Sparse Regression Screening Dual Polytope Projection Failure

Michael Rawson

Advisor: Prof. Carlos Fernandez-Granda

Abstract

Sparse regression screening with dual polytope projection (DPP) is an iterative process for computational efficiency. While the math is sound, using limited precision iteratively is hazardous, for example using double floating point precision in a computer. In analysis, synthetic data experiments, and real data experiments, we show DPP screening failure regions.

Introduction

Regression

Regression methods are a set of techniques used to infer information about systems from collected data. They are commonly used in engineering, science, and statistics. Linear regression creates a linear model to analyze the relationship between predictors and the sample values (1). If $n = p$ then a hyperplane through the data is calculated, if A is full rank. If $n > p$ then often the L_2 norm of the difference is minimized, $\min_{x \in \mathbb{R}^p} \|Ax - y\|_2$. If $n < p$ then, from Gaussian elimination, the solution has infinite solutions, if A is full rank.

$$A\ x = \begin{bmatrix} a_{11} & a_{12} & \dots & a_{1p} \\ a_{21} & a_{22} & \dots & a_{2p} \\ \vdots & \vdots & & \vdots \\ a_{n1} & a_{n2} & \dots & a_{np} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ \vdots \\ \vdots \\ x_p \end{bmatrix} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} = y \tag{1}$$

where a_{ij} is the value of predictor i for sample j
and x_i is the weight for predictor i
and y_i is the i^{th} sample value

Regularization is often used in data analysis to simplify a complex model and to avoid over-fitting. Overfitting happens when a model learns not just the underlying signal, but the noise too. If there is no noise, then a linear interpolation could be directly applied, to a linear problem. Ridge Regression dampens x by regularizing the L_2 norm of x in the objective function (2).

$$\min_{x \in \mathbb{R}^p} \|Ax - y\|_2^2 + \lambda \|x\|_2^2 \tag{2}$$

where $\lambda > 0$ is a hyper-parameter.

Often there is reason to believe that the relevant predictors to the sample value should be sparse. In order to find a sparse solution, often the L_1 norm of x is also minimized in the objective function. The least absolute shrinkage and selection operator (LASSO) adds the weighted L_1 norm of x to the objective function (3).

$$\min_{x \in \mathbb{R}^p} \|Ax - y\|_2^2 + \lambda \|x\|_1 \tag{3}$$

where $\lambda > 0$ is a hyper-parameter.

However, if predictors are identical, then the LASSO problem doesn't have a unique solution. There are infinite solutions. This does not achieve the objective of learning about the relationship between the

predictors and data. Strongly correlated predictors similarly create non-unique solutions due to limited machine precision. One way to avoid this uniqueness problem is to use an objective function that penalizes both the L_1 norm and L_2 norm of x, know as the Elastic Net (4).

$$\min_{x \in \mathbb{R}^p} \|Ax - y\|_2^2 + \lambda_1 \|x\|_1 + \lambda_2 \|x\|_2^2 \tag{4}$$

where $\lambda_1 > 0$ and $\lambda_2 > 0$ are hyper-parameters.

Screening

For large datasets, where the samples and/or predictors are large, computing time becomes a limitation. In order to be more computationally efficient many, screening methods have been designed. Screening methods remove predictors ahead of optimizing the objective function, allowing practitioners to solve systems with many more predictors. Two types of screening methods are the safe methods and the heuristic methods. The safe methods guarantee to not discard an important predictor. The heuristic methods might discard important predictors, though post-processing can be employed to identify and correct failures. The strong rule [2] is an unsafe method that discards all predictors that satisfy (5).

$$|a_j^T (y - Ax(\lambda'))| < 2\lambda - \lambda' \tag{5}$$

The strong rule branches off the "SAFE" rule [1]:

$$|a_j^T y| < \lambda - \|a_j\|_2 \|y\|_2 \frac{\lambda_{max} - \lambda}{\lambda_{max}}$$

And replaces $\|a_j\|_2 \|y\|_2 / \lambda_{max}$ with 1. Tibshirani et al. also propose sequential versions of "SAFE" and strong rules for efficiently solving the regression for many values of λ.

DPP

The Dual Polytope Projection (DPP) [3] is a theoretically safe method that discards all predictors that satisfy (6).

Let a_j is the j^{th} column of A
and $\lambda' = \gamma\lambda$ and $\gamma > 1$
and $x(\lambda')$ be the solution for λ'.

$$|a_j^T (y - Ax(\lambda'))| < \lambda' - \|a_j\|_2 \|y\|_2 \frac{\lambda' - \lambda}{\lambda} \tag{6}$$

This comes from the following. Start with the dual problem:

$$\theta^*(\lambda) = \texttt{argsup}_\theta \left\{ \frac{1}{2}\|y\|_2^2 - \frac{\lambda^2}{2} \left\| \theta - \frac{y}{\lambda} \right\|_2^2 : |a_i^T \theta| \leq 1, i = 1, 2, ..., p \right\} \tag{7}$$

The Karush-Kuhn-Tucker (KKT) conditions give:

$$y = A\beta^*(\lambda) + \lambda\theta^*(\lambda)$$

where $\beta^*(\lambda)$ is the solution of Eq. (3).
With some relaxation, $\sup_{\theta \in \Theta} |a_i^T \theta| < 1$ implies an inactive feature/predictor i.

The dual optimal solution can be expressed with a projection operator P_F to the feasible set of Eq. (7) by:

$$\theta^*(\lambda) = P_F(y/\lambda) = \texttt{argmin}_{\theta \in F} \|\theta - \frac{y}{\lambda}\|_2$$

From nonexpansiveness of projection operators defined in a Hilbert space with respect to a nonempty closed and convex set:

$$\|\theta^*(\lambda) - \theta^*(\lambda')\|_2 = \|P_F(y/\lambda) - P_F(y/\lambda')\|_2 \leq \|(y/\lambda) - (y/\lambda')\|_2 = \left|\frac{1}{\lambda} - \frac{1}{\lambda'}\right| \|y\|_2$$

Then $\theta^*(\lambda)$ is in the following ball:

$$\theta^*(\lambda) \in B\left(\theta^*(\lambda'), \left|\frac{1}{\lambda} - \frac{1}{\lambda'}\right| \|y\|_2\right)$$

From KKT relaxation above, we bound $|a_i^T \theta^*(\lambda)|$ with the ball from the information from λ'.

$$|a_i^T \theta^*(\lambda)| < 1 - \|a_i\|_2 \|y\|_2 \left|\frac{1}{\lambda} - \frac{1}{\lambda'}\right|$$

Then use KKT for θ^* and we have DPP.

Furthermore, Enhanced DPP (EDPP) [3] is an extension to DPP that discards all predictors that satisfy (8).

$$\left|a_j^T \left(\frac{y - Ax(\lambda')}{\lambda}\right)\right| < 1 - \frac{1}{2}\|a_j\|_2 \|v_2^\perp\|_2 \tag{8}$$

where $v_1(\lambda') = \frac{y}{\lambda'} - \theta(\lambda')$ for $\lambda' \in (0, \lambda_{\max})$
and $v_2(\lambda, \lambda') = \frac{y}{\lambda} - \theta(\lambda')$
and $v_2^\perp(\lambda, \lambda') = v_2(\lambda, \lambda') - \frac{\langle v_1(\lambda'), v_2(\lambda, \lambda')\rangle}{\|v_1(\lambda')\|_2^2} v_1(\lambda')$
and dual solution $\theta(\lambda') = \frac{y - Ax(\lambda')}{\lambda'}$ from KKT conditions

There are 'global' versions of DPP rules that use λ_{max} but we won't consider them since the rules quickly degrade as λ parts with λ_{max}.

Failure Analysis

The DPP and EDPP sequential rules utilize previously computed solutions x for larger λ. However, if only an approximate of the previous objective function's solution is known, then DPP and EDPP can fail.

Let $0 < \lambda < \lambda'$
Let the features/predictors, a_j, and response, y be normalized. $\|a_j\|_2 = 1$ and $\|y\|_2 = 1$
Let x be the solution for λ' and assume we only have access to an approximation, $x + \epsilon$ for $\epsilon \in \mathbb{R}^p$ and $0 < |\epsilon_i| << 1$.
Then for λ' we have $\|Ax - y\|_2 > 0$ and let $A(x + \epsilon) - y = \delta \in \mathbb{R}^p$ and $0 < |\delta_i| << 1$

DPP Failure

We examine the failure of the DPP rules.
From

$$|a_j^T (y - Ax(\lambda'))| < \lambda' - \|a_j\|_2 \|y\|_2 \frac{\lambda' - \lambda}{\lambda}$$

$$|a_j^T (y - A(x + \epsilon))| < \lambda' - \|a_j\|_2 \|y\|_2 \frac{\lambda' - \lambda}{\lambda}$$

$$|a_j^T \delta| < \lambda' - \frac{\lambda' - \lambda}{\lambda}$$

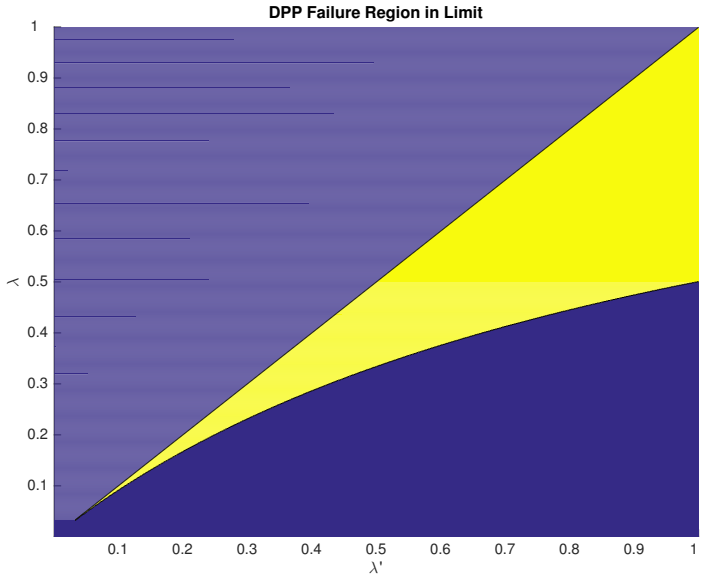

Figure 1: DPP failure region in limit of δ for λ' and λ in yellow.

Take the limit

$$\lim_{\delta \to 0} |a_j^T \delta| = 0$$

So

$$0 < \lambda' - \frac{\lambda' - \lambda}{\lambda}$$

$$0 < 1 - \frac{\lambda' - \lambda}{\lambda \lambda'}$$

In the limit, for certain λ' and λ the rule always evaluates to true, failing to safely screen out predictors. We plot the unsafe region in Fig. 1. By continuity, there is a neighborhood of failure about δ, λ', and λ.

EDPP Failure

We examine the failure of the EDPP rules.
Let

$$v_1^{perturted}(\lambda') = \frac{y}{\lambda'} - \frac{y - A(x+\epsilon)}{\lambda'}$$

$$v_2^{perturted}(\lambda, \lambda') = \frac{y}{\lambda} - \frac{y - A(x+\epsilon)}{\lambda'}$$

$$v_2^{\perp, perturted}(\lambda, \lambda') = v_2^{perturted}(\lambda, \lambda') - \frac{\left\langle v_1^{perturted}(\lambda'), v_2^{perturted}(\lambda, \lambda') \right\rangle}{\|v_1^{perturted}(\lambda')\|_2^2} v_1^{perturted}(\lambda')$$

Rule:

$$\left| a_j^T \left(\frac{y - A(x+\epsilon)}{\lambda} \right) \right| < 1 - \frac{1}{2} \|a_j\|_2 \left\| v_2^{\perp, perturted}(\lambda, \lambda') \right\|_2$$

Let $\gamma = \frac{\lambda'}{\lambda}$. So $\gamma > 1$.
Left Hand Side:

$$\left| a_j^T \left(\frac{y - A(x+\epsilon)}{\lambda} \right) \right| = \left| a_j^T \left(\frac{\delta}{\lambda} \right) \right|$$

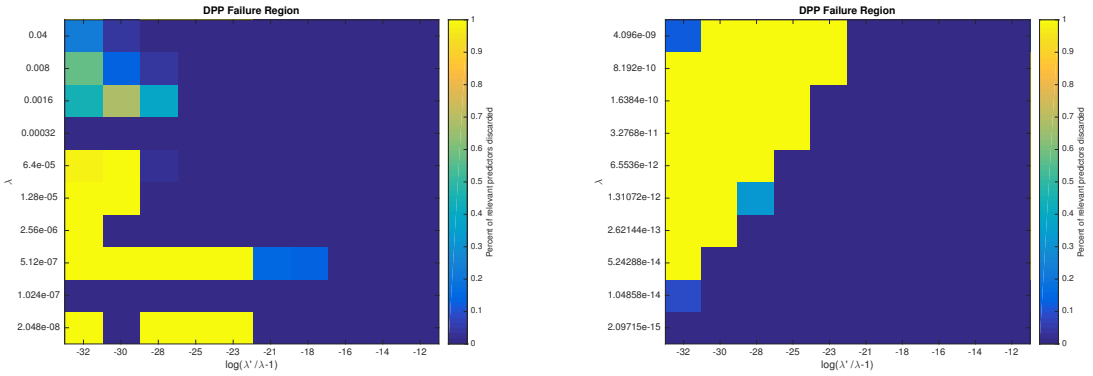

Figure 2: DPP failure region on synthetic data.

Right Hand Side:

$$1 - \frac{1}{2}\|a_j\|_2 \left\| v_2^{\perp,perturted}(\lambda, \gamma\lambda) \right\|_2 = 1 - \frac{1}{2} \left\| \frac{y}{\lambda} - \frac{\delta}{\gamma\lambda} - \frac{\left\langle \frac{y-\delta}{\gamma\lambda}, \frac{y}{\lambda} - \frac{\delta}{\gamma\lambda} \right\rangle}{\| \frac{y-\delta}{\gamma\lambda} \|_2^2} \frac{y-\delta}{\gamma\lambda} \right\|_2$$

Take the limit

$$\lim_{\delta \to 0} 1 - \frac{1}{2} \left\| \frac{y}{\lambda} - \frac{\delta}{\gamma\lambda} - \frac{\left\langle \frac{y-\delta}{\gamma\lambda}, \frac{y}{\lambda} - \frac{\delta}{\gamma\lambda} \right\rangle}{\| \frac{y-\delta}{\gamma\lambda} \|_2^2} \frac{y-\delta}{\gamma\lambda} \right\|_2 = 1 - \frac{1}{2} \left\| \frac{y}{\lambda} - \frac{\langle y, y \rangle}{\lambda \|y\|_2^2} y \right\|_2 = 1$$

Since

$0 < 1$

Then, in the limit, the rule always evaluates true and does not safely discard predictors. By continuity, there is a neighborhood of failure about δ, λ', and λ.

Experiment

Synthetic Data

We experiment with DPP and EDPP on synthetic data. We use a random Gaussian i.i.d. matrix, $A \in \mathbb{R}^{100 \times 500}$, and a random Gaussian i.i.d. sparse signal vector, $x \in \mathbb{R}^{500}$ and $\|\texttt{sign}(x)\|_1 = 100$. We compute the product, y, normalize predictors, a_j, and response, y, and then solve the inverse problem with LASSO to reconstruct the scaled, sparse signal x. We utilize a high precision optimizer set to continue as long as it can make progress. However, the result is still limited by double precision. We find and record where DPP and EDPP fails to safely discard predictors for a neighborhood of λ and λ'. See Fig. 2 for DPP and Fig. 3 for EDPP. We notice that the failure region for EDPP is much larger than DPP, which is not surprising since EDPP is a tighter bound than DPP. We also notice that the neighborhood of complete failure is not necessarily bordered by a neighborhood of complete success. This is not surprising since, for each predictor, the rules depends on the dot product with predictors, a_j, and since these vary, on the boundary, we expect some to cross while others do not. This explains the partial failure around the neighborhood of complete failure.

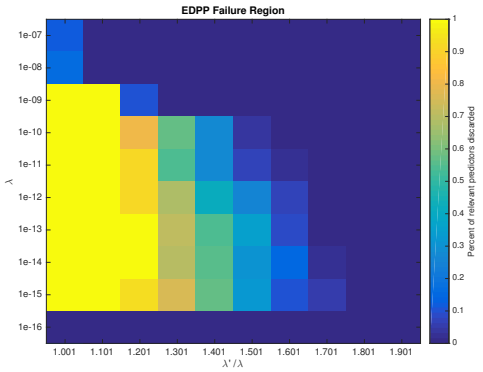

Figure 3: EDPP failure region on synthetic data.

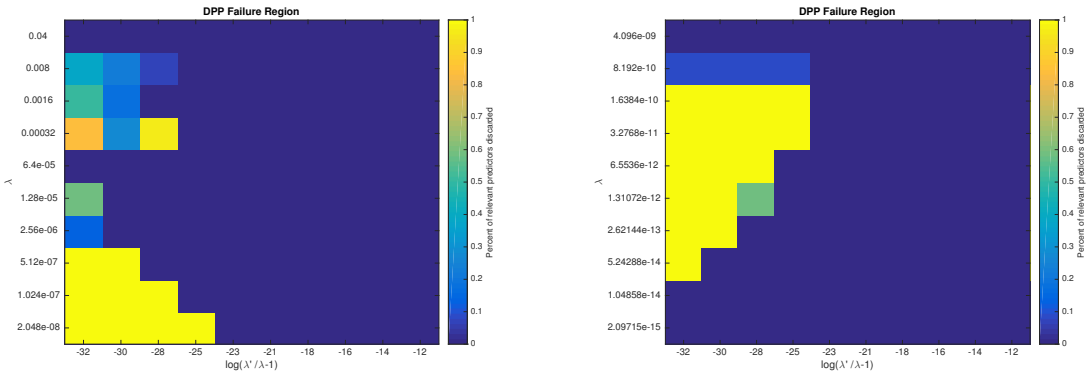

Figure 4: DPP failure region on the 'Gas sensor array under flow modulation Data Set' [4] after normalizing the features/predictors and response.

Gas sensor array under flow modulation Data Set

We experiment with DPP and EDPP on a real dataset from a gas sensor array [4]. We run a Lasso regression on each predictor against the remaining predictors. We choose the predictor with the lowest Lasso error. Then we test DPP and EDPP screening for failure regions. We plot DPP failure in Fig. 4 and EDPP failure in Fig. 5. The results are strikingly similar to the experiments with synthetic data.

Conclusion

We conclude that sequential DPP and EDPP rules are potentially useful screening rules but cannot be trusted to safely screen out predictors due to finite precision and limited optimizers. The KKT conditions can be checked on each predictor after an unsafe screening to identify mistakes [2]. If a mistake is found then adding the predictor back and rerunning the regression is required solve the inverse problem. Future directions might include finding a safe region with respect to λ' and λ for DPP and EDPP. Another future direction might be checking for failure in other safe screening rules like [1].

References

[1] L. El Ghaoui, V. Viallon, and T Rabbani. Safe feature elimination in sparse supervised learning. *Technical Report*, UC/EECS-2010-126, 2010.

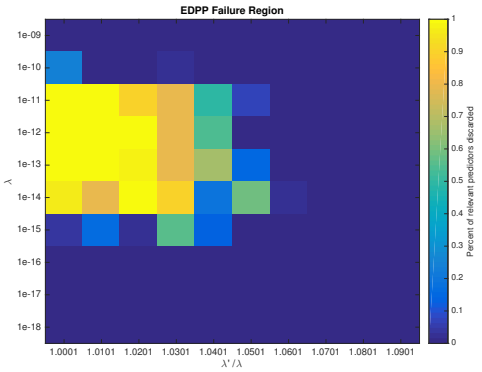

Figure 5: EDPP failure region on the 'Gas sensor array under flow modulation Data Set' [4] after normalizing the features/predictors and response.

[2] R. Tibshirani, J. Bien, J. Friedman, T. Hastie, N. Simon, J. Taylor, and R. J. Tibshirani. Strong rules for discarding predictors in lasso-type problems. *Journal of the Royal Statistical Society*, 74, 2012.

[3] Jie Wang, Peter Wonka, and Jieping Ye. Lasso screening rules via dual polytope projection. *CoRR*, abs/1211.3966, 2012.

[4] A Ziyatdinov, J Fonollosa, L Fernandez, A Gutierrez-Galvez, S Marco, and A Perera. Bioinspired early detection through gas flow modulation in chemo-sensory systems. *Sensors and Actuators*, B: Chemical 206:538–547, 2015.