Statistics of Big Data, Fall 2023-24

Class notes 5

Sources for today's material:
Chapters 2-3 of the Elements of Statistical Learning by Hastie et al.
Review by Rosset (2013) on ptactical sparse modeling
Candes et al. (2005)
Meinshausen and Yu (2009)

Introduction: the $p >> n$ conundrum

Predictive modeling setup: we have a data model $Pr(X, Y)$ with $X \in \mathbb{R}^p$ explanatory variables, $Y \in \mathbb{R}$ response. Our goal is to represent $Y \approx \hat{f}(X)$ so in the future we see X only and predict $\hat{Y} = \hat{f}(X)$.

Predictive modeling goals:

1. Predict well: choose$\hat{f}(X)$ so that indeed when we get a new prediction point x_0 we will have a good prediction, i.e. $\hat{Y}_0 = \hat{f}(x_0) \approx Y_0$ the true value of Y at this point. If we write $Y = E(Y|X) + \epsilon$, where by definition $\mathbb{E}(\epsilon) = 0$, then for squared loss the best possible model is $\hat{f}(X) = \mathbb{E}(Y|X)$.

2. Learn about the nature of the connection between X and Y, such as which coordinates of X are actually important for modeling the relationship. Typical example: assume a linear model $Y = X^t\beta + \beta_0 + \epsilon$, and we want to learn about β : which coordinates are non-zero, etc.

As mentioned, we write $Y = E(Y|X) + \epsilon$ and typically assume that $\epsilon \sim (0, \sigma^2)$ is independent of X. We sometime also assume normality: $\epsilon \sim N(0, \sigma^2)$.

We assume we are given a *training set* of n pairs $T = \{(X_i, Y_i)\}_{i=1}^n := (\mathbb{X}_{n\times p}, \mathbb{Y}_{n\times 1})$ sampled i.i.d from $Pr(X, Y)$, and want to use T to learn our model $\hat{f}(X)$ or properties of $\mathbb{E}(Y|X)$.

A generic description of the predictive modeling process is given in Fig. 1.

Sources of randomness in the data:

- The training set T is random $\Rightarrow$ $\hat{f}$ is a random function that depends on T.
- The new observation X^{new} for prediction

Note the prediction $\hat{Y}^{new} = \hat{f}(X^{new})$ has randomness from both sources.

Criteria for quality of model-building *black-box*:

1. Accuracy in prediction, for example:

   Expected pred. err.(EPE) $= \mathbb{E}_{\mathbb{X},\mathbb{Y},X^{new},Y^{new}}(Y^{new} - \hat{Y}^{new})^2$ or MSE $= \mathbb{E}_{\mathbb{X},\mathbb{Y},X^{new}}(\mathbb{E}(Y|X^{new}) - Y^{\hat{new}})^2$,

   with the simple connection $MSE = EPE - \sigma^2$ (the variance of ϵ).

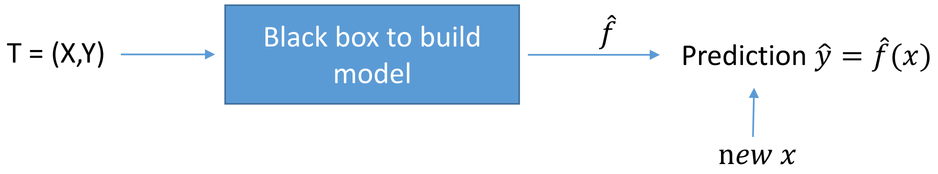


Figure 1: generic description

2. Quality of estimation of parameters or function properties: Assume again (as an example) the linear model $\mathbb{E}(Y|X) = X^T\beta$, and that the fitted model is also linear $\hat{f}(x) = x^T\hat{\beta}$. Then the quality of estimation can be measured in various ways:

    - MSE in estimation: $\mathbb{E}_{\mathbb{X},\mathbb{Y}}\|\beta - \hat{\beta}\|^2$, which is closely related to the prediction MSE.
    - Ability to identify the important coordinates in β. For example, we may assume $\|\beta\|_0 = k << p$, that only a small number of variables have non-zero coefficients, and may ask whether the modeling approach does well in identifying these coordinates through the symmetric difference:
$$\mathbb{E}\left|\{j : \beta_j \neq 0\}\triangle\{j : \hat{\beta}_j \neq 0\}\right|.$$

GWAS as a $p >> n$ problem

In GWAS, we have $p \approx 10^6$ SNPs measured for $n \approx 10^4$ people with $X_{ij} \in \{0, 1, 2\}$, and $Y_i \in \mathbb{R}$ a quantitative phenotype (say height). In our context we write $Y = \mathbb{E}(Y|X) + \epsilon$, where $\mathbb{E}(Y|X)$ can be thought of as the genetic component of the phenotype.

Adding a linear model *assumption* we get $\mathbb{E}(Y|X) = X^T\beta$. This assumes that all the genetic factors have a *linear and additive* effect (so $X_{ij} = 2$ has a double effect on height compared to $X_{ij} = 1$). We also invariably assume $\epsilon \sim (0, \sigma^2)$ is independent of X.

In this context we can think of the prediction problem: estimate $\hat{\beta}$, then given the *genotype* of a new individual X^{new}, predict them $\hat{Y}^{new} = X^{new,T}\hat{\beta}$, as accurately as possible.

Often more interesting is the estimation/interpretation problem: identify which are the important SNPs with big $|\beta_j|$ and/or estimate their effects accurately, or sometimes quantify the overall contribution of genetics to the phenotype: $\frac{Var(\mathbb{E}(Y|X))}{Var(Y)}$. This is called **heritability**.

Linear regression with $p >> n$

The simplest approach we know for building a linear model is least squares regression on the training data:
$$\hat{\beta}_{LS} = \arg\min_{\beta} \|\mathbb{Y} - \mathbb{X}\beta\|_2^2 = \left(\mathbb{X}^T\mathbb{X}\right)^{-1}\mathbb{X}^T\mathbb{Y} \;\Rightarrow\; \hat{f}_{LS}(x) = x^T\left(\mathbb{X}^T\mathbb{X}\right)^{-1}\mathbb{X}^T\mathbb{Y}.$$

A problem: If $p > n$ then $rank(\mathbb{X}^T\mathbb{X}) \leq n$ so $\mathbb{X}^T\mathbb{X}_{p\times p}$ is not invertible and the least squares solution is undefined. In other words $\mathbb{X}\beta = \mathbb{Y}$ is a collection of n equations with p unknowns that has many different *interpolating* solutions that give $\|\mathbb{Y} - \mathbb{X}\beta\|_2^2 = 0$.

Approaches to dealing with this $p >> n$ scenario:

1. **Marginal regression:** Instead of thinking of this as one big regression with p variables, think of it as p small regressions with one variable each (with or without intercept), as in the GFT example:

   $$\text{No intercept:} \qquad \hat{\beta}_j = \frac{\mathbb{X}_{\cdot,j}^T \mathbb{Y}}{\mathbb{X}_{\cdot,j}^T \mathbb{X}_{\cdot,j}}$$

   $$\text{With intercept:} \qquad \text{Define } \mathbb{X}_j = \{1_n,\ \mathbb{X}_{\cdot,j}\}\ ,\ \begin{pmatrix} \hat{\beta}_{0j} \\ \hat{\beta}_{1j} \end{pmatrix} = \left(\mathbb{X}_j^T \mathbb{X}_j\right)^{-1} \mathbb{X}_j^T \mathbb{Y}\ ,\ \text{ or: } \hat{\beta}_{1j} = \frac{\hat{cor}(\mathbb{X}_{\cdot,j}, \mathbb{Y})\hat{sd}(\mathbb{Y})}{\hat{sd}(\mathbb{X}_{\cdot,j})}.$$

   Important observation: if $p < n$ and $\mathbb{X}$ columns are orthonormal then $\mathbb{X}^T\mathbb{X} = \mathbb{I}_p$ and marginal regression (with no intercept) is identical to regular LS regression (with no intercept). In the $p > n$ setting we can still preserve the intuition coming out of this: if the columns of $\mathbb{X}$ are *weakly correlated* then also for $p > n$ marginal regression can give good results since it is not ignoring strong colinearity effects between the covariates.

   This can be considered quite relevant to GWAS, where under some assumptions, the correlation structure (linkage disequilibrium) is such that $cor(\mathbb{X}_{\cdot,j}, \mathbb{X}_{\cdot,k}) \approx 0$ for j, k that are not *too close* on the genome. For example, it is often reasonable to assume that out of the $p = 10^6$ SNPs, we have that if $|j - k| > 100$ then $cor(\mathbb{X}_{\cdot,j}, \mathbb{X}_{\cdot,k}) \approx 0$. Then one option is to do "hybrid" marginal regression where instead of building models for one SNP every time, we can do it for say 100 SNPs every time (still much less than all 10^6 simultaneously).

   A more striking equivalence can be demonstrated if we add another (reasonable) assumption to the GWAS analysis: that among every 100 consecutive SNPs there is at most one that has a real effect ($\beta_j \neq 0$ for the true model of $\mathbb{E}(Y|X)$). So in this setting our problem is reduced to: among every 100 consecutive SNPs find the one that is the true effect (or discover that there are none). We can think of this as a statistical testing/estimation problem. Let's add the normality assumption: $\epsilon \sim N(0, \sigma^2)$. Now we can employ a maximum likelihood approach to estimating the vector β of length 100 under the restriction that only one of its coordinates is non-zero:

   $$\hat{j} = \arg \max_{j \in \{1 \ldots 100\}, \beta_j} \mathcal{L}(\mathbb{X}, \mathbb{Y}; \beta_j) = \arg \min_{j, \beta_{0j}, \beta_{1j}} \|\mathbb{Y} - \mathbb{X}_j^T \beta_j\|_2^2,$$

   where we use the fact that the normal log-likelihood is just the (negative) sum of squares with additional terms that are fixed in this setting. In other words, in this setting the problem of finding the ML candidate for the associated SNP boils down to finding the best marginal regression solution among all 100.

   Conclusion: under these "reasonable" assumptions:

   - Linear model with normal errors
   - Only short term correlation
   - Sparsity: in a strongly correlated set there cannot be more than one true effect

   we get that marginal regression is the right solution for both estimation and prediction in GWAS! For prediction in this setting, we would choose the best j in each 100 set and simply build a model summing up the marginal regressions on these chosen SNPs (since now we have no correlation between them).

2. **Regularized regression:** In this approach we limit the complexity of the solutions $\hat{\beta}$ that we allow, achieving two goals:

   - Making sure the problem has a mathematical solution (unlike least squares for $p > n$)
   - Controlling and reducing the variance (recall the variance-bias tradeoff)

   The most common family of regularization approaches is based on norms of the coefficient vector $\|\beta\|_q$ (usually limited to $q \in \{0, 1, 2\}$, and can be formulated as either constrained or penalized optimization problems:

   Constrained: $\hat{\beta}^{con}(s) = \arg \min_{\|\beta\|_q \leq s} \|\mathbb{Y} - \mathbb{X}\beta\|^2$, Penalized:$\hat{\beta}^{pen}(\lambda) = \arg \min_{\|\beta\|} \|\mathbb{Y} - \mathbb{X}\beta\|^2 + \lambda\|\beta\|_q$.

   The two formulations are generally equivalent (Lagrange form of each other), in the sense that (roughly speaking) for each s there is λ such that $\hat{\beta}^{con}(s) = \hat{\beta}^{pen}(\lambda)$. We will use them exchangeably as convenience will dictate.

   We can interpret the different norms graphically, it is easiest in the constrained setting, with $s = 1, p = 2$:

   - $q = 0$ (variable selection) gives mass only along the axes — it selects exactly one variable of the two to have $\hat{\beta} \neq 0$
   - $q = 1$ (Lasso) gives constraint $|\beta_1| + |\beta_2| \leq 1$ — a diamond shape
   - $q = 2$ (Ridge) gives constraint $\beta_1^2 + \beta_2^2 \leq 1$ — a two-dimensional ball

   Once we add regularization, our regularized optimization problems generally have a solution also for the $p > n$ case, in particular the penalized form fpr $q \in \{1, 2\}$ always has an optimal solution. We will demonstrate it through ridge regression:

   $$\hat{\beta}(\lambda) = \arg \min_{\|\beta\|} \|\mathbb{Y} - \mathbb{X}\beta\|^2 + \lambda\|\beta\|_2^2 = \left(\mathbb{X}^T\mathbb{X} + \lambda\mathbb{I}_p\right)^{-1} \mathbb{X}^T\mathbb{Y},$$

   which is always defined since $rank\left(\mathbb{X}^T\mathbb{X} + \lambda\mathbb{I}_p\right)) = p$ for any $\lambda > 0$.

   However for dealing with sparsity in high dimension, ridge is less relevant, we will focus here on $q \in \{0, 1\}$:

   $q = 0$**: variable selection.** The constrained version explicitly tells us that we can have only s non-zero coefficients in our solution $\hat{\beta}(s)$. If we think or know that the true β is also sparse, it seems natural to apply this constraint to our calculated solution. However, the problem is computationally difficult because the constraint $\|\beta\|_0 \leq s$ is not a *convex* region, and so convex optimization approaches are not relevant, and instead we need to take a combinatorial approach to finding an optimal solution: Enumerate over all $\binom{p}{s}$ possible sets of s variables. Simple approximation using Stirling's formula gives us:

   $$\binom{p}{s} = \frac{p!}{s!(p-s)!} \approx \left(\frac{p}{s}\right)^s,$$

which diverges very quickly and is not practical already for $p = 10^6, s = 3$ in the GWAS context.

There are various approximations for the $q = 0$ case, like Forward Selection and Backward Elimination, but there are no guarantees that they will reach the optimal (or even a good) solution, especially in complex correlation scenarios.

$q = 1$**: Lasso.** This is a hugely important problem formulation in modern data analysis, and in particular in the $p > n$ and sparsity case. This formulation has several interesting properties:

- The problem is convex, and a key observation is that this is now a *quadratic programming* (QP) problem, with quadratic objective and linear constraints. This can be seen from the constrained version, which can equivalently be written as:

$$\begin{aligned} \min \quad & RSS(\beta^+ - \beta^-) \\ \text{s.t.} \quad & \sum_{j=1}^{p} \beta_j^+ + \beta_j^- \leq s \\ & \beta_j^-, \beta_j^+ \geq 0 \forall j, \end{aligned}$$

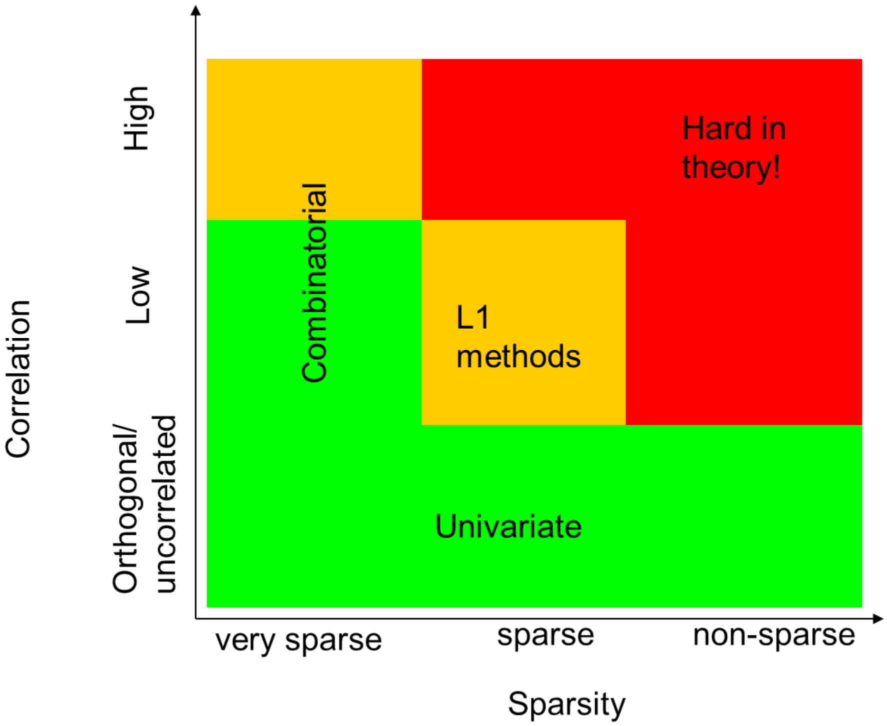
  (the problems are equivalent since in the optimal solution, it is guaranteed that either $\beta_j^+ = 0 \Rightarrow \beta_j = -\beta_j^-$ or $\beta_j^- = 0 \Rightarrow \beta_j = \beta_j^+$.)
  Since QP is a standard problem in convex optimization, standard solvers can be used for Lasso (and in fact the original paper by Tibshirani(1996) proposes a special QP variant that fits the structure of this problem).
  We will discuss the LARS algorithm which takes a different approach in a week or two.

- A key property of Lasso solutions is *sparsity*. This has several expressions:
  - In the high dimensional regime $p > n$, Lasso solutions always have at most n non-zero coefficients: $\|\hat{\beta}(\lambda)\|_0 \leq n$. This is in contrast to Ridge regression, where all coefficients are always non-zero.
  - In the low dimensional regime $p < n$, it is still true that for $\lambda >> 0$ heavy Lasso regularization, we will have $\|\hat{\beta}(\lambda)\|_0 << p$, many zero coefficients (again, in contrast to Ridge and other methods).

  In other words, Lasso is a different way to get variable selection and sparsity without explicitly requiring them with $q = 0$.

- **Compressed sensing:** The area of Compressed Sensing, which was extremely widely studied around 2005-2010, shows that under some assumptions:
  - The true model is linear $\mathbb{E}(Y|X) = X^t\beta$, and it is sparse $\|\beta\|_0 = k << p$.
  - The explanatory variables (columns of $\mathbb{X}$) have low correlation between them.
  - The amount of data we have is $n = O(k \log p)$ (which is still $<< p$ when the model is very sparse).

  then using Lasso we can find *with high probability* the true sparsity pattern (that is, $\hat{\beta}(\lambda)_j \neq 0 \Leftrightarrow \beta_j \neq 0$). However, due to the strong assumptions this intriguing area is not necessarily relevant to our settings of interest.


Hard in theory!
Combinatorial
L1 methods
Univariate
High
Low
Orthogonal/ uncorrelated
Correlation
very sparse
sparse
non-sparse
Sparsity