Open Access Medical Statistics

Dovepress
open access to scientific and medical research

Open Access Full Text Article

REVIEW

High-dimensional data and linear models: a review

This article was published in the following Dove Press journal:
Open Access Medical Statistics
6 August 2014
Number of times this article has been viewed

M Brimacombe

Department of Biostatistics,
University of Kansas Medical Center,
Kansas City, KS, USA

Abstract: The need to understand large database structures is an important issue in biological and medical science. This review paper is aimed at quantitative medical researchers looking for guidance in modeling large numbers of variables in medical research, how this relates to standard linear models and the geometry that underlies their analysis. Issues reviewed include LASSO-related approaches, principal-component based analysis, and issues of model stability and interpretation. Model misspecification issues related to potential nonlinearities are also examined, as is the Bayesian perspective on these issues.

Keywords: high dimensional data, quantitative medical research, database structures, linear models, LASSO

Introduction

As high-dimensional data structures have begun to be available and studied in many areas of medical research, the need for intuitive, geometric, and often linear model-based understanding of such data has grown. Genetic data, imaging data, health outcomes and clinical data, spatial positioning data, internet-based data: all are examples of settings where the flow of data is massive and the ability to analyze such a flow is typically restricted.

A very large number of variables and relatively few subjects (large p and small n) are often the mark of such data and the goal of the analysis is typically to detect various patterns within the overall dataset. In genomic settings these may be simple mean differences in gene expression levels across treatment groups, comprehensive correlated network clusters, or more detailed epigenetic patterns. Often there is limited theoretical modeling and much of the applied statistical research is empirically driven, falling under the hypothesis or model generating label, with the term "data science" sometimes used.

Standard methods of statistical analysis often do not hold up well in such settings. Multiple comparison issues where a large number of case-control comparisons are conducted require careful application and interpretation.[1] Indeed, multiple testing of one-at-a-time mean differences may be of limited use for understanding of genomic and epigenetic data structures or networks of genes relevant to specific cell and phenotypic structures. A more complex three-dimensional nonlinear aspect of chromosome structure may be relevant to such analyses.[2]

A geometric perspective is useful in understanding the properties of estimators and models developed in a linear model or associated analysis of variance (ANOVA) setting. These are often based on orthogonal projections onto a linear plane and the space

Correspondence: M Brimacombe
Department of Biostatistics,
University of Kansas Medical Center,
3901 Rainbow Blvd, Kansas City,
KS 66160, USA
Email mbrimacombe@kumc.edu

submit your manuscript | www.dovepress.com
Dovepress
http://dx.doi.org/10.2147/OAMS.S56499

orthogonal to it. The squared lengths of these projections can be compared, interpreted, and used to develop statistics for estimation and testing.[3] However, the basic geometric intuition of linear models is altered when $p > n$ and standard projections are restricted by limited dimensions. Here standard approaches must be applied carefully if the resulting models are to be interpretable.

Correlation and embedded nonlinear relationships also cause difficulty for linear models. The application of linear models to correlated structures may not be appropriate, with nonlinear functional relationships potentially going undetected and creating instabilities in the predictive model. Many developmentally related growth factors are nonlinear in pattern.[4] Scaling issues, gene clusters, and small embedded networks also affect the applicability of the linear model.

Statisticians have developed techniques for restricted or sparse situations, including least-angle regression (LARS) extended via application of the least absolute shrinkage and selection operator (LASSO),[5] and Dantzig[6] approaches, which extend older techniques such as restricted least squares, ridge regression, forward stagewise variable selection techniques, and principal components. Several earlier, more detailed reviews can be found in Johnstone and Titterington,[7] and in Bickel et al.[8] Note that these approaches do not always converge to a fitted model (in which case an all-subsets search is required to find a best fitting model, often impractical in terms of time)[9] or give a useful predictive model. Some methods have phase-threshold cutoff patterns that give insight into possible convergence.[10]

An area of application for high-dimensional methods is genetic data structures, which began with Southern blot electrogenesis technology and other fairly simple DNA-related technology and have developed into much more detailed approaches including: single-nucleotide polymorphisms, copy-number variation, gene splicing, and RNA-related deep sequencing.[11] These datasets often reflect specialized bioassays and there remains much to be done regarding standardizing platforms, alignment techniques, etc.[1] Recently, the rise of epigenetics, the chemical triggers governing gene expression (chromatin, histone, DNA methylation, for example), have lead to yet another level of complexity, as have the growing number of detected epigenetically triggered networks or clusters of genes that govern protein, cell, and other developmental and maintenance-related activities in the organism.[12,13]

Outside of genetics, the areas of systems and network-related biology, imaging, clinical data repositories, internet-based information, and other "large data" settings are all growing very quickly.[14] Data science or big data-based approaches to identifying patterns in these large sets of collected data are quite varied, often reflecting a mix of methods drawn from engineering, computer science, and mathematics.[15] The statistically based approaches reviewed here also apply to these areas of investigation.

Here we review and discuss several related statistical methods and tools of high-dimensional data analysis from a practical and geometric perspective, as they relate and extend existing standard methods such as ANOVA and principal components analysis (PCA). The stability of linear methods is examined and we investigate the effect of misspecification, especially where this is related to nonlinearity. We review practical interpretations of $p > n$ methods using the geometry of least squares, restricted least squares, correlation, and simulated examples, briefly mentioning Bayesian perspectives in this setting.

High-dimensional statistical approaches and linear models

A standard tool for understanding data and linking a response variable to various explanatory variables is the linear model

$$y = X\beta + \varepsilon$$

where y is a $(n \times 1)$ vector of responses, $X = [x_1, \ldots, x_n]$ a $(n \times p)$ matrix of p measured variables, and ε a $(n \times 1)$ vector of error components. If all variables x_i are thought to be relevant, the fitted least squares model is given by $\hat{y} = Xb$ where $b = (X'X)^{-1}X'y$.

Typically in the settings reviewed here, many variables x_i have been collected and only a few are thought to be relevant to predicting the outcome y. In such a setting, placing an explicit restriction on the model, for example, expecting only a few β_i values to differ significantly from zero, may be helpful in finding the "best" underlying linear model, which is typically carried out using stepwise or stagewise methods. Correlations among the explanatory variables and resulting rank deficiency in the X matrix may also require the use of modified or restricted linear model-based approaches such as ridge regression[16] that have a long history.

A sparseness restriction is often expressed as

$$\sum_{i=1}^{k} |\beta_i|^m < t$$

for relatively small chosen values t and k. Sparseness restrictions typically assume m =1 or 2. This is most useful in

settings where only a few x_i are thought to be significantly correlated with the response y_i but is also necessary when $p > n$ and the linear model suffers from less than full rank in the design matrix X. In these settings the usual least squares estimator $b = (X'X)^{-1}X'y$ will not exist.

From the perspective of the parameter space, such restrictions limit the set of possible β_i combinations that may be examined. The shape of the restricted parameter space depends typically on the value of m. From the perspective of the sample space they limit potential values for the estimators $\hat{\beta}_i$ affecting the application of standard least squares based geometry.

Ridge regression

In the case of highly correlated variables in the X design matrix, which affect the stability and existence of $(X'X)^{-1}$, the older and more commonly used ridge-regression approach can be applied and uses $m = 2$. It is worth examining ridge regression in the case $n > p$. Assuming centered data, the resulting estimator is given by

$$b^R = (X'X + \lambda I)^{-1}X'y$$

for scalar λ. Even with high correlation in the X design matrix this will exist, with λ chosen graphically or via Bayesian posterior calculation. The singular value decomposition (SVD) guides the development.

To apply SVD in general we write $X = UDV'$, where $U = (u_1, \ldots, u_p)$ is a n by p orthogonal matrix, the u_j form an orthonormal basis for the column space of X, and V is a similarly constructed orthogonal matrix for the row space of X. D is a diagonal matrix $(d_1, \ldots, d_p)$.

Geometrically, ridge regression is equivalent to projecting y onto the normalized principal components of X (ie, u_j) where the jth principal component of X is given by $d_j\ u_j$. Specifically,

$$b^R_j = [d^2_j/d^2_j + \lambda]\ u_j'\ y$$

which can be viewed as weighting the projection of y onto the principal component u_j by the relative weights of d_j and λ. The important role played by eigenvalues in the application of linear models in restricted settings and dimension reduction in general is further discussed below.

PCA LASSO procedures

The sparseness restriction itself can be applied directly as a further restriction on the calculation of eigenvalues underlying multivariate techniques such as cluster and factor analysis,[17] and these are referred to as PCA LASSO procedures. They are closely related to ridge-regression procedures.[18]

If the matrix A represents the transformation relating the original data X to the principal components $Y = AX$ then use of a sparseness restriction in this context gives the model

$$Y = Ax: \sum_{j=1}^{p} |a_{ij}| < t$$

where a_{ij} are the relevant coefficient elements of the $Y = Ax$ principal components transformation, subject to the usual PCA constraints. As PCA methods are themselves often an initial attempt to understand or explore the underlying dimensionality or structure of the data, and thus the degree of sparseness itself in large data settings, this may overly restrict an initially explorative approach. However, it does aid in the convergence and interpretability of the resulting PCA, especially where $p > n$. Indeed, the interplay of the LASSO restriction with the $p > n$ geometry described below is an interesting question, as the large p behavior identified in Hall et al[19] and Ahn et al[20] seems to suggest limited usefulness of the PCA approach in settings without the sparseness restriction.

LARS

The LARS algorithm underlying the basic LASSO approach to fitting models in standard $n > p$ linear models is very stable[21] and obtains a fitted model of size m in m steps. This is based on adding new variables in a forward stagewise search approach that uses equiangular bisectors to find the most correlated variables in the dataset, adding them sequentially. As such the standard linear geometry holds as we are only projecting y onto individual or small numbers of x_i vectors,

$$\hat{y} = x_i(x_i'x_i)^{-1}x_i'\ y$$

or related residual vectors, giving the LARS approach stability. In the $p > n$ case the LARS algorithm need only be slightly adjusted to accommodate a sparseness restriction. As shown in Efron et al,[21] the LASSO sparseness restriction does not greatly affect this type of forward stagewise search and projection, as it avoids much of the multidimensional geometric aspects discussed here.

The LARS algorithm is straightforward. Beginning with all $b_j = 0$ determine the x_j most correlated with y. Increase b_j in the direction of the sign of its correlation with y and obtain the residuals $(y - \hat{y})$ stopping when another x_k is found such that the $corr((y - \hat{y}), x_k) = corr((y - \hat{y}), x_j)$. Increase (b_j, b_k) in their joint least squares direction until another predictor

x_m has as much correlation with the new residual vector. Continue in this manner until all useful predictors are in the model. It can be shown that, if when a coefficient hits zero and is removed from the active set of predictors the joint direction is recomputed, this procedure gives the entire path of LASSO solutions, as t is varied from zero to infinity. The LARS algorithm with LASSO restriction is available as a package in R.

Extensions of the LASSO approach have been developed for logistic regression[22] and survival analysis[23] and many other settings. A Bayesian approach to these models can also be developed by assuming a modified Laplace prior distribution[24] to describe the sparseness restriction directly on the parameter space. The Bayesian approach is interesting as it views the data as fixed and probabilities as directly attached to the parameter space. This implies that some of the complexity of the $p > n$ geometry when viewed from this perspective is relevant only in as much as the likelihood function and prior elements are affected. That said, the use of normality in the likelihood links Bayesian and frequentist approaches to least squares geometric considerations. This is further discussed below.

Least squares geometry

To move from the standard $n > p$ geometry of least squares to the restricted $p > n$ setting it is useful to begin with the simple structures of ANOVA. Let $n > p$ and assume that the response and explanatory variables have been centered. In the standard setting of the linear model the least squares estimator is given by $b = (X'X)^{-1}X'y$ and the related predictive value given by $\hat{y} = Xb = X(X'X)^{-1}X'y = Hy$ where H is the p-dimensional orthogonal projection matrix onto the linear span of the column vectors of X, $L(X)$. The associated residual vector $e = [I - H]y$ is the $(n - p)$ dimensional orthogonal projection of y onto the linear span orthogonal to $L(X)$. The squared lengths of these vectors form the basis of standard ANOVA testing. The orthogonality of these spaces implies $Cov(y,e) = 0$ as the cosine of the angle between them is 0.

If $n = p$ we can show that $H = I$ and $\hat{y} = y$. The residual space within the standard geometry here has dimension 0.

In the $p > n$ setting, the loss of the usual residual space leads to different geometric considerations for analysis of the overall linear model. The usual orthogonal decomposition underlying the ANOVA breakdown of sums of squares is not applicable as the usual error degrees of freedom or dimension of the residual space, $(n - p)$ is negative. In this setting, the space within which the data vectors lie, seen as column vectors, is R^n, which is often relatively small and lies within R^p. Indeed it lies within the portion of these spaces, that agrees with the assumed linear structure of the model (assuming the model is correct), and typically is further subjected to a sparseness restriction, which superimposes a simplex structure on the data vectors. This is examined below.

Apart from limiting the region of the parameter space to be considered when estimating β the sparseness restriction in the $p > n$ setting reflects an interesting assumption on the data structure related to estimation. Assuming that all vectors in the analysis are continuous and centered, the sparsity restriction with small t and k can be interpreted as implying that relatively few of the data vectors x_i are expected to be correlated with the response vector y. Geometrically in the sample space this implies that y is approximately orthogonal to most of the x_i vectors, with only a few of them having relatively small cosine angles of departure from y and thus meaningful correlations with y. This restricts most of the x_i to $L^{\perp}(y)$, the space within R^n in which all vectors are orthogonal to y.

Assuming a large number of x_i vectors within $L^{\perp}(y)$, most are linearly dependent with each other. Thus the sparseness restriction, if correct, implies a structure of correlated x_i variables tied together typically through the assumption of a linear model. In a sense there are several sets of potential correlation structures: those x_i that are correlated with each other but not with y in $L^{\perp}(y)$, and those that are correlated with each other and y, thus potentially relevant to a linear model for y. These restrictions on the linear model and data imply that as p becomes very large relative to n, there are patterns and restrictions among the correlations (angular departures) for the variables in the model restricting the relative positions of y and the x_i vectors.

Hall et al[19] examined this situation generally, without assuming sparseness or a linear model, simply studying geometric structures among a normally distributed set of variables as p increased with $p > n$. Subject to further assuming the data follow a simple basic time series structure, they showed that for large p the data vectors in such a restricted linear model setting must cluster at the vertices of an n dimensional simplex. Further, these n directions lie approximately perpendicular to each other in forming the simplex structure.

A simplex can generally be written

$$S = \sum_{j=1}^{n} \beta_j x_j : \beta_j \geq 0 : \sum_{j=1}^{n} \beta_j = 1$$

having n vertices, and formally defines the convex hull of these vertices. A 2-simplex defines a triangle and a 3-simplex

defines a tetrahedron shape. Here it is the vertices that are of interest as the data vectors cluster at these points.

The proof uses the delta method applied to the distances between two arbitrary data vectors, letting p increase asymptotically with fixed n.[19] Assuming we have two standardized independent vectors z_1 and z_2 drawn from a $N_p(0,I)$ multivariate distribution, it follows that for z_1 its squared length can be expressed

$$\| z_1 \|^2 = \sum_{j=1}^{p} z_{1j}^2 \sim \chi_p^2$$

and has expected value p. Direct application of the delta method then implies $\| z_1 \| = \sqrt{p} + O(1)$ as $p \to \infty$.

It also follows for a pair of vectors that $\| z_1 - z_2 \| = \sqrt{2p} + O(1)$ as $p \to \infty$. Thus for large p any two vectors will be a deterministic distance apart and can be seen as each lying near the surface of an expanding sphere centered here at 0. The angle at the origin between any two vectors will also be given by $\pi/2 + O(p^{-1/2})$. This further implies that the vectors will be approximately perpendicular to each other. As all pairwise angles are approximately the same, a deterministic simplex structure will describe the overall shape of the data vectors (see Hall et al).[19]

This can further be interpreted as implying that as $p \to \infty$ any randomness in the data set is essentially generated by random rotations of the n vertices of this deterministic simplex structure. Note that in such a setting, the application of methods such as SVD will seem to work well when looking to classify or discriminate the large set of x_i variables as they already i) cluster at the vertices of the simplex and ii) lie perpendicular to each other.[19] These limitations also affect use of the bootstrap method, which uses random resampling of the data as the basis for significance assessments.

If a linear model structure is to be used to relate an outcome vector y to a set of best fitting x_i, using for example the LASSO approach, it will be among n perpendicular x_i vectors of this core clustering structure, providing a basis for $R^n \subset R^p$, that are not in $L^{\perp}(y)$ and that as a group satisfy the sparseness restriction, that stable fitted linear models may be found. There is of course no guarantee that the achieved fit will be useful. If the sparseness restriction is not appropriate and many x_i are not in $L^{\perp}(y)$ then this restriction will simply limit the number of fitted covariates, selecting those where the linear combination is closest to y and leaving many possible models that give similar goodness of fit values.

To summarize, in the case of $p > n$ with the linear model structure imposed with sparseness restriction, and x_i vectors assumed approximately normally distributed, we can expect the relative positioning of the response vector y in regard to the n dimensional simplex of the x_i to determine the set of potentially useful x_i values in predicting y. Indeed, the best fitting model can be seen as an exercise in choosing those vectors x_i in the simplex whose linear span lies most closely to the y variable for the actual observed rotation of the data simplex. Note use of the sparseness restriction expects that most of the x_i vectors are in $L^{\perp}(y)$ and thus not useful for prediction. Methods that detect this simplex in the case of large p and related summary set of x_i will be useful in guiding model fitting and assessment of significance for linear models in the $p > n$ setting.

Example one: mouse genetic data

A well-known example drawn from the genomics literature is given in Ghazalpour et al.[25] To give a sense of the interaction of eigenvalue structure with $p > n$ restrictions we apply PCA directly here, without LASSO restriction, looking at the results corresponding to selected p and n values for chromosome 11 where we begin with 100 genes and their expression levels for 255 subjects. We begin with $n > p$ and randomly remove subjects as shown in Table 1, carrying out

Table 1 Mouse data principal component analysis for values of n and p

n	n^*	p	e_1	e_2	e_3	e_4	e_5	e_6	e_7	e_8	e_9	e_{10}	e_{11}	e_{12}	Total variation
10	**7**	**100**	**0.380**	**0.239**	**0.145**	**0.127***	0.068	0.040	0.0	0.0	0.0	0.0	0.0	0.0	1.0
20	**15**	**100**	**0.267**	**0.202**	**0.145**	**0.105**	0.081*	0.048	0.040	0.028	0.024	0.022	0.015	0.010	0.984
30	**20**	**100**	**0.253**	**0.175**	**0.128**	**0.094**	0.069	0.063	0.038*	0.036	0.027	0.022	0.019	0.017	0.942
40	**28**	**100**	**0.251**	**0.156**	**0.115**	**0.085**	0.068	0.057	0.037	0.034*	0.030	0.025	0.019	0.018	0.896
50	**37**	**100**	**0.544**	**0.088**	**0.074**	**0.046**	0.038	0.033*	0.024	0.020	0.017	0.014	0.013	0.011	0.921
100	**80**	**100**	**0.365**	**0.107**	**0.079**	**0.064**	0.045	0.043	0.036	0.032	0.030*	0.020	0.017	0.014	0.853
150	121	100	0.353	0.100	0.070	0.067	0.042	0.039	0.037	0.033	0.027	0.022	0.017*	0.016	0.824
200	157	100	0.360	0.098	0.077	0.060	0.040	0.036	0.034	0.030	0.025	0.022	0.017	0.016	0.815
254	200	100	0.338	0.101	0.084	0.056	0.039	0.036	0.033	0.030	0.030	0.025	0.022	0.018	0.800

Notes: Proportion of total variation shown; *denotes 80% of variation explained. The e are the ordered principal components, n is the sample size, and p the number of variables. Results for the first four principal components with $p > n$ are highlighted in bold.

PCA for each set of subjects and variables, moving into the $p > n$ context. The 12 largest eigenvalues for each analysis are reported. Note that when $p > n$ there are only n possible nonzero eigenvalues.

For the initial PCA with $n = 255$ (n^* with missing values) 12 PCA variables account for 90% of the total variation implying potential structure in the set of gene expressions. As is common in most PCA analysis, the first eigenvalue can be seen as an overall mean value. Of greater interest in relation to the results discussed here is the overall structure of the remaining eigenvalues as p/n increases. As this increases, there are fewer relative sources of variation or degrees of freedom and a higher level of total variation explained. As this is real data, and p is large but not in the realm of large asymptotic values, the expected similarity of eigenvalues can be seen as slowly occurring, especially beyond the first or largest PCA, subject to random error. Further restricting our view here to the first four eigenvalues, we see a steady increase in their values and similarity as p/n increases when $p > n$, while the remaining values remain similar or trend to zero.

As noted in the results of Hall et al[19] and Ahn et al[20] above, as $p > n$ the information provided by the eigenvalues is less useful in regards to identifying clusters in the data. The results seem to indicate a growing convergence to a smaller subgroup of eigenvalues.

Correlations and eigenvalues

The relationships among a set of variables x_i subject to a sparsity restriction and $p > n$ can alternatively be examined from the perspective of the sample correlation matrix S itself and its relation to eigenvalues and principal components. This follows to some extent from the consideration of ridge regression discussed above. This is developed in Ahn et al,[20] where using the SVD of S it is shown that for $p \to \infty$ and n fixed, the eigenvalues of S can be viewed as being approximately equal, implying that the data are behaving as if the underlying distribution was spherical in nature. This is in agreement with the results in Hall et al[19] described above. This diffusion of eigenvalues implies that PCA may not be very useful in higher $p > n$ dimensions as a means of dimension reduction outside of the existing simplex structure.[20]

Model misspecification: nonlinearity

As nonlinear models may underlie many genomic data models, especially those related to developmental biology,[4] the immediate use of a linear model may lead to model misspecification issues. This may also be due to inappropriate scaling of specific variables or types of clustering. The linear stagewise orthogonal projection-based approach used to fit the LARS algorithm may not capture the relevant patterns in such data.

The sparseness limitation itself is based on the idea that only a small number of variables are in the end important to the response in question. If that small number are actually structured as a small set of variables sharing an embedded nonlinear model, there may be serious model misspecification. This is actually true of all linear models and this topic has been examined,[26,27] typically in $n > p$ linear model settings. The LARS algorithm uses correlation as the basis of a stagewise fitting approach. Correlation assumes linearity on some attainable scale. If there is a nonlinear pattern or model underlying the data, this approach may not be useful and may mislead.

To more formally address the misspecification issue express the linear model as a function of two sets of variables:

$$y = X\beta + \varepsilon = X_1\beta_1 + X_2\beta_2 + \varepsilon$$

where initially $n > p$. Let us assume that the key significant variables are grouped in the X_1 $(n \times p_1)$ matrix with p_1 variables, the X_2 $(n \times p_2)$ matrix has p_2 additional variables, where $p_1 << p_2$ and $p_1 + p_2 = p$. The error term ε $(n \times 1)$ is assumed to have the distribution $\varepsilon \sim N(0, \sigma^2 I)$.

Now assume the true aspect of interest is a nonlinear model underlying the X_1 set of variables. In this setting we wish to assess to what extent the LASSO or related technique may not detect the set of variables embedded in the nonlinear model. We re-express our initial model as

$$y = F(X_1\beta_1) + X_2\beta_2 + \varepsilon$$

where $F(X_1\beta_1)$ is a nonlinear model for the X_1 subset of variables. Replacing $F(X_1\beta_1)$ with its Taylor expansion about β_{10} to the first order we obtain

$$y = [X_1\beta_{10} + F'(X_1\beta_1)(\beta_1 - \beta_{10})] + X_2\beta_2 + \varepsilon$$

If we were to apply a linear model to this setting we would in essence be using a local linear approximation rather than the true model, giving

$$y = X_1\beta_{10} + X_2\beta_2 + \varepsilon^*$$

where $\varepsilon^* = \varepsilon + F'(X_1\beta_1)(\beta_1 - \beta_{10})$ and in fitting this, we will both potentially miss the nonlinear aspect of the data and apply an approach which has a biased error distribution $\varepsilon^* \sim N(F'(X_1\beta_1)(\beta_1 - \beta_{10}), \sigma^2 I)$. If the two sets of variables are approximately uncorrelated, we may end up with spurious

associations from patterns in the data that are not reflecting the true underlying model.[27]

Note also that if the underlying model is misspecified, the sparseness restriction may not make sense, as it applies a linear scale to the relative importance of the estimated parameter coefficients. In the context of the LARS algorithm, modified for sparseness, the projections of interest are all defined in relation to an underlying restricted but linear relationship among the variables. Nonlinear patterns are often not easily detected by correlation-based approaches and may be missed, even using the forward-fitting strategy of the LASSO. Note that incorporating nonlinear models directly into the $p > n$ setting further creates serious quadratic programming challenges, as typically restrictions are rewritten directly into the model, see, for example, Meier et al.[22]

Possible improvements in the forward stagewise fitting algorithm include examining relationships for possible curvature, graphically and otherwise. If knowledge of the type of nonlinearity present exists, implying local nonlinear models, local curvature corrections[28] at each stage of the forward stepwise procedure may be possible. These will be further discussed elsewhere.

Example two: simulations

Here we examine several simulations in relation to this issue. The goal is to detect a small number of useful explanatory variables, if possible. The C_p criteria is used in the LARS–LASSO based fitting procedures. We generate a set of p correlated explanatory variables x_i according to a multivariate normal distribution $N_n(0, \Sigma)$ with correlations $\rho(x_i, x_j) = 0.5^{|i-j|}$, which gives a set of correlated explanatory variables subject to random noise. Responses of various forms can then be generated, reflecting both highly and less correlated sets of explanatory variables.

An embedded gene correlation cluster was simulated for the $p > n$ case: a set of ten variables x_i were generated according to a multivariate normal distribution $N_n(0, \Sigma)$ where we set the correlations $\rho(x_i, x_j) = 0.5^{|i-j|}$ and $n = 7$. The related response is generated as $y = 5x_1 + 7.2x_2 + x_3 + 1.7x_4 + 1.3x_8$. This response was fit to the entire set of explanatory variables

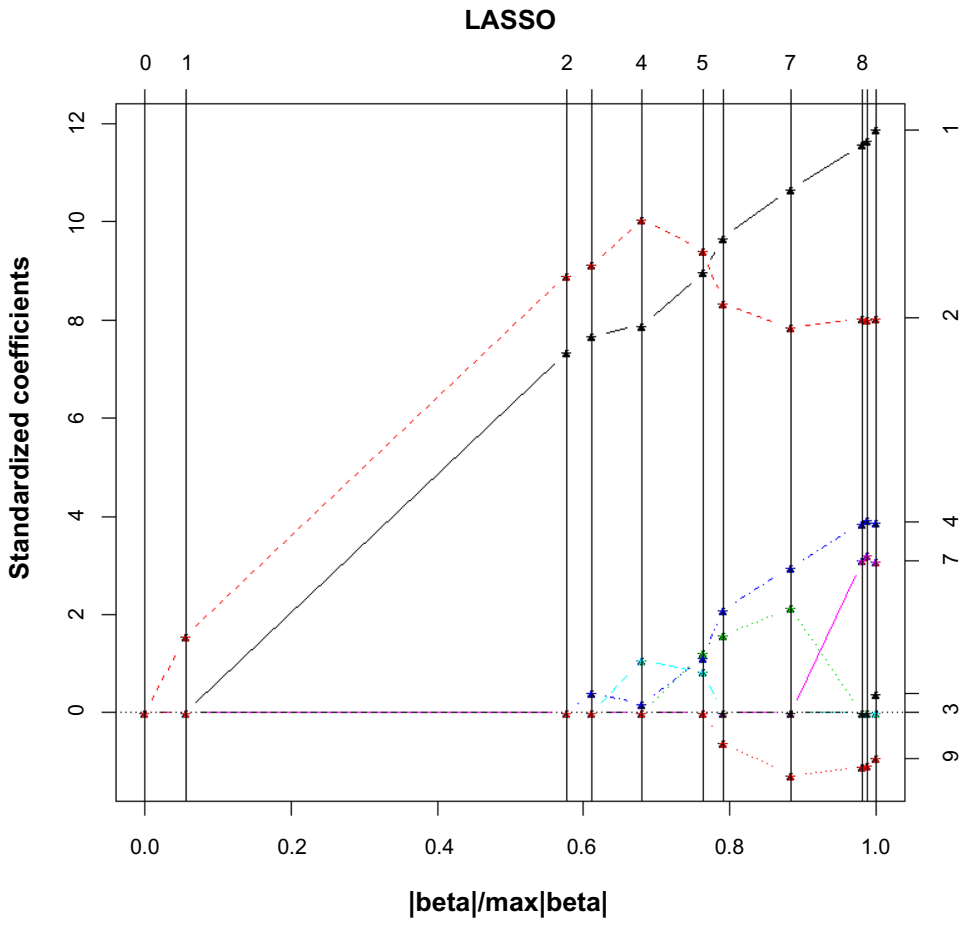

Figure 1 LASSO fit for simulated gene correlated cluster ($n = 7$, $P = 10$).
Note: Selected variables using C_p criteria suggested on right.
Abbreviation: LASSO, least absolute shrinkage and selection operator.

using a linear model and the LARS package in R with the LASSO option. For a single replication, Figure 1 shows x_1, x_2, x_4, x_7 as significant, close to the original model. To assess variation, 30 replications were carried out showing fairly good agreement with the underlying model (Table 2).

An embedded nonlinear function linking the response and only a few explanatory variables within a larger set of mildly correlated variables was also simulated. A set of ten centered, correlated explanatory variables x_i was generated according to a (centered) multivariate normal distribution $N_n(0, \Sigma)$ and we set the correlations $\rho(x_i, x_j) = 0.5^{|i-j|}$ and used a sample size of $n = 7$. The response here was generated as $y = -1.4exp(-4.5x_1 - 6x_2) + 2.2x_3 + 2.0x_4 + 1.3x_5 + 1.2x_6 + 1.1x_7 + 1.0x_8 + 0.8x_9 + 0.6x_{10}$ providing good control over the secondary correlation values. Table 3 is based on 30 replications and shows the selected variables having a weak relationship to the actual underlying model, with the most nonlinear and highly correlated of the variables selected only slightly more often than those variables with little correlation to y. The estimated coefficient values were highly variable throughout and the C_p criteria was often unstable. Figure 2 gives an example of the LARS package output (see also Table 3).

Phase thresholds for convergence

Donoho and Stodden[29] consider a toy example taking $y = X\beta + \varepsilon$ with $\varepsilon \sim N(0,16)$, setting $\beta = 0$ except for k values drawn from a uniform (0,100) distribution, giving a randomly generated set of sparseness values. For a LASSO-based model, a phase threshold plot was generated plotting k/n versus n/p. This displayed a better chance of convergence if k/n was smaller and n/p larger, (ie, as n increases). To get a practical sense of scale we display three approximate values for (k/n, n/p) supporting convergence in Table 1, and examine them.

In Table 4 the relationship between the sparse number of variables to be used (k) and the number of subjects (n) is nonlinear, altering from a factor of 1:10 to 1:5 to approximately 1:2 for the chosen set of values. From the first column, for example, we see that if n is one-tenth of p the original number of variables, then k must be approximately one one-hundredth of p, for convergence to occur. For example, with 2,000 variables and a sample size of 200, the sparsity restriction might be set near 20 to achieve convergence. This ratio varies across the columns in a nonlinear manner. This threshold further does not imply a highly predictive model, it simply reflects convergence, a consideration in the design of high-dimensional studies. The interaction between n, k, and p reflects both the restrictions on the model and the available degrees of freedom.

Bayesian perspective

The Bayesian approach presents a related but different perspective on the model–data combination and is very useful in these situations. The actual values of the data variables are conditioned upon and the probability, through application of Bayes theorem and the assumption of a prior density $p(\beta)$ for β, shifts the probabilities involved to the parameter space. Technically, the posterior density can be written

$$p(\beta|data) = c \cdot p(\beta) \cdot L(\beta|data)$$

where $L(\beta|data)$ is the likelihood function and c is a norming constant. Probability theory is then applied to determine marginal and conditional probability elements of interest. The relatively small number of n does not technically limit the obtaining of marginal posteriors for each of the overall p variables.

In the Bayesian LASSO setting, the high-dimensional $p > n$ geometry discussed above is relevant to the degree that i) the likelihood function is affected, ii) the parameters involved lack identifiability, and iii) the integration necessary to obtain marginal posteriors is affected. In the case of an assumed normal likelihood, least squares considerations are relevant as these affect calculation of posterior modes. The sparseness restriction is useful from the Bayesian perspective as it limits potential values for the parameter β and helps yield identifiability. It can be modeled through use of a specific prior density incorporating the sparseness restriction. Park and Cassela[24] suggest the following hierarchical description of the Bayesian LASSO model:

$$y \mid \mu, X, \beta, \sigma^2 \sim N_n(\mu + X\beta, \sigma^2 I)$$

$$\beta \mid \sigma^2, \tau_1^2, \ldots, \tau_p^2 \sim N_p(0, \sigma^2 D),\ D = diag(\tau_1^2, \ldots, \tau_p^2)$$

$$\sigma^2, \tau_1^2, \ldots, \tau_p^2 \sim \pi(\sigma^2) d\sigma^2 \prod_{j=1}^{p} \frac{\lambda^2}{2} e^{-\lambda^2 \tau_j^2} d\tau_j^2$$

Table 2 Results from LARS software in R-simulated linear model with correlation and LASSO restriction

Selected variables	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9	x_{10}
Proportion of time selected	28/30	29/30	20/30	15/30	15/30	8/30	6/30	14/30	6/30	7/30

Notes: X_i are simulated, centered variables with varying levels of correlation. The proportion of times each is selected into the fitted linear model is reported.
Abbreviations: LARS, least-angle regression; LASSO, least absolute shrinkage and selection operator.

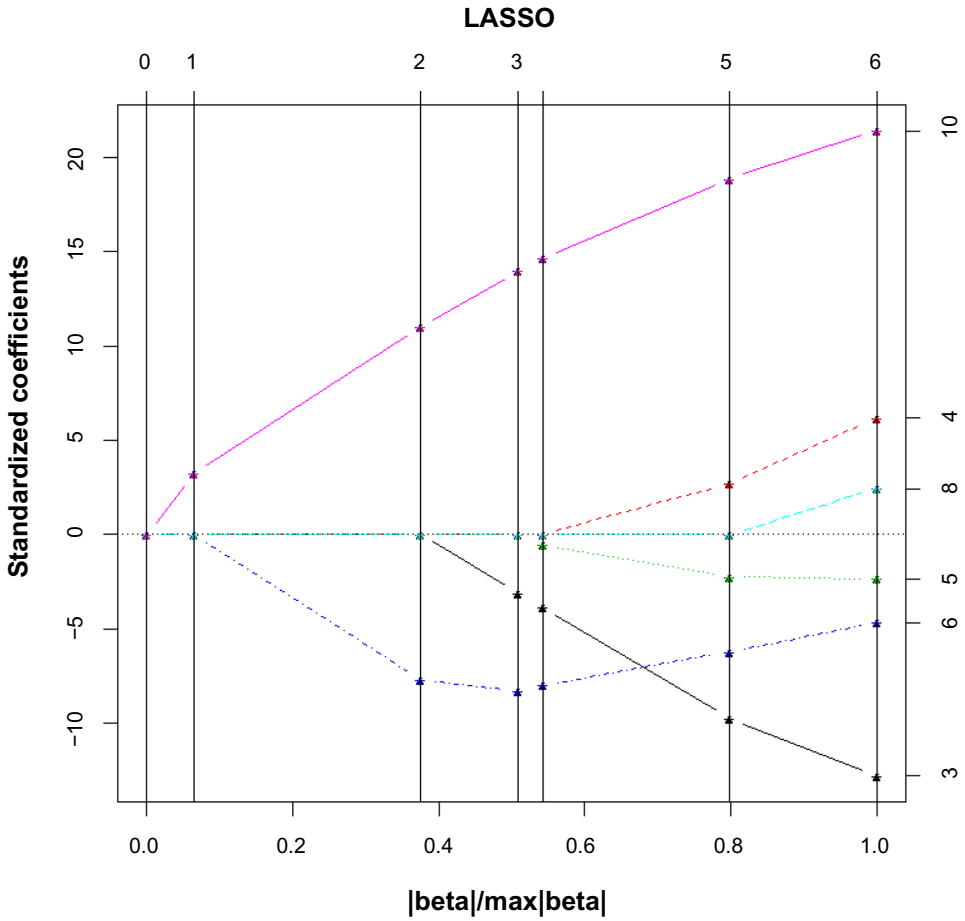

Figure 2 LASSO fit for simulated nonlinear model with gene correlated cluster ($n = 7$, $P = 10$).
Note: Selected variables using C_p criteria suggested on right.
Abbreviation: LASSO, least absolute shrinkage and selection operator.

$$\pi(\sigma^2) = 1/\sigma^2$$

where $\sigma^2, \tau_1^2, \ldots, \tau_p^2 > 0$.

Note that the Bayesian approach here reflects an implicit backward selection methodology. This is not possible in the frequentist setting, but is possible in the Bayesian setting due to the transferring of the probability element from sample space to parameter space and the assumption of a prior to support the many coefficient parameters β_i linked to the many collinear x_i. Thus the dimensional restrictions on the frequentist approach do not directly impede the Bayesian setting, but the assumption of much additional information regarding the β_i requires justification. That said, Bayesian approaches in many sparse settings have led to similar answers to those given by the LASSO and its variants. Ridge regression can be viewed similarly. Various Bayesian versions of LASSO modifications are given by Park and Casella.[24]

In the case of normal error the likelihood aspect justifies the least squares criteria and the prior provides the sparseness restriction, a very similar problem to the frequentist setting, though viewed as a function of the parameters. Thus the Bayesian approach, when comparable as there may be additional prior parameters to consider, often yields similar answers to the frequentist models, with the additional benefit of more stability in calculations due to assumed additional structure in the model to be analyzed.

Table 3 Results from LARS software in R-simulated nonlinear model with correlation and LASSO restriction

Selected variables	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9	x_{10}
Proportion of time selected	20/30	19/30	11/30	14/30	17/30	14/30	14/30	17/30	15/30	13/30

Abbreviations: LARS, least-angle regression; LASSO, least absolute shrinkage and selection operator.

Table 4 Example ranges of sample size, sparsity, and number of variables in relation to convergence

Values	(0.1, 0.1)	(0.2, 0.5)	(0.6, 0.9)
k	$0.01p$	$0.1p$	$0.54p$
n	$0.1p$	$0.5p$	$0.9p$
(k,n) (with p=100)	(1, 10)	(10, 50)	(54, 90)
(k,n) (with p=1,000)	(10, 100)	(100, 500)	(540, 900)

Notes: Typical values at which the model attains convergence are shown. n is sample size, p the overall number of variables and k the level of sparseness.

This similarity to frequentist approaches also implies that the misspecification issues discussed above will directly impact the stability of the Bayesian posterior mode as it reflects the least squares linear model-based sensitivity of the log-likelihood. Nonlinearity and model misspecification affect all statistical models.

The phase threshold issue from the Bayesian perspective addresses identifiability: an inability to clearly detect the mode of the joint posterior as a function of each potential coefficient β_i. The Bayesian approach provides more structure for assessing the $p > n$ setting with sparseness directly incorporated into a posterior density. The key aspects of the Bayesian analysis here are maximization of the joint posterior (identifiability) and integration of this posterior to achieve the required marginal posteriors, typically accomplished with Laplace approximation or an empirical Markov-chain Monte Carlo (MCMC) approach.[30] As with all Bayesian constructs, the issue of identifiability should be examined from a robustness perspective, altering the choice of prior and examining the stability of the threshold. If there are changes in the threshold behavior, relating these to properties of the prior itself may be useful. Changes from a baseline prior can be examined using the Kullback-Leibler distance of a posterior density $p(\Theta|x)$ from a prior $p(\Theta)$. This is given by:

$$KL(p(\Theta \mid x), p(\Theta)) = \int p(\Theta \mid x) \log \frac{p(\Theta \mid x)}{p(\Theta)} d\Theta$$

This is always non-negative and is zero when $p(\Theta|x)$ and $p(\Theta)$ are equal. It is nonsymmetric but can be adjusted. The change in KL due to altering the form of the prior $p(\Theta)$ itself or altering the sparseness restriction can be associated to changes in phase threshold behavior. The use of this strategy is considered by Barber.[31]

Bayesian results in these settings should be interpreted carefully. While the prior allows for model structure and analysis of all p parameters, there are only n sources of information in the problem outside of prior belief.

Discussion

The problems reviewed and discussed here focus on a fundamental set of difficulties arising in the fitting of linear models in the $p > n$ setting: restricted dimension, correlation, misspecification, identifiability, and the extension of the linear model using various sparseness restrictions. These apply across a wide variety of research areas where large databases are becoming more common. The methods reviewed here are surprisingly hopeful given the nature of the $p > n$ problem. Assuming there is indeed a true underlying model, the standard linear model structure can still be made to apply to many large datasets, assuming linearity is appropriate. Indeed, the underlying simplex-based nature provides hope that a core set of informative x_i variables can be detected and utilized to guide modeling and inference. As experience is gained in applying such large p, small n models, finding perspectives allowing researchers to move from an ANOVA-based intuition to perhaps a large-sample simplex-based reference model and resulting conditional sample space interpretation will grow in importance.

The Bayesian setting provides useful flexibility through the choice of prior and using the prior to incorporate restrictions. Many of the geometric concepts discussed here are not directly relevant to the Bayesian setting as the data is conditioned upon, though if the likelihood reflects an underlying normal distribution, the least squares aspect and correlation issues discussed above apply as they are relevant to the obtaining of identifiable posterior modal values. That said, Bayesian methods reflect backward fitting methods, incorporating the initial entire model, but requiring prior assessment on many model parameters. Related Bayesian phase threshold levels will reflect sparseness restrictions and the type of prior used to accommodate such restrictions.

From the perspective of the sample space, the restriction $p > n$ here can also be interpreted as inducing a type of large p versus n conditioning effect where conditional inference is seen in its broadest terms, assessing the properties of estimators for a model within a restricted portion of the original sample space, here given by the set of values defined by random rotations of the data simplex defining an (asymptotic) conditional sample space. This is essentially driven by the high degree of collinearity that restricts the model space. Sparseness, if correctly assumed, also restricts the sample space. These issues affect and motivate use of conditional procedures such as the bootstrap in assessing significance.

If there are potential nonlinearities in the underlying functional $y \sim x$ relationship implying model misspecification

as defined above, the methods discussed here need careful assessment and may not be helpful. Practical challenges also lie in the need to develop visual and analytic tools that can detect and allow researchers to observe the presence of high-dimensional correlation, spherical, and asymptotic simplex-related structures. A motivation for further work is that these geometries may become more common as large data science in medical research becomes practical and accessible.

Disclosure

The author has no conflict of interest, financial or otherwise, in regard to this article.

References

1. Lambert CG, Black LJ. Learning from our GWAS mistakes: from experimental design to scientific method. *Biostatistics*. 2012;13(2):195–203.
2. Tanizawa H, Iwasaki O, Tanaka A, et al. Mapping of long-range associations throughout the fission yeast genome reveals genome organization linked to transcriptional regulation. *Nucleic Acids Res*. 2010;38(22):8164–8177.
3. Christensen R. *Plane Answers to Complex Questions: The Theory of Linear Models*. New York: Springer-Verlag; 1987.
4. Bassingthwaighte JB, Liebovitch LS, West BJ. *Fractal Physiology*. New York: Oxford University Press; 1994.
5. Tibshirani R. Regression shrinkage and selection via the lasso. *J Royal Statist Soc B*. 1996;58(1):267–288.
6. James GM, Radchenko P, Lv J. DASSO: connections between the Dantzig selector and lasso. *J Royal Statist Soc B*. 2009;71(1):127–142.
7. Johnstone IM, Titterington DM. Statistical challenges of high-dimensional data. *Philos Trans A Math Phys Eng Sci*. 2009;367(1906):4237–4253.
8. Bickel PJ, Brown JB, Huang H, Li Q. An overview of recent developments in genomics and associated statistical methods. *Philos Trans A Math Phys Eng Sci*. 2009;367(1906):4313–4337.
9. Berger B, Peng J, Singh M. Computational solutions for omics data. *Nat Rev Genet*. 2013;14(5):333–346.
10. Donoho D, Tanner J. Observed universality of phase transitions in high-dimensional geometry, with implications for modern data analysis and signal processing. *Philos Trans A Math Phys Eng Sci*. 2009;367(1906):4273–4293.
11. Malone JH, Oliver B. Microarrays: deep sequencing and the true measure of the transcriptome. *BMC Biol*. 2011;9:34.
12. Vollmers C, Schmitz RJ, Nathanson J, Yeo G, Ecker JR, Panda S. Circadian oscillations of protein-coding and regulatory RNAs in a highly dynamic mammalian liver epigenome. *Cell Metab*. 2012;16(6):833–845.
13. van Noort V, Snel B, Huynen MA. The yeast coexpression network has a small-world, scale-free architecture and can be explained by a simple model. *EMBO Rep*. 2004;5(3):280–284.
14. Bickel P, JohnstoneI, Yu B. Discovery in Complex or Massive Datasets: Common Statistical Themes. Presented at National Science Foundation Workshop; October 16–17, 2007; Washington, DC.
15. Carlsson G. Topology and data. *Bull New Ser Am Math Soc*. 2009;46(2):255–308.
16. Hoerl AE, Kennard RW. Ridge regression: biased estimation for non-orthogonal problems. *Technometrics*. 1970;12(1):55–67.
17. Jolliffe IT, Trendafilov NT, Uddin M. A modified principal component technique based on the LASSO. *J Comput Graph Stat*. 2003;12(3):531–547.
18. Zou H, Hastie T, Tibshirani R. Sparse principal component analysis. *J Comput Graph Stat*. 2006;15(2):265–286.
19. Hall P, Marron JS, Neeman A. Geometric representation of high dimension, low sample size. *J R Stat Soc Series B Stat Methodol*. 2005;67(3):427–444.
20. Ahn J, Marron JS, Muller KM, Chi Y-Y. The high-dimension, low sample-size geometric representation holds under mild conditions. *Biometrika*. 2007;94(3):760–766.
21. Efron B, Hastie T, Johnstone I, Tibshirani R. Least angle regression. *Ann Stat*. 2004;32(2):407–451.
22. Meier L, Van De Geer S, Bühlmann P. The group lasso for logistic regression. *J R Stat Soc Series B Stat Methodol*. 2008;70(1):53–71.
23. Tibshirani R. The lasso method for variable selection in the Cox model. *Stat Med*. 1997;16(4):385–395.
24. Park T, Casella G. The Bayesian lasso. *J Am Stat Assoc*. 2008;103(482):681–686.
25. Ghazalpour A, Doss S, Zhang B, et al. Integrating genetic and network analysis to characterize genes related to mouse weight. *PLoS Genet*. 2006;2(8):e130.
26. White H. Consequences and detection of misspecified nonlinear regression models. *J Am Stat Assoc*. 1981;76(374):419–433.
27. Sarkar N. Comparisons among some estimators in misspecified linear models with multicollinearity. *Ann Inst Stat Math*. 1989;41(4):717–724.
28. Seber GAF, Wild CJ. *Nonlinear Regression*. New York: John Wiley; 1989.
29. Donoho DL, Stodden V. *Breakdown Point of Model Selection when the Number of Variables Exceeds the Number of Observations: Proceedings of the International Joint Conference on Neural Networks, Vancouver, Canada, July 16–21, 2006*. New York: IEEE; 2006.
30. Gelman A, Carlin JB, Stern HS, Rubin DB. *Bayesian Data Analysis*. 2nd ed. Boca Raton: Chapman and Hall; 2003.
31. Barber D. Identifying graph clusters using variational inference and links to covariance parameterization. *Philos Trans A Math Phys Eng Sci*. 2009;367(1906):4407–4426.