Submitted to the Annals of Statistics

REJOINDER

By Jiashun Jin* and Wanjie Wang†

*Carnegie Mellon University** *and National University of Singapore*†

"Screen first and estimate next" is a popular strategy for attacking many high dimensional problems. While the methods may vary from occurrence to occurrence, the high level ideas are all similar; below are some examples.

- *Screen and Clean.* Consider the variable selection problem where we have a large number of variables but most of them are 0. We may first screen out many variables and then focus on the small fraction of surviving variables. Success has been shown in e.g., Fan and Lv (2008); Jin, Zhang and Zhang (2014); Wasserman and Roeder (2009).
- *Screen and Classify.* Consider a classification problem where we have a large number of measured features but most of them are useless for the classification decision. We may first screen out many of them and use only the surviving ones for classification. Success has been shown in e.g., Donoho and Jin (2008); Efron (2009); Tibshirani et al. (2002).

Compared to popular penalization methods, the "screen first and estimate next" approach is usually computationally faster and sometimes easier to tune (e.g., Donoho and Jin (2008)). It may also have the optimality that penalization methods do not have (e.g., Jin, Zhang and Zhang (2014); Ke, Jin and Fan (2014)). Of course, we can always combine two approaches by applying some penalization methods to the post-screening data.

IF-PCA in our paper is one more example of the "screen first, estimate next" strategy: we first screen with the Kolmogorov-Smirnov (KS) statistic, and then cluster with the classical PCA. To set the threshold in the screening step, we use Efron's null correction and (Tukey's) Higher Criticism.

The focus of our paper is to find a balance between precise mathematical theory and practical feasibility, and to develop easy-to-use and yet effective methods that have minimum gaps between theory and real applications.

We would like to thank all the discussants for their very thoughtful and stimulating comments. Below are our responses.

1. Comparison with sparse clustering methods. Arias-Castro and Verzelen and Nadler pointed out several interesting literature works on sparse clustering, including COSA (Friedman and Meulman, 2004) and sparse

k-means (Witten and Tibshirani, 2012) (see also Dash et al. (2002); DeSarbo et al. (1984); He, Cai and Niyogi (2005)). These methods use a feature selection step implicitly or explicitly and are related to IF-PCA on a high level. It is of interest to compare COSA and sparse k-means with IF-PCA-HCT using the 10 microarray data sets studied in our paper (all data sets are normalized before we apply each of these methods). COSA has one tuning parameter, which we set "ideally" using the parameter in $\{.05, .10, \ldots, .5\}$ that has the the minimum clustering errors. Sparse k-means has a built-in step for selecting tuning parameters. The results are in Table 1, where IF-PCA-HCT has the lowest error rates for 7 of the data sets. We also find IF-PCA-HCT is computationally faster than COSA and sparse k-means.

TABLE 1
Clustering error rates of IF-PCA-HCT, COSA, and Sparse k-means (S-kmeans).

	Brn	Brst	Cln	Leuk	Lung1	Lung2	Lymp	Prst	SRB	Su
IF-PCA-HCT	.262	.406	.403	.069	.033	.217	.065	.382	.444	.333
COSA	.405	.359	.408	.167	.011	.350	371	.412	.587	.328
S-kmeans	.286	.442	.468	.278	.116	.448	.387	.422	.556	.477

We have also conducted a small-scale simulation study, using the same setting as in Experiment 1 of our paper, except for $p = 4000$ and $g_\sigma = U(.5, 2)$ (so the feature variances range from .5 to 2). The data are normalized before implementations and the tuning parameter of COSA is taken to be .2 (as recommended by Friedman and Meulman (2004)). The results are in Table 2, suggesting that IF-PCA-HCT outperforms the other two methods, especially when the signals are relatively strong.

TABLE 2
Clustering error rates of IF-PCA-HCT (left), COSA (middle), and Sparse k-means (right). The parameter r calibrates the signal strength (see Experiment 1 for details).

$r = .20$	$r = .35$	$r = .50$	$r = .65$
(.265, .444, .330)	(.202, 441, .287)	(.132, .438, .216)	(.127, .435, .237)

2. Correlation screening. Nadler proposes an interesting screening strategy called *CorrIF*, which evaluates the feature significances using the row-wise maximums of the empirical covariance matrix (excluding the diagonals). The strategy is more ambitious than the screening step of IF-PCA, and under the model in our paper, CorrIF is more efficient than IF-PCA in terms of screening: the critical order of the feature strengths required for the success of CorrIF is $(\log(p)/n)^{1/4}$, while that for IF-PCA is of $(\log(p)/n)^{1/6}$.

However, the success of CorrIF relies on the assumption that covariance matrix Σ in (1.2) is diagonal. When Σ is not diagonal, CorrIF faces chal-

lenges: when the (i, j)-th entry of the empirical covariance matrix is large, it could be (recall that we call a feature useful if the contrast mean is nonzero)

- either that both features (i and j) are useful,
- or that both features are useless but they are highly correlated.

We realize that in gene microarray data, many features are highly correlated. In such a difficult setting, it is preferable to use a marginal screening strategy, which is more conservative, but less vulnerable to the correlations.

This partially explains why (as reported by Nadler) CorrIF is successful for some simulated data, but not quite so for some of the microarray data sets. The point was further confirmed by our numerical study, where we compare IF-PCA-HCT with CorrIF-PCA using all the 10 microarray data sets. CorrIF-PCA has one tuning parameter α (see Page 2 of Nadler's discussion) and the recommended value is $\alpha = 1/p$. However, for the Brain data set, no feature is selected if we take $\alpha = 1/p$, so we raise α slightly to $\log(p)/p$ (for all other 9 data sets, there is only a negligible difference if we use $\alpha = \log(p)/p$ instead of $\alpha = 1/p$, both in the number of selected features and in the clustering error rates). For most of the data sets, IF-PCA screens out most of the features, but CorrIF-PCA only screens out less than half of the features. See Table 6 for the comparison of clustering error rates.

Note also that if we decide to use a marginal screening strategy, then our proposed approach is the right choice, and $(\log(p)/n)^{1/6}$ is the right order of critical signal strength required for success, because the diagonals of noise covariance matrix Σ are unknown and unequal.

TABLE 3
Clustering error rates of IF-PCA-HCT and CorrIF-PCA.

	Brn	Brst	Cln	Leuk	Lung1	Lung2	Lymp	Prst	SRB	Su
IF-PCA-HCT	.262	.406	.403	.069	.033	.217	.065	.382	.444	.333
CorrIF-PCA	.476	.438	.484	.264	.122	.434	.226	.422	.540	.489

In summary, we think CorrIF is a very nice idea, and in some cases, it may yield better screening results. It would be great to further develop the method so it can be more efficient in clustering microarray data sets. To do so, we may need to (a) extend the model to a more complicated form, (b) extend Efron's null correction idea to CorrIF, and (c) adapt the idea of threshold choice by Higher Criticism to the CorrIF. We very much hope to see interesting developments along these lines.

3. Minimax optimality. Several discussants, especially Stepanova and Tsybakov, have very interesting comments on the minimax optimality and the critical signal strength required for successful clustering. We wish to

clarify that, if the diagonals of the covariance matrix Σ in (1.2) are unequal and unknown, then marginal χ^2-screening may not work satisfactorily. Also, in this case, the order of the critical signal strength is $(\log(p)/n)^{1/6}$ if we limit our attention to marginal screening methods (e.g., the KS screening method). In cases where the diagonals are equal (since we have already assumed Σ is diagonal, the assumption of equal diagonals means that Σ is proportional to the identity matrix), marginal χ^2-screening works (as kindly pointed out by Stepanova and Tsybakov) and the critical signal strength is of the order $(\log(p)/n)^{1/4}$. In this case, the minimax optimality was studied in several recent papers, including Arias-Castro and Verzelen (2014); Collier, Comminges and Tsybakov (2015).

The clustering problem is closely related to two other problems.

- *Global testing.* When we can reliably test whether the sample vectors $X_1, X_2, \ldots, X_n$ are generated *iid* from $N(0, \Sigma)$, or generated from the model (1.1)-(1.2) in our paper.
- *Signal recovery.* When we can reliably estimate the contrast mean vectors $\mu_1, \mu_2, \ldots, \mu_K$ (see (1.3) of our paper).

In a companion paper Jin, Ke and Wang (2015a), we have considered all three problems in the case where Σ is the identity matrix. For each of them, we have found an interesting phase transition. Compared to the classical framework on minimax optimality, the phase transition can be viewed as a new (but closely related) criterion for assessing optimality. These results, we think, provide a more satisfactory answer to Stepanova and Tsybakov's question on the critical signal strength and optimality in general.

4. Robustness of the KS-screening method. Nadler, Verzelen, and Arias-Castro raised the concern that, when the noise is non-Gaussian or heavy-tailed, the KS screening may face challenges. This is a very interesting point, and we have already tried hard to alleviate the nonGaussian effects.

For example, out of many possible statistics, we choose the KS statistic for screening. For each feature, the KS statistic is defined as the maximum deviation between the empirical CDF of the normalized data samples and the CDF of $N(0,1)$. The maximum deviation is usually assumed near the barycenter of the distribution, and is reasonably far away from the tail. For this reason, the KS statistic is relatively robust to nonGaussianity.

Efron's null correction also helps alleviate the nonGaussian effects. To see the point, we use a small numerical experiment where we fix $(p, \theta) = (20000, 1/2)$ and a distribution F, and let $n = p^\theta \approx 141$ as in the paper. For each $1 \leq j \leq p$, we generate samples $X_i(j) \overset{iid}{\sim} F$, independent for different i,

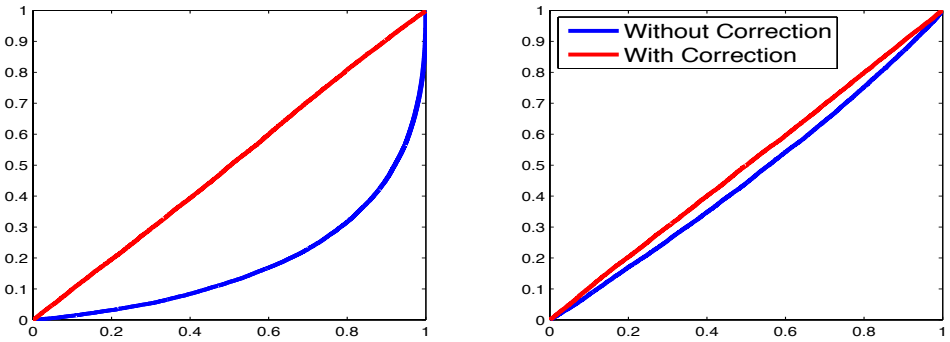

FIG 1. *Comparison of pp-plot for the case with (red) and without (blue) Efron's null correction.* $F = t_\nu(0)$ *with* $\nu = 5$ *(left) and* 20 *(right).*

and compute $\psi_{n,j}$ and $\psi^*_{n,j}$ according to (1.6) and (1.7) in the paper; these are the KS-scores with and without Efron's null correction, respectively.

We take F to be $N(0,1)$ and the central t-distribution $t_\nu(0)$ with $\nu = 5, 10, 15, 20$. Table 4 compares the empirical means and standard deviations of $\{\psi_{n,j}\}_{j=1}^p$ for different F, and there is a noticeable difference for the case of $F = N(0,1)$ and the case of $F = t_\nu(0)$, especially for relatively small ν.

However, the nonGaussian effects can be largely alleviated if we use Efron's null correction. Figure 1 compares the *pp*-plots for the case with and without Efron's null correction (left: $F = t_5(0)$; right: $F = t_{20}(0)$). For each j, the P-values of $\psi_{n,j}$ and $\psi^*_{n,j}$ are computed using 2×10^6 simulated KS-scores $\psi_{n,j}$ and $\psi^*_{n,j}$ with $F = N(0,1)$, respectively. With Efron's null correction, the *pp*-plot are almost a straight line, suggesting that the nonGaussian effect is largely alleviated.

TABLE 4
Comparison of the empirical means and standard deviations (SD) of $\{\psi_{n,j}\}_{j=1}^p$ *for different noise distribution* F.

Distribution F	$N(0,1)$	$t_5(0)$	$t_{10}(0)$	$t_{15}(0)$	$t_{20}(0)$
Mean/SD	.622/.146	.874/.250	.689/.172	.656/.158	.643/.146

Alternatively, one might choose a nonparametric approach for screening (e.g., the uni-modality test (Chan and Hall, 2010; Hartigan and Hartigan, 1985)). However, as Arias-Castro and Verzelen pointed out, while these methods may be more robust, they are clearly inferior in the normal (or approximately normal) settings considered in the paper.

Arias-Castro and Verzelen have also made an interesting observation that, if the null distribution and the alternative distribution have matching moments in the order of $1, 2, \ldots, (d-1)$ for some integer $d > 1$, then the

detection power of the KS statistic is asymptotically equivalent to that of the test based on the d-th empirical moment. However, as Arias-Castro and Verzelen kindly point out, it is preferable to use the KS statistic instead of a moment-based statistic, for in the former, it is not required to know d. In this sense, the KS statistic is adaptive to the unknown degree of matching moments between the null distribution and the alternative distribution. We note that the above observation by Arias-Castro and Verzelen is right and is justified in our paper ((Jin and Wang, 2015, Appendix B)).

5. Connection to SpetralGem and variants of the PCA step. We thank Arias-Castro and Verzelen for pointing out the subtle difference between SpectralGem and classical PCA. We agree that for the method we quoted as "SpectralGEM", it is more appropriate to call it "classical PCA". We now comment on SpectralGEM with more details.

Recall that $W \in \mathbb{R}^{n,p}$ denotes the normalized data matrix. Let $A = WW'$ and let A^* be the matrix defined by $A^*(i,j) = \sqrt{\max\{A(i,j), 0\}}$, $1 \leq i, j \leq n$. SpectralGem is basically the classical PCA algorithm (with some small differences, of course) applied to the the symmetric normalized Laplacian matrix $I_n - D^{-1/2}A^*D^{-1/2}$ corresponding to A^*, where D is a diagonal matrix defined by $D(i,i) = \sum_{j=1}^n A^*(i,j)$, $1 \leq i \leq n$; see Lee, Luca and Roeder (2010). Note that SpectralGem is originally proposed for low-dimensional case so a feature selection step is not required.

Alternatively, we may define A differently. One such example is to write $W' = [W_1, W_2, \ldots, W_n]$ and let A be the similarity matrix defined by $A(i,i) = 0$, and $A(i,j) = \exp(-\|W_i - W_j\|^2/(2\sigma^2))$, $1 \leq i, j \leq n$, where $\sigma > 0$ is a tuning parameter. We then apply the classical PCA to the symmetric normalized Laplacian matrix associated with A. This is basically the method proposed by Ng et al. (2002).

In our PCA step, if we replace the classical PCA by any of these methods (the IF-step remains unchanged), then we have a variant of the IF-PCA. However, our experience suggests that the success of IF-PCA depends more on the success of the IF-step, so it is unclear how much differences we may have if we use a different procedure in our post-screening clustering step. It would be very interesting to explore this in the future.

6. Comparison with sparse PCA. Cai and Zhang proposes an alternative "screen first, clustering next" approach where they first use a screening method to obtain an initial estimate of the principal subspace $\hat{V}_0$ (see Cai and Zhang for details), and then cluster by applying a sparse PCA approach to a matrix Y constructed from the data matrix and $\hat{V}_0$. Moreover,

- with elegant theory and careful analysis, they show that the algorithm leads to satisfactory clustering results, if we construct $\hat{V}_0$ by applying the regular SVD to the post-screening data matrix $\hat{W}^{\hat{S}}$ obtained in the PCA-1 step of IF-PCA.
- they also show that the algorithm leads to satisfactory results on 6 of the data sets in our paper (they do not study the other 4 data sets).

Cai and Zhang's results are very interesting and encouraging, both in theory and in applications. Their study supports one of our points aforementioned: we can always combine the screening step of IF-PCA with some other clustering approaches to form a new two-step method.

Table 5 compares IF-PCA-HCT with Cai and Zhang's sparse PCA (sPCA) approach with the 10 microarray data sets. Since both procedures are random, the error rates reported here are the average over 30 independent repetitions, so the numbers may be slightly different from that in Cai and Zhang's discussion, which are based on one repetition. The randomness of IF-PCA is due to the k-means algorithm in the PCA step, and the randomness of sPCA is due to an add-on random matrix the algorithm uses. For the 10 data sets we investigate, the standard deviations of the error rates of sPCA are about a few tens times larger than that of IF-PCA.

For IF-PCA-HCT, the data are normalized before implementation. For sPCA, the data may be normalized (sPCA (N)) (and we use the same screening method as in IF-PCA) or un-normalized (sPCA(U)) (where we use the χ^2-screening proposed by Cai and Zhang). From a practical viewpoint, it is usually preferable to normalize the data before implementation, so sPCA(N) is a more reasonable approach compared to sPCA(U).

As IF-PCA-HCT and sPCA(N) use the same screening step, their performances are more or less similar. This supports one of our points: in many cases, it is more critical to have an effective screening step than an effective post-screening clustering step: with a successful screening step, the difference between one (post-screening method) and another may be insignificant.

TABLE 5
Comparison of clustering error rates by IF-PCA, sPCA(N), and sPCA(U).

	Brn	Brst	Cln	Leuk	Lung1	Lung2	Lymp	Prst	SRB	Su
IF-PCA-HCT	.262	.406	.403	.069	.033	.217	.065	.382	.444	.333
sPCA(U)	.221	.440	.261	.032	.083	.276	.016	.422	.508	.484
sPCA(N)	.502	.411	.417	.040	.006	.352	.026	.422	.422	.477

There are several other reasons why we choose to use classical PCA in our (post-screening) clustering step: it is conceptually simple and easy to use, memory efficient, computationally relatively fast, and does not require

tuning (once K is given). The simplicity of classical PCA also allows us to derive an optimal threshold choice approach using the Higher Criticism (how to set the threshold in a data-driven fashion is a hard but very interesting problem; see e.g., Donoho and Jin (2009); Fan, Jin and Yao (2013)).

In summary, Cai and Zhang show that combining our IF-step with their sparse PCA gives rise to promising new approaches, and support their approaches with encouraging numerical results and elegant theory. Their study also suggests that IF-PCA is an adaptive two-step method, where we can modify either the IF-step or the PCA-step, depending on the situations. It would be very interesting to further investigate such an idea in the future.

7. Some comments on Higher Criticism Thresholding. Recall that in HCT, we define a HC function by

$$HC_{p,j} = \frac{\sqrt{p}(j/p - \pi_{(j)})}{\sqrt{\max\{\sqrt{n}(j/p) - \pi_{(j)}, 0\} + (j/p)}}, \qquad 1 \le j \le p,$$

and retain the first $\hat{j}$ features with smallest P-values, where

$$\hat{j} = \text{argmax}_{\{1 \le j \le p/2, \pi_{(j)} > \log(p)/p\}} HC_{p,j}.$$

Stepanova and Tsybakov have a few questions on HCT. First, they asked whether we should add an absolute sign in the numerator of $HC_{p,j}$. For a typical gene microarray data, the empirical distribution of the KS scores has two noticeable features: (a) the center of the distribution does not match well with its theoretical counterpart, and (b) near the right tail of the distribution, there is a small bump (many times noticeable) which we believe to be corresponding to useful features. When we apply Efron's null correction, the HC scores $HC_{p,j}$ may be negative for many j, but the corresponding features are usually those with large p-values, where the negativity is due to the error in Efron's null correction, not because these features are significant. A significant feature, however, will have a large and positive HC score $HC_{p,j}$. See Figure 2, where the $HC_{p,j}$ are positive for relatively small j (corresponding to significant features), and are negative for many large j (corresponding to non-significant features). If we add an absolute sign to the numerator of $HC_{p,j}$, then the HC function reaches its maximum at a relatively large j and thus provides a threshold choice that is much lower than desired. For this reason, we should not add an absolute sign as suggested.

Second, they asked the rationale of the HC function, where we think there is possible misunderstanding. They thought the remark in Section 1.3 "only arrives to the conclusion that the function $HC_p(t)$ is monotone between the

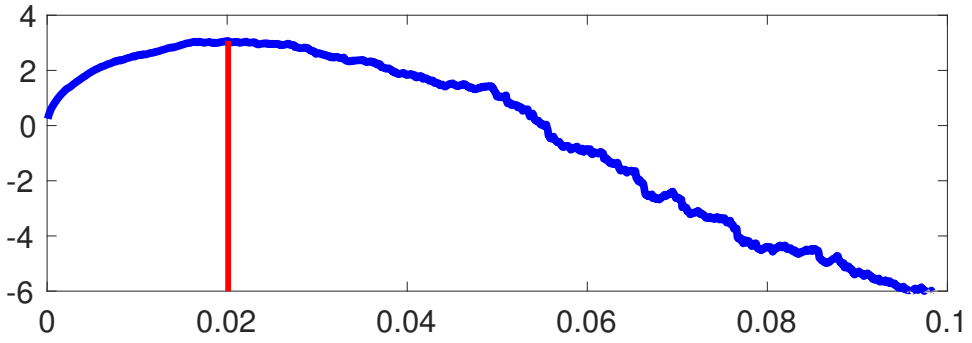

FIG 2. *Plot of $HC_{p,j}$ versus j/p (red line: $\hat{j}/p$), $1 \le j \le p/10$. Data: Lung Cancer (1).*

adjacent discontinuities". However, the remark concludes that HCT approximately maximizes the post-selection signal-to-noise ratio $\widetilde{snr}(t)$ and so is optimal, and the HC function is carefully designed to approximate $\widetilde{snr}(t)$. For example, in the denominator of $HC_{p,j}$, that we take the maximum over $\sqrt{n}((j/p) - \pi_{(j)})$ and 0 is not because we want to use a regularization, but rather because we wish to use the denominator of $HC_{p,j}$ to approximate that of $\widetilde{snr}(t)$. We understand that the current explanation on $\widetilde{snr}(t)$ can be more detailed, but as we mentioned in the paper, a full explanation is rather lengthy so we deter it to Jin, Ke and Wang (2015b).

Last, they asked a few questions about the constraint $\pi_{(j)} > \log(p)/p$. The constraint is not required for both our theoretical study and for the analysis of most of our date sets (out of 10 microarray data sets, the results remain the same for all except the Lymphoma data if we remove such a constraint).

On the other hand, for some data sets (e.g., the Lymphoma data), it could happen that $HC_{p,j}$ reaches the maximum at a very small j, and so our method ends up selecting very few features. This is not desirable for our experience suggests that we usually should select a few tens or a few hundreds of features (for clustering). This is why we impose such a constraint, with which we avoid to select fewer than $\log(p)$ (approximately) features.

For asymptotic theory, we could either remove such a constraint, or replace the term $\log(p)/p$ by $5\log(p)/p$ or $\log^2(p)/p$. For practice, it is preferable to stick to $\log(p)/p$; otherwise, our range of interest may fail to include the optimal j, simply because $5\log(p)$ and $\log^2(p)$ are too large for a typical p (e.g., $\log(p) \approx 9$, $5\log(p) \approx 46$ and $\log^2(p) \approx 85$ if $p = 10,000$).

Note that in all the simulations and real data analysis, we choose $\log(p)/p$ as the threshold in the constraint before we implement IF-PCA, and have never used the data to tune such a threshold.

The constraint is very different from those in Donoho and Jin (2004)

because the version of HC there is different from the version we have here, and because two papers have very different goals.

In principle, we can remove the constraint by adding some penalty in the denominator, however, choosing an appropriate penalty term requires very delicate analysis. We leave the study along this line to the future.

8. Threshold choice by controlling the (feature)-FDR. Benjamini and Hochberg's False Discovery Rate (FDR) control method (Benjamini and Hochberg, 1995) is a popular approach to threshold choice. Fixing an FDR control parameter $0 < q < 1$, let $\pi_{(1)} < \pi_{(2)} < \ldots < \pi_{(p)}$ be the sorted P-values as in Section 1.3, and let $k = k^{FDR}$ be the largest integer such that $\pi_{(k)} \leq q(k/p)$. The FDR threshold is the critical value for the KS-score to which the corresponding P-value is $\pi_{(k^{FDR})}$.

In our forthcoming manuscript Jin, Ke and Wang (2015b) (see also Donoho and Jin (2008, 2009), where we discuss the threshold choice in the context of classification), we find that the optimal FDR-control parameter q for IF-PCA (i.e., the q that yields the lowest clustering error rates for IF-PCA) critically depends on the rarity and strengths of the useful features.

- When the useful features are rare/weak, the optimal FDR control parameter q satisfies that $(1-q)$ tends to 0 algebraically fast as $p \to \infty$.
- When the useful features are rare/strong, the optimal q tends to 0 algebraically fast as $p \to \infty$.
- When the useful features are rare but moderately strong, the q may be bounded away from both 0 and 1.

In practice, all three cases can happen, and how to select the optimal q is a challenging problem. Therefore, the FDR approach is basically transforming the problem of selecting the best threshold (which is a tuning parameter) to the problem of selecting the best q (which is also a tuning parameter), but does not solve the problem of finding a way to set a threshold optimally and in a data-driven fashion.

Following the suggestion by Arias-Castro and Vezelen, we consider a variant of IF-PCA-HCT where we replace the HCT approach by the FDR approach above, and call the later by IF-PCA-FDR. Table 6 compares two methods with the 10 microarray data sets, where the FDR control parameter q is ideally selected using a grid search in $\{.01, .02, \ldots, .99\}$. The results suggest that, even if we select q ideally using a rather refined grid search, IF-PCA-FDR still underperforms IF-PCA-HCT for about 6 of the data sets. These suggest that, in practice, it is usually hard to pin down the ideal FDR control parameter q, so the FDR approach may face challenges.

TABLE 6
Comparison of the clustering error rates by IF-PCA-HCT and IF-PCA-FDR.

	Brn	Brst	Cln	Leuk	Lung1	Lung2	Lymp	Prst	SRB	Su
IF-PCA-HCT	.262	.406	.403	.069	.033	.217	.065	.382	.444	.333
IF-PCA-FDR	.119	.323	.371	.278	.116	.222	.194	.412	.286	.333

9. Connection to classification by Fisher's LDA. Cai and Zhang have some very interesting comments on the connection between the clustering problem and the (two-class) classification problem in the setting where the contrast mean vector is sparse and the samples from two classes share the same covariance matrix Σ that is not diagonal.

The (two-class) classification problem has been considered in our recent work Fan, Jin and Yao (2013); Hall and Jin (2010); Huang, Jin and Yao (2015) where we adapt Fisher's LDA to the high dimensional setting, assuming Σ^{-1} is sparse (also see related works in Efron (2009); Fan, Feng and Tong (2012); Donoho and Jin (2008, 2009); Hall, Pittelkow and Ghosh (2008); Jin (2009); Tibshirani et al. (2002)).

- In Fan, Jin and Yao (2013), we studied the case where Σ is known. We learned that, to adapt Fisher's LDA to such a high dimensional setting, it is desirable to do feature selection. We also learned that an appropriate way to do feature selection is as follows: (a) use the training sample to obtain a z-score for each feature, (b) transform the z-scores using the Innovated Transformation (proposed in Hall and Jin (2010) in the context of Innovated Higher Criticism), and (c) apply the feature selection to the transformed z-vector. See details therein.
- In Huang, Jin and Yao (2015), we further considered the case where Σ is unknown and proposed Partial Correlation Screening (PCS) as a new approach to estimating Σ^{-1}. The estimated matrix is then combined with the method in Fan, Jin and Yao (2013) to form a trained classifier.
- We also learned in Donoho and Jin (2008, 2009) that when the useful features are rare and weak, it may be impossible to identify all useful features, but is still possible to have successful classification. However, for optimal classification, we usually need to tolerate many false positives in our feature selection.

However, clustering when Σ is not diagonal is a more difficult problem, since the class labels are unknown: when two features are strongly correlated, it is hard to tell whether some of these features are useful, or that both features are useless but they are strongly correlated; see Section 2 of the Rejoinder.

Our finding is consistent with the points by Cai and Zhang, especially in that Σ plays an important role in feature selection.

Section 2 of Cai and Zhang also mentions that IF-PCA is *specifically* designed for the case where the noise covariance matrix Σ is diagonal, and that assuming Σ is diagonal is *essential* for the success of of IF-PCA. We would like to clarify that this is not true. Indeed, IF-PCA is successfully applied to 10 microarray data sets, where the measured features are highly correlated. The assumption that Σ is diagonal is mostly for simplicity in the presentation of our theoretical results, and can be largely relaxed.

10. Future directions. The discussants have many stimulating comments that are worthy of further study. We now discuss some of them.

In our theoretical study, we have assumed that (a) the sample noise is Gaussian, (b) samples in different classes share the same (noise) covariance matrix, (c) the covariance matrix is diagonal. While these assumptions may not hold for the microarray data, IF-PCA provides some encouraging clustering results. This suggests that IF-PCA may continue to work well in settings that are much broader than that considered in our theoretical study.

On the other hand, as several discussants pointed out, we may improve the screening step by incorporating the correlation structure. Successes have been shown in signal detection (e.g., Hall and Jin (2010)) and in classification (e.g., Fan, Jin and Yao (2013)), but how to extend ideas therein to clustering remains a challenging problem.

A closely related problem is how to estimate the correlation structures, which is also very challenging, because the class labels are unknown.

Cai and Zhang propose to combine our screening step with some of their sparse PCA approaches for a new clustering method, and present very interesting results. Given the recent interests on sparse PCA, it would be very interesting to study the connection between IF-PCA and sparse PCA.

The correlation screening by Nadler is a very interesting idea. In some cases (e.g., the noise covariance matrix is diagonal), the method seems to have advantages both in theory and in simulations. Unfortunately, when we apply the method to the microarray data, the results are relatively unsatisfactory. It is worthy to further study the idea of correlation screening, presumably with a more complex model than that in this paper.

Nadler also points out, it would be of great interest to study how to extend some of our ideas to the problem of *semi-supervised learning*, where out of many measured samples, most of them are unlabeled, while a few of them are labeled. The problem can be viewed as a hybrid of the classification problem and the clustering problem, so maybe both ideas in classification and those in clustering can be helpful.

A tribute to Peter G. Hall. We are saddened to know that Peter G. Hall, a monumental figure in statistics, probability, and applied mathematics and a legend of our time, has recently passed away (January 9, 2016). For over a decade, Peter Hall has been a great source of inspiration to the first author. Hall's work on classification and clustering (Chan and Hall, 2010; Hall, Jin and Miller, 2014; Hall, Pittelkow and Ghosh, 2008), on Higher Criticism (Delaigle and Hall, 2009; Delaigle, Hall and Jin, 2011; Hall and Jin, 2008, 2010), and on cosmology and astronomy (Bennett et al., 2012) are closely related to the current paper.

References.

Arias-Castro, E. and Verzelen, N. (2014). Detection and feature selection in sparse mixture models. *arXiv:1405.1478*.

Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. Roy. Statist. Soc. B* 289–300.

Bennett, F., Melatos, A., Delaigle, A. and Hall, P. (2012). Reanalysis of F-statistics gravitational-wave search with the higher criticism statistics. *Astrophysics J.* **766** 1–10.

Chan, Y.-b. and Hall, P. (2010). Using evidence of mixed populations to select variables for clustering very high-dimensional data. *J. Amer. Statist. Assoc.* **105**.

Collier, O., Comminges, L. and Tsybakov, A. B. (2015). Minimax estimation of linear and quadratic functionals on sparsity classes. *arXiv:1502.00665*.

Dash, M., Choi, K., Scheuermann, P. and Liu, H. (2002). Feature selection for clustering-a filter solution. In *Data Mining, 2002. ICDM 2003. Proceedings. 2002 IEEE International Conference on* 115–122. IEEE.

Delaigle, A. and Hall, P. (2009). Higher Criticism in the context of unknown distribution, non-independent and classification. *Perspectives in Mathematical Sciences I: Probability and Statistics* **2009** 109-138.

Delaigle, A., Hall, P. and Jin, J. (2011). Robustness and accuracy of methods for high dimensional data analysis based on Student's t-statistic. *J. Roy. Statist. Soc. B* **73** 283–301.

DeSarbo, W. S., Carroll, J. D., Clark, L. A. and Green, P. E. (1984). Synthesized clustering: A method for amalgamating alternative clustering bases with differential weighting of variables. *Psychometrika* **49** 57–78.

Donoho, D. and Jin, J. (2004). Higher criticism for detecting sparse heterogeneous mixtures. *Ann. Statist.* 962–994.

Donoho, D. and Jin, J. (2008). Higher criticism thresholding: Optimal feature selection when useful features are rare and weak. *Proc. Natl. Acad. Sci.* **105** 14790–14795.

Donoho, D. and Jin, J. (2009). Feature selection by higher criticism thresholding achieves the optimal phase diagram. *Phil. Trans. Roy. Soc. A* **367** 4449–4470.

Efron, B. (2009). Empirical Bayes estimates for large-scale prediction problems. *J. Amer. Statist. Assoc.* **104** 1015–1028.

Fan, J., Feng, Y. and Tong, X. (2012). A road to classification in high dimensional space: the regularized optimal affine discriminant. *J. Roy. Statist. Soc. B* **74** 745–771.

Fan, Y., Jin, J. and Yao, Z. (2013). Optimal classification in sparse Gaussian graphic model. *Ann. Statist.* **41** 2537–2571.

Fan, J. and Lv, J. (2008). Sure independence screening for ultrahigh dimensional feature space. *J. Roy. Statist. Soc. B* **70** 849–911.

Friedman, J. H. and Meulman, J. J. (2004). Clustering objects on subsets of attributes (with discussion). *J. Roy. Statist. Soc. B* **66** 815–849.

Hall, P. and Jin, J. (2008). Properties of Higher Criticism under strong dependence. *Ann. Statist.* **36** 381-402.

Hall, P. and Jin, J. (2010). Innovated Higher Criticism for detecting sparse signals in correlated noise. *Ann. Statist.* **38** 1686-1732.

Hall, P., Jin, J. and Miller, H. (2014). Feature selection when there are many influential features. *Bernoulli* **20** 1647–1672.

Hall, P., Pittelkow, Y. and Ghosh, M. (2008). Theoretical measures of relative performance of classifiers for high dimensional data with small sample sizes. *J. Roy. Statist. Soc. B* **70** 158–173.

Hartigan, J. A. and Hartigan, P. (1985). The dip test of unimodality. *Ann. Statist.* 70–84.

He, X., Cai, D. and Niyogi, P. (2005). Laplacian score for feature selection. In *Adv. Neural Inf. Process. Syst.* 507–514.

Huang, S., Jin, J. and Yao, Z. (2015). Partial correlation screening for estimating large precision matrices, with applications to classification. *Ann. Statist.* To appear.

Jin, J. (2009). Impossibility of successful classification when useful features are rare and weak. *Proc. Natl. Acad. Sci.* **106** 8859–8864.

Jin, J., Ke, Z. T. and Wang, W. (2015a). Phase transitions for high dimensional clustering and related problems. *arXiv:1502.06952*.

Jin, J., Ke, Z. T. and Wang, W. (2015b). Optimal spectral clustering by Higher Criticism Thresholding. *Manuscript*.

Jin, J. and Wang, W. (2015). Supplementary material for "Influential Features PCA for high dimensional clustering". *Manuscript*.

Jin, J., Zhang, C.-H. and Zhang, Q. (2014). Optimality of Graphlet Screening in high dimensional variable selection. *J. Mach. Learn. Res.* **15** 2723–2772.

Ke, T., Jin, J. and Fan, J. (2014). Covariate assisted screening and estimation. *Ann. Statist.* **42** 2202.

Lee, A. B., Luca, D. and Roeder, K. (2010). A spectral graph approach to discovering genetic ancestry. *Ann. Appl. Statist.* **4** 179–202.

Ng, A. Y., Jordan, M. I., Weiss, Y. et al. (2002). On spectral clustering: Analysis and an algorithm. In *Adv. Neural Inf. Process. Syst.* **2** 849–856. MIT; 1998.

Tibshirani, R., Hastie, T., Narasimha, B. and Chu, G. (2002). Diagnosis of multiple cancer types by shrunken centroids of gene expression. *Proc. Natl. Acad. Sci. USA* **99** 6567-6572.

Wasserman, L. and Roeder, K. (2009). High dimensional variable selection. *Ann. Statist.* **37** 2178–2201.

Witten, D. M. and Tibshirani, R. (2012). A framework for feature selection in clustering. *J. Amer. Statist. Assoc.*

J. Jin
Department of Statistics
Carnegie Mellon University
Pittsburgh, Pennsylvania, 15213
USA
E-mail: jiashun@stat.cmu.edu

W. Wang
Department of Statistics
and Applied Probability
National University of Singapore
Singapore 117546
E-mail: staww@nus.edu.sg