Practical Sparse Modeling: an Overview and Two Examples from Genetics

Saharon Rosset

March 15, 2013

1 Introduction: Sparse modeling roadmap

The sparse modeling assumption states that the true relationship of the response Y to the covariates $x_1, ..., x_p$ is a function of a small number of the covariates, i.e.,

$$E(Y|\mathbf{x}) = f(x_{j_1}, ..., x_{j_q})$$

with $q << p$. It is typically assumed further that the relationship is linear:

$$E(Y|\mathbf{x}) = \sum_{l=1}^{q} \beta_l x_{j_l},$$

though this assumption can sometime be relaxed.

This notion of sparsity is relevant and appropriate in many real-life domains, including signal processing [**???**] and others. Herein, we concentrate on applications of sparsity in genetics, in particular two major classes of problems described above, where sparsity is regularly assumed: genome wide association studies (GWASs) and gene microarray data analysis. As discussed earlier, in GWAS, the phenotype is measured for a large panel of individuals (typically several thousands) and a large number of single nucleotide polymorphisms (SNPs, typically hundreds of thousands) throughout the genome are genotyped in all these participants. The goal is to identify SNPs that are statistically associated with the phenotype, and ultimately to build statistical models to capture the effect of genetics on the phenotype. It is usually assumed (and invariably confirmed by GWAS results) that only a small number of SNPs are associated with any specific phenotype. Thus, the model based on GWAS which describes the dependence of the phenotype on SNP genotypes is expected to be sparse, usually extremely sparse. We further discuss this example below.

A second class of relevant problem is gene microarray modeling. Before the advent of GWAS, the major technology geared towards finding connections between genetic and phenotypic information is through measurement of gene expression levels in different individuals or different tissues. In this mode, the quantities being measured are the expression or activity level of actual proteins.

Proteins are encoded by genes, which are fragments of the genome. Hence gene expression experiments can be thought of as measuring the association between genomic regions and phenotypes, except that this is done through the actual biological mechanisms as expressed in proteins, rather than by direct inspection of genetic sequences as in GWAS. Not surprisingly, gene expression analysis also typically assumes that only a few genes are actually directly related to the phenotype of interest. Thus, this is also a sparse modeling situation, though the statistical setup has some major differences from the GWAS, as we discuss below.

Fundamentally, sparse recovery approaches are seeking two major goals:

- Correct recovery of the identities of the covariates that actually participate in the function f.
- Accurate estimation of the model f, both in terms of parameter estimates and prediction accuracy.

Traditional methods for sparse recovery can roughly be divided into two categories:

- Methods based on exact or approximate combinatorial enumeration over the space of possible "sparse" models and selection from this set based on model performance.
- Methods based on univariate modeling of the relationship between single covariates and the response y, then selection of a small set of covariates showing strong association with y for inclusion in the sparse model.

In feature selection nomenclature, the first approach is broadly termed the "wrapper" approach, while the second one is termed the "filter" approach [Guyon and Elisseeff, 2003].

Beyond wrappers and filters, the last few years have seen an outburst of interest in the use of convex optimization methods based on ℓ_1 norms for sparse recovery, including compressed sensing, Lasso, Dantzig selector and other methods [Tibshirani, 1996, Donoho, 2006, Candes and Tao, 2007]. These methods can provably succeed in situations where both wrapper and filter methods are unlikely to result in successful recovery. We omit a detailed technical review of this class of methods here, but for now we concentrate on a qualitative description of these approaches and their properties. ℓ_1-type methods all share some version of the same basic (and quite intuitive) conditions for success in sparse recovery:

- Sufficient sparsity — typically the number of covariates that participate in the solution is required to be $O(n)$, where n is the sample size, since the well-known results from the compressed sensing literature state that an accurate recovery of a sparse signal is possible when the number of samples is $O(k \log m)$, where k is the number of nonzeros and m is the dimensionality of a sparse vector.

- Low correlation among the "non-zero" covariates and between them and the other covariates. Different versions of this condition are termed "incoherence" [Candes and Plan, 2009], "irrepresentability" [Meinshausen and Yu, 2009] etc.

It is clearly true that these two dimensions – level of sparsity and degree of correlation between covariates – are the critical determinants of possibility or impossibility of recovery of sparse models from data, and the identity of approaches that are likely to be successful. Here we attempt to qualitatively divide the space of sparsity-correlation combinations into three regions:

- Situations that can be addressed by simple feature selection wrapper/filter approaches
- Situations that are appropriate for ℓ_1-based sparse recovery approaches
- Situations where sparse recovery is unlikely to be possible

We consider three qualitative level of sparsity: *very sparse* where the number of important variables is $O(1)$; *sparse*, where the number is $O(n)$, n being the number of samples, and *not sparse* otherwise. We also consider three qualitative levels of correlation between "non-zero" covariates and other covariates: *Uncorrlated/orthogonal*; *low correlation*, as defined in the ℓ_1 sparse recovery literature; and *high correlation*. We can characterize our genetic motivating applications above in terms of these dimensions: in the GWAS example we typically assume that the model is *very sparse* and the "non-zero" covariates (SNPs) almost uncorrelated between them and with almost all "zero" covariates; while in the gene expression modeling example we typically assume that large groups of covariates (genes) may have high correlation within them, but low correlations between groups, and that we are in the *sparse* situation [Leung and Cavalieri, 2003].

Considering which sparse recovery approaches fit which situation, some obvious observations are: (a) That in the very sparse situation combinatorial wrapper approaches are often likely to do well — in particular if we assume q is very small and $\binom{p}{q}$ is a manageable enumeration; (b) That in the uncorrelated/orthogonal situation, marginal univariate (filter) approaches are expected to do well in identifying important covariates. In fact, it can easily be shown that Lasso is equivalent to univariate regression when the covariates are orthogonal from a variable selection perspective [Tibshirani, 1996]. In that respect, the ℓ_1 type methods can be thought of as reducing to univariate modeling when there is no correlation.

Our conclusions from this qualitative discussion are summarized in Figure 1.

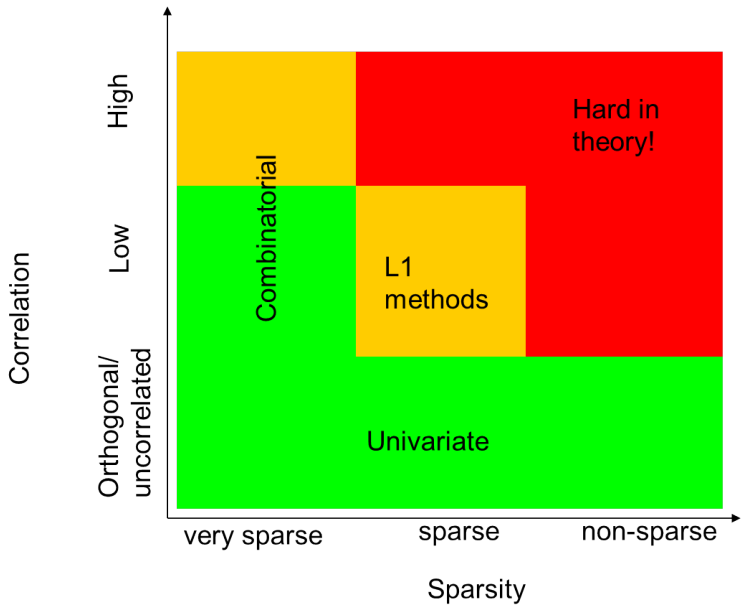

Figure 1: A schematic view of sparse modeling scenarios

2 Example 1: Genome-wide Association Studies (GWAS)

As previously discussed, in GWAS we can assume we have number of features (SNPs) p in the hundreds of thousands, number of observations (individuals) n in the thousands, and we also have the following statistical characteristics:

- Only a very small number of SNPs are associated with the phenotype y, typically ten or less. Thus, we are clearly in the *very sparse* scenario.
- The vast majority of SNP pairs are uncorrelated. This is due to the recombination process driving the SNP-SNP correlation in our genome. SNPs that are far from each other on the genome, and certainly SNPs on different chromosomes are in *linkage equilibrium*, meaning they are completely uncorrelated, due to being separated by many recombination events in the genetic history of the sample we are considering. Hence we can assume that each SNP is correlated only with a tiny fraction of all other SNPs, and typically all truly associated SNPs are uncorrelated between them. However, we should keep in mind that every SNP typically has some neighboring SNPs that are in high correlation with it.

The standard methodology in analyzing GWAS data is to perform p univariate tests of association between each SNP and the response [WTCCC, 2007,

Manolio, 2010]. Appropriate univariate models are chosen to accommodate the specific problem setting, like linear regression, logistic regression or chi-square tests of association . Each model is evaluated using the p-value for the effect being tested (SNP coefficient in the linear regression, chi-square statistic, etc.). The p-values from the univariate models or tests are ranked, and after appropriate multiple comparison corrections, as warranted, the top results are declared significant, therefore indicating likely true association. It should be noted that because of the correlation structure it is typical that each significant finding would actually be expressed as multiple neighboring SNPs that are significantly associated, and the typical policy is to select the "most associated" SNP in the region, consistent with the view that each region is likely to have only one true association.

The significant results from GWAS are usually used as motivation and guidance for follow-up studies, aimed at revalidating the findings and further examining the potential biological/genetic mechanisms underlying the discovered associations [Manolio, 2010].

When examining this process as a sparse modeling exercise, several obvious questions arise:

1. Is the approach of performing univariate tests instead of joint modeling justified? What can we gain from performing multivariate analysis?

2. Is the ranking and selection of SNPs based on p-values, rather than other commonly used model evaluation criteria (like likelihood) justified? Can a different approach give better results?

3. How should the methodology of selection be related to the nature of the follow-up studies to be performed?

We address here the first two questions, starting from the second: is selection by p-value justified? To frame the discussion theoretically, let's assume a standard univariate linear regression formulation, where:

$$y = \beta^T x + \epsilon \ , \ \epsilon \sim N(0, \sigma^2),$$

and assume for simplicity σ^2 is known and there is only one truly associated SNP. In other words, we assume all $\beta_j = 0$ except one. The coordinates x_j can be highly correlated and the dimensionality of the problem is not too high (i.e., assume we are concentrating in the genomic region around the true association). Our primary goal is to identify the SNP j_0 with the true association.

The obvious statistical approach for dealing with this situation is maximum likelihood (ML) estimation. Because of the normal noise model assumed, and the assumption that only one coefficient is non-zero, it is easy to see that ML estimation in this case amounts to finding the univariate model with the minimal residual sum of square (RSS):

$$\hat{j_0} = \arg\min_{j,\beta_j} \sum_{i=1}^{n} (y_i - \beta_j x_{ij})^2.$$

How does this compare to selecting $\hat{j_0}$ as the SNP attaining the minimal p-value in performing a z-test on the coefficient of the SNP (or, equivalently, a test for the univariate model against the null model)? As it turns out, the two are completely equivalent in this case, in the sense that ranking of the SNPs according to RSS is identical to their ranking according to z-test p-values. To see this, denote for SNP j the sum of squares of $x_{\cdot j}$ by $Sxx_j = \sum_i x_{ij}^2 - n\bar{x}_{\cdot j}$, and denote $Sxy_j = \sum_i x_{ij}y_i - n\bar{x}_{\cdot j}\bar{y}$ and $Syy = \sum_i y_i^2 - n\bar{y}$.

Then the coefficient of the regression of y on x_j is $b_j = Sxy_j/Sxx_j$ and the p-value of the z-test is

$$p_j = 2 * \Phi\left(\frac{-|b_j|}{\sqrt{\sigma^2/Sxx_j}}\right),$$

where $\Phi(\cdot)$ is the cumulative standard normal distribution function. Note that this expression is a monotone function of:

$$\frac{|b_j|}{\sqrt{\sigma^2/Sxx_j}} \propto \frac{Sxy_j}{\sqrt{Sxx_j}}.$$

From the standard theory of linear regression it follows that the best RSS for the univariate model with SNP j is:

$$RSS(\hat{\beta}_j) = Syy - \frac{nSxy_j^2}{Sxx_j} = Syy - n\left(\frac{Sxy_j}{\sqrt{Sxx_j}}\right)^2.$$

which is also clearly a monotone function of $Sxy_j/\sqrt{Sxx_j}$. Thus, selecting the lowest p value or using ML are mathematically equivalent.

This perfect equivalence breaks down once we move away from the simplest linear regression setting. For example, consider a logistic regression setup, where GWAS typically uses the Wald statistic for p value calculation [McCullagh and Nelder, 1989]. This is based on a quadratic approximation of the likelihood around the estimate. Selecting the SNP that gives the lowest p value is no longer equivalent to selecting the one that gives the best likelihood in a univariate model. We would intuitively expect that the maximum likelihood approach would be slightly better than the p-value based approach. To demonstrate that this is indeed the case, we present a simplistic simulation. Assume we have two SNPs, with $x_{i1} \sim N(0,1)$ and $x_{i2} = x_{i1} + r \cdot N(0,1)$, and $P(y_i = 1|x_i) = \exp(x_{i1})/(1+\exp(x_{i1}))$. Thus, SNP 1 is the true association, but the two SNPs are correlated with

$$\mathrm{cor}(x_{\cdot 1}, x_{\cdot 2}) = 1/\sqrt{1+r^2}$$

. We examine the rate of success of both approaches in identifying SNP 1 as the more highly associated, as a function of r. Results are given in Fig. 2. As expected, the success rate of both approaches if similar, but the approach based on likelihood is slightly better for all values of r.

To summarize our discussion of the use of p-values for model selection: this criterion is generally similar to using maximum likelihood, but could be inferior,

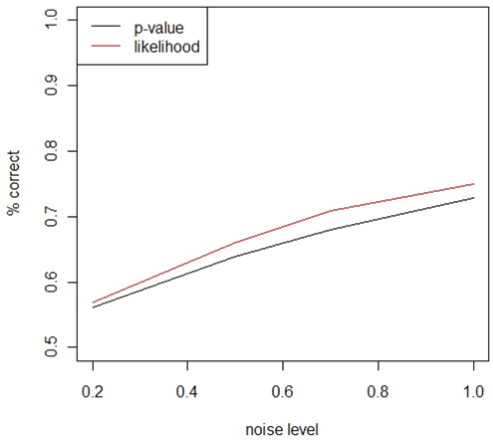

Figure 2: Percentage of cases the correct true association is identified by maximum likelihood (red) and Wald test p value (black) in a logistic regression setup, see text for details. The maximum likelihood criterion is slightly superior for all levels of correlation.

depending on the approximations used for calculating p-value, which may break down the equivalence.

The other question we wish to address pertains to the use of univariate models, as opposed to multivariate sparse modeling approaches like Lasso [Tibshirani, 1996]. Consider again a genomic region with correlated SNPs, where at most one SNP is associated, and we would like to compare the use of univariate models to find the associated SNP to the use of Lasso or similar methods. The Lasso formulation:

$$\hat{\beta}(\lambda) = \arg\min_{\beta} \sum_{i} (y_i - \beta^T \mathbf{x}_i)^2 + \lambda \|\beta\|_1, \tag{1}$$

includes a regularization parameter λ. At $\lambda = \infty$ the solution is all zeros, while as $\lambda \to 0$ the solution converges to the least squares solution. Specifically, at large enough λ the solution would contain only a single non-zero coefficient. It is easy to verify that this first variable is the maximizer of the empirical covariance, i.e., Sxy_j [Efron et al., 2004]. In other words, if all $\mathbf{x}_j$ are pre-standardized to have the same Sxx_j, then the univariate and Lasso approaches amount to selecting the same first covariate.

For lower values of λ, a reasonable approach using Lasso and assuming a single association is to select the largest absolute coefficient in $\hat{\beta}(\lambda)$ as $\hat{j_0}$. It is now relevant to enquire if this approach could prove superior to the univariate approach in identifying the correct association. To test this question, we performed a simulation study, this time with three covariates. We have

$x_{i1} \sim N(0,1)$, $y_i = 2 + 5x_{i1} + \epsilon_i$, $\epsilon_i \sim N(0,1)$ is the true association signal, and we define two correlated variables as: $x_{i2} = x_{i1} + \delta_{i2}$, $\delta_{i2} \sim N(0, 0.01)$ and similarly $x_{i3} = x_{i2} + \delta_{i3}$, $\delta_{i3} \sim N(0, 0.01)$. We examine the success of four approaches in detecting the first variable as the true association:

- The univariate regression approach in GWAS.
- Regular least squares, where the maximal coefficient is chosen.
- Lasso with standardized explanatory variables for various regularization levels, where the maximal Lasso coefficient is chosen.
- Lasso with non-standardized explanatory variables.

In Fig. 3 we present our results. The x-axis is the Lasso constraint (in its Lagrange-equivalent constrained form), and the y-axis is the percentage of correct identification of the first explanatory variable as the best association. The univariate approach and the standardized Lasso with small constraint (high penalty) are much better than the other two approaches. On our simulation data, there were a few examples where the standardized Lasso added the wrong variable first but then for higher constraint values the order of absolute coefficients reversed and the first variable was correctly chosen. Hence, there is a range of constraint around 0.4 where the Lasso does very slightly better than univariate. The generality of this phenomenon requires further research. Not surprisingly, the least squares approach and the non-standardized Lasso are far inferior in their model selection performance.

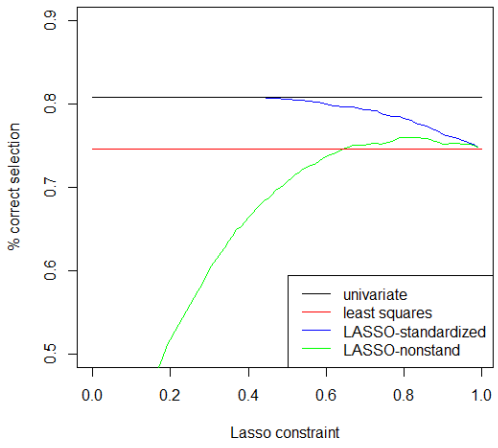

Figure 3: Success of different variable selection schemes on a simulated GWAS example. See text for details.

To summarize our analysis of univariate GWAS tests, we have shown that the common practice of using p-values for selection is generally similar to using maximum likelihood, although the latter may be slightly superior in some cases. We have also argued that because the problem is *very sparse* and because of its correlation structure, the univariate approach is appropriate and we are unlikely to gain from using multivariate approaches like Lasso for identifying the associated SNPs.

We have not discussed here further our third question, of how the selection should be affected by follow-up study design. As a simple example, if planned follow-up work is search for the biological mechanisms underlying statistical associations, then it may make sense to bias our modeling towards identification of associations in biologically plausible genomic regions (such as inside genes). This can be accomplished by using Bayesian priors or other intuitive weighting schemes [Cantor et al., 2010]. Further discussion of this aspect is outside the scope of our chapter.

3 Example 2: Gene Microarray Data Analysis

Microarray technology actually precedes the emergence of the GWAS approach [Schena et al., 1995]. The analyzed data comprises expression levels of genes – how much of each protein (equivalently, gene) is expressed in each sample. Different samples can be different individuals, different tissues, or even the same tissue under different environmental conditions. The most prevalent goal in analyzing gene expression data is to identify which genes are associated with the response of interest, which can be disease status as in GWAS (in which case, the same case-control design as in GWAS can be used), a measure of the environmental conditions being applied (such as concentration of sugar or temperature) etc. The number of samples (n) is usually in the tens or low hundreds, and the number of genes (p) is usually in the thousands or tens of thousands, hence we are in the $p >> n$ situation of "wide" data.

Like in GWAS, it is usually assumed that the true association relation between gene expression and the response is *sparse* or *very sparse*, in the sense that the true dependence (e.g. conditional expectation) of the response on the gene expression can be almost-fully modeled using few "true" genes. However, the correlation structure among genes' expression is much more complex than that among SNPs, since genes are organized in pathways and networks [Davidson and Levin, 2005], which interact and co-regulate in complex ways. It is usually not assumed that these interactions and the resulting correlation structure are known, hence we can consider this an example of a sparse modeling scenario with arbitrary complex correlations between the explanatory variables, in particular we cannot assume that the few true genes are uncorrelated as in the GWAS case. Hence univariate approaches are unlikely to properly address this situation, and although they had originally been used for gene expression analysis, in particular for identification of differentially expressed genes [Leung and Cavalieri, 2003], they have been surpassed in this task too by mutivariate approaches,

which have been demonstrated to be much more effective [Meinshausen, 2007, Wang et al., 2011]. It should be noted that combinatorial variable selection approaches are unlikely to be relevant, since with thousands of genes, even a very sparse model with several dozen genes is too prohibitive to enumerate.

Another important difference between GWAS and gene expression analysis is that in the latter case we are often interested in building an actual prediction model to describe the relation between gene expression and the response, rather than just identifying the associated genes for further study [Leung and Cavalieri, 2003]. This also affects the choice of models.

Since we are seeking a sparse prediction model in high dimension with limited samples, Lasso-type methods are a natural approach to consider. The standard Lasso has some major shortcomings in this situation:

- With $p >> n$, Lasso regularized models are limited to choosing at most n genes in the model [Efron et al., 2004]. This can become a problem in gene expression modeling with very few samples. Furthermore, Lasso typically selects one "representative" from each group of highly correlated explanatory variables (in gene expression, this could represent genes in a specific pathway). This is not necessarily desirable, as there could be multiple independent associations in the same path, or separating the true association from other genes that are highly correlated with it can be very difficult. Hence a selection of a single gene can be arbitrary or non-representative.

- If we are interested in prediction, then the shrinkage Lasso performs on its selected variables is likely to lead to sub-optimal predictive model [Meinshausen, 2007].

Several Lasso extensions have used gene expression as a motivating application and these specific problems as motivation for their proposed algorithmic extensions:

1. Elastic net [Zou and Hastie, 2005], which adds a second penalty to the Lasso formulation in (1), thus allowing solution with more than n distinct features, and similar coefficients for highly correlated features.

2. Adaptive Lasso [Zou, 2006], which adds weighting to the Lasso penalty of each feature, using the least square coefficients. This leads to favorable theoretical properties and has also shown improved empirical performance.

3. Relaxed Lasso [Meinshausen, 2007], which uses Lasso for variable selection, but then fits a less regularized model in these variables only, thus partially avoiding the excessive shrinkage behavior.

4. VISA [Radchenko et al., 2008], which implements a more involved version of the same idea, of performing less shrinkage on the "good" variables Lasso identifies than warranted by the Lasso solution.

5. Random Lasso [Wang et al., 2011]

We now describe Random Lasso in more detail, and demonstrate relative performance of these algorithms on simulated and real gene expression data, following Wang et al. [2011].

3.1 Random Lasso

When many highly correlated features are present, we want to consider the portion of them that is useful for our predictive modeling purposes. Lasso-type regularization would tend to pick one of them semi-arbitrarily, which can be considered a model-instability issue.

The statistics literature offers some recipes for dealing with instability, most popular among them Breiman's proposals of Bagging and Random Forest [Breiman, 2001]. The basic idea is to generate a variety of slightly modified versions of our data or modified versions of the model fitting algorithm, generating a variety of different prediction models which "approximately" fit our data. Then averaging these models has a stabilizing effect, as we hope that models that are not chosen for our original data would occasionally get chosen when the data is changed. Empirically, this tends to lead to much more accurate prediction models in many cases [Breiman, 2001].

As Breiman noted, linear modeling approaches are not subject to improvement from Bagging, but since Lasso is not a linear approach in this sense (because of the regularization), it can be subjected to Bagging-type modifications.

The first part of the Random Lasso (RLasso) is basically applying two-way bootstrap-aggregating, which can be considered a hybrid of Bagging and Random Forrest. The second part repeats the same exercise, but with variables weighted according to their importance in the first part, to accomplish stronger variable selection.

1. Iterate B_1 times:
   (a) Bootstrap-sample the data and sub-sample the features (two-dimensional sampling)
   (b) Fit a Lasso model to the current sample
2. Average the coefficients of all resulting models
3. Generate an important measure for each variable, typically proportional to its average coefficient
4. Perform a second iteration, this time B_2 times:
   (a) Bootstrap-sample the data and sub-sample the features according to their importance measure
   (b) Fit a Lasso model to the current sample
5. The final model is the average of the B_2 models from the second stage

Detailed discussion of the motivation behind the exact formulation of the algorithm is beyond the scope of this chapter, but we show here a comparison of the various Lasso extensions on simulation and real gene expression data.

In the simulation scenario there are $p = 40$ variables. The first 10 coefficients are nonzero. The correlation between each pair of the first 10 variables is set to be 0.9. The remaining 30 variables are independent with each other, and also independent with the first 10 variables. We let

$$\beta = (3, 3, 3, 3, 3, -2, -2, -2, -2, -2, 0, ..., 0),$$

and

$$y = \beta^T x + \epsilon, \ , \ \epsilon \sim N(0, 9).$$

The signal to noise ratio is about 3.2.

Table 1: Variable selection frequencies (%) of different methods for the simulation example. "IV": important variables; "UV": unimportant variables; "RME": relative model error (lower is better) .

	Lasso	**ALasso**	**Enet**	**Relaxo**	**VISA**	**RLasso**
$n = 50$						
IV	35	38	60	29	28	**98**
UV	20	11	13	9	**7**	17
RME	666	613	562	608	610	**299**
$n = 100$						
IV	69	82	76	62	62	**99**
UV	52	**21**	35	36	37	30
RME	505	*313*	471	487	487	**132**

Table 1 shows the performance of the various algorithms in selecting the important variables 1-10 (IV) and the unimportant variables 11-30 (UV), and also the relative model error (RME), as defined in [Wang et al., 2011]. The performance is averaged over 100 simulations. As can be seen, RLasso is far superior to all competing methods on both criteria. The paper contains many other simulation setups, including some where RLasso is inferior to some of the alternatives, and discussion of the underlying reasons. We note, however, that for most realistic simulation scenarios that are gene-expression motivated, RLasso performs best.

Finally, all methods were also applied to a famous real gene expression dataset, where the examined response is the log-survival time of Glioblastoma patients [Freije et al., 2004]. One dataset with $n = 50$ patients was used for training the models, and the other with $n = 61$ patients for comparing predictive performance. The number of genes is $p = 3600$, reduced to $p = 1000$ by initial filtering. Table 2 shows the results: number of genes selected and mean squared prediction error. As can be seen, RLasso chooses more genes than other methods (though still less than 6% of genes), and achieves the best predictive performance.

Table 2: Analysis of the glioblastoma data set.

Method	# of genes selected	Mean prediction error
Lasso	29	1.118 (0.205)
Adaptive Lasso	33	1.143 (0.211)
Relaxed Lasso	23	1.054 (0.194)
Elastic-net	28	1.113 (0.204)
VISA	15	0.997 (0.188)
Random lasso	58	0.950 (0.210)

4 Summary

As we have shown, practical sparse modeling is a wide area, and the selection of specific appropriate methods should strongly depend on the specific type of sparsity and correlation in the problem at hand, as well as on the desired performance metrics for the model: successful variable selection, favorable predictive performance, or both.

In the GWAS example, where the goal is mostly identification of associated SNPs for followup studies, the univariate *filter* approach commonly used is appropriate. In the gene expression example, with more complex correlation structure, and a second goal of good predictive performance, more complex methodology is required, and we surveyed variants of Lasso that aim to take the specifics of the problem into account and accomplish both goals.

References

Leo Breiman. Random forests. *Machine Learning*, 45:5–32, 2001.

Emmanuel Candes and Terence Tao. The dantzig selector: Statistical estimation when p is much larger than n. *Annals of Statistics*, 35(6):2313–2351, 2007.

Emmanuel J. Candes and Yaniv Plan. Near-ideal model selection by l1 minimization. *Annals of Statistics*, 37(5A):2145–2177, 2009.

Rita M. Cantor, Kenneth Lange, and Janet S. Sinsheimer. Prioritizing gwas results: A review of statistical methods and recommendations for their application. *Am J Hum Genet.*, 86(1):6–22, 2010.

Eric Davidson and Michael Levin. Gene regulatory networks special feature: Gene regulatory networks. 102(14):4935, 2005.

David Donoho. Compressed sensing. *IEEE Transactions on Information Theory*, 52(4):1289–1306, 2006.

Bradley Efron, Trevor Hastie, Iain Johnstone, and Robert Tibshirani. Least angle regression. *Annals of Statistics*, 32(2):407–499, 2004.

William A. Freije, F. Edmundo Castro-Vargas, Zixing Fang, Steve Horvath, Timothy Cloughesy, Linda M. Liau, Paul S. Mischel, and Stanley F. Nelson. Gene expression profiling of gliomas strongly predicts survival. *Cancer Res*, 64:6503–6510, 2004.

Isabelle Guyon and André Elisseeff. An introduction to variable and feature selection. *J. Mach. Learn. Res.*, 3:1157–1182, March 2003.

Yuk Fai Leung and Duccio Cavalieri. Fundamentals of cdna microarray data analysis. *Trends in Genetics*, 19(11):649 – 659, 2003.

Teri A. Manolio. Genomewide association studies and assessment of the risk of disease. *New England Journal of Medicine*, 363(2):166–176, 2010.

P. McCullagh and J.A. Nelder. *Generalized Linear Model*. Monographs on Statistics and Applied Probability. Chapman & Hall, 1989.

Nicolai Meinshausen. Relaxed lasso. *Computational Statistics and Data Analysis*, 52(1):374 – 393, 2007.

Nicolai Meinshausen and Bin Yu. Lasso-type recovery of sparse representations for high-dimensional data. *Annals of Statistics*, 37(1):246–270, 2009.

Peter Radchenko, Gareth, and M. James. Variable inclusion and shrinkage algorithms. *Journal of the American Statistical Association*, pages 1304–1315, 2008.

Mark Schena, Dari Shalon, Ronald W. Davis, and Patrick O. Brown. Quantitative monitoring of gene expression patterns with a complementary dna microarray. *Science*, 270(5235):467–470, 1995.

Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society*, 58(1):267–288, 1996.

Sijian Wang, Bin Nan, Saharon Rosset, and Ji Zhu. Random lasso. *Annals of Applied Statistics*, 5(1):468–485, 2011.

WTCCC. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature*, 447(7145):661–678, 2007.

Hui Zou. The adaptive lasso and its oracle properties. *Journal of the American Statistical Association*, 101(476):1418–1429, 2006.

Hui Zou and Trevor Hastie. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(2):301–320, 2005.