The Dantzig Selector*

Michael Hutchinson, Bryan Liu, Anna Menacher and Daniel Moss

March 20, 2020

1 Introduction

The Dantzig selector offers an approach to statistical estimation when p, the number of covariates, is much larger than n. This summary will follow the celebrated paper of Candes and Tao [3]. Consider the linear model

$$Y = X\beta + \epsilon$$

where $X \in \mathbb{R}^{n\times p}$ is the (known) *design matrix*, $\beta \in \mathbb{R}^p$ is (unknown) the vector of *coefficients*, $Y \in \mathbb{R}^n$ is the vector of *responses* (or *labels*) and $\epsilon \sim \mathcal{N}(0, \sigma^2 I_n)$ is the *noise*.

We can define a class of estimators $\hat{\beta}$ as solutions to the constrained optimization problem

$$\min_{\tilde{\beta}\in\mathbb{R}^p} \|\tilde{\beta}\|_1 \quad \text{subject to} \quad \|X^T r\|_\infty \leq \lambda_p \cdot \sigma \tag{1}$$

where r is the vector of *residuals*, defined for each $\tilde{\beta}$ by $r = Y - X\tilde{\beta}$ and where λ_p determines the class. When $\lambda_p = (1 + t^{-1})\sqrt{2\log p}$, for t to be specified later, we call this estimator the *Dantzig selector*.

The main results of [3] are in the spirit of high-dimensional non-asymptotics, where high probability statements are made regarding the performance of this estimator under a fixed n and p. These statements allow us to compare the procedure to some idealised 'oracle estimator' which relies on unknown information. An example of such an oracle estimator in the context of high-dimensional regression with sparsity is that obtained by performing OLS only on the $n \times s$ design matrix of 'signal' variables for which the respective entry in β is nonzero.

In order for such statements to be made, certain technical conditions must be met. One such condition, the *Uniform Uncertainty principle* (UUP), will be discussed in the next section.

2 Theoretical framework

2.1 The Uniform Uncertainty principle

Recall that a matrix $M \in \mathbb{R}^{p\times q}$ is said to be *orthogonal* if its columns $(M_i)_{i=1}^q$ define an orthonormal set in $\mathbb{R}^p$. This can only hold when $q \leq p$.

An immediate consequence of this definition is that, for any vector $v \in \mathbb{R}^q$, we have that

$$\|v\|_2 = \|Mv\|_2$$

and indeed orthogonal matrices are exactly those which define isometries.

*Submitted for the *Modern Statistical Theory* module in the StatML CDT.

Now, for our design matrix X, denoted by X_T, the $n \times |T|$ matrix is formed with the columns of X and indexed by $T \in \{1, \ldots, p\}$. We will assume that the columns of X are centred and scaled to have unit norm[1]. Define the *$s-$restricted isometry constant* δ_s of X to be the smallest number for which

$$(1 - \delta_S)\|c\|_2^2 \leq \|X_T c\|_2^2 \leq (1 + \delta_S)\|c\|_2^2$$

for all T with $|T| \leq S$ and for all $c \in \mathbb{R}^{|T|}$. This should be interepted as saying that all subsets of size at most s are 'almost' orthogonal in the sense that they are 'almost' isometries. Define also (again, in the spirit of 'almost an isometry') the *s, s'-restricted orthogonality constants* $\theta_{s,s'}$ for $s + s' \leq p$ to be the smallest numbers for which

$$|\langle X_T c, X_{T'} c' \rangle| \leq \theta_{ss'} \|c\|_2 \|c'\|_2$$

holds for all disjoint $T, T' \subset \{1, \ldots, p\}$ and all vectors $c \in \mathbb{R}^{|T|}$ and $c' \in \mathbb{R}^{|T'|}$.

The UUP was introduced in [4], although this was in the setting of random matrices. The basic summary of this condition, given in [2], is that $\delta_s \leq \frac{1}{2}$ for some particular choice of s, and with some high probability. In the absence of randomness, we simply require the inequality to hold in the usual sense. In our setting, we take s to be the sparsity level of the coeffecient vector, and replace the original UUP with the condition that

$$\delta_{2s} + \theta_{s,2s} < 1. \tag{2}$$

Note that this condition implicitly requires that $s \leq 3p$. Note also that, since the constants are increasing in s, s', when stating results which prescribe a sparsity level and which depend upon this version of the UUP, the results apply to any vector of coeffecients whose sparsity is less than that level as well.

It is perhaps not too surprising that asking the matrix to be 'almost orthogonal' is a useful condition. Indeed, it is reminiscent of conditions on the so-called *pairwise incoherence parameters* which lead to the satisfaction of the *restricted nullspace property*, which has implications for sparse inference when penalising with l_1 penalties. Intuitively, a matrix with orthogonal columns leads to easier inference under sparsity because it becomes easier to discern which entries of a noisy observation y are associated to a certain entry of β.

For concreteness, consider the case where $X = I_2$ and suppose $\beta = (2, 2)$. Then, even if there is some noise, the observations $y = (2 + \epsilon_1, 2 + \epsilon_2)$ should still point clearly to β_1 and β_2 being signal variables. Suppose instead that

$$X = \begin{pmatrix} 1 & 1 \\ 0 & \eta \end{pmatrix}$$

for some very small $\eta > 0$. Then, if $\beta = (2, 2)$, $y = (4 + \epsilon_1, 2\eta + \epsilon_2)$. But if the noise dominates η, as would be the case for η small enough, one could reasonably infer that $\beta \approx (4, 0)$ and thus incorrectly suggest that the true vector β was 1-sparse, as well as being quite far away in l_2 norm.

The authors in [3] further note that, in order for a particular $s-$sparse vector β to even be identifiable, we must at the very least assume that X has no rank-deficient $2s \times n$ submatrix, else it is possible to construct $s-$sparse β,β' such that $X\beta = X\beta'$. Thus, while we earlier noted that $3s \leq p$ is required, we can strengthen[2] this to $2s < n$.

In the next section, we will explore how, under this almost-orthogonal condition, we can indeed approach the performance of certain idealised estimators.

[1] This is so that the following definitions, designed to hold when the matrix is 'almost orthogonal' are sensitive only to the orthogonality (or otherwise) of the columns rather than their length as well.

[2] A strengthening assuming $p \gg n$.

2.2 Oracle estimators and inequalities

An *oracle estimator* is one which requires some additional knowledge of the problem which is not known to the practitioner. In certain contexts, they may be understood as estimators in ideal models, those which we would attempt to find using model selection techniques. Take, for instance, the Lasso estimator $\hat{\beta}_L^\lambda$. Let us take as our definition of an ideal estimator β^* within some class of possible estimators $\mathcal{E}$ as one which satisfies

$$\beta^* \in \arg\min_{\hat{\beta} \in \mathcal{E}} M(\beta, \hat{\beta})$$

where M is some deterministic loss function. Then finding the ideal Lasso estimator (by varying the tuning parameter) $\beta^* = \hat{\beta}_L^{\lambda^*}$ requires knowledge of $M(\cdot, \cdot)$ and thus knowledge of the quantity β which we are trying to estimate! However, an oracle with knowledge of $M(\cdot, \cdot)$ that could tell us the ideal tuning parameter λ^*. A typical choice of M is $M(\beta, \hat{\beta}) = \mathbb{E}[(\beta - \hat{\beta})^2]$, and we will assume this is the choice taken throughout.

An *oracle inequality* is then an inequality relating the performance of some usable estimator the the idealised oracle estimator. Two such inequalities are produced by the authors of [3], namely Theorem 1.1 and Theorem 1.2, and each compare the risk of the procedure they propose with the risk of some oracle estimator.

2.2.1 Knowledge of the nonzero entries

Suppose first of all that the Oracle knows exactly the support set $S = \{i \in \{1, \dots, p\} : \beta_i \neq 0\}$. Then an oracle estimate may be produced by performing OLS with design matrix X_S[3]. The risk of this estimate, call it β^*, is given by

$$\mathbb{E}\|\beta - \beta^*\|_2^2 = \sigma^2 \mathrm{tr}(X_S^T X_S)^{-1} \sigma^2$$

Since the trace is given by the sum of the eigenvalues, the restricted isometry constants provide lower bounds on this quantity and thus

$$\mathbb{E}\|\beta - \beta^*\|_2^2 = \sigma^2 \mathrm{tr}(X_S^T X_S)^{-1} \sigma^2 \geq (1 + \delta_{|S|}^{-1}) |S| \sigma^2$$

Theorem 1.1 of [3] then shows that a procedure similar to the Dantzig selector achieves this 'lower bound' up to multiplicative constants and $\log p$ factors, under the version of the UUP discussed previously. We state the result here.

Theorem 1 (Candes, Tao 2007)**.** *Let $\beta \in \mathbb{R}^p$ be supported on S, and $s = |S|$. Let the UUP (2) hold. Consider obtaining an estimator $\hat{\beta}$ from solving the procedure given by (1), with $\lambda_p = \sqrt{2 \log p}$. Then, except on an event of probability $O\left((\sqrt{\log p})^{-1}\right)$, we have*

$$\|\hat{\beta} - \beta\|_2^2 \leq C_1^2 (2 \log p) S \sigma^2$$

where $C_1 = \dfrac{4}{1 - \delta_s - \theta_{s,2s}}$.

2.2.2 Knowledge of entries below the noise threshold

In what precedes, we compared the proposed estimator to an oracle estimate where the oracle is aware of the support set of β. Theorem 1 the OLS estimate for that component (performing the OLS only on the signal covariates).

This oracle estimate, though unbiased, was constrained in its MSE by the variance of the noise. Suppose instead we kne, rather than whether entries of β were zero, whether they were instead below the noise threshold σ^2. Then, for those below to noise threshold, we could simply estimate the component by zero.

[3] In fact, we can further refine this by choosing among *all* subsets of $\{1, \dots, p\}$, which has the effect of also doing the model selection in the low-dimensional regime.

These estimates would (for components in the support set) be biased, but would have zero variance (for the estimated component) and thus would actually be better in the sense of MSE. We echo the sentiment of [3] who say that this is essentially a bias-variance tradeoff.

This actually produces a better oracle estimator (and thus a more powerful oracle inequality), despite the fact that the oracle no longer actually needs to know the support set. If we imagine a simple case where the design matrix is the identity, an estimator obtained from this process would have MSE

$$\sum_{i=1}^{p} \min(\beta_i^2, \sigma^2) \leq S\sigma^2$$

In the case of general design X, [3] also show that the analogous risk in this setting is lower bounded by this quantity, up to a multiplicative constant, whenever the UUP holds.

Theorem 2 then provides a comparison between the *Dantzig selector* (i.e. the specific case of (1) with $\lambda_p = (1 + t^{-1})\sqrt{2 \log p}$) and such oracle estimates.

Theorem 2 (Candes, Tao 2007)**.** *Let $\beta \in \mathbb{R}^p$ be supported on S and $s = |S|$. Let the UUP (2) hold. Choose $t > 0$ such that $\delta_s + \theta_{s,2s} < 1 - t$. Let $\lambda_p = (1 + t^{-1})\sqrt{2 \log p}$. Then, with high probability[4], the estimator $\hat{\beta}$ solving the procedure (1) satisfies*

$$\|\hat{\beta} - \beta\|_2^2 \leq C_2^2 \lambda_p^2 \left(\sigma^2 + \sum_{i=1}^{p} \min(\beta_i^2, \sigma^2) \right)$$

where the exact constant C_2 depends only on δ_s and $\theta_{s,2s}$ and, when these quantities are small, is approximately equal to 16.

3 Comparison to other methods

The Dantzig selector performs statistical estimation for linear models in addition to variable selection. Most importantly it aims to solves cases in which the number of covariates is far greater than the number of samples [3]. Generally, the variable selection problem poses a trade-off between the goodness-of-fit and the complexity of a model which is not only tackled by the Dantzig selector. Many other techniques also provide shrinkage regression models that aim to achieve the same trade-off through regularization. In the following section the benefits and disadvantages of the Dantzig selector are compared to other variable selection strategies, such as the LASSO, the DASSO and the LARS algorithm [8, 7, 5].

3.1 LASSO

The LASSO regression model proposes another parameter estimation technique for linear models in addition to performing variable selection. In general, the LASSO minimizes the residual sum of squares $RSS = \|Y - X\tilde{\beta}\|_2^2$ while subject to regularizing the sum of absolute values of coefficients $\|\tilde{\beta}\|_1$ as shown by

$$\min_{\tilde{\beta} \in \mathbb{R}^p} \|Y - X\tilde{\beta}\|_2^2 \quad \text{subject to} \quad \|\tilde{\beta}\|_1 \leq \lambda_L, \tag{3}$$

where λ_L is a tuning parameter [8]. On the other hand, the Dantzig selector regularizes the sum of absolute coefficients by minimizing the maximum component of the gradient of the residual sum of squares $X^T(Y - X\tilde{\beta})$ [6].

The relationship between the regression shrinkage methods, the LASSO and the Dantzig selector, can be defined in such a way that that under certain conditions of the design matrix X and with a given tuning

[4]The probability of failure is as in Theorem 1.

parameter λ the LASSO parameter estimates $\hat{\beta}_L$ and the estimated Dantzig selector coefficients $\hat{\beta}_{DS}$ are equivalent. Hence, the non-asymptotic bounds derived for the Dantzig selector can be extended to the LASSO in those scenarios [3, 7]. The proposed theoretical properties of the Dantzig selector are favorable as the non-asymptotic bounds are sharp on the L_2-error of the estimated parameter coefficients and furthermore constrain the L_2-error to a factor of $\log(p)$ of the error achieved by an oracle estimator which has knowledge of the nonzero entries of $\tilde{\beta}$ [7].

Parameter estimation by the Dantzig selector yields reliable estimates for the coefficients $\tilde{\beta}$ where the number of covariates p is higher than the sample size n [3]. However, comparing the Dantzing selector with the LASSO in terms of predictive accuracy has shown that the Dantzig selector results in consistently higher root mean squared error values for $\|\hat{\beta}\|_1$ and λ_L or λ_{DS} in a simulation setting where $p \gg n$ and the unknown parameter vector is sparse. This effect is not as severe for simulation studies where $p \ll n$ and the unknown parameter vector is sparse as well as for the scenario where $p \gg n$ and the unknown parameter vector is dense [6]. It should be noted that the authors of the Dantzig selector propose a bias correcting variation of the Dantzig selector which provides a higher statistical accuracy. The so-called Gaussian-Dantzig selector supposedly attenuates the soft-thresholding behavior of the Dantzig selector which is the reason for the persistent underestimation of the true values of the unknown parameters [3].

Lastly, the LASSO is a quadratic program that has a piecewise linear path in terms of computation of the parameter estimates. Furthermore, if the LASSO path is computed with the LARS (least-angle regression) algorithm, then the computational cost of calculating the entire path of all variables for all given λ_L is equivalent to solving a single least squares problem with all covariates included in the regression model [5, 6]. On the other hand, the Dantzig selector can be be formulated as a linear program with a piecewise linear path. However, the calculation of the entire path is computationally expensive as it requires the computation of a fine grid of values for λ_{DS}. It should be noted that solving a regression model via the Dantzig selector given λ_{DS} is computationally efficient [3].

3.2 DASSO

The DASSO aims to redefine the Dantzig selector with a computational efficiency similar to LARS. The entire piecewise linear coefficient path of the Dantzig selector is therefore calculated by a sequential simplex-like algorithm which eases the computational burden of having to compute the entire coefficient path for a fine grid of λ_{DS} by solving multiple linear programming problems. This modification of LARS identifies break points and hence solves less linear programs in the form of a sequential optimisation problem. Hence, the DASSO proposes an efficient method of computing a high-dimensional regression model which achieves variable selection as well as accurate parameter estimation in addition to a theoretical justification of error bounds on the L_2-error as previously described in the remarks of Section 3.1 [7].

Implementing DASSO

We implement the DASSO in Python from scratch, as we are unable to find a working implementation. The algorithm in our code is almost identical to that described in [7], which we omit for brevity, with two important distinctions:

Floating-point numerical error Similar to most numerical algorithms, the DASSO, which involves matrix multiplication and inversion, is subject to floating-point numerical errors [1]. The errors may lead to variables being unintentionally dropped from the set of active variables $\mathcal{A}$ (the maximum component(s) of the gradient of the residual sum of squares), or added to the set of variables with a non-zero coefficient $\mathcal{B}$. The DASSO relies on these two sets of variables to decide on which direction and how far to move, and tracking the incorrect sets may lead to the algorithm returning a suboptimal solution.[5] In our implementation, we

[5]Most of the time the program just crash due to dimension mismatch in the matrices being multiplied.

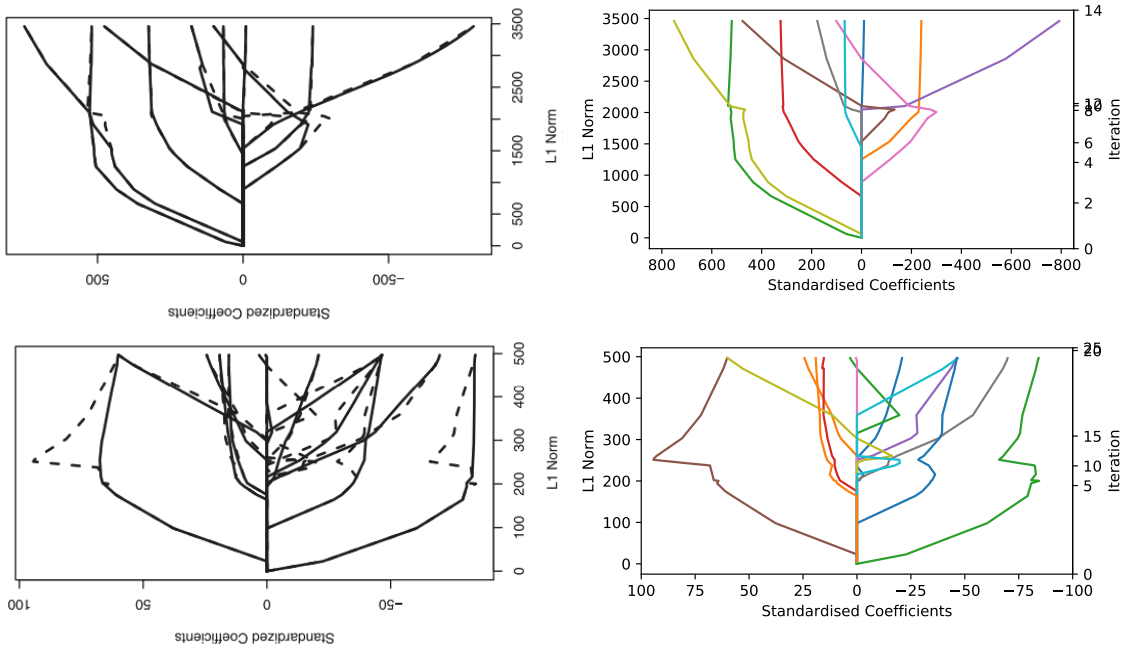

Figure 1: Coefficient paths generated by DASSO on (top) diabetes data and (bottom) Boston housing data. The paths for the datasets in our implementation (right) are comparable to that of the Dantzig selector presented in [7] (left, dashed lines).

specify that two values are equal if they are less than a pre-specified ε apart — we consider $x \neq 0$ if $|x| > \varepsilon$ and c as maximal if $c > \max_c(c) - \varepsilon \ \forall x, c$ featured in the algorithm.

Distance step for the final iterations The distance step calculates how far one should travel along the optimal direction of that iteration, which is the minimum distance required for any variable to either enter the active set $\mathcal{A}$ (γ_1) or have its coefficient crossing zero and hence leaving $\mathcal{B}$ (γ_2).[6] We notice γ_1 is not defined when every variable is in $\mathcal{A}$.[7] This is common in problems where $n > p$, where we expect a unique OLS estimator with all p variables present. A natural candidate for γ_1 in this case is $+\infty$. However, this means the distance is now purely determined by γ_2, which is constructed to reach the next point with a zero coefficient but not the optimal point. We observe that with $\gamma_1 = +\infty$ the algorithm simply cycles between a number of points with zero coefficients indefinitely once every variable is in $\mathcal{A}$.

We follow [5] and set $\gamma_1 = ||\mathbf{c}||_\infty$ when all variables are in $\mathcal{A}$, where $\mathbf{c} = X^T(Y - X\tilde{\beta})$ measures the correlation between the variables and the residual. Such value ensures the algorithm terminates by bringing $\mathbf{c}$ to zero in one iteration if $\gamma_1 < \gamma_2$. If $\gamma_1 > \gamma_2 \neq 0$, we note that γ_1 strictly decreases in successive iterations as the correlations are strictly reduced towards zero by construction, and eventually the $\gamma_1 < \gamma_2$ case applies.

We plot the coefficient paths (the L_1 norm vs. coefficient of each component in $\tilde{\beta}$) generated when we run DASSO on two datasets featured in [5, 7]. The first is a diabetes dataset with $p = 10$ variables containing measurements for $n = 442$ patient, where the aim is to predict a measure of progression of the disease in a year's time; and the second is a Boston housing data set with $p = 13$ predictors for $n = 506$ properties, where the aim is to predict the house value. We observe the paths for the two datasets are identical to the figures in [7] and conclude that our DASSO implementation is comparable to that in the original paper.

[6] See Appendix B of [7] for details.

[7] This makes sense as no variable will enter $\mathcal{A}$ if every variable is in $\mathcal{A}$, regardless of the distance travelled.

References

[1] David Bindel and Jonathan Goodman. *Principles of scientific computing.* 2009.

[2] Emmanuel Candes and Terence Tao. Decoding by linear programming. *IEEE transactions on information theory*, 51(12):4203–4215, 2005.

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

[4] Emmanuel J Candes and Terence Tao. Near-optimal signal recovery from random projections: Universal encoding strategies? *IEEE transactions on information theory*, 52(12):5406–5425, 2006.

[5] Bradley Efron, Trevor Hastie, Iain Johnstone, and Robert Tibshirani. Least angle regression. *The Annals of Statistics*, 32(2):407–451, 2004.

[6] Bradley Efron, Trevor Hastie, and Robert Tibshirani. Discussion: The dantzig selector: Statistical estimation when p is much larger than n. *Ann. Statist.*, 35(6):2358–2364, 2007.

[7] Gareth James, Peter Radchenko, and Jinchi Lv. Dasso: Connections between the dantzig selector and lasso. *Journal of the Royal Statistical Society Series B*, 71:127–142, 2009.

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