

Statistical Challenges of High-dimensional Data

BY IAIN M. JOHNSTONE¹ AND D. MICHAEL TITTERINGTON²

¹ *Department of Statistics, Stanford University, Stanford, CA 94305, USA*

² *Department of Statistics, University of Glasgow, Glasgow, G12 8QQ, UK*

Modern applications of statistical theory and methods can involve extremely large data-sets, often with huge numbers of measurements on each of a comparatively small number of experimental units. New methodology and accompanying theory have emerged in response: the goal of this theme issue is to illustrate a number of these recent developments. This overview article introduces the difficulties that arise with high-dimensional data in the context of the very familiar linear statistical model: we give a taste of what can nevertheless be achieved when the parameter vector of interest is sparse, that is, contains many zero elements. We describe other ways of identifying low-dimensional subspaces of the data space that contain all useful information. The topic of classification is then reviewed along with the problem of identifying, from within a very large set, the variables that help to classify observations. Brief mention is made of the visualization of high-dimensional data and ways to handle computational problems in Bayesian analysis are described. At appropriate points, reference is made to the other papers in the issue.

Keywords: Bayesian analysis, Classification, Cluster analysis, High-dimensional data, Regression, Sparsity

1. Introduction

While the origins of statistical investigation (e.g. Graunt 1662) predate even those of this eldest of extant scientific journals, the largest development of the science of statistics occurred in the twentieth century. The theory and practice of frequentist methods, the likelihood approach and the Bayesian paradigm all flourished, and informal graphical methods, computational algorithms and careful mathematical theory grew up together.

For most of the period, the primary motivating practical problems consisted of a comparatively large number of ‘experimental units’, on which a comparatively small number of features were measured. If, informally, we let p denote the dimension of what is ‘unknown’ and let n denote the cardinality of what is ‘known’, then traditional theory, and most practice, has until recently been largely limited to the ‘small p , large n ’ scenario. This scenario also naturally reflected the contemporary limitations of computers (the term meant *people* prior to 1950) and graphical display.

A natural mode for asymptotic approximation therefore imagines that $n \rightarrow \infty$ while p remains of smaller order than n , in fact usually fixed. Among the most familiar theoretical results of this type are the Laws of Large Numbers and the Central Limit Theorems. The former says that the sample mean of a random sample

of size n from a population has as a limit, in a well-defined sense, the population mean, as n tends to ∞ . The corresponding Central Limit Theorem shows that the limiting distribution of the sample mean about the population mean (when scaled up by \sqrt{n}) is of the Normal or Gaussian type. In statistics, such results are useful in deriving asymptotic properties of estimators of parameters, but their validity relies on there being, in theory at least, many ‘observations per parameter’.

In practice, n will generally correspond to the number of experimental units on which data are available; for p , however, there are at least two, albeit related, interpretations. The more basic interpretation is as the measure of complexity of the model to be fitted to the data. However, that is often determined by the dimension of the data as given by the number of items (variables) recorded for each experimental unit, and in our presentation we shall use p to represent either interpretation, as appropriate.

Over the last twenty years or so, however, the practical environment has changed dramatically, with the spectacular evolution of data acquisition technologies and computing facilities. At the same time, applications have emerged in which the number of experimental units is comparatively small but the underlying dimension is massive; illustrative examples might include image analysis, microarray analysis, document classification, astronomy and atmospheric science. Methodology has responded vigorously to these challenges, and procedures have been developed or adapted to provide practical results.

However, there is a need for consolidation in the form of a systematic and critical assessment of the new approaches as well as development of appropriate theoretical underpinning. In terms of asymptotic theory, the key scenarios to be investigated can be described as ‘large p , small n ’ or in some cases as ‘large p , large n ’; theory for the former scenario would assume that p goes to infinity faster than n and for the latter would assume that p and n go to infinity at the same rate.

The practical and theoretical challenges posed by the large p /small n settings, along with the ferment of recent research, formed the backdrop to the 2008 research programme ‘Statistical Theory and Methods for Complex, High-dimensional Data’ at the Isaac Newton Institute for Mathematical Sciences, which stimulated this theme issue. It is important to emphasize the breadth of the research community represented in that programme and this theme issue. From a theoretical/methodological point of view it is of relevance not only to statisticians but also to many in the growing population of machine-learning researchers. In addition increasingly many areas of application generate data the analysis of which requires the type of theory and methodology described in this issue.

Before setting the stage for the papers in this volume, we conclude this introduction to the introduction with some general remarks.

It should not, of course, be imagined that the ‘large p ’ scenarios are mere alternative cases to be explored in the same spirit as their ‘small p ’ forbears. A better analogy would lie in the distinction between linear and non-linear models and methods – the unbounded variety and complexity of departures from linearity is a metaphor (and in some cases a literal model) for the scope of phenomena that can arise as the number of parameters grows without limit.

Indeed, *a priori*, the enterprise seems impossible. Good data-analytical practice has always held that the number of data points n should exceed the number of

parameters p to be estimated by some solid margin: $n/p \geq 5$ is a plausible rule of thumb, mentioned for example by Hamilton (1970) and repeated in the classic text by Huber (1981).

The large p /small n world would therefore seem to depend on a certain statistical alchemy – the computational transformation of ignorance into parameter estimates by fearless specification and fitting of high-dimensional models.

Nevertheless, as will be indicated by example in the papers in this volume, in a variety of methodological settings, as well as in numerous scientific applications, there has been notable success with large p models and methods.

The key, of course, is that we are not always ignorant. Indeed, it seems clear that the enterprise can only have hopes of success if the actual number of influential parameters, k , say, is much smaller than the nominal number p . Thus, prior knowledge of the existence of a sparse representation, either of hypothesized form or to be discovered by exploration, is a *sine qua non*.

At the same time, statistical theory is challenged to provide heuristics, principles and results to help explain when sparse models can be expected to be well estimable, or alternatively when the enterprise is simply too ambitious without further reliable prior information.

One such theoretical construct that emerges in a couple of papers in this volume is the ‘phase diagram’. An asymptotic model of a large- p regression or classification problem is expressed in terms of parameters such as the data ratio n/p and the effective parameter sparsity k/n , and for which the diagram depicts sharp transitions between conditions in which estimation/classification is possible and those in which it must fail entirely.

2. Areas of application

Specific frontier-fields for development and application of methods for analysing complex, high-dimensional data include a wide variety of areas within bioinformatics, classification problems in astronomy, tool development for implementing Basel II finance proposals, weather prediction, and so on.

In this theme issue, the greatest emphasis in terms of applications is on aspects of biology. This was also the case in the Newton Institute research programme, with one of the workshops being explicitly so directed. Bickel *et al.* (2009b) provide an extensive review of genomics and the array of statistical techniques that have been recruited or developed to handle the required data analysis: the use of exploratory data analysis, cluster analysis and visualization methods to investigate patterns and structures; the use of modern approaches to classification and prediction to identify disease states or motifs, with the help of methods like the lasso, which we shall describe in §3, to handle scenarios involving very large numbers of explanatory variables or ‘predictors’; the application of latent-variable models such as hidden Markov models for DNA sequencing; and the need to overcome problems associated with multiple testing when investigation of a large number of variables requires the performance of a large number of statistical tests. This last issue is the main theme of Benjamini *et al.* (2009), the other biologically-oriented paper in the batch. They apply the highly influential false discovery rate approach, see §5, in the context of type 2 diabetes.

3. From simple linear regression to high dimension

In this section we attempt to indicate the nature of the general issue of high dimension by starting with a very elementary statistical model and showing how a straightforward process of evolution quickly takes us into realms where the difficulties of high dimension become clear.

(a) *The traditional scenario*

A very simple statistical model can be described as follows. We have data in the form of a pair of measurements on n individuals, $\{x_i, y_i; i = 1, \dots, n\}$, where x_i is a predictor and y_i is a response, and it is assumed that the two measurements are related by

$$y_i = \beta_1 + \beta_2 x_i + \epsilon_i,$$

for each i . It is also assumed that, independently for each i , $\epsilon_i \sim N(0, \sigma^2)$; that is, ϵ_i follows a Gaussian distribution with mean zero and variance σ^2 . Thus, the ‘average’ relationship between the predictor and the response follows a straight line with intercept β_1 and slope β_2 ; this is written

$$E(y_i|x_i) = \beta_1 + \beta_2 x_i,$$

and is referred to as the simple linear regression model for y on x , ‘simple’ because there is only one predictor. The two ‘parameters’, β_1 and β_2 are unknown constants, to be estimated. In one possible application, the predictor might be ‘age’ and the response might be ‘systolic blood pressure’. (Typically, σ^2 would also be unknown, but for simplicity here we take it as known.)

There is a convenient vector-matrix notation for the model for the complete set of n data-pairs:

$$y = X\beta + \epsilon, \tag{3.1}$$

where the $n \times 1$ vector y contains the responses, the vector β contains the 2 (in general p) parameters except for σ^2 , the $n \times 1$ vector ϵ contains the ‘noise’ and the $n \times p$ so-called design matrix X completes the model. In simple linear regression each element in the first column of X is 1.

The standard way of estimating the unknown slope and intercept in β is to use Gauss’s least squares approach and obtain

$$\hat{\beta} = \arg \min_{\beta} \sum_i (y_i - \beta_1 - \beta_2 x_i)^2,$$

which means that $\hat{\beta}$ is the minimizer of the sum of squares function on the right-hand side. In the general vector-matrix notation, this can be written in terms of the Euclidean or ℓ_2 norm, as

$$\hat{\beta} = \arg \min_{\beta} \|y - X\beta\|_2^2.$$

There is another important interpretation of $\hat{\beta}$. Our distributional assumption about ϵ implies that $y \sim N_n(X\beta, \sigma^2 I)$, in which N_n now denotes an n -variate multivariate Gaussian distribution, with $X\beta$ as the vector of means and $\sigma^2 I$ as the

covariance matrix, and where I is the $n \times n$ identity matrix. The probability density function for y is then

$$p(y|X, \beta) = \{\sqrt{(2\pi\sigma^2)}\}^{-n/2} \exp \{-\|y - X\beta\|_2^2 / (2\sigma^2)\}.$$

The available data provide y and X . When viewed as a function of the parameters, this is now called the likelihood function, and, clearly,

$$\hat{\beta} = \arg \max_{\beta} p(y|X, \beta).$$

Thus, $\hat{\beta}$ are the so-called ‘maximum likelihood estimators’ of β ; the use of maximum likelihood estimators is a very common paradigm for statistical inference.

In our problem $\hat{\beta}$ satisfies

$$\begin{aligned} X^{\top} X \hat{\beta} &= X^{\top} y \\ \hat{\beta} &= (X^{\top} X)^{-1} X^{\top} y, \end{aligned}$$

the explicit formula in the second equation being available *provided that $X^{\top} X$ can be inverted*. (Here X^{\top} is the matrix transpose of X .) Furthermore, if the model is correct,

$$\hat{\beta} \sim N_p(\beta, \sigma^2(X^{\top} X)^{-1}), \tag{3.2}$$

from which (frequentist) interval estimates for β can be obtained, with slight modification if, as is usually the case, σ^2 has to be estimated.

In general, for many maximum-likelihood scenarios concerning parametric models involving a fixed number of parameters β , for large n , approximately if rarely exactly,

$$\hat{\beta} \sim N_p(\beta, \Sigma_{\hat{\beta}}),$$

for a certain matrix $\Sigma_{\hat{\beta}}$. Thus, $\hat{\beta}$ is asymptotically unbiased, in that on average it does not overestimate or underestimate β , and Normally distributed.

(b) *A high-dimensional reality check and what to do about it*

For the above elegant and simple analysis we must have $p \leq n$, otherwise $(X^{\top} X)$ is singular and the parameters in the regression model cannot be uniquely estimated. Furthermore, in the general maximum-likelihood contexts, especially if p is not fixed, the asymptotic theory breaks down. What if $p > n$ or even $p \gg n$ in the regression problem, i.e. p is much greater than n ? One approach to side-stepping the singularity of $(X^{\top} X)$ is to use a method of *regularization*, otherwise known as *penalized least squares* or *penalized maximum likelihood*. An early example of this is *ridge regression* (Hoerl & Kennard 1970), in which we estimate β by

$$\hat{\beta}_R = S_{\lambda_2} X^{\top} y,$$

where $S_{\lambda_2} = (X^{\top} X + \lambda_2 I)^{-1}$, and the positive scalar λ_2 is called a ridge parameter or regularization constant. The frequency distribution of $\hat{\beta}_R$ is

$$\hat{\beta}_R \sim N_p(S_{\lambda_2} X^{\top} \beta, \sigma^2 S_{\lambda_2} (X^{\top} X) S_{\lambda_2}). \tag{3.3}$$

Thus the estimator $\hat{\beta}_R$ is now biased, but it can be calculated; as λ_2 increases, bias increases but ‘variance’ decreases, to compensate. There are a number of interpretations for $\hat{\beta}_R$:

- (i) $\hat{\beta}_R = \arg \min_{\beta} \{\|y - X\beta\|_2^2 + \lambda_2 \|\beta\|_2^2\}$.
- (ii) $\hat{\beta}_R$ minimizes $\|y - X\beta\|_2^2$ subject to $\|\beta\|_2^2 \leq c_2(\lambda_2)$, for some $c_2(\lambda_2)$, depending on λ_2 .
- (iii) $\hat{\beta}_R$ minimizes $\|\beta\|_2^2$ subject to $\|y - X\beta\|_2^2 \leq b_2(\lambda_2)$, for some $b_2(\lambda_2)$, depending on λ_2 .

The first interpretation shows that $\hat{\beta}_R$ corresponds to what is called ℓ_2 regularization, because the penalty function is given by the ℓ_2 or quadratic norm. (It also has an interpretation in the Bayesian approach to statistical analysis, as discussed in §7.) In the other interpretations λ_2 or its inverse is a Lagrange multiplier.

However, although invertible, $X^\top X + \lambda_2 I$ is $p \times p$ and still potentially a very large matrix! A dominant strategy in current approaches to this sort of difficulty is to try to exploit *sparsity*; in other words, to seek a solution for β in which many of the elements are zero. After all, if $n < p$ it is intuitively plausible that only sparse solutions can be obtained ‘reliably’. Furthermore, in practice, if there are vast numbers of predictors, it is often scientifically plausible that only a small proportion are likely to be influential as predictors. With this in mind we try changing the penalty function and consider what is called ℓ_0 regularization. Our three equivalent formulations are now as follows:

- (i) $\hat{\beta}_0 = \arg \min_{\beta} \{\|y - X\beta\|_2^2 + \lambda_0 \|\beta\|_0\}$, where $\|\beta\|_0$, the number of nonzero elements in β , is the ℓ_0 norm of β ;
- (ii) $\hat{\beta}_0$ minimizes $\|y - X\beta\|_2^2$ subject to $\|\beta\|_0 \leq c_0(\lambda_0)$;
- (iii) $\hat{\beta}_0$ minimizes $\|\beta\|_0$ subject to $\|y - X\beta\|_2^2 \leq b_0(\lambda_0)$.

The problem about implementing this approach is its combinatorial character: that there is no alternative to considering, individually, each configuration of zero and nonzero values in β and this leads to unacceptable computational complexity and a potential proliferation of local optima.

An intermediate strategy is to base the penalty function on the ℓ_1 norm,

$$\|\beta\|_1 = \sum_{j=1}^p |\beta_j|,$$

leading to the following equivalent formulations:

- (i) $\hat{\beta}_L = \arg \min_{\beta} \{\|y - X\beta\|_2^2 + \lambda_1 \|\beta\|_1\}$, for some λ_1 ;
- (ii) $\hat{\beta}_L$ minimizes $\|y - X\beta\|_2^2$ subject to $\|\beta\|_1 \leq c_1(\lambda_1)$;
- (iii) $\hat{\beta}_L$ minimizes $\|\beta\|_1$ subject to $\|y - X\beta\|_2^2 \leq b_1(\lambda_1)$.

The subscript L is chosen because this method is called the Lasso, for ‘least absolute shrinkage and selection operator’ (Tibshirani 1996). The method has the dual advantages that it is computationally feasible, fitting the paradigm of quadratic programming, and generally leads to sparse solutions. The latter can be explained informally in the context of the second formulation. The solution $\hat{\beta}_L$ will be at

the point of contact of a ‘smooth’ residual sum of squares function and a convex, piecewise-flat constraint surface. The point of contact is very likely to be at a vertex of the constraint surface and therefore at a point where elements of β are zero. As with ℓ_2 regularization, the Lasso has a Bayesian interpretation; see §7.

Some strong support for the Lasso strategy comes from considering noise-free versions of the problem: minimize $\|\beta\|_0$ or $\|\beta\|_1$ subject to the equality constraint $y = X\beta$. If the solution to the ℓ_0 problem is sufficiently sparse – having at most a sufficiently small number k non-zero entries, say – then the solutions to the ℓ_1 and ℓ_0 problems coincide; see for example Candès & Tao (2005) and Donoho (2006).

The main aims of any investigation of sparsity in the linear model can be described informally as follows:

- (i) to identify exactly which components of β are nonzero, i.e. the *support* of β ;
- (ii) to estimate ‘reliably’ the true values of the nonzero components, assuming that the model is correct;
- (iii) to do this ‘as well’ as one could be expected to do if the identities of the nonzero components were known from the beginning, the so-called *oracle* property.

We now describe briefly, at a very informal level, a few particular results along these lines and we give pointers to relevant papers later in this issue.

In their Dantzig Selector, Candès & Tao (2007) choose $\beta = \hat{\beta}_D$ to minimize $\|\beta\|_1$ subject to $\sup_i |X^\top(y - X\beta)|_i \leq c_D \sigma \sqrt{2 \log_e p}$ in which v_i denotes the i th element of a vector v . Note that the residuals, $y - X\beta$, enter into the formulation directly rather than being squared. Let the number of nonzero elements of the true β be k . Then, if $k < n/2$, a version of the oracle property obtains with overwhelming probability. The problem is amenable to algorithms based on linear programming, and the method was named in honour of the inventor of the Simplex Method for solving linear programs.

Bickel *et al.* (2009a) show that, under a sparsity scenario, the Lasso and Dantzig selector exhibit similar behaviour in both linear and nonparametric regression models, and satisfy parallel sparsity oracle inequalities.

For the Lasso itself, Wainwright (2006) investigates the probability, $\text{Prob}(\text{success})$, that the Lasso successfully identifies the correct support set of β , as a function of n , p and k , under certain assumptions about X . In the limit the results are, informally, as follows:

- (i) if $n > 2k \log(p - k)$ then $\text{Prob}(\text{success})$ tends to 1;
- (ii) if $n < (1/2)k \log(p - k)$ then $\text{Prob}(\text{success})$ tends to 0;
- (iii) as n increases from $(1/2)k \log(p - k)$ to $2k \log(p - k)$ the limiting value of $\text{Prob}(\text{success})$ increases from 0 to 1.

Donoho & Tanner (2009) show that sparse linear models can exhibit a transition in behaviour illustrated by a phase diagram. As one example, consider a noise-free setting, in which $y = X\beta$, and the matrix X is chosen at random, for example with independent Gaussian rows, or from the rows of a discrete Fourier transform matrix. Let $\rho = k/n$, the ratio of the cardinality of the support of β and the sample size, and

let $\delta = n/p$. Then, for a given value of $\delta < 1$, there is a threshold, $\rho^*(\delta)$, such that, for $\rho < \rho^*(\delta)$, one is virtually certain to be able to identify the nonzero elements of β and otherwise such identification will almost never occur. As δ increases then so does the threshold, $d(\delta)$, resulting in an intriguing phase diagram in the $\rho - \delta$ plane.

A fascinating property of such diagrams is that closely related versions of it appear in apparently unconnected settings. In geometrical probability, one can look at the convex hull of n points in d -dimensional space. In high-dimensional regression of the form (3.1) one can study the performance of traditional stepwise regression methods for entering variables. In high-dimensional geometry, one can ask what happens to the faces of a simplex under random projections A on to lower-dimensional subspaces. Donoho & Tanner (2009) review the deep connections between these and yet other settings in which a phase transition occurs. They use a large-scale computational experiment to support a conjectured ‘universality’ result that would establish such transition behaviour for a variety of sampling models for the random projection A .

4. More general subspace selection and dimension reduction

The identification of sparse β s in effect allows us to discard variables, leaving an informative k -dimensional linear subspace of the space of predictors, although of a rather special nature; if there are two predictors, $p = 2$, and if $k = 1$ the reduced space would be one of the coordinate axes. (Other modern approaches to the selection of variables, especially from a very large number of candidates, are described in the text and discussion of Fan & Lv (2008).) However, there are plenty of other one-dimensional linear subspaces of the plane, i.e. all straight lines are candidates. Why not, therefore, seek different types of informative k -dimensional linear subspace? So-called *sufficient dimension reductions* reduce the effective dimension of the data-space without loss of information about the distribution of $y|X$, in some sense, in which ‘|’ indicates conditioning. There are various ways of defining a sufficient reduction, $R(X)$. The first approach has the property that the distribution of $y|R(X)$ is the same as that of $y|X$; in other words no information is lost from the (forwards) regression of Y on X by restricting oneself to conditioning on $R(X)$. In the second approach, related to the idea of inverse regression, the aim is to choose $R(X)$ so that the distribution of $X|R(X)$ is the same as that of $X|\{y, R(X)\}$. Finally, in a ‘joint’ approach, $R(X)$ is sought such that y is independent of X given $R(X)$. These options are identified by Cook (2007), the leading figure in this area, who takes the inverse-regression approach, proposing a model of the form

$$X_y = \mu + \Gamma z_y + \epsilon,$$

with Γ a $p \times k$ matrix, with the goal that X_y has the same distribution as that of X given y . Here the vector z_y contains so-called ‘latent’ or ‘factor’ variables. In the present issue, Adragni & Cook (2009) rehearse and expand on these concepts in detail, as well as announcing new methodology for prediction.

All the discussion hitherto has been about the *linear* model (3.1) in which the part of the model that defines the means of the responses is *linear* in the parameters β . There is a vast array of other regression models that generalize the linear case,

including nonlinear parametric models in which the mean function is defined explicitly as a nonlinear function of parameters, and various types of semiparametric or nonparametric approach. The so-called ‘errors’, like ϵ in (3.1), need not be normally distributed and need not enter the model additively. Traditionally, as with the linear model, these other approaches have been applied in contexts where the number of experimental units has comfortably exceeded the number of explanatory variables, i.e. $n > p$, but the complementary scenario has increasingly attracted attention. Here we limit consideration to two particular manifestations.

Ravikumar *et al.*’s (2007) SpAM method, an abbreviation of ‘Sparse Additive Models’, aims to estimate a function rather than a vector, using so-called Generalized Additive Models (Hastie & Tibshirani 1990). For the i th observational unit, the model for the response is

$$y_i = \sum_{j=1}^p \beta_j f_j(x_{ij}) + \epsilon_i,$$

in which x_{ij} is the value of the j th explanatory variable for the i th observational unit. The $\{f_j\}$ are rather arbitrary functions, chosen to minimize the residual sum of squares but subject to constraints that both encourage sparsity and render the optimization problem convex. When p is very large but it is assumed that only a small subset of the explanatory variables truly contribute to the model, i.e. sparsity obtains, the objective is to identify those variables as well as possible.

Chipman *et al.* (2009) propose Bayesian Additive Regression Trees (BART), in which the regression function is the sum of ‘trees’:

$$y_i = \sum_{j=1}^m f_j(x_i; T_j, \beta_j) + \epsilon_i,$$

each T_j representing a tree structure involving a number, usually very small, of the explanatory variables.

We leave the regression setting and turn to some other traditional areas of multivariate statistical analysis. We describe briefly two active areas of development: investigation of the properties of classical methods when the number of variables is large, and development of new, typically nonlinear variants of these methods that respond to high dimension.

In the first direction, it is a striking fact that most of the standard techniques of classical multivariate analysis – principal components analysis, canonical correlation analysis, multivariate analysis of variance, discriminant analysis and so forth – are based on the eigenanalysis of sample covariance matrices. Under Gaussian assumptions for the distributions of the sampled data, and under the symmetry assumptions characteristic of null hypotheses, the sampling densities of the eigenvalues and generalized eigenvalues of these matrices have precisely the laws arising in the classical orthogonal polynomial ensembles of Random Matrix Theory. The natural asymptotics in Random Matrix Theory – reflecting its origins in the many-particle models of statistical and nuclear physics – allow the number of variables to become large. This leads to limiting distributions (Marčenko-Pastur, Tracy-Widom) that are new for multivariate statistics and to approximations for the distributions of extreme eigenvalues that can work well even the number of variables is small; see Johnstone (2007, 2008) and the references therein.

Ideas from Random Matrix Theory are also exerting active influence on a variety of other problems connected with large covariance matrices. A rich set of examples may be found in a recent special issue of the *Annals of Statistics* devoted to the topic (Bickel 2008). For example, an unknown covariance matrix for observations on p variables in principle has $p(p+1)/2$ unknown parameters and the sample covariance matrix is a poor estimate when p is at all large. Strategies to exploit sparsity are essential, for example by direct thresholding (Bickel & Levina 2008) or by exploiting sparse covariance models (El Karoui 2008). Estimation of the leading *eigenvectors* associated with large covariance matrices can encounter problems with consistency unless sparsity is assumed, and is attracting attention both in statistics, see for example Nadler (2008) and Johnstone & Lu (2009), and in econometrics (Onatski 2009).

We now focus on the second direction, methods of dimension reduction: given data x_1, \dots, x_n in \mathbb{R}^p , find representations y_1, \dots, y_n in \mathbb{R}^m for $m \ll p$ that preserve as much of the relevant information as possible. Of course, if m is three or less, one can then visualize the reduced data. The traditional approach to this problem is via principal components analysis, or its close relation, multidimensional scaling, which uses the leading eigenvectors of the sample covariance matrix of (x_i) (e.g. Jolliffe 2002). Geometrically, this corresponds to finding the m -dimensional *linear* subspace into which the projections of x_i have the largest variance.

In recent years, attention has been directed at settings in which a *nonlinear* manifold of low dimension might provide a better representation. For example, a collection of images of similar objects is nominally very high-dimensional (corresponding to the number of pixels in each image), but in fact potentially described by the variation of a much smaller number of parameters. Some influential approaches to this problem include ISOMAP (Tenenbaum *et al.* 2000), Local Linear Embedding (Roweis & Saul 2000), Laplacian Eigenmaps (Belkin & Niyogi 2003) and Hessian Eigenmaps (Donoho & Grimes 2003).

These nonlinear representations can also be described as spectral methods for dimension reduction, since each involves finding a small number of extreme eigenvalue/vector pairs for a suitable matrix derived from the local characteristics of the data (x_i) . These and other methods are reviewed by Belabbas & Wolfe (2009) in this volume.

The computational burden of solving an eigenproblem can be significant if $\min(n, p)$ is large. As one approach to reducing this burden, Belabbas & Wolfe (2009) compare algorithms for selecting a subset of the n data points (“landmark selection”), and performing an approximate spectral analysis known as the Nyström extension.

5. Classification

A variation of the regression problem is that of ‘Classification’, also known as ‘Discriminant Analysis’ or ‘Statistical Pattern Recognition’, in which the class label y_i is categorical or in particular binary: it might denote an image type or a disease category. Here the main objective is to estimate a classification rule for a future ‘patient’ on the basis of their x_i , which represents the patient’s medical history, given ‘training data’ from people whose disease categories and medical histories are known. This scenario is known as ‘supervised learning’, whereas any situation in

which the disease categories of even the training set are not available corresponds to ‘unsupervised learning’. There are many approaches to the classification problem, from both statistics and machine learning, the latter including the so-called ‘support vector machine’(SVM); see for example Hastie *et al.* (2009). The ‘large p , small n ’ problem arises here too and the method can also be described in terms of constrained optimization. In the simplest version of the problem there are two disease categories and the training data consist, geometrically, of two clouds of points that can be separated by at least one hyperplane. The SVM simply identifies a hyperplane such that the resulting degree of separation is, informally speaking, as large as possible. Many variations of the method have been developed, to cope with non-planar separating surfaces, cases in which the two point-clouds overlap, and so on.

So far as probabilistic modelling in classification is concerned there are two approaches, corresponding to two possible factorizations of $p(x, y)$:

$$\begin{aligned} p(x, y) &= p(x|y, \theta_1)p(y|\theta_2), \\ p(x, y) &= p(x|\phi_1)p(y|x, \phi_2), \end{aligned}$$

where (θ_1, θ_2) and (ϕ_1, ϕ_2) denote two parameterizations, with θ_1 distinct from θ_2 and ϕ_1 distinct from ϕ_2). In the modern machine-learning literature these approaches are called the ‘generative’ and ‘discriminative’ approaches, respectively. Much earlier, Dawid (1976) called them the ‘sampling’ and ‘diagnostic’ paradigms. Which approach is selected is crucial in so-called ‘semisupervised’ situations in which the available database contains both diagnosed and undiagnosed cases. From the point of view of diagnosing new patients, the conditional distribution $p(y|x)$ is the key tool. If the θ version is adopted then the undiagnosed cases do contribute information about $p(y|x)$ but if one uses the ϕ version they do not. There is much current work re-visiting this issue in the machine-learning literature, under key words such as ‘domain adaptation’ as well as ‘semisupervised learning’.

The selection of good class-predictors from among many raises problems incurred in ‘multiple testing’. In microarray analysis, for example, we could be comparing expression levels of $p = 20000$ genes obtained from people from 2 populations, e.g. diseased and healthy, and wish to decide which genes respond differently in the two populations. It is natural to examine $|\bar{X}_{1j} - \bar{X}_{2j}|$, the difference in the average values of the expression levels of the j th gene, in the members of the two disease-classes in the available database, for $j = 1, \dots, 20000$, and to select those genes for which there is a ‘significant difference’. The following problems arise: first, with a conventional significance level of 5% we should expect about 1000 significant results even when there is no real difference; secondly, results for different genes may well be correlated. There is much recent work under this banner of ‘multiple testing’ on the same data-set, including methods for controlling the so-called ‘false discovery rate’ (Benjamini & Hochberg 1995). The actual false discovery rate is the number of ‘null hypotheses’ wrongly rejected divided by the total number of ‘null hypotheses’ rejected. In practice, of course, this proportion is not knowable, so one aims to control its expectation. As mentioned in §2, this topic is the subject, in this issue, of Benjamini *et al.* (2009).

The problem of distinguishing between two classes on the basis of noise-filled, high-dimensional data is one that lends itself to the theoretical challenge described

earlier of delimiting the limits of the possible. Perhaps the simplest idealization runs as follows. The n labels y_i are either -1 or $+1$ and the feature vectors x_i have p components x_{ij} . The components x_{ij} are assumed to be independently normally distributed, with means $y_i\delta_j$ and variance 1. The co-ordinate separations δ_j are mostly zero, that is for all but a small number of components j whose identities are unknown – more precisely, a proportion $p^{-\beta}$ for $0 < \beta < 1$. In those rare separated co-ordinates, the dependence on n and p is assumed to be given by $\delta_j = \sqrt{\{(2r \log p)/n\}}$.

Two papers in this volume (Donoho & Jin 2009; Ingster *et al.* 2009) independently study this problem asymptotically. Whether or not successful classification is possible at all in this model turns out to depend in a precise way on the sparsity β and separation constant r . Both papers show that there is a ‘detection boundary’ $r = r(\beta)$, or ‘phase diagram’, such that above the boundary – corresponding to greater separation r and less sparsity β – correct classification is possible, while below the boundary it is impossible, whatever be the method used.

Donoho & Jin (2009) in this volume, and Jin (2009) elsewhere, focus on a particular method for feature selection, based on the *higher criticism* approach to simultaneous inference associated with J. W. Tukey, while Ingster *et al.* (2009) consider a somewhat more general setting.

While this theoretical setting is certainly highly idealized, it is worth noting that a simple classifier based on higher-criticism feature selection turns out to be completely competitive, if not superior, to standard classification methods including SVMs and Random Forests (Breiman 2001), at least on some standard test datasets as reported in Dettling (2004) and Donoho & Jin (2008).

6. Visualization and informal inference

The facility to *look* at data is vital, partly with a view to exploring the data and suggesting questions to be followed up more formally and partly as a way of checking assumptions that underlie a formal procedure being implemented, often after that procedure has been carried out and a model fitted; Buja *et al.* (2009) regard these two activities respectively as cases of ‘exploratory data analysis’ and ‘model diagnostics’. At a trivial level, in the context of the simple linear regression model of §3, a two-dimensional plot of responses against predictors will give an immediate warning if the straight-line model is implausible. It is hard to achieve such a direct look in high dimensions. However, the power of modern computers is there to be exploited. Dynamic graphics allow the display of a multitude of two-dimensional projections of three-dimensional data-clouds by spinning the clouds, with the axis of rotation being chosen by the user. The so-called ‘grand tour’ generalizes this notion to the exploration of p -dimensional data-clouds for larger values of p . The idea is to generate a space-filling tour of low-dimensional projections of the data, to be visualized as a movie-like presentation. Modifications of the grand tour allow the user to interact with the tour, for example concentrating on classes of projections of particular interest. A different medium for the two-dimensional display of high-dimensional data is that of ‘parallel coordinates’. In the most basic form of this procedure for p -dimensional data the p typically orthogonal axes are replaced by a set of p parallel axes, displayed in two dimensions. For a given observational unit the values of the variables are marked off on the axes and joined up in a piecewise-

linear way. This allows, for instance, ‘similar’ points to be recognized through their similar piecewise-linear plots. There are many refinements and variations of the basic idea.

For algorithmic and practical discussion of these and other visualization techniques see Buja *et al.* (2005) and Wegman (2003), and for treatments of the underlying mathematics see Buja *et al.* (2004) and Wegman & Solka (2002). These papers also provide links to relevant software.

In most statistical practice, the consequence of data visualization, be it for exploratory analysis or model diagnostics, is typically an informal, non-quantitative assessment of the plausibility of some ‘hypothesis’, to be followed up by the performance of an appropriate formal test or construction of a confidence interval. In this theme issue, Buja *et al.* (2009) pursue the goal of adding quantification of the level of statistical significance of the sort of ‘visual discovery’ that plots and diagrams may provide

It is in the spirit of data exploration that Banks *et al.* (2009) develop their data-analytic approach designed to exploit what they call *mixture sparsity*. For example, in the context of cluster analysis they envisage that a dataset comprises a number of clusters, but that cluster membership indicators are missing, and that the different clusters may be best characterized by possibly different (small) subsets of variables. By ‘cherry-picking’ small numbers of variables in a way designed to identify mutually similar data-points they construct clusters one-by-one. Similarly they tease out mixtures of regression models where, again, the different (sparse) regression models may involve different (small) sets of predictors.

7. A brief encounter with Bayesian statistics

We return to §3*b* and the discussion of the linear model as expressed in equation (3.1). Frequentist distributional results for estimators of β were given in equation (3.2) for the least-squares estimators and in equation (3.3) for the ridge-regression estimators. In addition, the ridge-regression estimators could be interpreted as penalized maximum likelihood estimators. Here we provide another interpretation in the context of so-called Bayesian inference, in which probability distributions are developed for unknown parameters. Bayes’ Theorem is used to relate the distribution $p(\beta)$, assumed for β ‘prior’ to the data-gathering process, to the distribution $p(\beta|y)$ ‘posterior’ to that process. To be specific,

$$p(\beta|y)p(y) = p(\beta, y) = p(y|\beta)p(\beta),$$

so that, as a function of β ,

$$p(\beta|y) \propto p(y|\beta)p(\beta);$$

that is,

$$\text{posterior for } \beta \propto \text{likelihood} \times \text{prior for } \beta. \quad (7.1)$$

In the Bayesian approach, inference about β is based on $p(\beta|y)$, historically a great source of controversy, essentially because of the combination within (7.1) of probability distributions of quite different natures, the likelihood term being frequentist-based and the prior not so. In addition adoption of the approach often led to practical (computational) difficulties, to which we allude later. However, these problems

do not arise in the context of the linear model in equation (3.1), if we are prepared to act as if, *a priori*,

$$\beta \sim N_p(0, \sigma^2 \lambda_2^{-1} I),$$

and that σ^2 is known, for then the posterior distribution for β is a Gaussian distribution with mean, and indeed mode, given by $\hat{\beta}_R$. Thus, the ridge-regression estimator has the Bayesian interpretation of being a *Maximum a Posteriori* (MAP) estimator of β . Similarly, the Lasso has a MAP interpretation provided that the prior for β is

$$p(\beta) \propto \exp \{-\lambda_1 \|\beta\|_1 / (2\sigma^2)\}.$$

In other contexts the analysis is not so straightforward, especially if there are missing or ‘latent’ (unobservable) variables, z , as well as observed data y . Often (not always!), analysis would be comparatively easy were z known, in that simple formulae would exist for $p(y, z|\beta)$ and $p(\beta|y, z)$. With z unknown, the key quantity of interest is $p(z, \beta|y)$, the distribution of everything that is unknown given what is known, but this distribution is typically complicated. We briefly indicate two popular ways of handling this problem.

Stochastic method: Generate a large number of realizations from $p(z, \beta|y)$, where

$$p(z, \beta|y) \propto p(y, z, \beta) = p(y, z|\beta)p(\beta).$$

The corresponding realizations of β are a sample from $p(\beta|y)$ and, for instance, their average would be a good estimator of the true posterior mean of β . During the last twenty-five years or so a large repertoire of so-called Markov chain Monte Carlo methods, some based on antecedents from the statistical physics literature, has been developed for simulating the required realizations.

Variational method: Here the approach is to choose a deterministic approximation $q_y(z, \beta)$ that is as close as possible to $p(z, \beta|y)$ but that is of a more manageable structure than $p(z, \beta|y)$. Here ‘closeness’ is measured by the Kullback-Leibler directed divergence, $\text{KL}(q_y, p)$, where p denotes $p(z, \beta|y)$:

$$\text{KL}(q_y, p) = \int_z \int_\beta q_y(z, \beta) \log \left(\frac{q_y(z, \beta)}{p(z, \beta|y)} \right) d\beta dz.$$

Key properties are that $\text{KL}(q_y, p) \geq 0$, with equality essentially if and only if q_y and p are the same. The typical way of choosing a ‘more manageable structure’ is to assume a factorized form for q_y , that is,

$$q_y(z, \beta) = q_y^{(z)}(z) \times q_y^{(\beta)}(\beta),$$

choose the factors so as to minimize $\text{KL}(q_y, p)$, and use the resulting $q_y^{(\beta)}(\beta)$ to approximate $p(\beta|y)$.

This variational approach also has links to statistical physics, including concepts such as mean-field approximations. Its recent application to many implementations of the Bayesian approach has mainly developed in machine learning, see Bishop (2006) and references therein, rather than in the mainstream statistics literature, although not exclusively so (Titterton 2004; Beal & Ghahramani 2006).

In theory the stochastic method is superior in that, if enough realizations are generated, one can get arbitrarily close to the target distribution. However, especially in problems on a very large scale, practical considerations render the stochastic approach non-viable and the variational method is a good pragmatic alternative.

Among the papers in this issue, Barber (2009) describes a particular application to the identification of clusters of books about U.S. politics. In general, variational approximations have proved to be extremely useful in the analysis of graphical models, which involve structures of nodes and connecting edges with a wide range of architectures. As Barber (2009) indicates, such models are essential in contexts such as social networks, web analysis and bioinformatics, the models often necessarily display a high degree of complexity and efficient approximate methodology is essential; roughly speaking, the more connectivity there is in the graphical model, the more intractable the analysis thereof becomes.

This research was supported in part by the Isaac Newton Institute for Mathematical Sciences, Cambridge, U.K. and by grants NSF DMS 0505303 and NIH RO1 EB 001988.

References

- Adraghi, K. P. & Cook, R. D. 2009 Sufficient dimension reduction and prediction in regression. *Phil. Trans. R. Soc. A*, to appear.
- Banks, D. L., House, L. & Killhoury, K. 2009 Cherry-picking for complex data: robust structure recovery. *Phil. Trans. R. Soc. A*, to appear.
- Barber, D. 2009 Identifying graph clusters using variational inference and links to covariance parameterisation. *Phil. Trans. R. Soc. A*, to appear.
- Beal, M. J. & Ghahramani, Z. 2006 Variational Bayesian learning of directed graphical models with hidden variables. *Bayesian Statist.* **1**, 793-822.
- Belabbas, M-A. & Wolfe, P. J. 2009 On landmark selection and sampling in high-dimensional data analysis. *Phil. Trans. R. Soc. A*, to appear.
- Belkin, M. & Niyogi, P. 2003 Laplacian eigenmaps for dimensionality reduction and data representation. *Neural Computation* **15**, 1373-1396.
- Benjamini, Y., Heller, R. & Yekutieli, D. 2009 Selective inference in complex research. *Phil. Trans. R. Soc. A*, to appear.
- Benjamini, Y. & Hochberg, Y. 2005 Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Statist. Soc. B* **57**, 289-300.
- Bickel, P. 2008 Random matrix theory: A program of the Statistics and Applied Mathematical Sciences Institute (SAMSI). *Ann. Statist.* **36**, 2551-2552.
- Bickel, P. J., Brown, J. B., Huang, H. & Li, Q. 2009b An overview of recent developments in genomics and associated statistical methods. *Phil. Trans. R. Soc. A*, to appear.
- Bickel, P. J. & Levina, E. 2008 Covariance regularization by thresholding. *Ann. Statist.* **36** 2577-2604.
- Bickel, P. J., Ritov, Y. & Tsybakov, A. B. 2009a Simultaneous analysis of Lasso and Dantzig selector. *Ann. Statist.* to appear.
- Bishop, C. M. 2006 *Pattern recognition and machine learning*, ch. 10. Springer.
- Breiman, L. 2001 Random forests. *Machine Learning* **45**, 5-32.
- Buja, A., Cook, D., Hofmann, H., Lawrence, M., Lee, E.-K., Swayne, D. F. & Wickham, H. 2009 Statistical analysis for exploratory data analysis and model diagnostics. *Phil. Trans. R. Soc. A*, to appear.
- Buja, A., Cook, D., Asimov, D. & Hurley, D. 2004 Theory of dynamic projections in high-dimensional data visualization. Technical report.
- Buja, A., Cook, D., Asimov, D. & Hurley, D. 2005 Computational methods for high-dimensional rotations in data visualization. In *Handbook of statistics, Volume 24: Data mining and computational statistics* (ed. C. R. Rao, E. J. Wegman & J. L. Solka), pp. 391-414. Amsterdam: North Holland.

- Candès, E. J. & Tao, T. 2005 Decoding by linear programming. *IEEE Trans. Inform. Theory* **51** 4203–4215.
- Candès, E. & Tao, T. 2007 The Dantzig selector: statistical estimation when p is much larger than n (with discussion). *Ann. Statist.* **35**, 2313–2404.
- Chipman, H. A., George, E. I. & McCulloch, R. 2009 BART: Bayesian additive regression trees. Preprint NI09002-SCH. Isaac Newton Institute, Cambridge.
- Cook, R. D. 2009 Fisher lecture: dimension reduction in regression (with discussion). *Statist. Sci.* **22**, 1–43.
- Dawid, A. P. 1976 Properties of diagnostic data distributions. *Biometrics* **32**, 647–658.
- Dettling, M. 2004 BagBoosting for tumor classification with gene expression data. *Bioinformatics* **20** 3583–3593.
- Donoho, D. L., & Grimes, C. 2003 Hessian eigenmaps: locally linear embedding techniques for high-dimensional data. *Proc. Natl. Acad. Sci. USA* **100** 5591–5596.
- Donoho, D. L. 2006 For most large underdetermined systems of linear equations the minimal ℓ_1 norm is also the sparsest solution. *Comm. Pure Appl. Math.* **59** 797–829.
- Donoho, D. L. & Jin, J. 2008 Higher criticism thresholding: optimal feature selection when useful features are rare and weak *Proc. Nat. Acad. Sci.* **105** 14790–14795.
- Donoho, D. & Jin, J. 2009 Feature selection by higher criticism thresholding: optimal phase diagram. *Phil. Trans. R. Soc. A*, to appear.
- Donoho, D. & Tanner, J. 2009 Observed universality of phase transitions in high-dimensional geometry with broad implications for 21st-century data analysis and signal processing. *Phil. Trans. R. Soc. A*, to appear.
- El Karoui, N. 2008 Operator norm consistent estimation of large-dimensional sparse covariance matrices. *Ann. Statist.* **36** 2717–2756.
- Fan, J. & Lv, J. 2008 Sure independence screening for ultrahigh dimensional feature space (with discussion). *J. R. Statist. Soc. B* **70**, 849–911.
- Graunt, J. 1662 *Natural and political observations made upon the bills of mortality*. London: Martyn.
- Hamilton, W. C. 1970 The revolution in crystallography. *Science* **169** 133–141.
- Hastie, T. & Tibshirani, R. 1990 *Generalized additive models*. London: Chapman & Hall/CRC.
- Hastie, T., Tibshirani, R. & Friedman, J. H. 2009 *The elements of statistical learning*, ch. 12, 2nd edn. New York: Springer.
- Hoerl, A. E. & Kennard, R. W. 1970 Ridge regression: biased estimation for nonorthogonal problems. *Technometrics* **12**, 55–82.
- Huber, P. J. 1981 *Robust statistics*. New York: Wiley.
- Ingster, Y. I., Pouet, C. & Tsybakov, A. B. 2009 Classification of sparse high-dimensional vectors. *Phi. Trans. R. Soc. A*, to appear.
- Jin, J. 2009 Impossibility of successful classification when useful features are rare and weak. *Proc. Natl. Acad. Sci.* in press.
- Johnstone, I. M. 2007 High dimensional statistical inference and random matrices. *International Congress of Mathematicians, Vol. I.* 307–333. Zürich: Eur. Math. Soc.
- Johnstone, I. M. 2008 Multivariate analysis and Jacobi ensembles: Largest eigenvalue, Tracy-Widom limits and rates of convergence. *Ann. Statist.* **36**, 2638–2716.
- Johnstone, I. M. & Lu, A. Y. 2009 On consistency and sparsity for principal components analysis in high dimensions (with discussion). *J. Am. Statist. Assoc.* **104**, 682–703.
- Jolliffe, I. T. 2002 *Principal component analysis*, 2nd edn. New York: Springer.
- Lindsay, B. G., Kettenring, J. & Siegmund, D. O. 2004 A report on the future of statistics (with discussion). *Statist. Sci.* **19**, 387–413. (DOI 10.1214/088342304000000404.)

- Nadler, B. 2008 Finite sample approximation results for principal component analysis: A matrix perturbation approach. *Ann. Statist.* **36**, 2791–2817.
- Onatski, A. 2009 Asymptotics of the principal components estimator of large factor models with weak factors. Working Paper, Department of Economics, Columbia University.
- Ravikumar, P., Liu, H., Lafferty, J. & Wasserman, L. 2007 SpAM: sparse additive models. In *Advances in Neural Information Processing Systems* (ed. J. C. Platt, D. Koller, Y. Singer & S. Roweis), **20**.
- Roweis, S. T. & Saul, L. K. 2000 Nonlinear dimensionality reduction by locally linear embedding. *Science* **290**, 2323–2326.
- Tenenbaum, J., deSilva, V. & Langford, J. 2000 A global geometric framework for nonlinear dimensionality reduction. *Science* **290**, 2319–2323.
- Tibshirani, R. 1996 Regression shrinkage and selection via the lasso. *J. R. Statist. Soc. B* **58**, 267–288.
- Titterton, D. M. 2004 Bayesian methods for neural networks and related models. *Statist. Sci.* **19**, 128–139.
- Wainwright, M. J. 2006 Sharp thresholds for noisy and high-dimensional recovery of sparsity using l_1 -constrained quadratic programming. Technical Report 709, Department of Statistics, University of California, Berkeley.
- Wegman, E. J. 2003 Visual data mining. *Statist. Med.* **22**, 1383–1397.
- Wegman, E. J. & Solka, J. L. 2002 On some mathematics for visualizing high dimensional data. *Sankhya A* **64**, 429–452.