Bernoulli **10**(6), 2004, 939–942

High-dimensional data: $p \gg n$ in mathematical statistics and bio-medical applications

SARA A. VAN DE GEER[1] and HANS C. VAN HOUWELINGEN[2]

[1]*Mathematical Institute, University of Leiden, P.O. Box 9512, 2300 RC Leiden, The Netherlands. E-mail: geer@math.leidenuniv.nl*
[2]*Department of Medical Statistics, Leiden University Medical Center, P.O. Box 9604, 2300 RC Leiden, The Netherlands. E-mail: jcvanhouwelingen@lumc.nl*

The workshop 'High-dimensional data: $p \gg n$ in mathematical statistics and bio-medical applications' was held at the Lorentz Center in Leiden from 9 to 20 September 2002. This special issue of *Bernoulli* contains a selection of papers presented at that workshop.

The introduction of high-throughput micro-array technology to measure gene-expression levels and the publication of the pioneering paper by Golub *et al.* (1999) has brought to life a whole new branch of data analysis under the name of micro-array analysis. Some aspects of micro-array data are quite new and typical of the data-extraction technique, but the issue of using high-dimensional data as explanatory variables in classification or prediction models has been recognized as a scientific problem in its own right in chemometrics, machine learning and mathematical statistics. The aim of the workshop was to bring together researchers from the more theoretical side (mathematical statistics, chemometrics, machine learning) and the applied side (biostatistics) to have a cross-disciplinary discussion on the analysis of high-dimensional data and to be more than just another workshop on micro-arrays. The first lesson learned is that quite different languages are spoken in the different fields and that the communication between hardcore mathematical statistics and practical data analysis on micro-arrays is far from easy. Further meetings of this sort will be beneficial because they improve interdisciplinary communication.

This special issue contains papers on different issues of micro-array data analysis and papers on statistical models for high-dimensional data. There are different statistical challenges in micro-array data analysis. A major problem with the micro-array technology (and similar high-throughput techniques) is that the outcomes obtained within one experiment (array) are of a relative nature. The outcomes of one single array can be normalized by comparing them with the (geometric) mean or the median of all values. In the common two-colour (red–green) experiment this problem is partly solved by measuring two samples on the same array in different colours. Relative measures can directly be obtained by comparing red with green. Even then, there appears to be a need for normalization because the relation between red and green can be distorted. Developing

1350–7265

proper normalization methods is an important statistical challenge in micro-array data analysis. The paper by Lee and Whitmore gives a nice insight in the normalization debate. It is interesting to observe that normalization is made possible by the abundance of data. Having tens of thousands of gene expressions measured on the same array makes it possible to use the variation over the genes within one array to construct a reasonable normalization. Also, the high dimension is in this case a blessing, not a curse.

The next step in gene-expression data analysis is to determine which genes are differentially expressed, which means that they show differences between subgroups of individuals. This can be a case of supervised learning if the individuals are characterized as normal/abnormal or of unsupervised learning if there is no further information on the individuals available. The paper of Garrett and Parmigiani discusses an interesting mix of unsupervised and supervised learning. Using latent class modelling, they manage to reduce the gene-expression information to a trichotomous outcome 'underexpressed', 'normal' or 'overexpressed'. This reduction is helpful to reduce the noise and to select the genes that could be of interest for further data analysis. In this search for differentially expressed genes, the large number of genes is again more of a blessing than a curse. Similarities between genes can be used in a multi-level (or empirical Bayes) setting to find the cut-off values for being under- or overexpressed per gene. The large number of genes only becomes cumbersome if one wants to test each gene for differential expression between normals and abnormals or any similar grouping of individuals. Controlling the studywise error rate by Bonferroni or more sophisticated corrections can be detrimental, but after switching to false discovery rates as introduced by Benjamini and Hochberg (1995), the large number of genes can be helpful to establish the prevalence of truly expressed genes.

The curse of dimensionality $p \gg n$ comes into play when micro-array data are used for diagnosis/classification or prediction. It is this application of gene-expression data in the paper by Golub *et al*. (1999) that excited a lot of interest in micro-array data among machine learners and statisticians. The remaining papers in this issue all address preventing overfitting in classification/regression models on a high-dimensional predictor.

Early papers on classification using micro-array data exploited rather simple classification rules that appeared hard to beat by more sophisticated classification rules. The paper by Bickel and Levina, inspired by analysing high-dimensional texture data, discusses and explains why the so-called naive Bayes classifier, which ignores the dependencies between the predictors, behaves so well. To put some structure in the high-dimensional explanatory variable, they view the sequence of predictors as a stochastic process and assume stationarity of the covariance function. It is not quite clear how this carries over to the unstructured micro-array data.

From the theoretical point of view, it is interesting to understand why simple rules are hard to beat, but from a more practical point of view it is disappointing that the wealth of data cannot be more efficiently analysed. The lesson is that we need more biological understanding of the relations between genes if we want to get more out of gene-expression data. If $p \gg n$, it is impossible to discover the relevant relations from the data and use these in an efficient way for classification or prediction.

The paper by Greenshtein and Ritov approaches a problem very similar to the one in

Bickel and Levina's paper, but from a different angle, with the emphasis on linear prediction. They offer a theoretical framework for the popular lasso of Tibshirani (1996), which is closely related to soft-thresholding (Donoho 1995). The lasso restricts the ℓ_1-norm when fitting a linear regression model using least squares, or adds an ℓ_1-penalty to the sum of squares. The finding of Greenshtein and Ritov is that persistent procedures (as good as the best procedure under the same restrictions) can be obtained under quite liberal conditions on the restriction. They conclude that there is 'asymptotically no harm' in introducing many more explanatory variables than observations as far as prediction is concerned. It is implicit in their paper that finding the best predictor is different from estimating the vector of regression coefficients. The latter is hopeless if $p \gg n$. The message for the practitioner should be that the lasso (and also some other penalized methods leading to sparse representations) can be safely used in combination with proper cross-validation for the purpose of prediction, but that one should avoid any (biological) interpretation of the set of explanatory variables that are thus selected and their regression coefficients. The link with Bickel and Levina might be that penalization by the ℓ_1-norm of the regression vector has the effect of undoing multi-collinearity and acting as if the predictors were independent.

The paper by Keleș, van der Laan and Dudoit is in the same spirit of finding the best predictor, but in the setting of right-censored survival data. The first problem they deal with is the estimation of prediction error in censored data. They show that the problem of censoring can be handled by IPCW, that is, weighting by the inverse probability of censoring. Secondly, they use cross-validation to estimate the prediction error. They do not use the terminology of Greenshtein and Ritov, but their main result basically states that their procedure is 'persistent'. They show that, asymptotically, the rule that minimizes the empirical cross-validated prediction error behaves as well as the rule that minimizes the expected cross-validation error (the benchmark in their terminology). They do not explicitly address the issue of high-dimensional data. The class of prediction rules is left open and the practitioner has to make sure he/she uses a class of predictors that is rich enough to give cross-validation a chance.

The papers by Birgé and by Kerkyacharian and Picard discuss the problem of estimating an unknown regression function $f(X)$ from a sample from (X, Y) with random X. Birgés paper is theoretical in nature. He defines model selection as selecting a small number of basis functions of which the unknown f is supposed to be a linear combination. Results about optimal selection in L_2-norm are available for designed experiments in which X can be chosen by the observer. Life is more complicated when X is random. Birgé argues that for random pairs (X, Y) the Hellinger distance is more natural and the usual rates can be obtained for this distance, but might not hold for the L_2-norm. The paper by Kerkyacharian and Picard is more practical (but also highly technical). Their starting point is the use of shrunken wavelets in the case of a designed experiment with equidistant observations. In the case of random observations they combine shrunken wavelets with warping of the x-axis induced by the distribution function G of X. If G is not known, it can be estimated by the empirical distribution function. They show that under certain regularity conditions, the behaviour of the new basis is quite similar to the behaviour of the regular wavelet basis.

References

Benjamini, Y. and Hochberg, Y. (1995) Controlling the false discovery rate – a practical and powerful approach to multiple testing. *J. Roy. Statist. Soc. Ser. B*, **57**, 289–300.

Donoho, D.L. (1995) De-noising via soft-thresholding. *IEEE Trans. Inform. Theory*, **41**, 613–627.

Golub, T.R., Slonim, D.K., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J.P., Coller, H., Loh, M.L., Downing, J.R., Caligiuri, M.A., Bloomfield, C.D. and Lander, E.S. (1999) Molecular classification of cancer: Class discovery and class prediction by gene expression monitoring. *Science*, **286**, 531–537.

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