L_0 and Genetic Risk Prediction

Robert M. Porsch, PhD

October 16, 2018

Centre for Genomic Science

About Me

- PhD in Statistical Genetics
- Now working at the Centre for Genomic Science at HKU
- My focus is on genetic risk prediction for diabetes, height, IQ and many others

Goals for this presentation

- Give some insight into my work
- Introduce the L_0-norm
- Show some results
- Have some fun

The Data

About the data

Lets look at a DNA sequence from two alleles of a single subject:
TCACTAGGATTTACGCCGCGAGTCCCACTTGGGCACCT
TCACTATGATTTACGGCGCHAGTCCCACTTGHGTACCT
We can also write this in a look-up table:

	Chromosome	StartPosition	ReferenceAlternativeAllele	rsID
0	1	865545	G/A	exm1916089
1	1	865694	C/T	exm55
2	1	874762	C/T	exm106
3	1	878423	C/T	exm145
4	1	878667	G/T	Asn_Vand_chr1_878667

- Reference allele: Commonly the most frequent allele in the sample
- There are also insertion/deletions (Start / End position)

Binary Representation

So how do we represent individual DNA changes of n subjects and p DNA changes?

Subject	rsID	Reference/Alternative	Observed	Coding
Sub1	rs1	G/A	G/G	0
Sub1	rs2	C/T	C/T	1
Sub1	rs3	G/T	T/T	2

These codings are called *genotypes* and are usually stored in a *genotype matrix* X in which $X \in \mathbb{Z}^{p \times n}$.

- Commonly we deal with $> 1 million$ variants
- Only a few thousand subjects

Classical $p >> n$ problem.

A further complication

Furthermore, positions are not independent but highly correlated with each other. We say that those positions are in high linkage disequlibrium (LD) with each other.

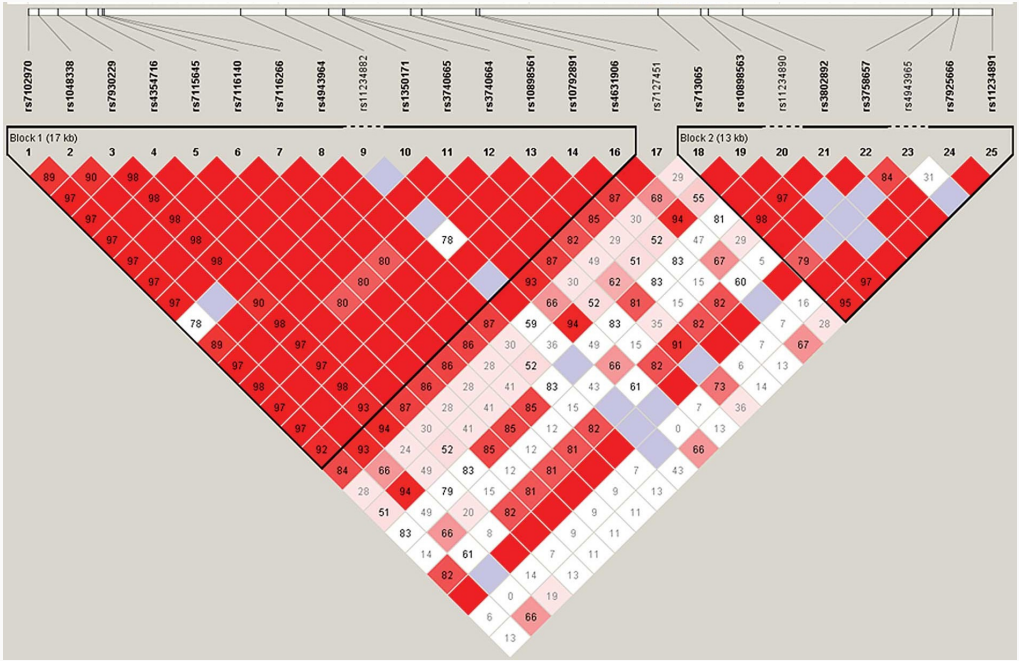


Figure 1: Example of two LD blocks [1]

[1] Qu, L., Jin, F., Guo, Y., Liu, T., Xue, R., Huang, X. ... Shen, X. (2016). Nine susceptibility loci for hepatitis B virus-related hepatocellular carcinoma identified by a pilot two-stage genome-wide association study. Oncology

The goal

Let's say we have the genotype matrix X from n subjects as well as some trait y.
Problem of predicting y from X:

- tons of features
- low sample size
- low effect sizes

Currently used approach ($n \leq 100{,}000$):

(a) Estimate effect of each position ($\hat{\beta}$)
(b) Reduce the number of features by removing $\hat{\beta}$ above a certain p-value threshold
(c) Apply p-value clumping (generate independent $\hat{\beta}$)
(d) Compute $\hat{y} = \hat{\beta}^T X^*$

Hypothesis

The availability of more and more data makes it possible to estimate the conditional effect sizes ($n \geq 500{,}000$).

Let $\mathcal{D}$ the dataset consisting of N input-output pairs $\{(x_1, y_1), \ldots, (x_N, y_N)\}$ and consider the following regularized minimization procedure

$$\mathcal{R}(\theta) = \frac{1}{N}(\sum_{i=1}^{N} \mathcal{L}(h(x_i; \theta), y_i)) + \lambda \mathcal{P}(\theta) \tag{1}$$

With $\theta^* = \underset{\theta}{\text{argmin}}\{\mathcal{R}(\theta)\}$.

$\mathcal{P}(\theta)$ is a penalization function for the parameters θ.

Forms of $\mathcal{P}(\theta)$

L_1-norm:

$$\mathcal{P}_{L_1}(\theta) = ||\theta||_1 = \sum_{i=1}^{|\theta|} |\theta_i| \tag{2}$$

or L_2-norm:

$$\mathcal{P}_{L_2}(\theta) = ||\theta||_2 = \sum_{i=1}^{|\theta|} \theta_i^2 \tag{3}$$

When being more general. Let $p \geq 1$

$$\mathcal{P}_{L_p}(\theta) = ||\theta||_p = (\sum_{i=1}^{|\theta|} |\theta_i|^p)^{1/p} \tag{4}$$

What about $p = 0$?

There are **two** L_0-norms:

- L_0-norm (Established by Banach and not relevant here)
- L_0-'norm' (Established by David Donoho)

The Ł$_0$-'norm' is not a real norm, we just call it L_0 because of the limit

$$P_{L_0}(\theta) = ||\theta||_0 = \lim_{p \to 0} (\sum_{i=1}^{|\theta|} |\theta_i|^p) \tag{5}$$

In general we define it as

$$P_{L_0}(\theta) = ||\theta||_0 = \sum_{j=1}^{|\theta|} \mathcal{I}[\theta_j \neq 0] \tag{6}$$

So its just the count of non-zero parameters.

Some graphical representation always helps

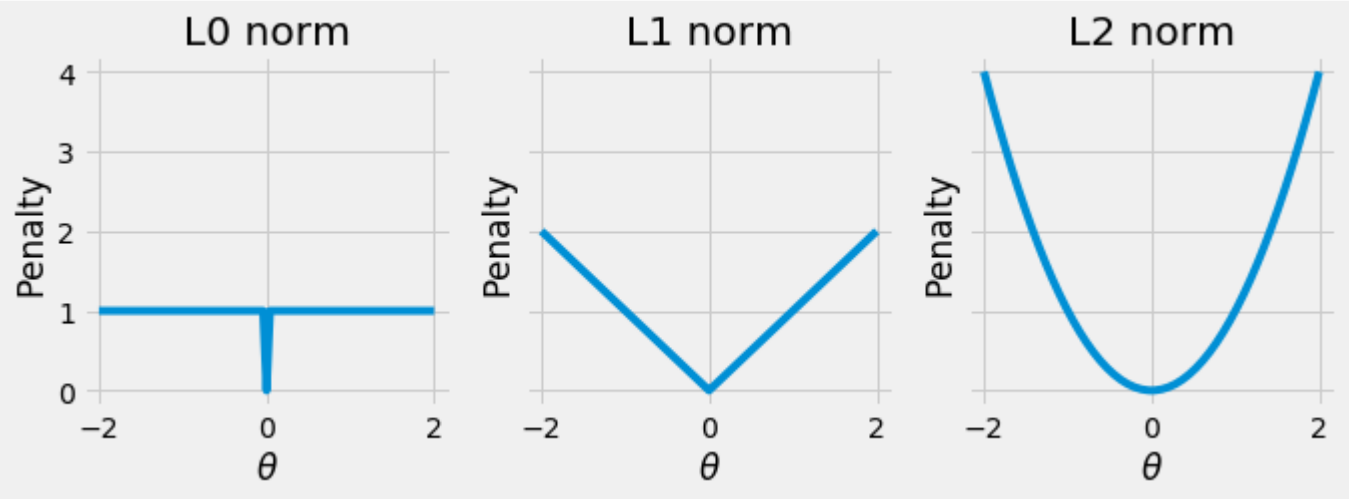


Figure 2: Various Norms

Furthermore,

- same as the other norms, it encourages sparsity in the parameters
- L_0 does not induce any shrinkage in the parameters

There is a problem however . . .

Optimization is computationally intractable under the L_0 penalty,

- non-differentiable
- $2^{|\theta|}$ possible states
- NP-hard problem

So there is a need to relax the discrete nature of L_0 to allow for efficient optimizations.

Literature

Most of the material in this presentation is from:
Learning Sparse Neural Networks through L_0 Regularization
Louizos, Max Kingma

and to a lesser extend from:
The Variational Garrote
Kappen, Gomez

The General Recipe

The first step is to reformulate the L_0 norm under the parameters θ. Hence let,

$$\theta_j = \tilde{\theta}_j z_j, \quad z_j \in \{0, 1\}, \quad \tilde{\theta}_j \neq 0 \tag{7}$$

Therefore z_j can be considered as binary gates (parameter has an effect).

Then we can reformulate the minimization from Eq. 1 by letting $q(z_j|\pi_j) = Bern(\pi_j)$

$$\mathcal{R}(\tilde{\theta}, \pi) = \mathbb{E}_{q(z|\pi)}[\frac{1}{N}(\sum_{i=1}^{N} \mathcal{L}(h(x_i; \tilde{\theta} \otimes z), y_i)] + \lambda \sum_{j=1}^{|\theta|} \pi_j \tag{8}$$

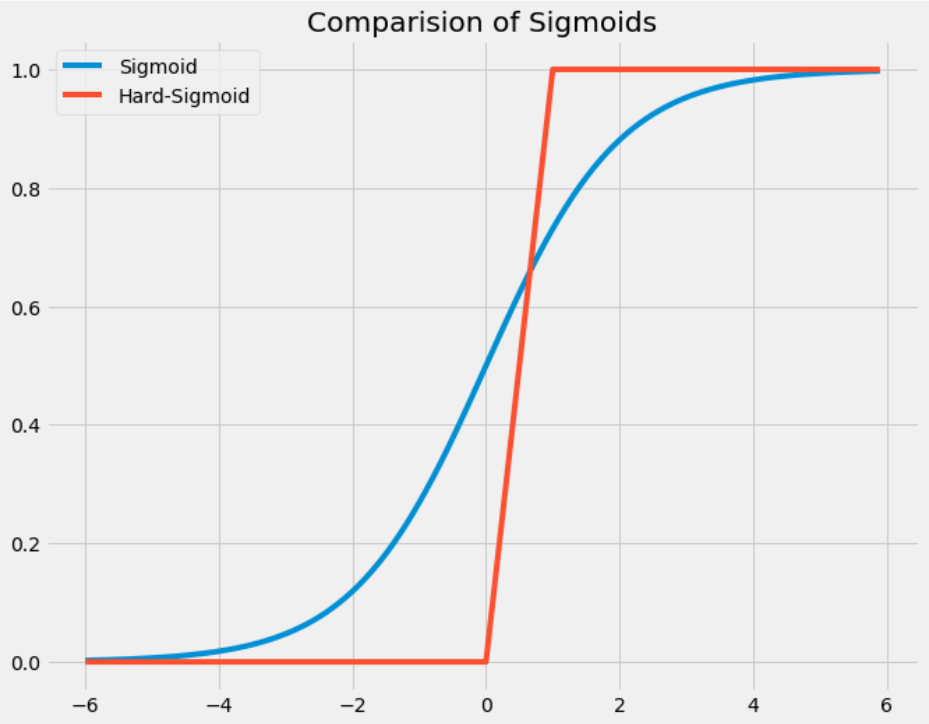
with $\tilde{\theta}^*, \pi^* = \underset{\tilde{\theta}, \pi}{\text{argmin}}\{\mathcal{R}(\tilde{\theta}, \pi)\}$

However, the discrete nature of z makes it still difficult to minimize π. A solution is to give z a **smoothing** function.

Smoothing z

Let s be a continuous random variable with a distribution $q(s)$ with parameter ϕ. Then we let the gates z be given by a hard sigmoid function:

$$\begin{aligned} s \sim & q(s|\phi) \\ z = & \min(1, \max(0, s)) \end{aligned} \tag{9}$$



Updating $\mathcal{R}$

We can then define the probability of the gates being non-zero as

$$q(z \neq 0|\phi) = 1 - Q(s \leq 0|\phi) \tag{10}$$

in which $Q(\cdot)$ is the cumulative distribution function of s. Hence we update our original optimizations function as

$$\begin{aligned}\mathcal{R}(\tilde{\theta}, \phi) = \mathbb{E}_{q(s|\phi)}[\frac{1}{N}(\sum_{i=1}^{N} \mathcal{L}(h(x_i; \tilde{\theta} \otimes g(s)), y_i)] + \\ \lambda \sum_{j=1}^{|\theta|}(1 - Q(s_j \leq 0|\phi_j))\end{aligned} \tag{11}$$

with $\theta^*, \phi^* = \underset{\tilde{\theta}, \phi}{\text{argmin}}\{\mathcal{R}(\tilde{\theta}, \phi)\}$ and $g(\cdot) = \min(1, \max(0, \cdot))$

The Hard Concrete Distribution

The literature suggests to use a hard concrete distribution as a smoothing function $q(s)$. The parameters of the distribution are $\phi = (\log \alpha, \beta)$ and can be stretched to (γ, ζ) intervals.

$$
\begin{aligned}
u &\sim \mathcal{U}(0,1) \\
s &= \mathit{Sigmoid}((\log u - \log(1-u) + \log \alpha)/\beta) \\
\bar{s} &= s(\zeta - \gamma) + \gamma
\end{aligned}
\tag{12}
$$

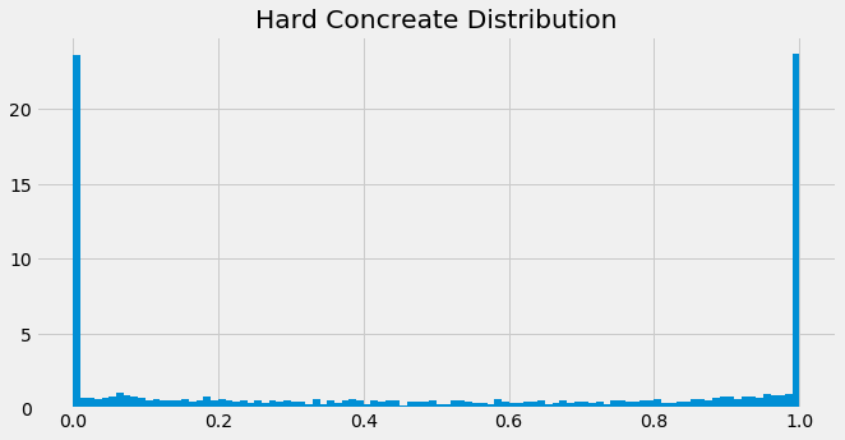


Figure 3: Sample from the Hard Concrete Distribution

Varying the hard concrete distribution

interactive example (link)

Some Results

Simulations

I simulated some y from 10,000 features of which only 0.1% were causal. Combined variance explained by these causal features was set to $R^2 = 0.5$. Effect sizes were drawn from $N(0, 1)$. Results are shown after cross-validation.

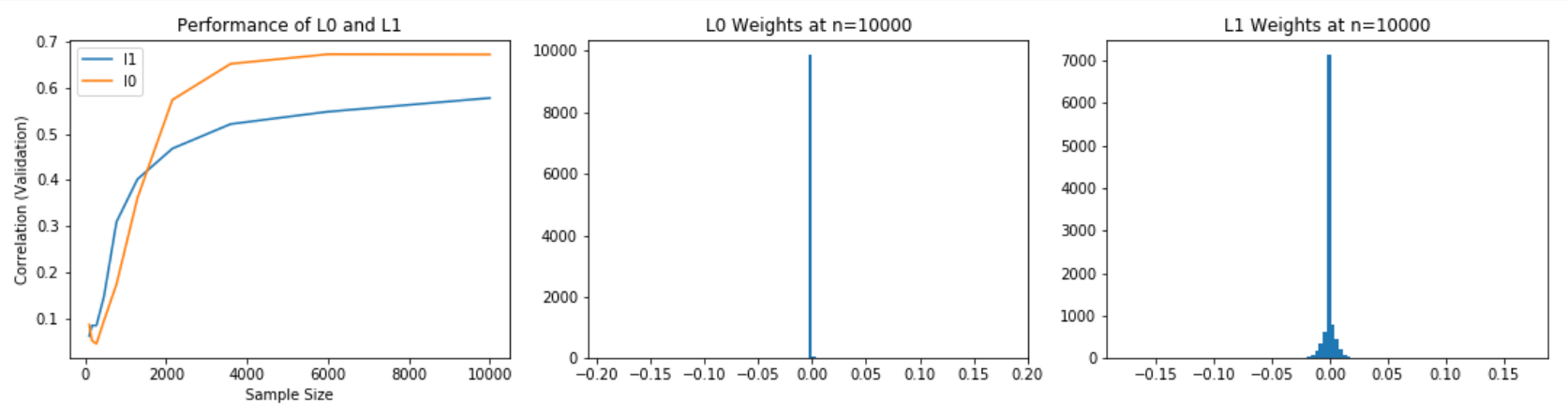
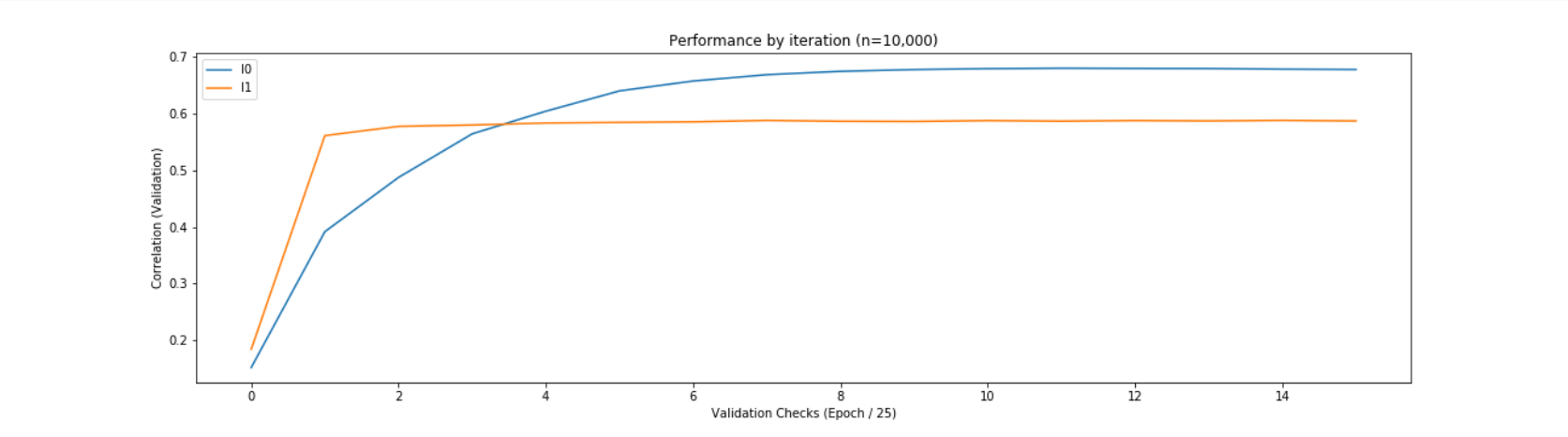
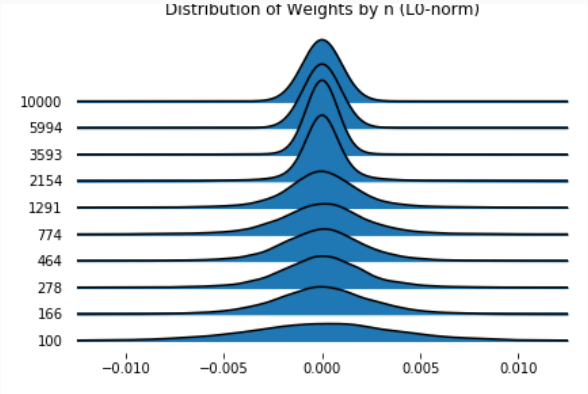


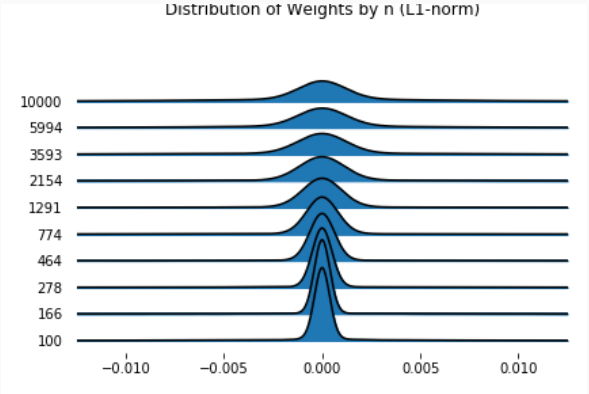
Figure 4: Performance

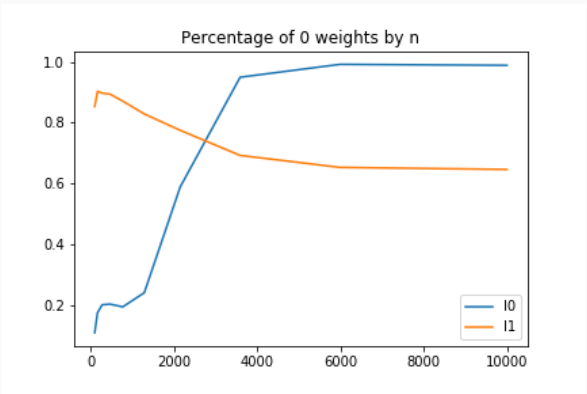
As larger the sample size as better the performance of L_0, since $\hat{\beta}$ becomes closer to real value β.

Simulations









Some reflections and conclusions

- Overall this method is rather simple and easy to implement
- L_0-norm essentially performs p-value thresholding while predicting y
- Remains a valuable alternative not only for penalized regressions but also NN
- Only more effective when sample size is significantly large enough
- Shortcomings of L_1 can be addressed (inconsistent oracle property)

Thank you for your time