview. BJS 2016; 103:1039-1047.2.
2. Cook JA, Ranstam J. Overfitting. BJS 2016. DOI: 10.1002/bjs.10244
3. Tibshirani R. Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society, Series B 1996;58:267–288.
4. Zou H, Hastie T. Regularization and variable selection via the elastic net. J R Statist Soc B 2005;67:301–320.