Breakdown Point of Model Selection When the Number of Variables Exceeds the Number of Observations

David Donoho and Victoria Stodden

Abstract **The classical multivariate linear regression problem assumes p variables $X_1, X_2, \ldots, X_p$ and a response vector y, each with n observations, and a linear relationship between the two: $y = X\beta + z$, where $z \sim N(0, \sigma^2)$. We point out that when $p > n$, there is a *breakdown point* for standard model selection schemes, such that model selection only works well below a certain critical complexity level depending on n/p. We apply this notion to some standard model selection algorithms (Forward Stepwise, LASSO, LARS) in the case where $p \gg n$. We nd that 1) the breakdown point is well-de ned for random X-models and low noise, 2) increasing noise shifts the breakdown point to lower levels of sparsity, and reduces the model recovery ability of the algorithm in a systematic way, and 3) below breakdown, the size of coef cient errors follows the theoretical error distribution for the classical linear model.**

I. Introduction

The classical multivariate linear regression problem postulates p variables $X_1, X_2, \ldots, X_p$ and a response vector y, each with n observations, and a linear relationship between the two: $y = X\beta + z$, where X is the $n \times p$ model matrix and $z \sim N(0, \sigma^2)$. The vector of coef cients, β, is estimated by $(X'X)^{-1}X'y$. Hence the classical model requires $p \leq n$. Developments in many elds, such as genomics, nance, data mining, and image classi cation, have pushed attention beyond the classical model, to cases where p is dramatically larger than n.

II. Estimating the Model when $p >> n$

George Box coined the term *Effect Sparsity* [1] to describe a model where the vast majority of factors have zero effect

only a small fraction actually affect the response. If β is *sparse*, ie. containing a few nonzero elements, then $y = X\beta + z$ can still be modeled successfully by exploiting sparsity, even when the problem is underdetermined in the classical sense.

A number of strategies are commonly used to extract a sparse model from a large number of potential predictors: all subsets regression (t all possible linear models for all levels of sparsity), forward stepwise regression (greedily adds terms to the model sequentially, by signi cance level), LASSO [8] and Least Angle Regression [7] ('shrinks' some coef cient estimates to zero).

David Donoho is with the Department of Statistics, Stanford University, Stanford, CA 94305-4065, USA (email: donoho@stanford.edu).

Victoria Stodden is with the Department of Statistics, Stanford University, Stanford, CA 94305-4065, USA (email: vcs@stanford.edu).

A. Ideas from Sparse Representation

We mention some ideas from signal processing that will allow us to see that, in certain situations, statistical solutions such as LASSO or Forward Stepwise, are just as good as all subsets regression.

When estimating a sparse *noiseless* model, we would ideally like to nd the sparsest solution:

$$min_\beta ||\beta||_0 \ s.t. \ y = X\beta. \tag{1}$$

This is intuitively compelling. Unfortunately it is not computationally feasible since it requires an all-subsets search. Basis Pursuit [3], was pioneered for sparse representation in signal processing, and solves:

$$min_\beta ||\beta||_1 \ s.t. \ y = X\beta; \tag{2}$$

this is a convex optimization problem. Surprisingly, under certain circumstances the solution to (2) is also the solution to (1) [4]. More precisely, there is a threshold phenomenon or breakdown point such that, provided the sparsest solution is suf ciently sparse, the ℓ_1 minimizer is precisely that sparsest solution.

The LASSO model selection algorithm solves:

$$min_\beta ||y - X\beta||_2^2 \ s.t. \ ||\beta||_1 \leq t \tag{3}$$

for a choice of t. (The Least Angle Regression (LARS) algorithm provides a stepwise approximation to LASSO.) For an appropriate choice of t, problem (3) describes the same problem as equation (2). This suggests a course of investigation: we may be able to cross-apply the equivalence of equations (1) and (2) to the statistical model selection setting. Perhaps we can observe a threshold in behavior such that for suf ciently sparse models, traditional model selection works, while for more complex models the algorithm's ability to recover the underlying model breaks down.

B. The Phase Transition Diagram

In the context of ℓ_0/ℓ_1 equivalence, Donoho [4] introduced the notion of a *Phase Diagram* to illustrate how sparsity and indeterminacy affect success of ℓ_1 optimization. One displays a performance measure of the solution as a function of the level of underdeterminedness n/p and sparsity level k/n, often with very interesting outcomes.

Figure 1 has been adapted from [4]. Each point on the plot corresponds to a statistical model for certain values of

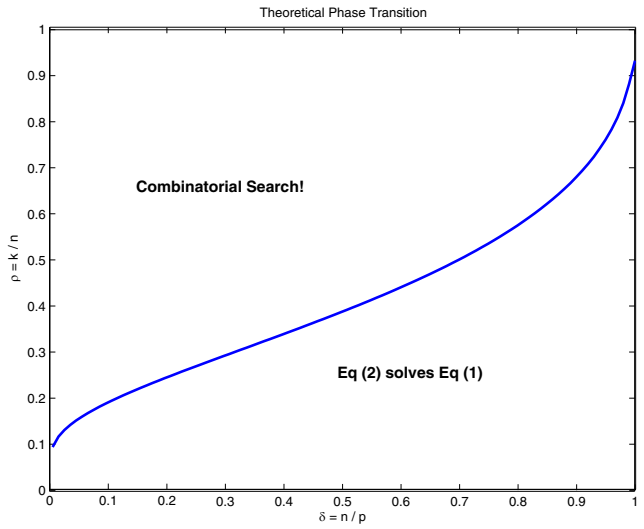

Fig. 1. The theoretical threshold at which the l_1 approximation to the l_0 optimization problem no longer holds. The curve delineates a Phase Transition from the lower region where the approximation holds, to the upper region, where we are forced to use combinatorial search to recover the optimal sparse model. Along the x-axis the level of underdeterminedness decreases, and along the y-axis the level of sparsity of the underlying model increases.

n, p, and k. The abscissa runs from zero to one, and gives values for $\delta = \frac{n}{p}$. The ordinate is $\rho = \frac{k}{n}$, measuring the level of sparsity in the model. Above the plotted phase transition curve, the ℓ_1 method fails to find the sparsest solution; below the curve the solution of (2) is precisely the solution of (1).

Inspired by this approach, we have studied several statistical model selection algorithms by the following recipe:

1) Generate underlying model, $y = X\beta + z$ where β is sparse i.e. has $k < p$ nonzeros.
2) Run a model selection algorithm, obtaining $\hat{\beta}$,
3) Evaluate performance: $\frac{||\hat{\beta}-\beta||_2}{||\beta||_2} \leq \gamma$.

Averaging the results over numerous realizations, we get a picture of the performance of model selection as a function of the given sparsity and indeterminacy. In this note we present phase diagrams for LASSO, LARS, Forward Stepwise, and Forward Stepwise with False Discovery Rate threshold.

III. Main Results

A. Phase Diagrams

For our experiments we specified the underlying model to be $y = X\beta + z$ where $z \sim N(0, 16)$, β is zero except for k entries drawn from $unif(0, 100)$, and each $X_{ij} \sim N(0, 1)$ with columns normalized to unit length.

We display the median of the normalized l_2 errors for 30 estimated models for each (underdeterminedness, sparsity) combination, $p = 200$ fixed.

Forward Stepwise enters variables into the model in a sequential fashion, according to greatest t-statistic value [9]. A $\sqrt{2\log(p)}$ threshold, or *Bonferroni* threshold, will admit terms provided the absolute value of their t-statistic is not less than $\sqrt{2\log(p)}$, about 3.25 when $p = 200$. Using Forward Stepwise with a False Discovery Rate threshold

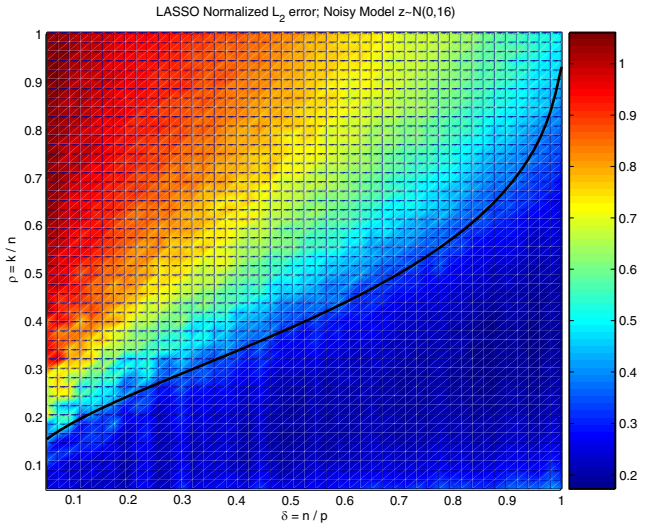

Fig. 2. Phase transition diagram when the sparse model is recovered using the LASSO Algorithm [8], where the number of variables, p, is fixed at 200. The theoretical phase transition curve from Fig 1 has been superimposed. The dark blue area, below the curve, indicates where the algorithm recovered the underlying model with near zero error, but above the curve in the colored area, the algorithm was unable to recover the correct model. As you proceed further above the curve, the ability of the algorithm to recover the model progressively drops. As with the theoretical phase transition diagram in Fig 1, along the x-axis the level of underdeterminedness decreases, and along the y-axis the level of sparsity of the underlying model increases. Each color indicates a different median normalized l_2 error of the coefficients $\frac{||\hat{\beta}-\beta||_2}{||\beta||_2}$ over 30 realizations.

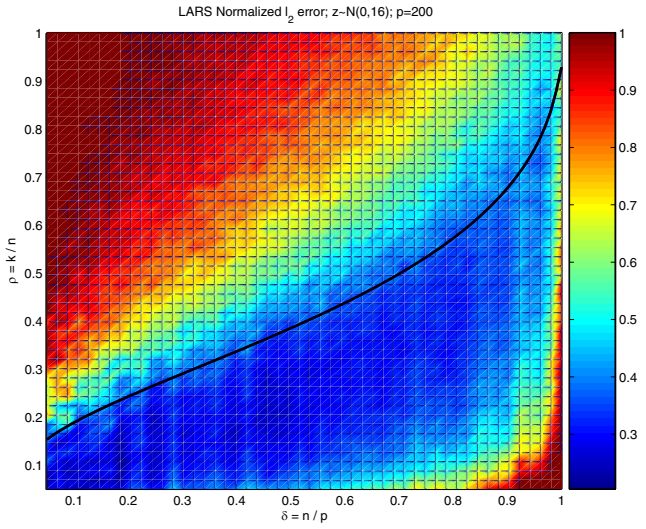

Fig. 3. This diagram displays the median error rates when the LARS Algorithm [7] is used to recover the underlying sparse model, with the number of variables, p, fixed at 200. The algorithm does not find the correct model well above the threshold, and seems to have trouble when there are as many predictors as observations. Each color indicates a different median normalized l_2 error of the coefficients $\frac{||\hat{\beta}-\beta||_2}{||\beta||_2}$ over 30 realizations.

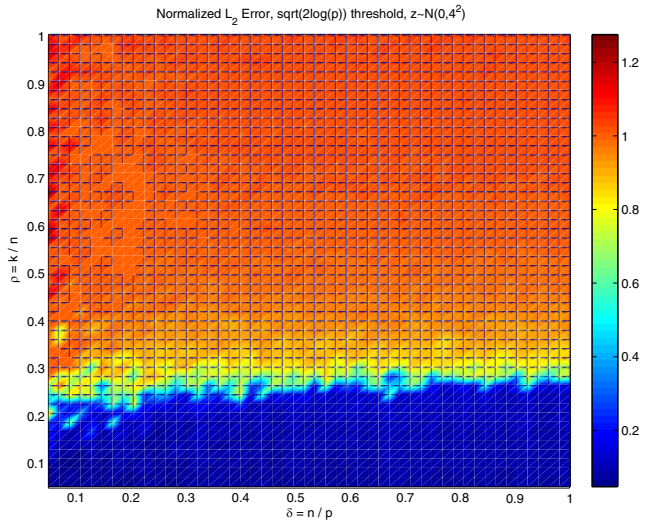

Fig. 4. Phase diagram when the underlying sparse model is recovered using the Forward Stepwise Algorithm, with the number of variables, p, xed at 200. Variables were greedily added to the model until no remaining t-statistic was greater than $\sqrt{2\log(p)}$. The phase transition is striking here: there is a very sharp dropoff below which the algorithm recovers the model with near zero error, and above which the model is unrecoverable. As with the theoretical phase transition diagram in Fig 1, along the x-axis the level of underdeterminedness decreases, and along the y-axis the level of sparsity of the underlying model increases. Each color indicates a different median normalized l_2 error of the coef cients $\frac{||\hat{\beta}-\beta||_2}{||\beta||_2}$ over 30 realizations.

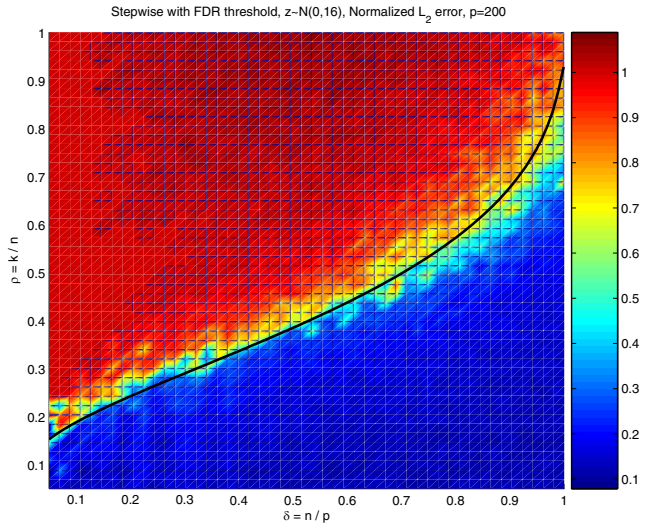

Fig. 5. This phase diagram shows the implementation of the Forward Stepwise Algorithm, but with a False Discovery Rate threshold: a term is added to the model if it has the largest t-statistic of all candidate terms and its corresponding p-value is less than the FDR value, de ned as ($.25\times$(number of terms currently in the model)/(total number of variables)). The number of variables is xed at 200. This version of Forward Stepwise has a Phase Transition similar to the theoretical curve from Fig 1 (overlaid) rather than the steep dropoff of classical Forward Stepwise. Each color indicates a different median normalized l_2 error of the coef cients $\frac{||\hat{\beta}-\beta||_2}{||\beta||_2}$ over 10 realizations.

allows more terms to enter the model, provided the p-value associated with the t-statistic is less than the FDR threshold. The FDR threshold is de ned as $q\frac{m}{p}$, where $q = .25$ (the FDR parameter [5]), m is the number of terms in the current estimated model, and p is the total number of variables available for potential inclusion.

Each of the phase diagrams shows a transition from accurate model recovery to very inaccurate as the underlying complexity of the model increases. Both Stepwise and Stepwise with FDR Threshold show a clear boundary below which the algorithm recovers the correct model, and above which it does not. The predicted curve from Figure 1 has been overlaid on the LASSO and Stepwise FDR plots: interestingly, both algorithms break down in line with the predicted curve. The behavior of LARS is very similar to LASSO although it has trouble when $p = n$. Bonferroni Forward Stepwise appears uniformly worse than either LASSO or LARS, with a constant sparsity level beyond which it fails dramatically, irregardless of the level of underdeterminedness. The FDR threshold depends on the number of terms currently in the model implying that as the sparsity level increases so does the FDR parameter, and more terms are included.

B. Examining a Slice

To understand what drives these phase transitions, we can zero in by xing $\delta = \frac{1}{2}$, i.e. twice as many variables as observations. If p is xed at 200, this means we examine a vertical slice of the phase diagram with $n = 100$.

For the case of Forward Stepwise with $\sqrt{2\log(p)}$ threshold we varied both p and the model noise level, z. Figures 6 and 7 show $\delta = \frac{1}{2}$ slices for $p = 200$ and 500, and noise levels $\sigma = \{0, 1, 2, 4, 6, 9, 12, 16\}$.

The breakdown in the error level shows where the algorithm stops recovering the correct underlying model. Notice how increased model noise drives the breakdown point lower, at sparser models. The noise level also limits the accuracy even when the correct model type is recovered, the coef cients are noisily estimated, and the coef cients are estimated with more noise when the noise in the underlying model increases. As the number of variables increases from $p = 200$ to $p = 500$, the breakdown point occurs earlier, i.e. as the number of spurious predictors increases. In order to understand the effect of model noise, we expect that when the algorithm estimates the model with small error (before the breakdown), the error will follow the *oracle MSE*:

$$tr((X'X)^{-1})\sigma^2 \tag{4}$$

where X is an $n \times k$ matrix where each column is a variable included in the nal model. Suppose (contrary to the usual case in regression modeling!) the existence of an all-knowing oracle, in particular with knowledge of the correct underlying model in each instance. We can compare the performance of Forward Stepwise Selection with the oracle model as a baseline. Figures 8 and 9 show error results for this oracle model, under the same circumstances as Forward

Stepwise: $\delta = \frac{1}{2}$ slices for $p = 200$ and 500, and noise levels $\sigma = \{0, 1, 2, 4, 6, 9, 12, 16\}$. For both $p = 200$ and $p = 500$ the median oracle MSE increases proportionally to the model noise level.

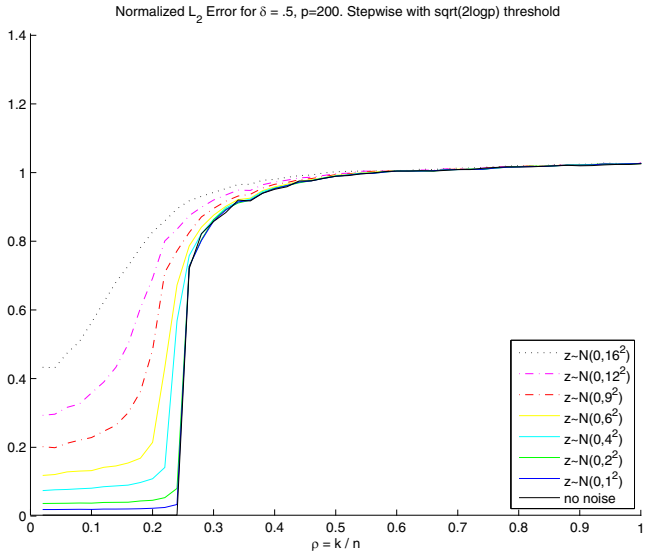

Fig. 6. This shows a single vertical slice of the Phase Diagram for Forward Stepwise (Fig 4), with varying noise levels, with $\delta = \frac{n}{p}$ xed at .5 and the number of variables xed at 200. Each increase in model noise (from no noise to $N(0, 16^2)$), causes the algorithm to break down at higher sparsity levels. The median of the normalized l_2 error for the coef cient estimates is shown, over 1000 replications.

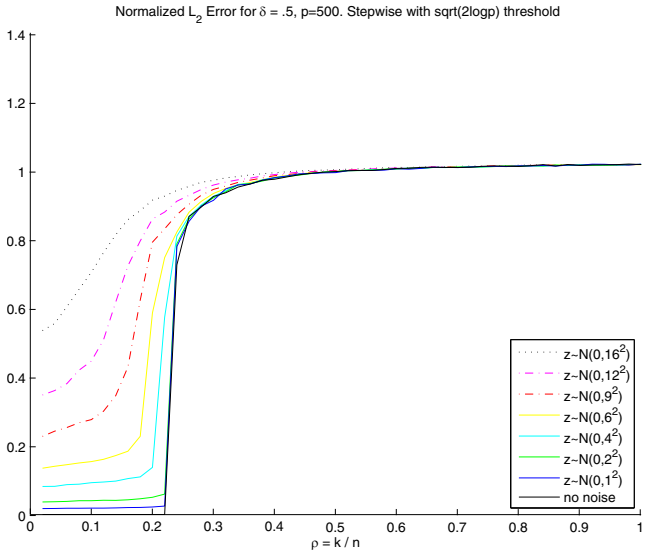

Fig. 7. This shows vertical slices at $\delta = \frac{n}{p} = .5$ through the Forward Stepwise Phase Diagram (Fig 4), with the number of variables now xed at 500, and the number of replications at 300. As the noise level is increased from $\sigma = 0$ to $\sigma = 16$ the breakdown point occurs earlier, ie. for sparser and sparser models. Notice also that with the increase in p from 200 to 500 the breakdown point, for the same level of noise, occurs at sparser underlying models.

When the Forward Stepwise algorithm *does* nd the correct model, we expect the errors to be distributed according to equation (4); exhibiting similar behavior as the oracle MSE. Figures 10 and 11 plot the ratio of the median Forward Stepwise MSE *before the breakdown point* to the oracle MSE. The breakdown point was determined to be the point at which the rst difference was maximized. The gures show the earlier breakdown point, at sparser underlying models, for the higher noise models. In both plots the noiseless model gives very nearly the same zero error rate using both algorithms, but as we add increasing levels of noise to the model, the median MSE for Forward Stepwise increases at a greater rate than that for the median Oracle MSE; roughly 1-2 times that of the oracle MSE. The effect of increasing the number of spurious variables can be seen in two ways: the Forward Stepwise MSE increases relative to the oracle MSE (for a given noise level), and it implies a breakdown point at lower sparsity levels.

IV. Software and Reproducible Research

All the code used to create the gures appearing in this paper is publicly available as a Matlab toolbox *SparseLab* downloadable at http://sparselab.stanford.edu . The toolbox contains scripts which reproduce all the calculations of this paper. SparseLab is designed to provide the research community with open source tools for sparse representation, supplemented with detailed examples and demonstrations. For more about *reproducible research* see [2], [6].

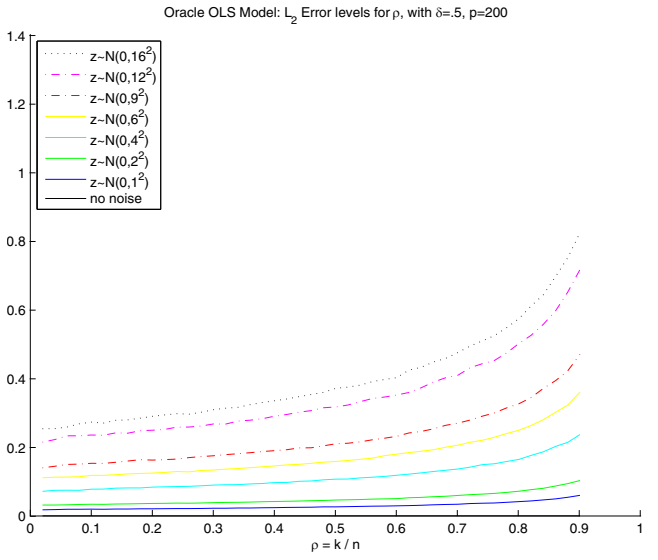

Fig. 8. Median normalized l_2 error rates over 500 replications when the true underlying model is known and estimated directly using ordinary least squares. The number of variables is xed at 200, and the number of observations xed at 100, $\delta = \frac{n}{p} = .5$. The oracle MSE is directly proportional to the model noise, and increases sharply as the underlying model becomes less sparse. The data are right truncated at ρ=.9 because the OLS MSE approaches in nity as ρ approaches 1. (i.e. the number of observations equals the number of variables)

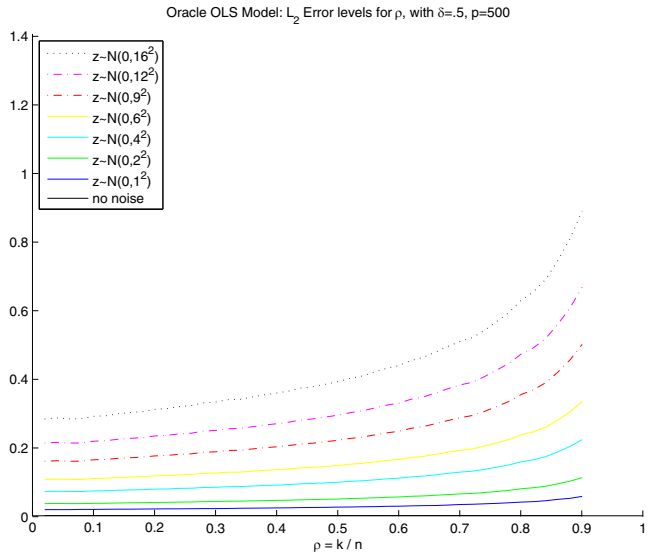

Fig. 9. Median normalized l_2 error rates over 500 replications when the true underlying model is known and estimated directly with ordinary least squares. The number of variables is xed at 500, and the number of observations xed at 250, $\delta = \frac{n}{p} = .5$. The oracle MSE is directly proportional to the model noise, and increases sharply as the underlying model becomes less sparse. As the number of variables increases from 200 (Fig 8) to 500, the error rate increases at slightly lower sparsity levels. The data are right truncated at ρ=.9 because the OLS MSE approaches in nity as ρ approaches 1 (i.e. the number of observations equals the number of variables).

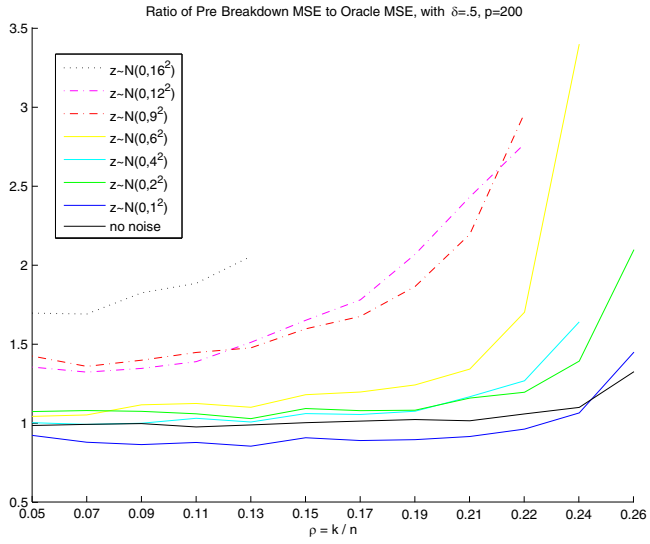

Fig. 10. Ratio of median Forward Stepwise MSE to the median oracle MSE. The number of variables is xed at 200, the number of observations at 100, ie. $\delta = \frac{n}{p} = .5$, and the median was taken over 1000 replications. The error rates were truncated at the maximum rst difference, to isolate the region in which Forward Stepwise does recover the underlying model correctly. At higher noise levels, the breakdown point occurs at lower sparsity levels, i.e. the model must be more sparse for Forward Stepwise to recover it with small error. The errors increase *more* for Forward Stepwise with increasing noise levels, than for the oracle MSE.

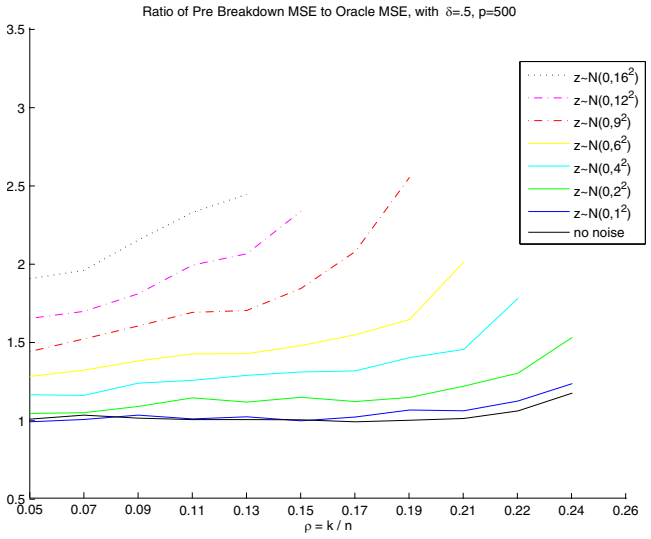

Fig. 11. Ratio of median Forward Stepwise MSE to the median oracle MSE. The number of variables is xed at 500, the number of observations at 250, maintaining $\delta = \frac{n}{p} = .5$, and the median was taken over 300 replications. The error rates were truncated at the maximum rst difference, isolating the region in which Forward Stepwise does recover the underlying model correctly. For Forward Stepwise, the median MSE is roughly 1-2 times that of the median oracle MSE, and at higher noise levels, the breakdown point occurs at lower sparsity levels. The errors increase *more* for Forward Stepwise with increasing noise levels, than for the oracle MSE. The change in the number of variables from 200 to 500 causes the Forward Stepwise MSE to increase and implies a breakdown point at lower sparsity levels.

V. Conclusions

We borrowed the idea of the phase diagram from studies of sparse underdetermined equations and used it to document the existence of a well-de ned breakdown point for model selection algorithms, for linear regression in the $p > n$ case. When the true underlying model is suf ciently sparse less than the breakdown point Forward Stepwise regression, LASSO, and LARS can nd a good model. When the true underlying model uses a number of terms close to the number of observations, such model selection methods do not work well. We nd that the Forward Stepwise algorithm exhibits different breakdown behavior depending on the threshold-to-enter used: for Bonferroni ($\sqrt{2\log(p)}$) thresholding, the algorithm ceased to recover the underlying model correctly at a particular sparsity level, regardless of the level of underdetermindness. Using the False Discover Rate threshold gave breakdown results similar to the theoretical threshold. Forward Stepwise with the Bonferroni threshold breaks down earlier (at sparser models) with an increase in model noise, and an increase in the total number of variables. Before the breakdown point for this algorithm, the size of coef cient errors follows the theoretical error distribution for the classical linear regression model.

We introduce a Matlab toolbox for sparse representation, SparseLab, available at http://sparselab.stanford.edu; it contains the code used to create the gures in this paper.

ACKNOWLEDGMENT

The authors would like to thank Iddo Drori and Yaakov Tsaig for useful discussions. This work was supported in part by the National Science Foundation under grant DMS-05-05303.

REFERENCES

[1] G. Box, D. Meyer, An Analysis for Unreplicated Fractional Factorials *Technometrics* Vol. 28, No. 1, pp 11-18, 1986.

[2] J. Buckheit, D. Donoho, WaveLab and Reproducible Research, in A. Antoniadis, Editor, *Wavelets and Statistics*, Springer, 1995.

[3] S. Chen, D. Donoho, M. Saunders, Atomic Decomposition by Basis Pursuit, *SIAM Review*, Vol. 43, Issue 1, pp. 129-159, 2001.

[4] D. Donoho, High-Dimensional Centrosymmetric Polytopes with Neighborliness Proportional to Dimension *Discrete and Computational Geometry*, online rst edition, December 22, 2005.

[5] D. Donoho, J. Jin, Asymptotic Minimaxity of FDR Thresholding for Sparse Mixtures of Exponentials *Ann. Stat.*, to appear.

[6] D. Donoho, X. Huo, Beamlab and Reproducible Research, *International Journal of Wavelets, Multiresolution and Information Processing*, vol. 2, No. 4, pp. 391-414, 2004.

[7] B. Efron, T. Hastie, I. Johnstone, R. Tibshirani, Least Angle Regression, *Ann. Statist.*, vol. 32, No. 2, pp. 407-499, 2004.

[8] R. Tibshirani, Regression Shrinkage and Selection via the Lasso, *J. Royal. Statist. Soc B.*, vol. 58, No. 1, pp. 267 288, 1996.

[9] S. Weisberg, *Applied Linear Regression* 2nd ed., John Wiley & Sons, 1985.