Bernoulli **19**(4), 2013, 1484–1500
DOI: 10.3150/13-BEJSP14

Stability

BIN YU

Departments of Statistics and EECS, University of California at Berkeley, Berkeley, CA 94720, USA.
E-mail: binyu@stat.berkeley.edu

Reproducibility is imperative for any scientific discovery. More often than not, modern scientific findings rely on statistical analysis of high-dimensional data. At a minimum, reproducibility manifests itself in stability of statistical results relative to "reasonable" perturbations to data and to the model used. Jacknife, bootstrap, and cross-validation are based on perturbations to data, while robust statistics methods deal with perturbations to models.

In this article, a case is made for the importance of stability in statistics. Firstly, we motivate the necessity of stability for interpretable and reliable encoding models from brain fMRI signals. Secondly, we find strong evidence in the literature to demonstrate the central role of stability in statistical inference, such as sensitivity analysis and effect detection. Thirdly, a smoothing parameter selector based on estimation stability (ES), ES-CV, is proposed for Lasso, in order to bring stability to bear on cross-validation (CV). ES-CV is then utilized in the encoding models to reduce the number of predictors by 60% with almost no loss (1.3%) of prediction performance across over 2,000 voxels. Last, a novel "stability" argument is seen to drive new results that shed light on the intriguing interactions between sample to sample variability and heavier tail error distribution (e.g., double-exponential) in high-dimensional regression models with p predictors and n independent samples. In particular, when $p/n \to \kappa \in (0.3, 1)$ and the error distribution is double-exponential, the Ordinary Least Squares (OLS) is a better estimator than the Least Absolute Deviation (LAD) estimator.

Keywords: cross-validation; double exponential error; estimation stability; fMRI; high-dim regression; Lasso; movie reconstruction; robust statistics; stability

1. Introduction

In his seminal paper "The Future of Data Analysis" (Tukey, 1962), John W. Tukey writes:

> "It will still be true that there will be aspects of data analysis well called technology, but there will also be the hallmarks of stimulating science: intellectual adventure, demanding calls upon insight, and a need to find out 'how things really are' by investigation and the confrontation of insights with experience" (p. 63).

Fast forward to 2013 in the age of information technology, these words of Tukey ring as true as fifty years ago, but with a new twist: the ubiquitous and massive data today were impossible to imagine in 1962. From the point of view of science, information technology and data are a blessing, and a curse. The reasons for them to be a blessing are many and obvious. The reasons for it to be a curse are less obvious. One of them is well articulated recently by two prominent biologists in an editorial Casadevall and Fang (2011) in *Infection and Immunity* (of the American Society for Microbiology):

> "Although scientists have always comforted themselves with the thought that science is self-correcting, the immediacy and rapidity with which knowledge disseminates today means that incorrect information can have a profound impact before any corrective process can take place" (p. 893).

> "A recent study analyzed the cause of retraction for 788 retracted papers and found that error and fraud were responsible for 545 (69%) and 197 (25%) cases, respectively, while the cause was unknown in 46 (5.8%) cases (31)" (p. 893).

The study referred is Steen (2011) in the *Journal of Medical Ethics*. Of the 788 retracted papers from PubMed from 2000 to 2010, 69% are marked as "errors" on the retraction records. Statistical analyses are likely to be involved in these errors. Casadevall and Fang go on to call for "enhanced training in probability and statistics," among other remedies including "reembracing philosophy." More often than not, modern scientific findings rely on statistical analyses of high-dimensional data, and reproducibility is imperative for any scientific discovery. Scientific reproducibility therefore is a responsibility of statisticians. At a minimum, reproducibility manifests itself in the stability of statistical results relative to "reasonable" perturbations to data and to the method or model used.

Reproducibility of scientific conclusions is closely related to their reliability. It is receiving much well-deserved attention lately in the scientific community (e.g., Ioannidis, 2005; Kraft et al., 2009, Casadevall and Fang, 2011; Nosek et al., 2012) and in the media (e.g., Naik, 2011; Booth, 2012). Drawing a scientific conclusion involves multiple steps. First, data are collected by one laboratory or one group, ideally with a clear hypothesis in the mind of the experimenter or scientist. In the age of information technology, however, more and more massive amounts of data are collected for fishing expeditions to "discover" scientific facts. These expeditions involve running computer codes on data for data cleaning and analysis (modeling and validation). Before these facts become "knowledge," they have to be reproduced or replicated through new sets of data by the same group or preferably by other groups. Given a fixed set of data, Donoho et al. (2009) discuss reproducible research in computational hormonic analysis with implications on computer-code or computing-environment reproducibility in computational sciences including statistics. Fonio et al. (2012) discuss replicability between laboratories as an important screening mechanism for discoveries. Reproducibility could have multitudes of meaning to different people. One articulation on the meanings of reproducibility, replication, and repeatability can be found in Stodden (2011).

In this paper, we advocate for more involvement of statisticians in science and for an enhanced emphasis on stability within the statistical framework. Stability has been of a great concern in statistics. For example, in the words of Hampel et al. (1986), "...robustness theories can be viewed as stability theories of statistical inference" (p. 8). Even in low-dimensional linear regression models, collinearity is known to cause instability of OLS or problem for individual parameter estimates so that significance testing for these estimates becomes unreliable. Here we demonstrate the importance of statistics for understanding our brain; we describe our methodological work on estimation stability that helps interpret models reliably in neuroscience; and we articulate how our solving neuroscience problems motivates theoretical work on stability and robust statistics in high-dimensional regression models. In other words, we tell an interwinding story of scientific investigation and statistical developments.

The rest of the paper is organized as follows. In Section 2, we cover our "intellectual adventure" into neursocience, in collaboration with the Gallant Lab at UC Berkeley, to understand human visual pathway via fMRI brain signals invoked by natural stimuli (images or movies) (cf. Kay et al., 2008, Naselaris et al., 2009, Kay and Gallant, 2009, Naselaris et al., 2011). In particular, we describe how our statistical encoding and decoding models are the backbones of "mind-

reading computers," as one of the 50 best inventions of 2011 by the Time Magazine (Nishimoto et al., 2011). In order to find out "how things really are," we argue that reliable interpretation needs stability. We define stability relative to a data perturbation scheme. In Section 3, we briefly review the vast literature on different data perturbation schemes such as jacknife, subsampling, and bootstrap. (We note that data perturbation in general means not only taking subsets of data units from a given data set, but also sampling from an underlying distribution or replicating the experiment for a new set of data.)

In Section 4, we review an estimation stability (ES) measure taken from Lim and Yu (2013) for regression feature selection. Combining ES with CV as in Lim and Yu (2013) gives rise to a smoothing parameter selector ES-CV for Lasso (or other regularization methods). When we apply ES-CV to the movie-fMRI data, we obtain a 60% reduction of the model size or the number of features selected at a negligible loss of 1.3% in terms of prediction accuracy. Subsequently, the ES-CV-Lasso models are both sparse and more reliable hence better suited for interpretation due to their stability and simplicity. The stability considerations in our neuroscience endeavors have prompted us to connect with the concept of stability from the robust statistics point of view. In El Karoui et al. (2013), we obtain very interesting theoretical results in high-dimensional regression models with p predictors and n samples, shedding light on how sample variability in the design matrix meets heavier tail error distributions when p/n is approximately a constant in $(0, 1)$ or in the random matrix regime. We describe these results in an important special case in Section 5. In particular, we see that when $p/n \to \kappa$ and $1 > \kappa > 0.3$ or so, the Ordinary Least Squares (OLS) estimator is better than the Least Absolute Deviation (LAD) estimator when the error distribution is double exponential. We conclude in Section 6.

2. Stable models are necessary for understanding visual pathway

Neuroscience holds the key to understanding how our mind works. Modern neuroscience is invigorated by massive and multi-modal forms of data enabled by advances in technology (cf. Atkil, Martone and Van Essen, 2012). Building mathematical/statistical models on this data, computational neuroscience is at the frontier of neuroscience. The Gallant Lab at UC Berkeley is a leading neuroscience lab specializing in understanding the visual pathway, and is a long-term collaborator with the author's research group. It pioneered the use of natural stimuli in experiments to invoke brain signals, in contrast to synthetic signals such as white noise and moving bars or checker boards as previously done.

Simply put, the human visual pathway works as follows. Visual signals are recorded by retina and through the relay center LGN they are transmitted to primary visual cortex areas V1, on to V2 and V4, on the "what" pathway (in contrast to the "where" pathway) (cf. Goodale and Milner, 1992). Computational vision neuroscience aims at modeling two related tasks carried out by the brain (cf. Dayan and Abbott, 2005) through two kinds of models. The first kind, the encoding model, predicts brain signals from visual stimuli, while the second kind, the decoding model recovers visual stimuli from brain signals. Often, decoding models are built upon encoding models and hence indirectly validate the former, but they are important in their own right. In the September issue of *Current Biology*, our joint paper with the Gallant Lab, Nishimoto et al.

(2011) invents a decoding (or movie reconstruction) algorithm to reconstruct movies from fMRI brain signals. This work has received intensive and extensive coverage by the media including The Economist's Oct. 29th 2011 issue ("Reading the Brain: Mind-Goggling") and the National Public Radio in their program "Forum with Michael Krasny" on Tue, Sept. 27, 2011 at 9:30 am ("Reconstructing the Mind's Eye"). As mentioned earlier, it was selected by the Time Magazine as one of the best 50 inventions of 2011 and dubbed as "Mind-reading Computers" on the cover page of the Time's invention issue.

What is really behind the movie reconstruction algorithm?

Can we learn something from it about how brain works?

The movie reconstruction algorithm consists of statistical encoding and decoding models, both of which employ regularization. The former are sparse models so they are concise enough to be viewed and are built via Lasso + CV for each voxel separately. However, as is well-known Lasso + CV results are not stable or reliable enough for scientific interpretation due to the L_1 regularization and the emphasis of CV on prediction performance. So Lasso + CV is not estimation stable. The decoding model uses the estimated encoding model for each voxel and Tiknohov regularization or Ridge in covariance estimation to pull information across different voxels over V1, V2 and V4 (Nishimoto et al., 2011). Then an empirical prior for clips of short videos is used from movie trailers and YouTube to induce posterior weights on video clips in the empirical prior database. Tiknohov or Ridge regularization concerns itself with the estimation of the covariance between voxels that is not of interest for interpretation. The encoding phase is the focus here from now on.

V1 is a primary visual cortex area and the best understood area in the visual cortex. Hubel and Wiesel received a Nobel Prize in Physiology or Medicine in 1981 for two major scientific discoveries. One is Hubel and Wiesel (1959) that uses cat physiology data to show, roughly speaking, that simple V1 neuron cells act like Gabor filters or as angled edge detectors. Later, using solely image data, Olshausen and Field (1996) showed that image patches can be sparsely represented on Gabor-like basis image patches. The appearance of Gabor filters in both places is likely not a coincidence, due to the fact that our brain has evolved to represent the natural world. These Gabor filters have different locations, frequencies and orientations. Previous work from the Gallant Lab has built a filter-bank of such Gabor filters and successfully used them to design encoding models with single neuron signals in V1 invoked by static natural image stimuli (Kay et al., 2008, Naselaris et al., 2011).

In Nishimoto et al. (2011), we use fMRI brain signals observed over 2700 voxels in different areas of the visual cortex. fMRI signals are indirect and non-invasive measures of neural activities in the brain and have good spatial coverage and temporal resolution in seconds. Each voxel is roughly a cube of 1 mm by 1 mm by 1 mm and contains hundreds of thousands of neurons. Leveraging the success of Gabor-filter based models for single neuron brain signals, for a given image, a vector of features is extracted by 2-d wavelet filters. This feature vector has been used to build encoding models for fMRI brain signals in Kay et al. (2008) and Naselaris et al. (2011). Invoked by clips of videos/movies, fMRI signals from three subjects are collected with the same experimental set-up. To model fMRI signals invoked by movies, a 3-dim motion-energy Gabor filter bank has been built in Nishimoto et al. (2011) to extract a feature vector of dimension of

26K. Linear models are then built on these features at the observed time point and lagged time points.

At present sparse linear regression models are favorites of the Gallant Lab through Lasso or ε-L2Boost. These sparse models give similar prediction performance on validation data as neural nets and kernel machines on image-fMRI data; they correspond well to the neuroscience knowledge on V1; and they are easier to interpret than neural net and kernel machine models that include all features or variables.

For each subject, following a rigorous protocol in the Gallant Lab, the movie data (how many frames per second?) consists of three batches: training, test and validation. The training data is used to fit a sparse encoding model via Lasso or e-L2Boost and the test data is used to select the smoothing parameter by CV. These data are averages of two or three replicates. That is, the same movie is played to one subject two or three times and the resulted fMRI signals are called replicates. Then a completed encoding determined model is used to predict the fMRI signals in the validation data (with 10+ replicates) and the prediction performance is measured by the correlation between the predicted fMRI signals and observed fMRI signals, for each voxel and for each subject. Good prediction performance is observed for such encoding models (cf. Figure 2).

3. Stability considerations in the literature

Prediction and movie reconstruction are good steps to validate the encoding model in order to understand the human visual pathway. But the science lies in finding the features that might drive a voxel, or to use Tukey's words, finding out "how things really are."

It is often the case that the number of data units is easily different from what is in collected data. There are some hard resource constraints such as that human subjects can not lie inside an fMRI machine for too long and it also costs money to use the fMRI machine. But whether the data collected is for 2 hours as in the data or 1 hours 50 min or 2 hours and 10 min is a judgement call by the experimenter given the constraints. Consequently, scientific conclusions, or in our case, candidates for driving features, should be stable relative to removing a small proportion of data units, which is one form of reasonable or appropriate data perturbation, or reproducible without a small proportion of the data units. With a smaller set of data, a more conservative scientific conclusion is often reached, which is deemed worthwhile for the sake of more reliable results.

Statistics is not the only field that uses mathematics to describe phenomena in the natural world. Other such fields include numerical analysis, dynamical systems and PDE and ODE. Concepts of stability are central in all of them, implying the importance of stability in quantitative methods or models when applied to real world problems.

The necessity for a procedure to be robust to data perturbation is a very natural idea, easily explainable to a child. Data perturbation has had a long history in statistics, and it has at least three main forms: jacknife, sub-sampling and bootstrap. Huber (2002) writes in "John W. Tukeys Contribution to Robust Statistics:

"[Tukey] preferred to rely on the actual batch of data at hand rather than on a hypothetical underlying population of which it might be a sample" (p. 1643).

All three main forms of data perturbation rely on an "actual batch of data" even though their theoretical analyses do assume hypothetical underlying populations of which data is a sample. They all have had long histories.

Jacknife can be traced back at least to Quenouille (1949, 1956) where jacknife was used to estimate the bias of an estimator. Tukey (1958), an abstract in the Annals of Mathematical Statistics, has been regarded as a key development because of his use of jacknife for variance estimation. Miller (1974) is an excellent early review on Jacknife with extensions to regression and time series situations. Hinkley (1977) proposes weighted jacknife for unbalanced data for which Wu (1986) provides a theoretical study. Künsch (1989) develops Jacknife further for time series. Sub-sampling on the other hand was started three years earlier than jacknife by Mahalanobis (1946). Hartigan (1969, 1975) buids a framework for confidence interval estimation based on subsampling. Carlstein (1986) applies subsampling (which he called subseries) to the time series context. Politis and Romano (1992) study subsampling for general weakly dependent processes. Cross-validation (CV) has a more recent start in Allen (1974) and Stone (1974). It gives an estimated prediction error that can be used to select a particular model in a class of models or along a path of regularized models. It has been wildly popular for modern data problems, especially for high-dimensional data and machine learning methods. Hall (1983) and Li (1986) are examples of theoretical analyses of CV. Efron's (1979) bootstrap is widely used and it can be viewed as simplified jacknife or subsampling. Examples of early theoretical studies of bootstrap are Bickel and Freedman (1981) and Beran (1984) for the i.i.d. case, and Künsch (1989) for time series. Much more on these three data perturbation schemes can be found in books, for example, by Efron and Tibshirani (1993), Shao and Tu (1995) and Politis, Romano and Wolf (1999).

If we look into the literature of probability theory, the mathematical foundation of statistics, we see 5 that a perturbation argument is central to limiting law results such as the Central Limit Theorem (CLT).

The CLT has been the bedrock for classical statistical theory. One proof of the CLT that is composed of two steps and is well exposited in Terence Tao's lecture notes available at his website (Tao, 2012). Given a normalized sum of i.i.d. random variables, the first step proves the universality of a limiting law through a perturbation argument or the Lindebergs swapping trick. That is, one proves that a perturbation in the (normalized) sum by a random variable with matching first and second moments does not change the (normalized) sum distribution. The second step finds the limit law by way of solving an ODE.

Recent generalizations to obtain other universal limiting distributions can be found in Chatterjee (2006) for Wigner law under non-Gaussian assumptions and in Suidan (2006) for last passage percolation. It is not hard to see that the cornerstone of theoretical high-dimensional statistics, concentration results, also assumes stability-type conditions. In learning theory, stability is closely related to good generalization performance (Devroye and Wagner, 1979, Kearns and Ron, 1999, Bousquet and Elisseeff, 2002, Kutin and Niyogi, 2002, Mukherjee et al., 2006, Shalev-Shwartz et al., 2010).

To further our discussion on stability, we would like to explain what we mean by statistical stability. We say statistical stability holds if statistical conclusions are robust or stable to appropriate perturbations to data. That is, statistical stability is well defined relative to a particular aim and a particular perturbation to data (or model). For example, aim could be estimation, prediction or limiting law. It is not difficult to have statisticians to agree on what are appropriate data

perturbations when data units are i.i.d. or exchangeable in general, in which case subsampling or bootstrap are appropriate. When data units are dependent, transformations of the original data are necessary to arrive at modified data that are close to i.i.d. or exchangeable, such as in parametric bootstrap in linear models or block-bootstrap in time series. When subsampling is carried out, the reduced sample size in the subsample does have an effect on the detectable difference, say between treatment and control. If the difference size is large, this reduction on sample size would be negligible. When the difference size is small, we might not detect the difference with a reduced sample size, leading to a more conservative scientific conclusion. Because of the utmost importance of reproducibility for science, I believe that this conservatism is acceptable and may even be desirable in the current scientific environment of over-claims.

4. Estimation stability: Seeking more stable models than Lasso + CV

For the fMRI problem, let us recall that for each voxel, Lasso or e-L2Boost is used to fit the mean function in the encoding model with CV to choose the smoothing parameter. Different model selection criteria have been known to be unstable. Breiman (1996) compares predictive stability among forward selection, two versions of garotte and Ridge and their stability increases in that order. He goes on to propose averaging unstable estimators over different perturbed data sets in order to stabilize unstable estimators. Such estimators are prediction driven, however, and they are not sparse and thereby not suitable for interpretation.

In place of bootstrap for prediction error estimation as in Efron (1982), Zhang (1993) uses multi-fold cross-validation while Shao (1996) uses m out of n bootstrap samples with $m \ll n$. They then select models with this estimated prediction error, and provide theoretical results for low dimensional or fixed p linear models. Heuristically, the m out of n bootstrap in Shao (1996) is needed because the model selection procedure is a discrete (or set) valued estimator for the true model predictor set and hence non-smooth (cf. Bickel, Götze, and van Zwet, 1997).

The Lasso (Tibshirani, 1996) is a modern model selection method for linear regression and very popular in high-dimensions:

$$\hat{\beta}(\lambda) = \underset{\beta \in R^p}{\arg} \left\{ \|Y - X\beta\|_2^2 + \lambda \|\beta\|_1 \right\},$$

where $Y \in R^n$ is the response vector and $X \in R^{n \times p}$ is the design matrix. That is, there are n data units and p predictors. For each λ, there is a unique L_1 norm for its solution that we can use to index the solution as $\hat{\beta}(\tau)$ where

$$\tau = \tau(\lambda) = \left\| \hat{\beta}(\lambda) \right\|_1.$$

Cross-validation (CV) is used most of the time to select λ or τ, but Lasso + CV is unstable relative to bootstrap or subsampling perturbations when predictors are correlated (cf. Meinshausen and Bühlmann, 2010, Bach, 2008).

Using bootstrap in a different manner than Shao (1996), Bach (2008) proposes BoLasso to improve Lasso's model selection consistency property by taking the smallest intersecting model

of selected models over different bootstrap samples. For particular smoothing parameter sequences, the BoLasso selector is shown by Bach (2008) to be model selection consistent for the low dimensional case without the irrepresentable condition needed for Lasso (cf. Meinshausen and Bühlmann, 2006, Zhao and Yu, 2006; Zou, 2006; Wainwright, 2009). Meinshausen and Bühlmann (2010) also weaken the irrepresentable condition for model selection consistency of a stability selection criterion built on top of Lasso. They bring perturbations to a Lasso path through a random scalar vector in the Lasso L_1 penalty, resulting in many random Lasso paths. A threshold parameter is needed to distinguish important features based on these random paths. They do not consider the problem of selecting one smoothing parameter value for Lasso as in Lim and Yu (2013).

We would like to seek a specific model along the Lasso path to interpret and hence selects a specific λ or τ. It is well known that CV does not provide a good interpretable model because Lasso + CV is unstable. Lim and Yu (2013) propose a stability-based criterion that is termed Estimation Stability (ES). They use the cross-validation data perturbation scheme. That is, n data units are randomly partitioned into V blocks of pseudo data sets of size $(n-d)$ or subsamples where $d = \lfloor n/V \rfloor$.[1]

Given a smoothing parameter λ, a Lasso estimate $\hat{\beta}_v(\lambda)$ is obtained for the vth block $v = 1, \ldots, V$. Since the L_1 norm is a meaningful quantity to line up the V different estimates, Lim and Yu (2013)[2] use it, denoted as τ below, to line up these estimates to form an estimate $\hat{m}(\tau)$ for the mean regression function and an approximate delete-d jacknife estimator for the variance of $\hat{m}(\tau)$:

$$\hat{m}(\tau) = \frac{1}{V}\sum_v X\hat{\beta}_v(\tau),$$

$$\hat{T}(\tau) = \frac{n-d}{d}\frac{1}{V}\sum_v \left(\left\| X\hat{\beta}_v(\tau) - \hat{m}(\tau)\right\|^2\right).$$

The last expression is only an approximate delete-d jacknife variance estimator unless $V = \binom{n}{n-d}$ when all the subsamples of size $n-d$ are used. Define the (estimation) statistical stability measure as

$$ES(\tau) = \frac{1/V\sum_v \|X\hat{\beta}_v(\tau) - \hat{m}(\tau)\|^2}{\hat{m}^2(\tau)} = \frac{d}{n-d}\frac{\hat{T}(\tau)}{\hat{m}^2(\tau)} = \frac{d}{n-d}\frac{1}{Z^2(\tau)},$$

where $Z(\tau) = \hat{m}(\tau)/\sqrt{\hat{T}(\tau)}$.

For nonlinear regression functions, ES can still be applied if we take an average of the estimated regression functions. Note that ES aims at estimation stability, while CV aims at prediction stability. In fact, ES is the reciprocal of a test statistic for testing

$$H_0 : X\beta = 0.$$

[1] $\lfloor x \rfloor$ is the floor function or the largest integer that is smaller than or equal to x.

[2] It is also fine to use λ to line up the different solutions, but not a good idea to use the ratio of λ and its maximum value for each pseudo data set.

Since $Z(\tau) = \hat{m}(\tau)/\sqrt{\hat{T}(\tau)}$ is a test statistic for H_0, $Z^2(\tau)$ is also a test statistic. $ES(\tau)$ is a scaled version of the reciprocal $1/Z^2(\tau)$.

To combat the high noise situation where ES would not have a well-defined minimum, Lim and Yu (2013) combine ES with CV to propose the *ES-CV selection criterion* for smoothing parameter τ:

Choose the largest τ that minimizes ES (τ) and is smaller or equal to the CV selection.

ES-CV is applicable to smoothing parameter selection in Lasso, and other regularization methods such as Tikhonov or Ridge regularization (see, for example, Tikhonov, 1943, Markovich, 2007, Hoerl, 1962, Hoerl and Kennard, 1970). ES-CV is well suited for parallel computation as CV and incurs only a negligible computation overhead because $\hat{m}(\tau)$ are already computed for CV. Moreover, simulation studies in Lim and Yu (2013) indicate that, when compared with Lasso + CV, ES-CV applied to Lasso gains dramatically in terms of false discovery rate while it loses only somewhat in terms of true discovery rate.

The features or predictors in the movie-fMRI problem are 3-d Gabor wavelet filters, and each of them is characterized by a (discretized) spatial location on the image, a (discretized) frequency of the filter, a (discretized) orientation of the filter, and 4 (discrete) time-lags on the corresponding image that the 2-d filter is acting on. For the results comparing CV and ES-CV in Figure 1,

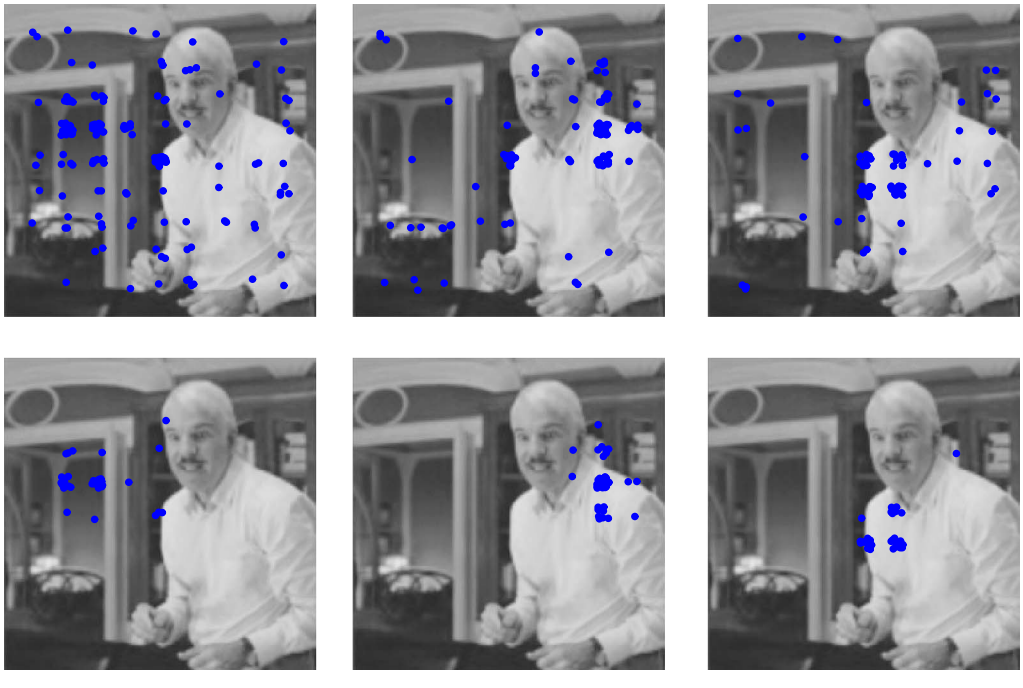

Figure 1. For three voxels (one particular subject), we display the (jittered) locations that index the Gabor features selected by CV-Lasso (top row) and ESCV-Lasso (bottom row).

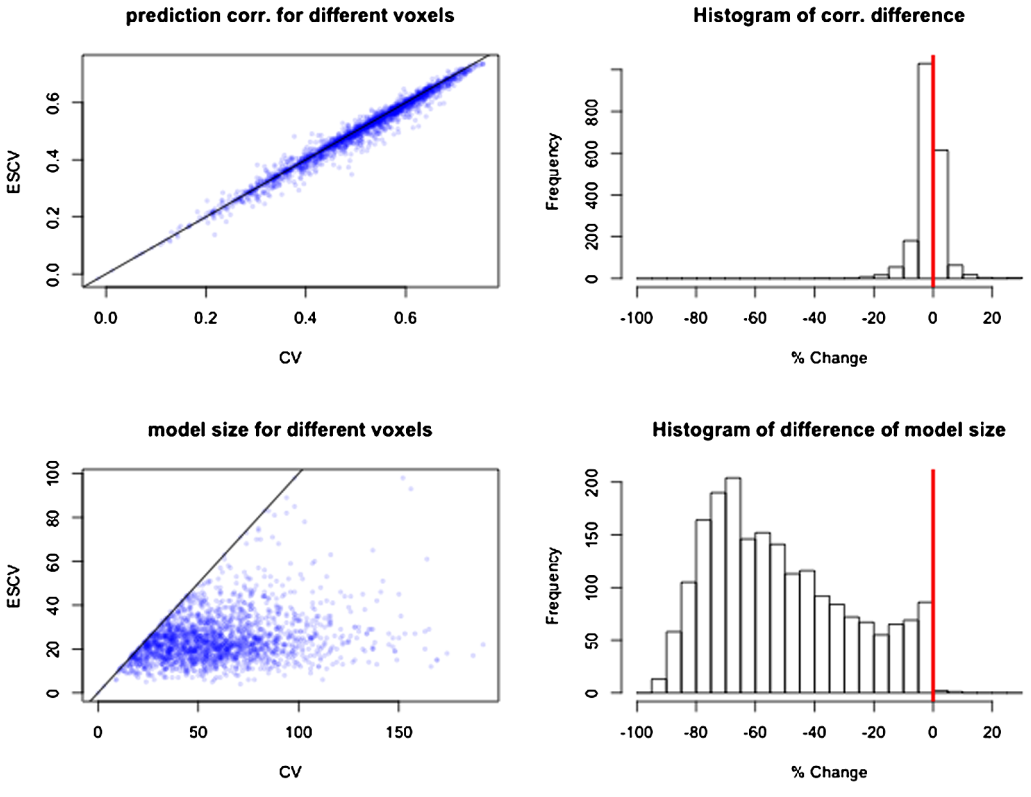

Figure 2. Comparisons of ESCV(Lasso) and CV(Lasso) in terms of model size and prediction correlation. The scatter plots on the left compare ESCV and CV while the histograms on the right display the differences of model size and prediction correlation.

we have a sample size $n = 7{,}200$ and use a reduced set of $p = 8{,}556$ features or predictors, corresponding to a coarser set of filter frequencies than what is used in Nishimoto et al. (2011) with $p = 26{,}220$ predictors.

We apply both CV and ES-CV to select the smoothing parameters in Lasso (or e-L2Boost). For three voxels (and a particular subject), for the simplicity of display, we show the locations of the selected features (regardless of their frequencies, orientations and time-lags) in Figure 1. For these three voxels, ES-CV maintains almost the same prediction correlation performances as CV (0.70 vs. 0.72) while ES-CV selects many fewer and more concentrated locations than CV. Figure 2 shows the comparison results across 2088 voxels in the visual cortex that are selected for their high SNRs. It is composed of four sub-plots. The upper two plots compare prediction correlation performance of the models built via Lasso with CV and ES-CV on validation data. For each model fitted on training data and each voxel, predicted responses over the validation data are calculated. Its correlation with the observed response vector is the "prediction correlation" displayed in Figure 2. The lower two plots compare the sparsity properties of the models or model size. Because of the definition of ES-CV, it is expected that the ES-CV model are always smaller

than or equal to the CV model. The sparsity advantage of ES-CV is apparent with a huge overall reduction of 60% on the number of selected features and a minimum loss of overall prediction accuracy by only 1.3%. The average size of the ES-CV models is 24.3 predictors, while that for the CV models is 58.8 predictors; the average prediction correlation performance of the ES-CV models is 0.499, while that for the CV models is 0.506.

5. Sample variability meets robust statistics in high-dimensions

Robust statistics also deals with stability, relative to model perturbation. In the preface of his book "Robust Statistics," Huber (1981) states:

"Primarily, we are concerned with *distributional robustness*: the shape of the true underlying distribution deviates slightly from the assumed model."

Hampel, Rousseeuw, Ronchetti and Stahel (1986) write:

"Overall, and in analogy with, for example, the stability aspects of differential equations or of numerical computations, robustness theories can be viewed as stability theories of statistical inference" (p. 8).

Tukey (1958) has generally been regarded as the first paper on robust statistics. Fundamental contributions were made by Huber (1964) on M-estimation of location parameters, Hampel (1968, 1971, 1974) on "break-down" point and influence curve. Further important contributions can be found, for example, in Andrews et al. (1972) and Bickel (1975) on one-step Huber estimator, and in Portnoy (1977) for M-estimation in the dependent case.

For most statisticians, robust statistics in linear regression is associated with studying estimation problems when the errors have heavier tail distributions than the Gaussian distribution. In the fMRI problem, we fit mean functions with an L2 loss. What if the "errors" have heavier tails than Gaussian tails? For the L1 loss is commonly used in robust statistics to deal with heavier tail errors in regression, we may wonder whether the L1 loss would add more stability to the fMRI problem. In fact, for high-dimensional data such as in our fMRI problem, removing some data units could severely change the outcomes of our model because of feature dependence. This phenomenon is also seen in simulated data from linear models with Gaussian errors in high-dimensions.

How does sample to sample variability interact with heavy tail errors in high-dimensions?

In our recent work El Karoui et al. (2013), we seek insights into this question through analytical work. We are able to see interactions between sample variability and double-exponential tail errors in a high-dimensional linear regression model. That is, let us assume the following linear regression model

$$Y_{n\times 1} = X_{n\times p}\beta_{p\times 1} + \varepsilon_{n\times 1},$$

where

$$X_i \sim N(0, \Sigma_p), \text{i.i.d.}, \varepsilon_i \text{i.i.d.}, E\varepsilon_i = 0, E\varepsilon_i^2 = \sigma^2 < \infty.$$

An M-estimator with respect to loss function ρ is given as

$$\hat{\beta} = \underset{\beta\in R^p}{\operatorname{argmin}} \sum_i \rho\big(Y_i - X_i'\beta\big).$$

We consider the random-matrix high-dimensional regime:

$$p/n \to \kappa \in (0, 1).$$

Due to rotation invariance, WLOG, we can assume $\Sigma_p = I_p$ and $\beta = 0$. We cite below a result from El Karoui et al. (2013) for the important special case of $\Sigma_p = I_p$:

Result 1 (El Karoui et al., 2013). *Under the aforementioned assumptions, let* $r_\rho(p, n) = \|\hat{\beta}\|$, *then* $\hat{\beta}$ *is distributed as*

$$r_\rho(p, n)U,$$

where $U \sim \text{uniform}(S^{p-1})(1)$, *and*

$$r_\rho(p, n) \to r_\rho(\kappa),$$

as $n, p \to \infty$ *and* $p/n \to \kappa \in (0, 1)$.

Denote

$$\hat{z}_\varepsilon := \varepsilon + r_\rho(\kappa)Z,$$

where $Z \sim N(0, 1)$ and independent of ε, and let

$$\text{prox}_c(\rho)(x) = \underset{y \in R}{\text{argmin}}\left[\rho(y) + \frac{(x-y)^2}{2c}\right].$$

Then $r_\rho(\kappa)$ satisfies the following system of equations together with some nonnegative c:

$$E\left\{\left[\text{prox}_c(\rho)\right]'\right\} = 1 - \kappa,$$
$$E\left\{\left[\hat{z}_\varepsilon - \text{prox}_c(\hat{z}_\varepsilon)\right]^2\right\} = \kappa r_\rho^2(\kappa).$$

In our limiting result, the norm of an M-estimator stabilizes. It is most interesting to mention that in the proof a "leave-one-out" trick is used both row-wise and column-wise such that one by one rows are deleted and similarly columns are deleted. The estimators with deletions are then compared to the estimator with no deletion. This is in effect a perturbation argument and reminiscent of the "swapping trick" for proving the CLT as discussed before. Our analytical derivations involve prox functions, which are reminiscent of the second step in proving normality in the CLT. This is because a prox function is a form of derivative, and not dissimilar to the derivative appearing in the ODE derivation of the analytical form of the limiting distribution (e.g normal distribution) in the CLT.

In the case of i.i.d. double-exponential errors, El Karoui et al. (2013) numerically solve the two equations in Result 1 to show that when $\kappa \in (0.3, 1)$, L_2 loss fitting (OLS) is better than L_1 loss fitting (LAD) in terms of MSE or variance. They also show that the numerical results match very well with simulation or Monte Carlo results. At a high level, we may view that $\hat{z}_\varepsilon$ holds the key to this interesting phenomenon. Being a weighted convolution of Z and ε, it embeds the interaction between sample variability (expressed in Z) and error variability (expressed in ε)

and this interaction is captured in the optimal loss function (cf. El Karoui et al., 2013). In other words, $\hat{z}_\varepsilon$ acts more like double exponential when the influence of standard normal Z in $\hat{z}_\varepsilon$ is not dominant (or when $\kappa < 0.3$ or so as we discover when we solve the equations) and in this case, the optimal loss function is closer to LAD loss. In cases when $\kappa > 0.3$, it acts more like Gaussian noise, leading to the better performance of OLS (because the optimal loss is closer to LS).

Moreover, for double exponential errors, the M-estimator LAD is an MLE and we are in a high-dimensional situation. It is well-known that MLE does not work in high-dimensions. Remedies have been found through penalized MLE where a bias is introduced to reduce variance and consequently reduce the MSE. In contrast, when $\kappa \in (0.3, 1)$, the better estimator OLS is also unbiased, but has a smaller variance nevertheless. The variance reduction is achieved through a better loss function LS than the LAD and because of a concentration of quadratic forms of the design matrix. This concentration does not hold for fixed orthogonal designs, however. A follow-up work (Bean et al., 2013) addresses the question of obtaining the optimal loss function. It is current research regarding the performance of estimators from penalized OLS and penalized LAD when the error distribution is double-exponential. Preliminary results indicate that similar phenomena occur in non-sparse cases.

Furthermore, simulations with design matrix from an fMRI experiment and double-exponential error show the same phenomenon, that is, when $\kappa = p/n > 0.3$ or so, OLS is better than LAD. This provides some insurance for using L2 loss function in the fMRI project. It is worth noting that El Karoui et al. (2013) contains results for more general settings.

6. Conclusions

In this paper, we cover three problems facing statisticians at the 21st century: figuring out how vision works with fMRI data, developing a smoothing parameter selection method for Lasso, and connecting perturbation in the case of high-dimensional data with classical robust statistics through analytical work. These three problems are tied together by stability. Stability is well defined if we describe the data perturbation scheme for which stability is desirable, and such schemes include bootstrap, subsampling, and cross-validation. Moreover, we briefly review results in the probability literature to explain that stability is driving limiting results such as the Central Limit Theorem, which is a foundation for classical asymptotic statistics.

Using these three problems as backdrop, we make four points. Firstly, statistical stability considerations can effectively aid the pursuit for interpretable and reliable scientific models, especially in high-dimensions. Stability in a broad sense includes replication, repeatability, and different data perturbation schemes. Secondly, stability is a general principle on which to build statistical methods for different purposes. Thirdly, the meaning of stability needs articulation in high-dimensions because it could be brought about by sample variability and/or heavy tails in the errors of a linear regression model. Last but not least, emphasis should be placed on the stability aspects of statistical inference and conclusions, in the referee process of scientific and applied statistics papers and in our current statistics curriculum.

Statistical stability in the age of massive data is an important area for research and action because high-dimensions provide ample opportunities for instability to reveal itself to challenge reproducibility of scientific findings.

As we began this article with words of Tukey, it seems fitting to end also with his words:

"What of the future? The future of data analysis can involve great progress, the overcoming of real difficulties, and the provision of a great service to all fields of science and technology. Will it? That remains to us, to our willingness to take up the rocky road of real problems in preferences to the smooth road of unreal assumptions, arbitrary criteria, and abstract results without real attachments. Who is for the challenge?" – Tukey (p. 64, 1962).

Acknowledgements

This paper is based on the 2012 Tukey Lecture of the Bernoulli Society delivered by the author at the 8th World Congress of Probability and Statistics in Istanbul on July 9, 2012. For their scientific influence and friendship, the author is indebted to her teachers/mentors/colleagues, the late Professor Lucien Le Cam, Professor Terry Speed, the late Professor Leo Breiman, Professor Peter Bickel, and Professor Peter Bühlmann. This paper is invited for the special issue of *Bernoulli* commemorating the 300th anniversary of the publication of Jakob Bernoulli's Ars Conjectandi in 1712.

The author would like to thank Yuval Benjamini for his help on generating the results in the figures. She would also like to thank two referees for their detailed and insightful comments, and Yoav Benjamini and Victoria Stodden for helpful discussions. Partial supports are gratefully acknowledged by NSF Grants SES-0835531 (CDI) and DMS-11-07000, ARO Grant W911NF-11-1-0114, and the NSF Science and Technology Center on Science of Information through Grant CCF-0939370.

References

Allen, D.M. (1974). The relationship between variable selection and data augmentation and a method for prediction. *Technometrics* **16** 125–127. MR0343481

Andrews, D.F., Bickel, P.J., Hampel, F.R., Huber, P.J., Rogers, W.H. and Tukey, J.W. (1972). *Robust Estimates of Location*: *Survey and Advances*. Princeton, NJ: Princeton Univ. Press. MR0331595

Atkil, H., Martone, M.E. and Essen, D.C.V. (2012). Challenges and opportunities in mining neuroscience data. *Science* **331** 708–712.

Bach, F. (2008). Bolasso: Model consistent lasso estimation through the bootstrap. In *Proc. of ICML*. Helsinki, Finland.

Bean, D., Bickel, P.J., El Karoui, N. and Yu, B. (2013). Optimal M-estimation in high-dimensional regression. *Proc. Natl. Acad. Sci. USA*. To appear.

Beran, R. (1984). Bootstrap methods in statistics. *Jahresber. Deutsch. Math.-Verein*. **86** 14–30. MR0736625

Bickel, P.J. (1975). One-step Huber estimates in the linear model. *J. Amer. Statist. Assoc*. **70** 428–434. MR0386168

Bickel, P.J. and Freedman, D.A. (1981). Some asymptotic theory for the bootstrap. *Ann. Statist*. **9** 1196–1217. MR0630103

Bickel, P.J., Götze, F. and van Zwet, W.R. (1997). Resampling fewer than n observations: Gains, losses, and remedies for losses. *Statist. Sinica* **7** 1–31. MR1441142

Booth, B. (2012). Scientific reproducibility: Begley's six rules. *Forbes* September 26.

Bousquet, O. and Elisseeff, A. (2002). Stability and generalization. *J. Mach. Learn. Res*. **2** 499–526. MR1929416

Breiman, L. (1996). Heuristics of instability and stabilization in model selection. *Ann. Statist*. **24** 2350–2383. MR1425957

Carlstein, E. (1986). The use of subseries values for estimating the variance of a general statistic from a stationary sequence. *Ann. Statist.* **14** 1171–1179. MR0856813

Casadevall, A. and Fang, F.C. (2011). Reforming science: Methodological and cultural reforms. *Infection and Immunity* **80** 891–896.

Chatterjee, S. (2006). A generalization of the Lindeberg principle. *Ann. Probab.* **34** 2061–2076. MR2294976

Dayan, P. and Abbott, L.F. (2005). *Theoretical Neuroscience: Computational and Mathematical Modeling of Neural Systems*. Cambridge, MA: MIT Press. MR1985615

Devroye, L.P. and Wagner, T.J. (1979). Distribution-free inequalities for the deleted and holdout error estimates. *IEEE Trans. Inform. Theory* **25** 202–207. MR0521311

Donoho, D.L., Maleki, A., Shahram, M., Rahman, I.U. and Stodden, V. (2009). Reproducible research in computational harmonic analysis. *IEEE Computing in Science and Engineering* **11** 8–18.

Efron, B. (1979). Bootstrap methods: Another look at the jackknife. *Ann. Statist.* **7** 1–26. MR0515681

Efron, B. (1982). *The Jackknife, the Bootstrap and Other Resampling Plans. CBMS-NSF Regional Conference Series in Applied Mathematics* **38**. Philadelphia, PA: SIAM. MR0659849

Efron, B. and Tibshirani, R.J. (1993). *An Introduction to the Bootstrap. Monographs on Statistics and Applied Probability* **57**. New York: Chapman & Hall. MR1270903

El Karoui, N., Bean, D., Bickel, P.J., Lim, C. and Yu, B. (2013). On robust regression with high-dimensional predictors. *Proc. Natl. Acad. Sci. USA*. To appear.

Fonio, E., Golani, I. and Benjamini, Y. (2012). Measuring behavior of animal models: Faults and remedies. *Nature Methods* **9** 1167–1170.

Goodale, M.A. and Milner, A.D. (1992). Separate visual pathways for perception and action. *Trends Neurosci.* **15** 20–25.

Hall, P. (1983). Large sample optimality of least squares cross-validation in density estimation. *Ann. Statist.* **11** 1156–1174. MR0720261

Hampel, F.R. (1968). Contributions to the theory of robust estimation. Ph.D. thesis, Univ. California, Berkeley. MR2617979

Hampel, F.R. (1971). A general qualitative definition of robustness. *Ann. Math. Statist.* **42** 1887–1896. MR0301858

Hampel, F.R. (1974). The influence curve and its role in robust estimation. *J. Amer. Statist. Assoc.* **69** 383–393. MR0362657

Hampel, F.R., Ronchetti, E.M., Rousseeuw, P.J. and Stahel, W.A. (1986). *Robust Statistics: The Approach Based on Influence Functions. Wiley Series in Probability and Mathematical Statistics: Probability and Mathematical Statistics*. New York: Wiley. MR0829458

Hartigan, J.A. (1969). Using subsample values as typical values. *J. Amer. Statist. Assoc.* **64** 1303–1317. MR0261737

Hartigan, J.A. (1975). Necessary and sufficient conditions for asymptotic joint normality of a statistic and its subsample values. *Ann. Statist.* **3** 573–580. MR0391346

Hinkley, D.V. (1977). Jacknifing in unbalanced situations. *Technometrics* **19** 285–292. MR0458734

Hoerl, A.E. (1962). Application of ridge analysis to regression problems. *Chemical Engineering Progress* **58** 54–59.

Hoerl, A.E. and Kennard, R.W. (1970). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics* **42** 80–86.

Hubel, D.H. and Wiesel, T.N. (1959). Receptive fields of single neurones in the cat's striate cortex. *Journal of Physiology* **148** 574–591.

Huber, P.J. (1964). Robust estimation of a location parameter. *Ann. Math. Statist.* **35** 73–101. MR0161415

Huber, P.J. (1981). *Robust Statistics*. New York: Wiley. MR0606374

Huber, P.J. (2002). John W. Tukey's contributions to robust statistics. *Ann. Statist.* **30** 1640–1648. MR1969444

Ioannidis, J.P.A. (2005). Why most published research findings are false. *PLoS Med.* **2** 696–701.

Kay, K.N. and Gallant, J.L. (2009). I can see what you see. *Nat. Neurosci.* **12** 245.

Kay, K.N., Naselaris, T., Prenger, R.J. and Gallant, J.L. (2008). Identifying natural images from human brain activity. *Nature* **452** 352–355.

Kearns, M. and Ron, D. (1999). Algorithmic stability and sanity-check bounds for leave-one-out cross-validation. *Neural Comput.* **11** 1427–1453.

Kraft, P., Zeggini, E. and Ioannidis, J.P.A. (2009). Replication in genome-wide association studies. *Statist. Sci.* **24** 561–573. MR2779344

Künsch, H.R. (1989). The jackknife and the bootstrap for general stationary observations. *Ann. Statist.* **17** 1217–1241. MR1015147

Kutin, S. and Niyogi, P. (2002). Almost-everywhere algorithmic stability and generalization error. In *Proc. of UAI: Uncertainty in Artificial Intelligence 18*.

Li, K.C. (1986). Asymptotic optimality of C_L and generalized cross-validation in ridge regression with application to spline smoothing. *Ann. Statist.* **14** 1101–1112. MR0856808

Lim, C. and Yu, B. (2013). Estimation stability with cross-validation (ES-CV). Available at arXiv.org/abs/1303.3128.

Mahalanobis, P. (1946). Sample surveys of crop yields in India. *Sankhyā*, *Series A* **7** 269–280.

Markovich, N. (2007). *Nonparametric Analysis of Univariate Heavy-Tailed Data*: *Research and Practice*. *Wiley Series in Probability and Statistics*. Chichester: Wiley. MR2364666

Meinshausen, N. and Bühlmann, P. (2006). High-dimensional graphs and variable selection with the lasso. *Ann. Statist.* **34** 1436–1462. MR2278363

Meinshausen, N. and Bühlmann, P. (2010). Stability selection. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **72** 417–473. MR2758523

Miller, R.G. (1974). The jackknife—A review. *Biometrika* **61** 1–15. MR0391366

Mukherjee, S., Niyogi, P., Poggio, T. and Rifkin, R. (2006). Learning theory: Stability is sufficient for generalization and necessary and sufficient for consistency of empirical risk minimization. *Adv. Comput. Math.* **25** 161–193. MR2231700

Naik, G. (2011). Scientists' elusive goal: Reproducing study results. *Wall Street Journal (Health Industry Section)* December 2.

Naselaris, T., Prenger, R.J., Kay, K.N. and Gallant, M.O.J.L. (2009). Bayesian reconstruction of natural images from human brain activity. *Neuron* **63** 902–915.

Naselaris, T., Kay, K.N., Nishimoto, S. and Gallant, J.L. (2011). Encoding and decoding in fmri. *Neuroimage* **56** 400–410.

Nishimoto, S., Vu, A.T., Naselaris, T., Benjamini, Y., Yu, B. and Gallant, J.L. (2011). Reconstructing visual experiences from brain activity evoked by natural movies. *Current Biology* **21** 1641–1646.

Nosek, B.A., Spies, J.R. and Motyl, M. (2012). Scientific utopia: II. Restructuring incentives and practices to promote truth over publishability. In *Proc. of CoRR*.

Olshausen, B.A. and Field, D.J. (1996). Emergence of simple-cell receptive field properties by learning a sparse code for natural images. *Nature* **381** 607–609.

Politis, D.N. and Romano, J.P. (1992). A general theory for large sample confidence regions based on subsamples under minimal assumptions. Technical Report 399. Dept. Statistics, Stanford Univ.

Politis, D.N., Romano, J.P. and Wolf, M. (1999). *Subsampling*. New York: Springer. MR1707286

Portnoy, S.L. (1977). Robust estimation in dependent situations. *Ann. Statist.* **5** 22–43. MR0445716

Quenouille, M.H. (1949). Approximate tests of correlation in time-series. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **11** 68–84. MR0032176

Quenouille, M.H. (1956). Notes on bias in estimation. *Biometrika* **43** 353–360. MR0081040

Shalev-Shwartz, S., Shamir, O., Srebro, N. and Sridharan, K. (2010). Learnability, stability and uniform convergence. *J. Mach. Learn. Res.* **11** 2635–2670. MR2738779

Shao, J. (1996). Bootstrap model selection. *J. Amer. Statist. Assoc.* **91** 655–665. MR1395733

Shao, J. and Tu, D.S. (1995). *The Jackknife and Bootstrap*. New York: Springer. MR1351010

Steen, R.G. (2011). Retractions in the scientific literature: Do authors deliberately commit fraud? *J. Med. Ethics* **37** 113–117.

Stodden, V. (2011). Trust your science? Open your data and code. *AMSTATNEWS*. Available at http://magazine.amstat.org/blog/2011/07/01/trust-your-science/.

Stone, M. (1974). Cross-validatory choice and assessment of statistical prediction. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **36** 111–147.

Suidan, T. (2006). A remark on a theorem of Chatterjee and last passage percolation. *J. Phys. A* **39** 8977–8981. MR2240468

Tao, T. (2012). Lecture notes on the central limit theorem. Available at http://terrytao.wordpress.com/2010/01/05/254a-notes-2-the-central-limit-theorem/.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **58** 267–288. MR1379242

Tikhonov, A.N. (1943). On the stability of inverse problems. *Doklady Akademii Nauk SSSR* **39** 195–198.

Tukey, J.W. (1958). Bias and confidence in not quite large samples. *Ann. Math. Statist.* **29** 614.

Tukey, J.W. (1962). The future of data analysis. *Ann. Math. Statist.* **33** 1–67. MR0133937

Wainwright, M.J. (2009). Sharp thresholds for high-dimensional and noisy sparsity recovery using ℓ_1-constrained quadratic programming (Lasso). *IEEE Trans. Inform. Theory* **55** 2183–2202. MR2729873

Wu, C.F.J. (1986). Jackknife, bootstrap and other resampling methods in regression analysis (with discussion). *Ann. Statist.* **14** 1261–1295. MR0868303

Zhang, P. (1993). Model selection via multifold cross validation. *Ann. Statist.* **21** 299–313. MR1212178

Zhao, P. and Yu, B. (2006). On model selection consistency of Lasso. *J. Mach. Learn. Res.* **7** 2541–2563. MR2274449

Zou, H. (2006). The adaptive lasso and its oracle properties. *J. Amer. Statist. Assoc.* **101** 1418–1429. MR2279469