Sparsity

Rajen D. Shah, Statistical Laboratory, University of Cambridge
Talk at workshop on multivariate analysis today (WOMAT), Open University

18 May, 2015

Classical statistics was designed for datasets with a large (but not overwhelming) number of observations of a few carefully chosen variables. In recent years, however, as statisticians we've been challenged with an increasing number of datasets of very different shape and size, for which we have needed to develop, and indeed still need to invent, radically different new tools to perform data analysis. Sparsity has been at the centre of much of this statistical research. The word is used to mean different ideas in different contexts, but perhaps the two most important forms of sparsity are what one might term *signal sparsity* and *data sparsity*.

Signal sparsity

Consider a regression context with n observations $(Y_i, x_i) \in \mathbb{R} \times \mathbb{R}^p$, $i = 1, \ldots, n$ with Y_i the response and x_i a p-dimensional covariate vector. Signal sparsity refers to an assumption that most of the p predictors are unrelated to the response. Specialising to a linear model for the data,

$$Y_i = \mu + x_i^T \beta^0 + \varepsilon_i,$$

this translates to the coefficient vector β^0 being sparse i.e. most of its components are 0.

Though such sparse models are unlikely to be exactly true, they are nevertheless useful for a great number of modern datasets, particularly in the so-called high-dimensional setting where the number of predictors p greatly exceeds the number of observations n. Microarray datasets, where the number of variables (gene expression values) may number in the tens of thousands, and the number of observations (tissue samples) could be a few hundred at best, are a prototypical example.

When $p > n$ (so X does not have full column rank), the estimate of β^0 from ordinary least squares (OLS) will not be unique and it will overfit to the extent that the fitted values will equal those observed. The famous Lasso estimator (Tibshirani, 1996) addresses this problem by adding an ℓ_1 penalty term to the least squares objective criterion, with the estimated regression coefficients, $\hat{\beta}$, and estimated intercept $\hat{\mu}$, satisfying

$$(\hat{\mu}, \hat{\beta}) = \underset{\mu, \beta}{\arg\min} \, \{ \tfrac{1}{2n} \|Y - \mu \mathbf{1} - X\beta\|_2^2 + \lambda \|\beta\|_1 \}, \tag{0.1}$$

where $\mathbf{1}$ denotes an n-vector of 1's and $\|\beta\|_1 := \sum_{k=1}^p |\beta_k|$. The *tuning parameter*, $\lambda > 0$, trades off the complexity of the fitted model with the fidelity of the fit to Y. Due to the form of the penalty on β, a large number of components of $\hat{\beta}$ will be exactly 0. This sparsity property of the estimator helps it deliver sensible estimates even when $p \gg n$.

An enormous amount of research has been directed at extending the Lasso in various ways, and studying its theoretical properties. See Bühlmann and van de Geer (2011) for some of these developments, and references therein. In terms of theory, asymptotic arguments where $n \to \infty$ whilst p remains fixed cannot usually be relied upon to be relevant in finite samples where $p \gg n$. Instead finite sample results are often obtained. An important result concerning the Lasso

establishes that with λ chosen appropriately and independent errors $\varepsilon_i \sim N(0, \sigma^2)$, one has with high probability

$$\|\hat{\beta} - \beta^0\|_1 \leq \text{constant} \times s\,\sigma\,\sqrt{\frac{\log(p)}{n}}$$

under conditions on the design matrix that in particular disallow certain variables from being too correlated with one another. More recently, inferential procedures based on the Lasso have been produced which, for example, can give confidence intervals for coefficients β_k^0.

Data sparsity

Many modern datasets fall within the high-dimensional "large p, small n" setting. However, primarily from industry, we are also seeing datasets where both the numbers of observations and predictors can number in the millions or (much) more. Where in high-dimensional data the main challenge is a statistical one was more parameters must be estimated than there are observations available, in this "big data" setting there are also serious computational issues. Here OLS may be infeasible for computational, rather than statistical, reasons.

An important feature of many of these large-scale datasets, such as those arising from a bag-of-words representation of a corpus of documents, is that the overwhelming majority of entries in the design matrices are exactly zero: the data matrices are sparse. Successful methods in this setting exploit this sparsity for computational and statistical gains.

Given the scale of the data, a sensible way to proceed is by first performing dimension reduction, that is mapping the original design matrix $X \in \mathbb{R}^{n \times p}$ to $S \in \mathbb{R}^{n \times L}$ with $L \ll p$. This dimension reduction step must be computationally fast and ideally should scale linearly with the number of non-zeroes in the design matrix. Developing methods that perform well under this computational constraint is a rather young area of statistical research. However, the computer science literature does have a variety of what are known sometimes known as sketching or hashing algorithms for performing this step.

One of the most prominent approaches is to form S via random projections of X. The simplest version of this technique effectively forms S through $S = XA$ where the entries in A are i.i.d. standard normals. Another interesting approach that is particularly suited to the sparse setting and is applicable when the design matrix is binary is b-bit min-wise hashing (Li and König, 2011), which is based on an earlier technique called min-wise hashing (Broder *et al.*, 1998). This technique is studied from a statistical perspective in Shah and Meinshausen (2015), where an extension to continuous data is also introduced.

References

Broder, A., Charikar, M., Frieze, A. and Mitzenmacher, A. (1998) Min-wise independent permutations. *Proceedings of the thirtieth annual ACM symposium on theory of computing*, 327–336.

Bühlmann, P. and van de Geer, S. (2011) Statistics for High-Dimensional Data: Methods, Theory and Algorithms. Springer, Springer Series in Statistics.

Li, P. and König, A.C. (2011) Theory and applications of b-bit min-wise hashing. *Communications of the ACM*, **54**, 101–109.

Shah, R.D. and Meinshausen, N. (2015) Min-wise hashing for large-scale regression and classification with sparse data. arXiv preprint.

Tibshirani, R. (1996) Regression shrinkage and selection via the Lasso *J. Roy. Statist. Soc., Ser. B*, **58**, 267–288.