Regularised LS

Jack Lucchetti

26th May 2025 [preliminary and incomplete]

Disclaimer: this is just a set of lecture notes that follow more or less what we did in class. For the real thing, you may want to have a look at Efron and Hastie (2016) or Hastie et al. (2015).

1 Prediction and the bias-variance tradeoff

In most of statistical inference, the chief quality of a statistic, such as an estimator, is that its distribution should be centred around the target quantity. This is why we value properties such as unbiasedness and consistency. Sometimes, however, we may want to adopt a differenty take. This typically happens in prediction problems, when we may want to trade bias in exchange for smaller variance. For example: consider a situation where you have two predictors, $\hat{y}$ and $\tilde{y}$, for the same unobservable quantity, whose real value is 2. Figure 1 shows their densities. The first one, $\hat{y}$, is unbiased, but has a large variance; the other one, $\tilde{y}$, is biased but less dispersed. Which one would you pick?

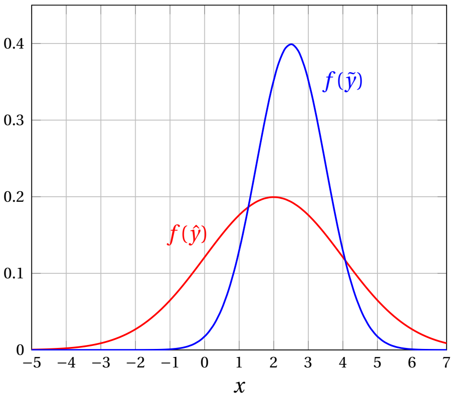


Figure 1: Two predictors

In most cases, you'd want to use $\tilde{y}$, for the very simple reason that your prediction errors will be biased (not 0 on average), but the probability of making big mistakes (say, predicting y to be outside the $[0,4]$ range) is much smaller than the one you'd get by using $\hat{y}$.

Let's formalise this: suppose you have a predictor $\hat{y}$ of a quantity y, with $\mathrm{E}\left[\hat{y}\right] = m$. Of course it would be nice if $m = y$, but its variance must also be considered. What we really care about is *how large the prediction error is going to be*. We measure this via the expected value of the squared prediction error, also known as MSE (Mean Square Error):

$$\begin{aligned} MSE &= \mathrm{E}\left[(y-\hat{y})^2\right] = \\ &= \mathrm{E}\left[[(y-m)-(\hat{y}-m)]^2\right] = \mathrm{E}\left[y-m\right]^2 + \mathrm{V}\left[\hat{y}\right] = \\ &= \text{Bias}^2 + \text{Variance} \end{aligned}$$

which is the sum of the square of the bias and the variance.[1] In many cases in practice, one uses its square root, the "Root Mean Square Error":

$$RMSE = \sqrt{\mathrm{E}\left[(y-\hat{y})^2\right]},$$

but the principle remains the same.

When prediction is about something that can be represented via a DGP, one would think that the best predictor would be a function of the most efficient among all consistent estimators. In fact, it needn't be so: in some cases, we may accept some bias in return for a smaller variance and end up with a predictor with a smaller RMSE. In many cases, this can be achieved via *shrinkage*.

2 The precursor: James-Stein

One of the most egregious examples of shrinkage in practice was provided in James and Stein (1961). The problem seems somewhat artificial, but provides a surprising example for a seemingly innocuous and straightforward problem.

Imagine your m-dimensional DGP is $\mathbf{y} \sim \mathcal{N}\left(\boldsymbol{\mu}, \sigma^2 I\right)$, with σ known, and you have only one observation, so $n = 1$. How do you estimate $\boldsymbol{\mu}$? Of course, the most natural estimator of $\boldsymbol{\mu}$ is the sample average, which coincides with your only observation, $\mathbf{y}$, if $n = 1$, so it would seem natural to use that.[2] However, the statistic

$$\tilde{\boldsymbol{\mu}}_{JS} = \left(1 - \frac{(m-2)\sigma^2}{\mathbf{y}'\mathbf{y}}\right)\mathbf{y}.$$

can be proven to dominate the ML in terms of RMSE for $m > 3$, which may come a bit as a surprise. Note that $\tilde{\boldsymbol{\mu}}_{JS}$ is just the "obvious" estimator multiplied by a scalar that, in most cases, is between 0 and 1. Hence the idea of *shrinkage*.

The problem above sounds somehow artificial: however it can be made more realistic by conceding that the variance is unknown or that you have more than

[1] If the third equality leaves you perplexed, remember that $\mathrm{E}\left[\hat{y}-m\right] = 0$ by definition.

[2] Using the sample average has an obvious justification as a method-of-moments estimator, but it's also the maximum likelihood estimator.

one observation, but the result still stands.[3] This gives you the idea that shrinkage towards zero could actually help in terms of RMSE.

In the context of linear models like

$$y_i = \mathbf{x}_i \boldsymbol{\beta} + \varepsilon_i, \tag{1}$$

an extension of the James-Stein idea was proposed in Copas (1983), but is not particularly popular. The shrinkage estimators that practitioners use most often are based on optimizing an objective function that balances (a) the sum of squared residuals and (b) the "size" of the elements of the estimator vector $\tilde{\boldsymbol{\beta}}$. Since there are different ways to define "size", you get different estimators. When the dimension of $\mathbf{x}_i$ is possibly *very* large, shrinkage can be very effective in forecasting.

3 Model selection via classical methods

In a way, one could think that a crude way to perform shrinkage is by data-based model selection. Given the usual linear model

$$y_i = \mathbf{x}_i \boldsymbol{\beta} + \varepsilon_i,$$

where $\mathbf{x}_i$ is a k-element vector, it is easy to see that the number of different model one can have by selectively excluding a subset of regressors (that is, fixing the corresponding coefficient to 0) equals 2^k . For example, if $\mathbf{x}'_i = [a_i, \quad b_i, \quad c_i]$, so $k = 3$, the number of different models you can select is $2^3 = 8$:

	1	2	3	4	5	6	7	8
a_i		✓			✓	✓		✓
b_i			✓		✓		✓	✓
c_i				✓		✓	✓	✓

When you choose one of the possible 2^k alternatives, you're implicitly shrinking some of the coefficients to 0 (the ones for the excluded variables), while leaving the others unrestricted. The idea is to make this decision on the basis of the available data.

It is quite obvious that, in a real-life problem, a complete exploration of the model space is unfeasible. For example, suppose that we'd like to choose via the "best" model among the possible alternatives by selecting the one with the lowest BIC. Suppose also that the CPU time to compute the BIC for a given model is one thousandth second, so you can evaluate 3600 × 1000 models in one hour. If $k = 30$, the CPU time to evaluate all possible models is

$$\text{time} = \frac{2^{30}}{3600 \times 1000} = 298.26,$$

[3] See the Wikipedia page for more details.

which is nearly two weeks of non-stop computation. You wouldn't want to waste all that electricity, would you?

There are techniques for speeding up the process, which are known as "stepwise" regressions. The main two variants are the so-called "backward" and "forward" regressions. With the former, you start with the full model and keep dropping the least significant regressor until all the remaining ones are significant at a given level (usually, 10%). With the latter, you start from a minimal model (typically, constant only), and then you select the explanatory variable whose inclusion reduces the SSR the most;[4] then you keep going until none of the remaining regressors enhances the fit of your model significantly.

These two techqniques may or may not arrive at the same model, but are generally quite effective and perform rather well in out-of-sample forecasting. You must be careful on using them as model selection tools if you need hypothesis testing, since it can be proven that the usual OLS formulae for models whose regressors have been selected via a stepwise procedure need some adjustments. However, this is not a big problem in a prediction scenario.

4 Ridge regression

The so-called **ridge** estimator has a long history in econometrics: it was used, originally, to to avoid collinearity and handle cases when "reasonable data collection results in an $\mathbf{X}'\mathbf{X}$ with one or more small eigenvalues" (Hoerl and Kennard, 1970, p. 56). This means, in practice, in cases when the sample size n is not much larger than the number of regressors k and/or the columns of $\mathbf{X}$ are very collinear. In these cases, the smallest eigenvalue of $\mathbf{X}'\mathbf{X}$ is near zero, so $(\mathbf{X}'\mathbf{X})^{-1}$ becomes a matrix with huge numbers.

The idea is to boost $\mathbf{X}'\mathbf{X}$ away from near-singularity by adding a positive scalar λ to its diagonal:

$$\tilde{\beta} = \left[\mathbf{X}'\mathbf{X} + \lambda I\right]^{-1} \mathbf{X}'\mathbf{y} \tag{2}$$

In fact, it's a minimiser: it can be proven that $\tilde{\beta}$ can be defined as

$$\tilde{\beta} = \underset{\beta'\beta = t}{\text{Argmin}}\ \mathbf{e}'\mathbf{e}, \tag{3}$$

which is equivalent to[5]

$$\tilde{\beta} = \underset{\beta \in \mathbb{R}^k}{\text{Argmin}}\ \mathbf{e}'\mathbf{e} + \lambda \cdot ||\beta||_2^2 \tag{4}$$

where the scalar λ controls the amount of shrinkage and $||\beta||_2^2$ is a fancy way of writing $\beta'\beta = \sum_{i=1}^k \beta_i^2$ (see Section A.1 for more details). The $||\beta||_2^2$ term is often called an "ℓ_2 penalty" term.

[4] There is a clever algorithm for computing this quickly.

[5] If you don't believe me, write the Lagrangean for the problem (3).

Note that in a linear model each element of the vector β is related to the unit of measurement of the corresponding regressor. Hence, the solution to the problem above changes if we decide to measure a certain regressor in metres, centimetres, feet or inches. To get rid of this potential ambiguity, data in $\mathbf{X}$ are usually standardised, and we will assume from here on that the all elements on the diagonal of $\mathbf{X}'\mathbf{X}$ are equal to n. Re-tranforming $\hat{\beta}$ to any unit you want is clearly trivial.

Of course, when $\lambda = 0$ you get OLS: $\hat{\beta} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$. For $\lambda \to \infty$, $\tilde{\beta} \to \mathbf{0}$ (but is never actually 0). By rewriting (2) as

$$\tilde{\beta} = \left[\mathbf{X}'\mathbf{X} + \lambda I\right]^{-1} (\mathbf{X}'\mathbf{X})\hat{\beta} = T\hat{\beta}$$

you see that $\tilde{\beta}$ is a linear transformation of $\hat{\beta}$, in which the matrix T is $(\mathbf{X}'\mathbf{X})$ "divided by" a matrix that's "bigger" than $(\mathbf{X}'\mathbf{X})$, and therefore is shrunk towards $\mathbf{0}$ (in fact, if $\mathbf{X}'\mathbf{X}$ is a scalar matrix, $\tilde{\beta}$ is a *scalar* multiple of $\hat{\beta}$).

The ridge estimator could be also given an "added observations" interpretation:[6]

$$\tilde{\mathbf{X}} = \begin{bmatrix} \mathbf{X} \\ \sqrt{\lambda} I \end{bmatrix} \qquad \tilde{\mathbf{y}} = \begin{bmatrix} \mathbf{y} \\ \mathbf{0} \end{bmatrix}$$

and $\tilde{\beta}$ is just OLS on the above.

Note that shrinkage is primarily used as a means to improve the efficiency (in terms of RMSE) of *predictions*. We know that $\tilde{\beta}$ is biased and inconsistent and using it for inference purposes is rather complicated. This is the reason why, although it is possible in principle to calculate the covariance matrix for $\tilde{\beta}$, this is almost never done.

Moreover, a quick and neat numerical solution is available using the singular value decomposition (SVD for short):[7]

$$\mathbf{X} = U\langle \mathbf{d}\rangle V,$$

where I am using the notation $\langle \mathbf{d}\rangle$ to indicate a diagonal matrix, with the vector $\mathbf{d}$ on the diagonal. Therefore,

$$\begin{aligned} \mathbf{X}'\mathbf{X} + \lambda I &= V'[\langle \mathbf{d}\rangle^2 + \lambda I]V \\ (\mathbf{X}'\mathbf{X} + \lambda I)^{-1} &= V'[\langle \mathbf{d}\rangle^2 + \lambda I]^{-1}V \end{aligned}$$

and therefore

$$\tilde{\beta} = V'[\langle \mathbf{d}\rangle^2 + \lambda I]^{-1}VV'\langle \mathbf{d}\rangle U'\mathbf{y} = V'\langle \mathbf{h}\rangle U'\mathbf{y}$$

where $h_i = \frac{d_i}{d_i^2+\lambda}$. This is useful because, given U and V (whose computation is fast) you can compute $\tilde{\beta}$ for any λ with no matrix inversion (nice).

This is also useful for computing the "effective degrees of freedom", which are then used as an ingredient in optimising λ via AIC, BIC or Mallows' C_p. This

[6] Bayesian, really, but I digress.

[7] See section A.2.

number can be seen as a generalisation of the concept of "number of parameters". Of course the number of elements of $\tilde{\beta}$ is k, but we have to take into account that $\tilde{\beta}$ is the outcome of an optimisation strategy in which the optimal value for $\tilde{\beta}$ is shrunk towards 0, so it is somewhat constrained: in the limiting case when $\lambda \to \infty$, $\tilde{\beta} = \mathbf{0}$, so the number of "effective" parameters would actually be 0.

A neat way to see this is by considering the model's fitted values:

$$\tilde{\mathbf{y}} = \mathbf{X}\tilde{\beta}.$$

If we take the extreme case $\lambda = 0$ then ridge is OLS, and therefore $\tilde{\mathbf{y}} = \mathbf{P_X y}$. In this case, we can think of the fitted value as the projection of $\mathbf{y}$ onto the space spanned by the columns of $\mathbf{X}$.

A neat result in matrix algebra (which I'm not proving here) is that the dimension of the space pertaining to a given projection matrix can be recovered by simply computing its trace, that is the sum of its diagonal elements. The dimension of $\mathrm{Sp}(\mathbf{X})$ is therefore equal to[8]

$$\mathrm{tr}(\mathbf{P_X}) = \mathrm{tr}\left(\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\right) = \mathrm{tr}\left((\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X}\right) = \mathrm{tr}(I) = k.$$

For $\lambda > 0$, the fitted values can be computed by extending slightly the usual notation and defining

$$\mathbf{P}_{\mathbf{X},\lambda} = \mathbf{X}\left[\mathbf{X}'\mathbf{X} + \lambda I\right]^{-1}\mathbf{X}'$$

so that

$$\tilde{\mathbf{y}} = \mathbf{X}\left[\mathbf{X}'\mathbf{X} + \lambda I\right]^{-1}\mathbf{X}'\mathbf{y} = \mathbf{P}_{\mathbf{X},\lambda}\mathbf{y}.$$

Note that, for $\lambda > 0$, $\mathbf{P}_{\mathbf{X},\lambda}$ is symmetric but *not* idempotent. We define the "effective number of parameters" as the trace of $\mathbf{P}_{\mathbf{X},\lambda}$, so the case $\lambda = 0$ (that is, OLS) would just be a special case. By using the singular value decomposition, this is easily computed as

$$\mathrm{tr}\left(\mathbf{X}(\mathbf{X}'\mathbf{X} + \lambda I)^{-1}\mathbf{X}'\right) = \mathrm{tr}\left(U\langle\mathbf{d}\rangle VV'\langle\mathbf{h}\rangle U'\right) = \mathrm{tr}(\langle\mathbf{d}\rangle\langle\mathbf{h}\rangle) = \mathrm{tr}\left(\langle\psi\rangle\right) = \boldsymbol{\iota}'\psi$$

where $\psi_i = \frac{d_i^2}{d_i^2+\lambda}$: hence,

$$edf = \sum_{i=1}^{k} \frac{d_i^2}{d_i^2 + \lambda}. \tag{5}$$

Clearly, this number is k for $\lambda = 0$, and it's a decreasing function of λ, that tends to 0 as $\lambda \to \infty$.

The practical consequence of this result is that if you choose λ via criteria, such as the BIC, in which the number of parameters of the model counts, you must make sure you're using the quantity above instead of k.

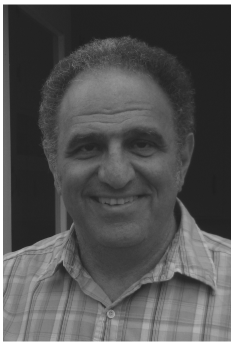
[8] The key algebra trick needed here is $\mathrm{tr}(AB) = \mathrm{tr}(BA)$.

5 LASSO

The LASSO is an acronym for “least absolute shrinkage and selection operator”, and was first proposed in Tibshirani (1996): the basic idea is similar to the one used in ridge regression, with the difference that the penalty term uses the ℓ_1 metric instead.

$$\tilde{\beta} = \underset{\sum_j |\beta_j| = t}{\text{Argmin}} \ \mathbf{e}'\mathbf{e} = \underset{\beta \in \mathbb{R}^k}{\text{Argmin}} \ \mathbf{e}'\mathbf{e} + \lambda ||\beta_j||_1. \tag{6}$$

Like for the ridge estimator, the columns of **X** must be standardised for obvious reasons.



ROBERT TIBSHIRANI

Typically, in the optimal solution some of the elements of $\tilde{\beta}$ are 0 when λ is away from 0, so the advantage is that you perform shrinkage *and* model selection at the same time. This feature has helped the LASSO become enormously popular in the past 20 years.

In fact, contrary to what happens with the ridge estimator, with the ℓ_1 penalty term you have a (data-dependent) value of λ, above which $\tilde{\beta} = \mathbf{0}$. Therefore, if we call this $\bar{\lambda}$, it only makes sense to consider values of λ between 0 (where you get $\tilde{\beta} = \hat{\beta}$) and $\bar{\lambda}$ (where you get $\tilde{\beta} = \mathbf{0}$).

Another difference from the ridge estimator is that you don’t have a closed-form solution, so $\tilde{\beta}$ has to be found by numerical methods. One of the most popular is ADMM. These methods are usually quite fast and robust, so this is unlikely to be a problem in practice.

It would be nice if the LASSO had the oracle property, which is basically the property by which the procedure

- selects the non-zero coefficients correctly, and
- has an asymptotic distribution that is the same you’d have if the true structure of the model was known in advance.

Unfortunately, it doesn’t; however, there is a variant, called the adaptive LASSO by Zou (2006), that does.

5.1 Choice of λ in LASSO estimation

The existence of an upper bound for λ in a LASSO model makes it very natural to think in terms of $s = \lambda/\bar{\lambda}$, which is between 0 and 1, and try a grid of values. Then, you choose one either via the BIC or via cross-validation (fold-wise).

In other words, you estimate β by solving problem (6) for different values of s, usually arranged in a logarithimic grid, eg

$$s = [\quad 0.001, \quad 0.01, \quad 0.1, \quad 1 \quad]$$

and you compute your criterion of choice for each.

You then choose s (and hence λ) by picking the values that yield the best criterion, or by a somewhat more conservative approach known as the "one standard deviation rule" (see Hastie et al., 2009, sec. 7.10.1).

6 Elastic net

$$\tilde{\beta} = \underset{\beta \in \mathbb{R}^k}{\text{Argmin}}\ \mathbf{e}'\mathbf{e} + \lambda \left[\frac{1-\alpha}{2} ||\beta||_2^2 + \alpha ||\beta||_1 \right]$$

Thus $\alpha = 1$ gives LASSO, $\alpha = 0$ gives Ridge, and anything between gives a combination. The elastic net is often seen as a very good approximation to the problem

$$\tilde{\beta} = \underset{||\beta||_p = t}{\text{Argmin}}\ \mathbf{e}'\mathbf{e}$$

for $1 \leq p \leq 2$, that is, something in between tthe ridge and lasso estimators. In practice, for $\alpha = 0.5$, which is a value used quite often, you get an estimator that contains quite a few zeros, but not as many as you would get with the lasso.

Quoting from (Efron and Hastie, 2016, p. 316):

> When the predictors are excessively correlated, the LASSO performs somewhat poorly, since it has difficulty in choosing among the correlated cousins. Like ridge regression, the elastic net shrinks the coefficients of correlated variables toward each other, and tends to select correlated variables in groups.

So the elastic net seems to perform rather well in real-life problems.

A Assorted results

A.1 Norms

How long is a vector? In other words, given a point in a k-dimensional space, how do we define its distance from the origin? The usual Euclidean norm is defined as

$$||\mathbf{x}|| = \sqrt{\mathbf{x}'\mathbf{x}} = \left[\sum_{i=1}^{k} x_i^2 \right]^{1/2}.$$

It turns out that in some contexts it is useful to generalise the expression above by considering

$$||\mathbf{x}||_p = \left[\sum_{i=1}^{k} |x_i|^p \right]^{1/p}, \tag{7}$$

where $p \geq 0$. It is possible to verify that the expression above, known as ℓ_p-norm, satisfies all the formal requirements that a distance must have. Of course, when $p = 2$ you have Euclidean distance, but there are two more interesting cases.[9]

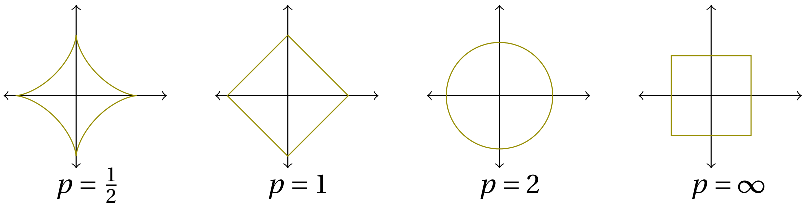


Figure 2: Shape of a circle for various values of p

When $p = 1$, the norm is simply the sum of the absolute values. The limiting case when $p \to \infty$ is also interesting, because it gives you the the maximum element (in absolute value) of $\mathbf{x}$. So, for example, if $\mathbf{x} = [-3, 0, 4, 0]'$

$$||\mathbf{x}||_1 = |-3| + |4| = 7, \quad ||\mathbf{x}||_2 = \sqrt{3^2 + 4^2} = 5, \quad ||\mathbf{x}||_\infty = 4.$$

If you take the definition of a circle as "the set of points at the same distance from the origin", then Figure 2 shows you what a circle looks like for various values of p.

A.2 The SVD

Consider an $r \times c$ matrix A with rank k. With no loss of generality, let's say that

$$0 \leq k \leq c \leq r,$$

so A is a "tall" (but possibly square) matrix, whose rank could be non-full. Then, it is always possible to express A as

$$A = U \langle \mathbf{d} \rangle V \tag{8}$$

where

1. U is a $r \times c$ matrix, with $U'U = I$.
2. $\langle \mathbf{d} \rangle$ is a $c \times c$ diagonal matrix; the vector on the diagonal $\mathbf{d}$ contains k positive entries and $c - k$ zeros.
3. V is a $c \times c$ matrix, with $VV' = V'V = I$.

[9] Readers with a *penchant* for microeconomics will doubtlessly recognise the similarity with the CES production funciton.

A few fun facts: the generalised inverse of A equals as $A^+ = U\langle \mathbf{d}^+ \rangle V$, where $\mathbf{d}^+$ is a a vector such that

$$d_i^+ = \begin{cases} 0 & \text{if } d_i = 0 \\ 1/d_i & \text{if } d_i > 0 \end{cases}.$$

Additionally, it can be proven quite easily that the scalars d_i are the square roots of the eigenvalues of $A'A$. Finally, the OLS statistic can be written as $\hat{\beta} = \mathbf{X}^+\mathbf{y}$. Cool, huh?

References

COPAS, J. B. (1983): "Regression, Prediction and Shrinkage," *Journal of the Royal Statistical Society: Series B (Methodological)*, 45, 311–335.

EFRON, B. AND T. HASTIE (2016): *Computer Age Statistical Inference: Algorithms, Evidence, and Data Science*, Cambridge University Press, 1st ed.

HASTIE, T., R. TIBSHIRANI, AND J. FRIEDMAN (2009): "The elements of statistical learning: data mining, inference, and prediction," .

HASTIE, T., R. TIBSHIRANI, AND M. WAINWRIGHT (2015): *Statistical learning with sparsity*, 143, CRC press.

HOERL, A. E. AND R. W. KENNARD (1970): "Ridge regression: Biased estimation for nonorthogonal problems," *Technometrics*, 12, 55–67.

JAMES, W. AND C. STEIN (1961): "Estimation with Quadratic Loss," in *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Contributions to the Theory of Statistics*, Berkeley, Calif.: University of California Press, 361–379.

TIBSHIRANI, R. (1996): "Regression Shrinkage and Selection via the Lasso," *Journal of the Royal Statistical Society. Series B (Methodological)*, 58, 267–288.

ZOU, H. (2006): "The adaptive lasso and its oracle properties," *Journal of the American statistical association*, 101, 1418–1429.