Discussion on 'Confidence Intervals and Hypothesis Testing for High-Dimensional Regression'

Yen-Chi Chen and Yu-Xiang Wang
Department of Statistics and Machine Learning
Carnegie Mellon University

This is a discussion on the work in Javanmard and Montanari (2013). We begin with a short introduction to statistical inference and then briefly summarize the basic ideas in this paper. We discuss how they design the novel algorithm to de-biase LASSO to arrives limiting distributions and what type of tests they have considered in this paper. In the last section, we compare several common methods in high dimensional inference such as de-baised methods, selective methods and resampling methods.

1 Introduction to Statistical Inference

This paper deals with statistical inference. Two major topics for statistical inference are hypothesis test and parameter inference (at some situations like this paper, these two topics coincide). In what follows we briefly introduce these two topics and show that the limiting distributions for test statistics and estimated parameter are essential for statistical inference.

1.1 Hypothesis Test

Hypothesis test is the main inferential approach used on the scientific research. Hypothesis test is essentially a proof by contradiction. We want to prove the alternative hypothesis is correct so we construct the null hypothesis H_0, consists of situations that complement to alternative, and we try to argue that the null hypothesis contradicts to the data we observed.

The standard approach to conduct hypothesis test is first to find a test statistics T_n that summarize information from the data. Then we study the behavior of T_n under null hypothesis. When we observed data $x_1, \cdots, x_n$, we can compute the value of this test statistics, called it $t_n = T_n(x_1, \cdots, x_n)$. If t_n is far away from what usually T_n under H_0 is, then H_0 is unlikely to be true so that we reject H_0. The region that we reject H_0 is called rejection region, often denoted as $\mathcal{R}$. However, given a rejection rule ($T_n \in \mathcal{R}$), we may falsely reject H_0. That is, H_0 is correct but due to the randomness of sample, our test statistics behaves like H_0 is not true. This error rate (falsely reject H_0) is called the type-1 error. Since the hypothesis test is mainly for scientific inference and we are trying to prove the alternative hypothesis H_A. We need to provide 'sufficient' evidence to conclude H_0 is incorrect. Controlling the type-1 error provides a quantitative way to require the 'sufficiency of evidence' to reject H_0. If we require a smaller type-1 error to reject H_0, this means that we need 'more sufficient' evidence to do so. The requirement for the sufficiency of evidence, or the level for type-1 error we wish to constrol is call 'significance level', often denoted as α. Thus, given the significance level α, the rejection region $\mathcal{R} = \mathcal{R}_\alpha$ depends on α since we need to make sure we have controlled type-1 error. Note that the significance level α in practice should be determined by the scientists. Some common example is $\alpha = 0.1, 0.05, 0.01$ but at some cases we may require $\alpha = 10^{-5}$ or even smaller.

The limiting distribution for T_n is vital to the above standard approach for hypothesis test since we need to find the rejection region $\mathcal{R}_\alpha$ under significance level α. However, in general it is hard to obtain such a nice region unless we know the distribution of test statistics T_n. In many cases, T_n has a very complicated distribution for finite n. Thus, we are allowed to use the knowledge of limiting distribution for T_n to construct rejection region $\mathcal{R}_\alpha$. At least the $\mathcal{R}_\alpha$ from limiting distribution asymptotically controls the type-1 error rate. Therefore, we need limiting distribution to T_n

1.2 P-value

The p-value is an alternative idea that directly measures the 'support/contradiction' of null hypothesis H_0 given the data and test statistics. In short, p-value is the measure of 'support' for H_0 by the test statistics under the data. A smaller p-value indicates less support for H_0 and equivalently, more contradiction to H_0. Note that p-value depends on the test statistics; p-value will be different for two test statistics under the same H_0 and data. Computing p-value does not require significance level α. However, if we want to conduct the test under α, all we need to do is to compare p-value and α and we reject H_0 if p-value is less than α. P-value is like an universal quantity for all the test statistics. Even we have different test statistics for different hypothesis using different data, the p-value between all of them are comparable to each other since it is a measure of support for H_0. Thus, this allows us to compare p-values while doing multiple test which is why the False Discovery Rate (Benjamini and Hochberg, 1995) uses the rank of p-value.

The definition of p-value is from the rejection region $\mathcal{R}_\alpha$ so that the computation of p-value requires the limiting distribution of test statistics. To be short, the p-value from a given data is defined as follows. Given the data, we realize the test statistics. The p-value is the smallest α that we the realized test statistics falls within $\mathcal{R}_\alpha$ (Wasserman, 2004).

1.3 Confidence Interval/Set

Another topic of statistical inference is parameter inference which concerns how can we construct a 'random region' in the parameter space based on the data such that this random region has certain probability to cover the true parameter. In usual parametric model, parameters are like population mean μ or the regression coefficient β. In general, parameter can be a function like density function p or a set e.g. density level set $L_\lambda = \{x : p(x) \geq \lambda\}$ or even a manifold $\mathcal{M}$ (see Chen et al. (2014) for an example on constructing confidence set for manifolds). Let θ be the true parameter (parameter of interest, some characteristics for the population). A valid/proper $(1-\alpha)\%$ confidence set $C_{n,\alpha} = C(X_1, \cdots, X_n; \alpha)$ is a random set such that

$$\mathbb{P}(\theta \in C_{n,\alpha}) \geq 1 - \alpha + o(1).$$

Note that the randomness for the above is from the confidence set C_n rather than the parameter θ; θ is an unknown but fixed quantity. The quantity $(1-\alpha)\%$ is called coverage/confidence, a quality for the 'procedure' (for constructing $C_{n,\alpha}$), rather than the significance level although in many cases they coincide. The $(1-\alpha)\%$ confidence/coverage is intepreted as follows. Imagine that we can resample a size n sample from the true population for several times. (Each time while we are sampling, we allow the parameter changes.) For every size n sample, we use the same procedure to construct a confidence set. If this procudure has $(1-\alpha)\%$ confidence, then in the long run we will have $(1-\alpha)\%$ ratio of total confidence set that cover the corresponding parameter.

We can (and often) invert the parameter estimate $\widehat{\theta}_n$ to construct confidence set C_n and this requires the knowledge about the distribution of $\widehat{\theta}_n$. Let θ_0 be the true parameter. The simplest way to build C_n is first find $\widehat{\theta}_n$ and then study $\widehat{\theta}_n - \theta_0$ (or more general, the metric $d(\widehat{\theta}, \theta)$), and C_n is the collection of θ such that $|\widehat{\theta}_n - \theta_0| \leq a_{n,\alpha}$ where $a_{n,\alpha}$ is picked from the distribution of $\widehat{\theta}_n$. The analytical form of $a_{n,\alpha}$ requires the distribution (limiting distribution is allowed) of $\widehat{\theta}_n - \theta_0$ and this is why we need the limiting distribution.

In this paper, we are working on a regression problem and the parameters of interest are the regression coefficients θ_0. The null hypothesis is whether each coefficient is 0 or not. In this case, a typical test statistics is the estimate for the regression coefficient so that all the above three topics require the same limiting distribution of $\widehat{\theta}_n - \theta_0$. This is a non-trivial problem since most of the asymptotic theories for regression are valid only for dimension p being fixed and $n \to \infty$. In the next section, we will show how to find the limiting distribution in high dimensional settings.

2 Basic Ideas for the Paper

Let $X_1, \cdots, X_n \in \mathbb{R}^p$ be the feature and $Y_1, \cdots, Y_n \in \mathbb{R}$ be the response. For simplicity, let $\mathbb{X} = [X_1, \cdots, X_n] \in \mathbb{R}^{n \times p}$ and $Y = (Y_1, \cdots, Y_n) \in \mathbb{R}^n$. We consider the linear model such that
(Assumption 1) $Y = \mathbb{X}\theta_0 + W, \quad W \sim N(0, \sigma^2 \mathbb{I}_n)$.
Note we allow $p = p_n \to \infty$ with sparsity. i.e. $\mathsf{supp}(\theta_0) = s << p$.

The usual method to attack this problem is by LASSO. Namely,

$$\widehat{\theta}^n(\lambda) = \underset{\theta \in \mathbb{R}^p}{argmin} \frac{1}{2n} \|Y - \mathbb{X}\theta\|_2^2 + \lambda \|\theta\|_1. \tag{1}$$

In this paper, they propose a de-biased method modified from the LASSO estimate:

$$\widehat{\theta}^*(\lambda, M) = \widehat{\theta}^n(\lambda) + \frac{1}{n} M \mathbb{X}^T (Y - \mathbb{X}\widehat{\theta}^n(\lambda)), \tag{2}$$

where $M \in \mathbb{R}^{p \times p}$ is an arbitrary matrix. This de-biased method has been proposed by van de Geer et al. (2013); Zhang and Zhang (2014); in these papers, they used surrogate inverse covariance matrix (via a method called nodewise regression (Meinshausen and Bühlmann, 2006)) for the matrix M.

The main breakthrough for this paper is that under suitable choice of M, we can weaken the coherence assumption on $\mathbb{X}$ (for the common conditions in LASSO literatures, see Van De Geer et al. (2009) for a comparison). Let $\Sigma = \mathsf{Cov}(X_i)$. We assume the following:
(Assumption 2) each row of $\mathbb{X}\Sigma^{-1/2}$ is independent and has subgaussian norm.

How do we weaken the coherence assumption? The key is on the quantity

$$\mu_*(\mathbb{X}; M) = |M\widehat{\Sigma}_n - \mathbb{I}_p|_\infty,$$

where $\widehat{\Sigma}_n = \frac{1}{n}\mathbb{X}^T\mathbb{X}$ is the sample covariance matrix. Theorem 2.3 shows that under assumption 1 and 2 and suitable choice of $\lambda = \lambda_n$,

$$\sqrt{n}(\widehat{\theta}^* - \theta_0) = Z + \Delta, \quad Z \sim N(0, \sigma^2 M\widehat{\Sigma}_n M^T), \quad \Delta = \sqrt{n}(M\widehat{\Sigma}_n - \mathbb{I}_p)(\theta_0 - \widehat{\theta}^n). \tag{3}$$

Thus, once we can show $\Delta = o_P(1)$, then by Slutsky theorem we obtain limiting distribution for $\sqrt{n}(\widehat{\theta}^* - \theta_0)$.

Someone may ask: in general, do we have $\Delta = o_P(1)$?

- For the simple case, $p < n$, it is not hard to see $|M\widehat{\Sigma}_n - \mathbb{I}_p|_\infty = 0$ so that we have the desire result.
- Under RIP condition, this is also true.
- In more general case, when we assume Assumption 2, this is also true.

In the what follows we derive how Assumption 2 helps us to obtain consistency for Δ.

A deeper look at Δ shows that the key element is $M\widehat{\Sigma}_n - \mathbb{I}_p$ since the other term involving LASSO solution is fixed. Thus, this motivates us to pick M minimizing $|M\widehat{\Sigma}_n - \mathbb{I}_p|_\infty$ and this is what algorithm 1 is doing. Moreover, also in Theorem 2.3, they show a concentration for Δ:

$$\mathbb{P}\left(\|\Delta\|_\infty \geq \frac{4c\mu_*\sigma s_0}{\phi_0^2}\sqrt{\log p}\right) \leq 2p^{-c_0}, \tag{4}$$

where c, c_0 are constants. This concentration depends on other 5 terms, namely, noise level σ, dimension of data p, sparsity s_0 and two other terms $\mu_* = \mu_*(\mathbb{X}, M)$ and compatibility ϕ_0^2. By applying algorithm 1, μ_* can be replaced by $\mu_{\min}$ and they show in Theorem 4 that both ϕ_0 and $\mu_{\min}$ have a concentration bound. ϕ_0 is bounded from belowed by a constant and $\mu_{\min}$ is at rate $O(\sqrt{\frac{\log p}{n}})$. Thus, putting altogether, we obtain

$$\|\Delta\|_\infty = O_P\left(\frac{s_0 \log p}{\sqrt{n}}\right). \tag{5}$$

A remark on the proof. The key idea on obtaining the decomposition (3) is as follows:

$$\begin{aligned}
\sqrt{n}(\widehat{\theta}^* - \theta_0) =& \sqrt{n}\left(\widehat{\theta}^n(\lambda) + \frac{1}{n}M\mathbb{X}^T(Y - \mathbb{X}\widehat{\theta}^n(\lambda)) - \theta_0\right) \\
=& \sqrt{n}\left(\widehat{\theta}^n(\lambda) + \frac{1}{n}M\mathbb{X}^T(\mathbb{X}\theta_0 + W - \mathbb{X}\widehat{\theta}^n(\lambda)) - \theta_0\right) \\
=& \sqrt{n}\left(\widehat{\theta}^n(\lambda)\left[I - \frac{1}{n}M\mathbb{X}^T\mathbb{X}\right] - \left[I - \frac{1}{n}M\mathbb{X}^T\mathbb{X}\right]\theta_0\right) \\
=& \sqrt{n}\left[I - \frac{1}{n}M\mathbb{X}^T\mathbb{X}\right]\left(\widehat{\theta}^n(\lambda) - \theta_0\right) + \frac{1}{\sqrt{n}}M\mathbb{X}^T W
\end{aligned}$$

The first term is what we called Δ and it is bounded using H older's inequality via

$$\sqrt{n}\left\|I-\frac{1}{n}M\mathbb{X}^T\mathbb{X}\right\|_\infty\left\|\widehat{\theta}^n(\lambda)-\theta_0\right\|_1.$$

The 1-norm term is bounded by the existing results for lasso and the entrywise ∞-norm is bounded by inventing a so-called "generalized incoherence condition". Since W is normal, the second term is also normal with covariance matrix $\sigma^2 M\left[\frac{1}{n}\mathbb{X}^T\mathbb{X}\right]M^T$. This is what they call Z. Under their assumption, Δ goes to 0 faster than Z, therefore we say the estimator is asymptotically unbiased.

3 Statistical Inference

Now we dicuss the statistical inference presented in this paper. We are working on high dimensional regression and the parameters of interest are the regression coefficients β. In this case, conducting hypothesis test and constructing confidence sets are equivalent.

For high dimension regression, the hypothesis test has two features

1. Multiple test. i.e. $H_{0,j}:\beta_j=0$, for each $j=1,\cdots,p$.
2. Number of test increases with sample size n.

In this paper, they study three testing scheme:

1. Individual test. We test only one hypothesis $H_{0,j}:\beta_j=0$. Note that $j=j(n)$ may change according to the sample size n.
2. Subset test. We test a fixed size multiple parameters. In particular, they construct a confidence set for the subset of parameters.
3. Multiple test. They constrol FWER (familywise error rate) for multiple testing.

From what we have derived, there are good news and bad news.

- Good news. By the above derivation, we have obtained a limiting distribution.
- Bad news. The limiting distribution involves unknow quantity σ^2.

If we somehow get a consistent estimator to σ^2 such that $|\frac{\widehat{\sigma}}{\sigma}-1|=o_P(1)$, by Lemma 3.2 we have

$$\lim_{n\to\infty}\sup_{\theta_0\in\mathbb{R}^p;\|\theta_0\|_0\le s_0}\left\|\mathbb{P}\left(\frac{\sqrt{n}(\widehat{\theta}_i^*-\theta_{0,i})}{\widehat{\sigma}[M\widehat{\Sigma}_nM^T]_{i,i}^{1/2}}\le x\right)-\Phi(x)\right\|=0\quad a.s. \tag{6}$$

- This is a pointwise convergence. i.e. convergence in distribution, not Berry-Eseen type bound.

- This is for i being fixed.
- However, this is uniform in sparsity settings (different combination for θ_0).

We can invert this for the confidence interval or do the hypothesis test. They show consistency for both cases. In particular, they conduct a power analysis for a signal strength $\gamma > 0$ case. Let

$$\beta_{i,n}(\gamma) = \sup\{\mathbb{P}(\mathsf{NotReject} H_0) : \theta_0 \in \mathbb{R}^p, \|\theta_0\|_0 \leq s_0, |\theta_{0,i}| \geq \gamma\} \tag{7}$$

be the type 2 errors. Then the power function, which is $1 - \beta_{i,n}(\gamma)$, can be shown

$$\liminf_n \frac{1 - \beta_{i,n}(\gamma)}{1 - \beta^*_{i,n}(\gamma)} \geq 1, \quad 1 - \beta^*_{i,n}(\gamma) = G\left(\alpha, \frac{\sqrt{n}\gamma}{\sigma[\Sigma^{-1}_{i,i}]^{1/2}}\right) = 1 - o(1),$$

where $G(\alpha, \mu) = 2 - \Phi(\Phi^{-1}(1 - \frac{\alpha}{2}) + \mu) - \Phi(\Phi^{-1}(1 - \frac{\alpha}{2}) - \mu)$ is the power function for testing standard normal with mean 0 (null) under the true mean equals μ. Thus, the power is kind of lower bounded by the power to a known problem and they further show that this power is nearly optimal (differ by a constant factor).

Subset Test. They do not explicitly frame this problem in terms of hypothesis test but instead they look for the confidence set for a subset of parameter. Let $R \subset \{1, \cdots, p\}$ be a subset of parameter and $|R| = k$. i.e. the number of elemtents in R is fixed to be k. Then by simple generalization from the limiting distribution, we can construct $C_R(\alpha)$ such that

$$\mathbb{P}(\theta_{0,R} \in C_R(\alpha)) \geq 1 - \alpha + o(1),$$

where $\theta_{0,R}$ is the element of θ_0 within the subset R.

Multiple Test. For multiple test, they consider two settings: a fixed size simultaneous test and a test based on FWER (family wise error rate). For the fixed size simultaneous test, the basic idea is to generalize (6) to multivariate case. Note that they fixed the number of parameter to be tested when $n \to \infty$. For the FWER test, just simply control FWER to carry out the test. Let $\widehat{T}^F$ be their test and $\mathsf{FWER}(T, n)$ be the FWER for a test T under sample size n. They show that

$$\limsup_{n\to\infty} \mathsf{FWER}(\widehat{T}^F, n) \leq \alpha.$$

The basic idea for this result follows from the fact that the error decays at exponential rate due to normality. Even if we have p number of tests to carry out, this is a polynomial factor and is negligible compared to the exponential rate.

Non-Gaussian Case. The assumption 1 may not be necessary for individual statistical inference (testing for a given i parameter). The main observation is from the normality term Z in (3). In the proof of theorem 2.3, they show that $Z = M\mathbb{X}^T W/\sqrt{n}$. Thus, when we focus on one particular parameter, say $\theta_{0,i}$, we have

$$\begin{aligned}\sqrt{n}(\widehat{\theta}^*_{0,i} - \theta_{0,i}) &= \frac{1}{\sqrt{n}} m_i^T \mathbb{X}^T W + o_P(1)\\ &= \frac{1}{\sqrt{n}} m_i^T \sum_{j=1}^n X_j W_j + o_P(1),\end{aligned}$$

where m_i is the i-th column of M. Thus, if the random variable $A_j = X_j W_j$ satisfies Lindeberg-Feller's condition (in CLT), we still have asymptotic normality even if W_j is not normal.

4 Other High Dimensional Inferences

- **De-bias method.** A common approach to do inference in regression under high dimensional setting is to 'debias' the LASSO. That is,

  $$\widehat{\theta}_n = \widehat{\theta}_{LASSO,n} + A_n, \qquad A_n = \frac{1}{n} S \mathbb{X}^T (Y - \mathbb{X}\widehat{\theta}_{LASSO,n}(\lambda)),$$

  where S is some matrix. Note that this is exact the same as in (2). The main issue is how to construct a data-driven S such that it eliminate the bias in LASSO. For related articles, see van de Geer et al. (2013); Zhang and Zhang (2014); Caner and Kock (2014). Note that van de Geer has a series of work on the de-biased method; see Jankova and van de Geer (2014) for confidence interval and van de Geer (2014) for a summary for their works.

- **Selective method.** An alternative method to do high dimensional inference by LASSO is a sequential and selective method called covariance test (Lockhart et al., 2014). Note that this is sequentially conditional test and the null hypothesis is random (what to test depends on what passed in previous test). For instance, at step k ($\lambda = \lambda_k$), we are testing

  $$H_0 : \mathsf{supp}(\theta_0) \subseteq A_k,$$

  where A_k is the active set (parameters selected by LASSO) at step k. The covariance test can be applied to graphical LASSO (G'Sell et al., 2013) and a modified method called spacing test is recently proposed (Taylor et al., 2014).

- **Resampling method.** We use resampling method (including data splitting, bootstrap) to conduct hypothesis test in high dimension. For the sample splitting, Wasserman and Roeder (2009) splits the data and conduct a screening (using the first two subsample) and cleaning (use the remaining subsample) to carry out variable selection based on t-test. The screening and cleaning procedure can also be used in high dimensional estimation; see Jin et al. (2012); Huang et al. (2014) for example.

  The bootstrap can also be applied in inference under high dimensional settings. The theoretical foundation for this idea is given in Chernozhukov et al. (2013). Essentially, they prove that for two sequences of random vectors $X_1, \cdots, X_n \in \mathbb{R}^p$ and $Y_1, \cdots, Y_n \in \mathbb{R}^p$ such that each X_i is iid with mean 0 and covariance matrix Σ and each Y_i is iid 'multivariate normal' with mean 0 and covariance matrix Σ. Let $\bar{X}_n = \frac{1}{n}\sum_{i=1}^n X_i \in \mathbb{R}^p$ be the mean of X_i's and $\bar{Y}_n$ is defined similarly. Denote $\bar{X}_{n,j}$ be the j-th component of $\bar{X}_n$ for $j = 1, \cdots, p$. Then

  $$\sup_t \left\| \mathbb{P}\left(\sqrt{n} \max_{1\leq j\leq p} \bar{X}_{n,j} \leq t\right) - \mathbb{P}\left(\sqrt{n} \max_{1\leq j\leq p} \bar{Y}_{n,j} \leq t\right) \right\| \to 0.$$

  Note this is a Berry-Esseen type bound. With this result, they further show the bootstrap consistently approaches the limiting distribution. Wasserman et al. (2013) applied this idea to estimating undirected graph in high dimension and Zhang and Cheng (2014) extends this result to high dimensional time series.

- Other testing problems

- **Testing independence.** See Han and Liu (2014) for a non-parametric rank test for testing independence.
- **Two sample test.** See Hall and Tajvidi (2002) for a permutation test and Reddi et al. (2014) for a test based on Maximum Mean Discrepancy.
- **Testing influential observation.** We can also test if an observation is influential or not in high dimensional settings; see Zhao et al. (2013).

References

Y. Benjamini and Y. Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 289–300, 1995.

M. Caner and A. B. Kock. Asymptotically honest confidence regions for high dimensional parameters by the desparsified conservative lasso. *arXiv preprint arXiv:1410.4208*, 2014.

Y.-C. Chen, C. R. Genovese, and L. Wasserman. Asymptotic theory for density ridges. *arXiv preprint arXiv:1406.5663*, 2014.

V. Chernozhukov, D. Chetverikov, K. Kato, et al. Gaussian approximations and multiplier bootstrap for maxima of sums of high-dimensional random vectors. *The Annals of Statistics*, 41(6): 2786–2819, 2013.

M. G. G'Sell, J. Taylor, and R. Tibshirani. Adaptive testing for the graphical lasso. *arXiv preprint arXiv:1307.4765*, 2013.

P. Hall and N. Tajvidi. Permutation tests for equality of distributions in high-dimensional settings. *Biometrika*, 89(2):359–374, 2002.

F. Han and H. Liu. Distribution-free tests of independence with applications to testing more structures. *arXiv preprint arXiv:1410.4179*, 2014.

S. Huang, J. Jin, and Z. Yao. Partial correlation screening for estimating large precision matrices, with applications to classification. *arXiv preprint arXiv:1409.3301*, 2014.

J. Jankova and S. van de Geer. Confidence intervals for high-dimensional inverse covariance estimation. *arXiv preprint arXiv:1403.6752*, 2014.

A. Javanmard and A. Montanari. Confidence intervals and hypothesis testing for high-dimensional regression. *arXiv preprint arXiv:1306.3171*, 2013.

J. Jin, C.-H. Zhang, and Q. Zhang. Optimality of graphlet screening in high dimensional variable selection. *arXiv preprint arXiv:1204.6452*, 2012.

R. Lockhart, J. Taylor, R. J. Tibshirani, R. Tibshirani, et al. A significance test for the lasso. *The Annals of Statistics*, 42(2):413–468, 2014.

N. Meinshausen and P. Bühlmann. High-dimensional graphs and variable selection with the lasso. *The Annals of Statistics*, pages 1436–1462, 2006.

S. J. Reddi, A. Ramdas, B. Póczos, A. Singh, and L. Wasserman. Kernel mmd, the median heuristic and distance correlation in high dimensions. *arXiv preprint arXiv:1406.2083*, 2014.

J. Taylor, R. Lockhart, R. J. Tibshirani, and R. Tibshirani. Exact post-selection inference for forward stepwise and least angle regression. *arXiv preprint arXiv:1401.3889*, 2014.

S. van de Geer. Statistical theory for high-dimensional models. *arXiv preprint arXiv:1409.8557*, 2014.

S. van de Geer, P. Bühlmann, Y. Ritov, and R. Dezeure. On asymptotically optimal confidence regions and tests for high-dimensional models. *arXiv preprint arXiv:1303.0518*, 2013.

S. A. Van De Geer, P. Bühlmann, et al. On the conditions used to prove oracle results for the lasso. *Electronic Journal of Statistics*, 3:1360–1392, 2009.

L. Wasserman. *All of statistics: a concise course in statistical inference.* Springer, 2004.

L. Wasserman and K. Roeder. High dimensional variable selection. *Annals of statistics*, 37(5A): 2178, 2009.

L. Wasserman, M. Kolar, and A. Rinaldo. Estimating undirected graphs under weak assumptions. *arXiv preprint arXiv:1309.6933*, 2013.

C.-H. Zhang and S. S. Zhang. Confidence intervals for low dimensional parameters in high dimensional linear models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 76(1):217–242, 2014.

X. Zhang and G. Cheng. Bootstrapping high dimensional time series. *arXiv preprint arXiv:1406.1037*, 2014.

J. Zhao, C. Leng, L. Li, H. Wang, et al. High-dimensional influence measure. *The Annals of Statistics*, 41(5):2639–2667, 2013.