A local dependence measure and its application to screening for high correlations in large data sets

Kumar Sricharan*, Alfred O. Hero III*, Bala Rajaratnam†,

*Department of EECS, University of Michigan, Ann Arbor, MI 48109
†Department of Statistics, Stanford University, Stanford, CA 97331
Email: {kksreddy, hero}@umich.edu, brajarat@stanford.edu

Abstract—Correlation screening is frequently the only practical way to discover dependencies in very high dimensional data. In correlation screening a high threshold is applied to the matrix of sample correlation coefficients of the multivariate data. The variables having coefficients that exceed the threshold are called discoveries and are classified to be dependent. The mean number of discoveries and the number of false discoveries in correlation screening problems depend on a information-theoretic measure $J$, a novel type of information divergence that is a function of the joint density of pairs of variables. It is therefore important to estimate $J$ in order to determine screening thresholds for desired false alarm rates. In this paper, we propose a kernel estimator for $J$, establish asymptotic consistency and determine the asymptotic distribution of the estimator. These results are used to minimize the MSE of the estimator and to determine confidence intervals on $J$. We use these results to test for dependence between variables in both simulated data sets and also between email spam harvesters. Finally, we use the estimate of $J$ to determine screening thresholds in correlation screening problems involving gene expression data.

Keywords: Dependence measure, correlation screening, Information theory, estimation, CLT.

I. INTRODUCTION

Consider the problem of screening among a large number $p$ of variables for those having significant correlations. Examples of such high dimensional data sets include gene expression arrays, traffic over the internet and multivariate financial time series. Screening can be used to discover a small number of such high dimensional data sets. In many practical examples $J$ is unknown and must be estimated empirically.

In this paper, we propose an asymptotically consistent estimator for $J$ and determine the limiting distribution of our estimator to be normal. We use our estimate of $J$ to test for dependence between variables and to determine suitable screening thresholds for achieving desired false alarm rates in correlation screening problems. The rest of the paper is organized as follows: In Section 2, we introduce preliminaries and describe the correlation screening procedure in Section 3. We introduce the estimator of $J$ in Section 4 and provide asymptotic analysis of bias and variance of the estimator. We also establish a central limit theorem for the estimator. We experimentally validate the results in Section 5. Furthermore, we use the theoretical results to obtain confidence intervals on the measure $J$ and subsequently use the confidence intervals to test for dependence between variables and to determine thresholds in correlation screening. We conclude in Section 6.

II. PRELIMINARIES

The asymptotic expressions for the mean number of discoveries in correlation screening will be a function of several quantities introduced below.

A. Relevant definitions

Define the Spherical cap probability to be

$$P_o(r, n) = a_n \int_0^1 (1 - u^2)^{\frac{n-1}{2}} du,$$

where $a_n$ is

$$a_n = |S_n-2| = \frac{2\Gamma((n-1)/2)}{\sqrt{\pi}\Gamma((n-2)/2)}.$$

The quantity $P_o(r, n)/2$ is equal to the proportional area of the spherical cap of radius $r = \sqrt{2(1-\rho)}$ on $S_n-2$. It is the probability that a uniformly distributed point $U$ on the sphere lies in the pair of hyper spherical cones symmetric about the origin [8]. Application of the mean value theorem to the integral in the definition of spherical cap probability yields the relation

$$P_o(r, n) = \frac{a_n}{n-2}(1 - \rho^2)^{(n-2)/2}(1 + O(1 - \rho^2)).$$
For random variables $U$ and $V$ with joint density $f_{U,V}$ on $S_{n-2} \times S_{n-2}$ with marginals $f_U$ and $f_V$ define
\[
J_\rho(f_{U,V}) = \frac{Pr\left(||U - V|| \leq \sqrt{2(1 - \rho)}\right)}{P_\rho(\rho, n)},
\]
and the limit
\[
J(f_{U,V}) = \lim_{\rho \rightarrow 1} J_\rho(f_{U,V}) = |S_{n-2}| \int_{S_{n-2}} f_{U,V}(u, u) du.
\]
The authors of [4] have shown that the number of variables who are highly correlated will likely be discovered.

We give an intuitive interpretation of $J(f_{U,V})$ as a measure of dependence between $U$, $V$. It is equal to the Bhattacharyya affinity between the product of the marginal distributions $m_{U,V}(w) = f_U(w)f_V(w)$ and the product of the conditional distributions $c_{U,V}(w) = f_{U|V}(w|u)f_{V|U}(w|u)$:
\[
J(f_{U,V}) = |S_{n-2}| \int \sqrt{m_{U,V}(w)c_{U,V}(w)} dw,
\]
\[
\leq |S_{n-2}| \left( \int m_{U,V}(w) dw \right)^{1/2} \left( \int c_{U,V}(w) dw \right)^{1/2},
\]
where equality occurs iff $f_{U,V}(u, u) = f_U(u)f_V(v)$ and therefore measures local dependence along the diagonal $u = v$. In contrast, the Shannon MI captures global dependence information over the entire range of $u, v$. In correlation screening with a high threshold global dependency (MI) is not as relevant as local dependency $(J)$ since only those variables who are highly correlated will likely be discovered.

**B. Z-score representation**

Let $X = [X_1, \ldots, X_p]^T$ be a vector of random variables with mean $\mu$ and $p \times p$ covariance matrix $\Sigma$. Denote the correlation matrix $\Gamma = D_{\Sigma}^{-1/2} \Sigma D_{\Sigma}^{-1/2}$ where $D_{\Sigma} = diag(\Sigma_{ii})$ is the diagonal matrix of variances of components of $X$. Assume that $n$ independent identically distributed (i.i.d.) samples of $X$ are available and arrange these samples in a $n \times p$ data matrix
\[
X = [X_1, \ldots, X_n] = [X_1^T, \ldots, X_n^T]^T,
\]
where $X_i = [X_{1i}, \ldots, X_{pi}]^T$ and $X_{(i)} = [X_{1i}, \ldots, X_{pi}]$ denote the $i$-th column and row, respectively, of $X$.

Define the sample mean of the $i$-th column $X_i = \frac{1}{n} \sum_{j=1}^{n} X_{ij}$, the vector of sample means $\bar{X} = [\bar{X}_1, \ldots, \bar{X}_p]$, the $p \times p$ sample covariance matrix $S = (n-1)^{-1}(X_{(i)} - \bar{X})^T(X_{(i)} - \bar{X})$, and the $p \times p$ sample correlation matrix
\[
R = D_S^{-1/2}SD_S^{-1/2},
\]
where $D_S = diag(S_{ii})$ is the diagonal matrix of component sample variances. Let the $ij$-th entry of the ensemble covariance $I$ be denoted $\gamma_{ij}$ and the $ij$-th entry of the sample covariance $R$ be $r_{ij}$.

We now define the Z-scores, $U = [U_1, \ldots, U_p]$, $U_i \in \mathbb{R}^{n-1}$ as follows. The Z-scores are constructed to lie on the $(n-2)$-sphere $S_{n-2} \subset \mathbb{R}^{n-1}$ and are determined by projecting away the components of $X_i$ orthogonal to the $n-1$ dimensional hyperplane $u \in \mathbb{R}^n : U^T u = 0, i = 1, \ldots, p$. The Z-scores can be computed from $X$ using a Gramm-Schmidt procedure. The Z-scores satisfy the relation
\[
R = U^TU.
\]
Furthermore, the sample correlation between $X_i$ and $X_j$ can be computed using the inner product or the Euclidean distance between associated Z-scores
\[
r_{ij} = U_i^TU_j = 1 - \frac{||U_i - U_j||^2}{2}.
\]
Pairs of Z-scores that are close to each other, as measured by Euclidean distance, have high associated sample correlations.

Figure 1. The Z-scores associated with $n = 4$ realizations of 1000 variables are $n-1$-element vectors that lie on the unit $n-2$ dimensional sphere $S_{n-2}$. Shown are Z-scores for a multivariate normal sample with diagonal and non-diagonal covariance structure. Pairs of Z-scores that are close to each other, as measured by Euclidean distance, have high associated sample correlations.
In the case of non-diagonal Σ elliptical distributions the distribution of the Z-scores over the sphere $S_{n-2}$ will generally be far from uniform (Fig. 1). The Z-score representations of the sample correlation will be a key ingredient for deriving the asymptotic results in this paper.

### III. Correlation Screening

Consider an experiment to compare $p$ variables under a treatment, called $X$. This experiment produces the data matrix $X$, which is a $n \times p$ matrix. From this data matrix, extract the Z-score matrix $U$. We then construct the sample correlation matrix $R = UTU^T$.

We are primarily interested in the case $n << p$ so that the matrix will be rank deficient. Let the $ij$-th element of this matrix be denoted as $r_{ij}$. The objective is to screen the $p$ variables for those whose maximal magnitude correlation exceeds a given threshold $\rho$. Specifically, for $i,j = 1, \ldots, p$, the $i$-th variable passes the screen if $\max_{j \neq i} |r_{ij}| > \rho$.

For this test a discovery is declared if an index $i$ passes the screen and we denote by $N$, the total number of discoveries. In previous work done by one of the authors, it was shown that the number of discoveries satisfy the following properties.

**IV. Estimation of $J$**

Consider a $p$ dimensional joint density $f_{U_1,\ldots,U_p}$ defined on $S_{n-2} \times \ldots \times S_{n-2}$. Now define the average pairwise density

$$ f_{U_i, U_j} (u, v) = \frac{1}{p(p-1)} \sum_{i=1}^{p} \sum_{j \neq i}^{p} \left( f_{U_i, U_j}^2 (u, v) \right). $$

We are interested in estimating $J(f_{U_i, U_j})$ from a single sample $[U_1, \ldots, U_p]$ which is drawn from the joint density $f_{U_1,\ldots,U_p}$. Note that

$$ J(f_{U_i, U_j}) = \frac{1}{p(p-1)} \sum_{i=1}^{p} \sum_{j \neq i}^{p} J(f_{U_i, U_j}). $$

In order to estimate $J(f_{U_i, U_j})$, we build an uniform kernel estimator $J$, which we describe below. We will show that the estimator $\hat{J}$ is asymptotically unbiased and consistent. Finally, we will show that the estimator, when suitable normalized, converges weakly to a normal distribution.

**A. Estimate of $J$**

For variable size $p$ and corresponding screening threshold $\rho$, let $\rho_s$ be the estimation threshold which satisfies (i) $\lim_{p \rightarrow \infty} \rho_s = 1$, and (ii) $\lim_{p \rightarrow \infty} (1 - \rho)/(1 - \rho_s) = 0$.

Define $A(r,v) = C(r,v) \cup C(r,-v)$ to be the union of spherical cap regions centered at $v$ and $-v$ with radius $r = \sqrt{2(1 - \rho_s)}$. Let $b_{ij}$ denote the event that $U_j \in A(r, U_i)$. Note that there are $p(p-1)/2$ distinct combinations of $\{i,j\}$. Therefore the cardinality of the set $B = \{b_{ij}, i < j\}$ is $p(p-1)/2$.

We sample from the set $B$ without replacement to obtain a finite sequence of random variables $V = \{v_1, v_2, \ldots, v_m\}$, where $m = p(p-1)/2$. We now define our estimator of $J(p, U_i, U_j)$ to be

$$ \hat{J} = \frac{1}{P_v(\rho_s, n)} \left( \frac{1}{m} \sum_{i=1}^{m} v_i \right). $$

**B. Intuition**

Before stating the theoretical results, we provide some intuition behind the requirement $(1 - \rho)/(1 - \rho_s) \rightarrow 0$ which translates to the condition that the estimation threshold $\rho$ approach 1 faster than the screening threshold $\rho_s$. The screening problem is a detection problem where we want the number of discoveries to converge to a non-zero limit given by the RHS of Eq. 1 and hence choose the threshold $\rho$ according to the equation $p(p-1)(1 - \rho^2)^{(n-2)/2} \sim d_n$.

On the other hand, we are concerned with an estimation problem. If the number of discoveries $\sum_{i=1}^{m} v_i$ at threshold $\rho_s$ converges to a non-zero limit, then the variance of our estimator will also converge to a non-zero limit which then implies that the estimator $\hat{J}$ will not be consistent. Hence we choose a threshold $\rho_s$ satisfying $(1 - \rho)/(1 - \rho_s) \rightarrow 0$ which then implies that the mean number of discoveries $\left( \sum_{i=1}^{m} J(f_{U_i, U_j})/(n-2) \right)p(p-1)(1 - \rho^2)^{(n-2)/2}$ at threshold $\rho_s$ converges to 0.
C. Bias

Define the maximal gradient of the average pairwise density to be
\[ \nabla M = \sup_{u,v \in S_{n-2}} \| \nabla_v f_{U,U_{-}}(u,v) \|_{u=v}. \]
The bias of \( \hat{J} \) is bounded by
\[ | E[\hat{J}] - J(f_{U,U_{-}}) | \leq 2a_n^2 \nabla M \sqrt{2(1 - p_s)} = \sqrt{2a_n^2 n^{(2-n)} \nabla M} (\frac{1}{p_s}) (1 + o(1)) \]
which implies that the estimator \( \hat{J} \) is asymptotically unbiased.

D. Variance

Define
\[ M = \frac{\text{avg}}{i \neq j \neq k \neq l} \sqrt{\nabla[f_{U,U}(U_i|U_l)] \nabla[f_{U,U}(U_i|U_l)]}. \]
Also define,
\[ \delta_{ijkl} = f_{U,U,U,U}(u_i,u_j,u_k,u_l) - f_{U,U,U,U}(u_i,u_j,u_k,u_l), \]
and
\[ \delta_p = \text{avg}_{i \neq j \neq k \neq l} \left( \sup_{u_i,u_j,u_k,u_l} | \delta_{ijkl} | \right), \]
and further assume \( \delta_p \to 0 \) as \( p \to \infty \). Specifically, let \( \delta_p = O(1/p^\alpha) \) for some \( \alpha > 0 \). The variance of \( \hat{J} \) is then bounded by
\[ \text{Var} \{ \hat{J} \} = (P_o(p_s,n))^{-2} \sqrt{\frac{1}{m} \sum_{i=1}^{m} v_i}, \]
\[ = \frac{((P_o(p_s,n))^{-2}/m) \text{Var} \{ v_1 \} + (P_o(p_s,n))^{-2}(1 - 1/m) \text{Cov} \{ v_1, v_2 \}}{\kappa_n} \]
\[ \leq \frac{2J(f_{U,U_{-}})}{p_s} \left( \frac{p_s}{p} \right)^2 (1 + o(1)) \]
\[ + \frac{2M}{3} \left( \frac{1}{p} \right) + a_n^2 \delta_p (1 + o(1)), \]
and the variance of our estimator tends to 0 as \( p \to \infty \).

E. Optimization of MSE

Note that our intuition behind choosing \( (1 - \rho)/(1 - p_s) \) is verified by our expressions for the MSE, \( p_s \) approaches 1 faster than \( \rho \), which then implies that \( p_s/p \to 0 \), which in turn implies that our estimator is consistent. We will now optimize the choice of \( p_s \) for minimum MSE.

1) General case: When all the variates are significantly correlated, \( \nabla M = O(1) \) which implies that the bias is \( O(p_s^{-2}(n-2)) \). We note that the overall MSE is then given by
\[ O(p_s^{-4}(n-2)) + O(p_s^2/p^2) + O(1/p) + O(1/p^\alpha). \]
Optimizing this expression over \( p_s \) gives the relation \( p_s = O(p^{(n-2)/n}) \) which then gives the optimized M.S.E to be \( O(p^{-4/n}) + O(1/p) + O(1/p^\alpha). \)

2) \( r \)-sparse case: We consider the more realistic scenario when only a small fixed number \( r \) of the \( p \) variates are strongly correlated and the rest are weakly correlated. This implies that the corresponding Z-scores of the \( p - r \) uncorrelated variables are fairly uniformly distributed on the sphere \( S_{n-2} \), which in turn implies that \( \nabla M = O(1/p) \). In this case, the squared bias is \( O(p_s^{-4}/p^{-1}) \) and is clearly dominated by the variance. Also notice that this \( r \)-sparse setting implies \( \alpha >> 1 \). This implies that the overall M.S.E. for the \( r \)-sparse case is dominated by the variance and is given by \( O(p_s^2/p^2) + O(1/p) \).

In this scenario, our analysis shows that \( p_s \) should be chosen to be as small as possible.

F. Asymptotic distribution

Since the set of random variables \( V = \{ v_1, v_2, \ldots, v_m \} \) are exchangeable [6] for every \( p \), we can assume that \( V = \{ v_1, v_2, \ldots, v_m \} \) are a finite segment of an infinite length exchangeable sequence. Consider the normalized version of our estimator \( \hat{J} \) given by
\[ \hat{J} = \frac{J(f_{U,U_{-}})}{(p_s/p) \sqrt{2J(f_{U,U_{-}})/\kappa_n}}. \]
We will show that \( \hat{J} \) is asymptotically normal in distribution. Set \( X_i = P_o(p_s,n)^{-1} v_i \) and consider the sum
\[ S_p = \sqrt{m} \sum_{i=1}^{m} X_i - \frac{\text{E} \{ X_i \}}{\sqrt{\text{Var} \{ X_i \}}} \]
where \( Y_i \) are the normalized exchangeable random sequence \( \{ X_i \} \). Because \( |J(f_{U,U_{-}}) - J(p_s(f_{U,U_{-}}))| \to 0 \) and \( \text{Var} \{ X_i \}/m - (p_s/p)^2 J(f_{U,U_{-}})/\kappa_n \to 0 \), it follows that \( \hat{J} \) has the same asymptotic distribution as \( S_p \). We will now show that \( S_p \) converges weakly to a standard normal distribution.

From our analysis of the variance of \( \hat{J} \) and using the fact that \( b_{ij} \) are binomial, we see that the corresponding \( Y_i \) satisfy
\[ C[Y_1, Y_2] = (p_s^{-2} \kappa_n)/(p_s^2 J(f_{U,U_{-}})/\kappa_n)) \] \[ \to 0, \] as \( p \to \infty \). It is similarly possible to show that \( C[Y_1^2, Y_2^2] \to 0 \) as \( p \to \infty \). Then, using our central limit theorem for asymptotically uncorrelated interchangeable processes (Theorem 3. [10]), it follows that
\[ \lim_{p \to \infty} Pr \{ \hat{J} \leq \alpha \} = \lim_{p \to \infty} Pr \{ S_p \leq \alpha \} = \phi(\alpha), \]
where \( \phi(\cdot) \) is the distribution function of a Gaussian random variable with mean 0 and variance 1. Bounds on the Wasserstein distance between the distribution of \( \hat{J} \) and the standard normal distribution can be found in [10].
V. Experiments

The first set of experiments deal with verifying the theoretical results on bias, variance and asymptotic distribution for the estimator $\hat{J}$. In the second set of experiments, we use these results to test if the hypothesis $H_0 : J = 1$ holds, i.e., if the variates are independent. If $J = 1$, we can then use the theory in [4] to set appropriate thresholds for false alarm rates. The third set of experiments deal with the $r$-sparse scenario, i.e. $J \neq 1$. In this case, we show that using our estimate of $J$ to determine screening thresholds for desired false alarm rates works better than using the approximation $J = 1$.

A. Verification of theory

We simulate $n = 10$ i.i.d. samples from a multivariate Gaussian distribution with $p = 1000$ where all but $r = 5$ variates are uncorrelated. Clearly, this situation corresponds to the $r$-sparse scenario in Sec. IV-E2. In the first experiment, we vary $p_s$ from 1 to 1000 and compare the experimentally observed M.S.E. and the theoretically predicted bound on the M.S.E. This is shown in Fig. 2.

As predicted by our theoretical bound, the MSE decreases with decreasing $p_s$ in this $r$-sparse case (Sec. IV-E2). Furthermore, we see that there is close agreement between the theoretical bound and the experimental MSE in the regime $p_s > 10$. This can be explained by observing that for large values of $p_s$, the term $(2J(f_0, U_{(1)}))/(\kappa_0) (p_s/p)^2$ dominates the MSE and this term is exact. On the other hand, for small values of $p_s$, the MSE is dominated by the upper bound $2M/3p$.

*Alternative estimator:* From Eq. 1 (proposition 1 in [4]), one can propose the following alternative estimator for $\hat{J}$: $\hat{J} = 2N/\alpha d_n$, where $N$ is the number of discoveries at threshold $p$. We note that the estimator $\hat{J}$ for $p_s = p$ is identical to $\tilde{J}$. From our theory in Sec. IV-E2 and the corresponding simulation results in Fig. 2, we see that the MSE performance of $\tilde{J}$ is indeed significantly worse when compared to $\hat{J}$ for small values of $p_s$.

In our next experiment, we fix $p_s$ and obtain repeated estimates of $\hat{J}$ and use these to obtain a q-q plot of the quantiles of $\hat{J}$. The q-q plot in Fig. 3 shows that the asymptotic distribution of $\tilde{J}$ is indeed standard normal.

B. Test for independence

In this set of experiments, we are interested in verifying if the variates in a given data set are independent or not, i.e. to test if $J = 1$. We first consider two training data sets: (i) a multivariate normal sample where all the variates are independent, and (b) a multivariate normal sample where a fraction of 5% of the variates are dependent. We show the 95% confidence intervals on $J$ for increasing dimension $p$ for these two cases in Fig. 4(a) and Fig. 4(b) respectively. As expected, the confidence intervals tightly sandwich around $J = 1$ in the first case and a value $J > 1$ in the second case.

Given the above result, we use the central limit theorem of $\hat{J}$ to determine the p-value that the hypothesis $H_0 : J = 1$ holds. We generate 2 sets of 100 training samples each of $n = 10$ i.i.d. samples from a multivariate Gaussian distribution with $p = 1000$ with all the variates being independent in the first set and $r = 5$ variates being dependent in the second set. The resulting p-values are shown in histograms in Fig. 5(a) and Fig. 5(b) respectively. In the case of independent variates, 1% of the training sets have p-value less than 0.01. On the other hand, 90% of the training sets have p-value less than 0.01 in the dependent variates scenario. The estimate $\hat{J}$ can therefore be used as a test statistic to determine if all the variates in a given sample are independent or not.

C. Correlations between spam harvesters

In this experiment, we analyze email spam data collected via Project Honey Pot [7], a web-based network for monitoring harvesting and spamming activity by using trap email addresses. Project Honey Pot is a distributed system for monitoring harvesting and spamming activity via a network of decoy web pages with trap email addresses, known as honey pots. For every spam email received at a trap email address,
the Project Honey Pot data set provides us with the IP address of the harvester that acquired the recipient’s email address.

We are specifically interested in studying the correlation between harvesters in a given month. Using data from Project Honey Pot, for each of the 24 months between January 2006 and December 2007, we build a coincidence matrix where the columns correspond to harvesters and the rows correspond to each day of the month. Each entry in the matrix therefore corresponds to the number of emails collected by a particular harvester in a given day of the month.

We treat the number of emails harvested on each day to be independent and identically distributed, i.e. we treat each coincidence matrix as a $n \times p$ data matrix $\mathcal{X}$, where the dimension $p$ is the number of harvesters and $n$ is the number of days in a given month. The average number of harvesters each month is approximately $p_{avg} \approx 1500$. We now estimate the dependence measure $J$ for each of the 24 months along with 95% confidence intervals. This is shown in Fig. 6.

A stronger dependence measure $J$ indicates greater correlation among the harvesters, which can be viewed as an indicator of more co-ordinated harvesting activity. Our premise is partly corroborated by the fact that there is an increase in harvester dependence in October 2006 (corresponding to time point 10 in Fig. 6) which coincides with media reports [2] suggesting that there was a spam outbreak in October 2006.

D. Correlation screening for dependent variates

In [4], the authors used the Poisson-type limit in Eq. 2 to set screening thresholds for desired false positive rates. To do so, the authors use examples where the data sets are sparsely dependent and subsequently approximate the value of $J$ by 1.
In our first experiment, we use this limit in Eq. 2 and the value of $J = 1$ to determine screening thresholds in a $n = 10$, $p = 1000$ multivariate normal sample where all the variates are independent. There is very good agreement between the desired and observed false alarm rates as shown in Table I. Next, we consider two sets of a $n = 10$, $p = 1000$ multivariate normal samples where $r = 5$ of the $p = 1000$ variates are dependent. In previous experiments, we have shown that the value of $J$ significantly differs from 1 in this case. The first data set is used as training sample to estimate $J$. We now repeat the correlation screening experiment on the second data set, with two sets of thresholds - the first set of thresholds are determined with the value $J = 1$ and the second set of values are determined by the upper 95% confidence interval on $J$ as determined previously in Fig. 4(b). Using the upper 95% confidence interval $J_u$ should guarantee that the observed false rate is always less than or equal to the desired false alarm rate. The desired and corresponding observed false alarm rates are shown in Table II. We see that the observed false alarm rates are significantly higher than the desired rates when setting thresholds using the approximation $J = 1$. On the other hand, there is very good agreement between the desired and observed false alarm rates when setting the screening threshold using the upper 95% confidence interval $J_u$. Using dependency adjusted thresholds determined by using $J = J_u$ helps remove additional false positives that are otherwise discovered when screening at thresholds determined by using the approximation $J = 1$. This result underlines the contribution of this paper.

**E. Correlation screening on gene expression data**

We applied the correlation screening theory to a dataset downloaded from the public Gene Expression Omnibus (GEO) NCBI web site [3]. The dataset consists of 108 Affymetrix HU133 Genechips containing $p = 22,283$ gene probes hybridized from peripheral blood samples taken from 6 individuals at 5 time points (0,1,2, 4 and 12 hours) on four independent days under $m = 4$ treatments: intake of alcohol, grape juice, water, or red wine. After removing samples taken at pretreatment baseline (time 0) there remained $n = 87$ samples distributed over the treatments as: $n_1 = 20$ (alcohol), $n_2 = 22$ (grape juice), $n_3 = 23$ (water), and $n_4 = 22$ (wine).

Fig. 7 gives a visualization of the Z-scores for each treatment in the training data. Observe that the Z-scores display non-uniformity on the sphere $S_2$. Using the results in this paper, we are able to characterize this non-uniformity by estimating $J_i$, $i = 1,2,3,4$ from the data and subsequently determining the p-value that the hypothesis $H_1 : J_i = 1$ holds. Under the r-sparse assumption in Sec. IV-E2, the lower and upper 95% confidence intervals $J_i$, $J_u$ for each of the 4 treatments is shown in Table III below.

For comparison, we also compute the length $l$(MST) of the minimal spanning tree graphs constructed on the Z-scores, which is a test of randomness [5] of the Z-scores. The length of the MST graphs were normalized by the corresponding MST lengths on Z-scores of independent variates on $S_{n_1-2}$ to account for the different dimension $n_i - 2$ of the 4 treatments. From the results in Table III, we can infer a monotonic relationship between the MST lengths and the confidence intervals on $J_i$. Furthermore, we found the p-values that the hypothesis $H_1 : J_i = 1$ is true and the p-values based on minimal spanning tree lengths that the hypothesis: ‘the Z-scores are a uniform draw on $S_{n_i-2}$’ holds to be extremely small ($\sim 10^{-15}$). It is clear from these results that the Z-scores are not uniformly distributed on the spheres $S_{n_i-2}$.

### Table I

<table>
<thead>
<tr>
<th>Desired</th>
<th>.20</th>
<th>.10</th>
<th>.05</th>
<th>.02</th>
<th>.01</th>
</tr>
</thead>
<tbody>
<tr>
<td>Observed</td>
<td>.18</td>
<td>.11</td>
<td>.043</td>
<td>.016</td>
<td>.01</td>
</tr>
</tbody>
</table>

**Table I**

**DESIRED AND CORRESPONDING OBSERVED FALSE ALARM RATES IN CORRELATION SCREENING EXPERIMENT FOR INDEPENDENT VARIATES.** Thresholds are determined using the value $J = 1$. There is good agreement between desired and observed false alarm rates.

### Table II

<table>
<thead>
<tr>
<th>Desired</th>
<th>.20</th>
<th>.10</th>
<th>.05</th>
<th>.02</th>
<th>.01</th>
</tr>
</thead>
<tbody>
<tr>
<td>Observed ($J = 1$)</td>
<td>.30</td>
<td>.18</td>
<td>.063</td>
<td>.033</td>
<td>.020</td>
</tr>
<tr>
<td>Observed ($J = J_u$)</td>
<td>.216</td>
<td>.123</td>
<td>.043</td>
<td>.026</td>
<td>.013</td>
</tr>
</tbody>
</table>

**Table II**

**DESIRED AND CORRESPONDING OBSERVED FALSE ALARM RATES IN CORRELATION SCREENING EXPERIMENT FOR R-SPARSE VARIATES.** Thresholds are determined using (i) $J = 1$ and (ii) $J = J_u$. While there is good agreement between desired and observed false alarm rates using $J = J_u$, the observed false alarm rates with thresholds using $J = 1$ are significantly higher than the desired rates. Using the dependency adjusted thresholds with $J = J_u$ therefore helps remove spurious false positive discoveries.

### Table III

<table>
<thead>
<tr>
<th>Treatments</th>
<th>MST length</th>
<th>Lower level ($J_l$)</th>
<th>Upper level ($J_u$)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Alcohol</td>
<td>0.81</td>
<td>1.87</td>
<td>1.88</td>
</tr>
<tr>
<td>Grape Juice</td>
<td>0.76</td>
<td>1.81</td>
<td>1.83</td>
</tr>
<tr>
<td>Water</td>
<td>0.73</td>
<td>2.01</td>
<td>2.03</td>
</tr>
<tr>
<td>Wine</td>
<td>0.69</td>
<td>2.50</td>
<td>2.52</td>
</tr>
</tbody>
</table>

**Table III**

**NORMALIZED MST LENGTHS AND 95% CONFIDENCE INTERVALS ON DEPENDENCE MEASURE $J$ FOR EACH OF THE 4 TREATMENTS.** There is good agreement between the MST length statistic and the estimates of $J$.

In our experiment, we are interested in discovering dependencies between the genes under the alcohol treatment. Because we do not have training data to estimate $J_1$ for the alcohol treatment, we take the following approach. We test the hypothesis that the effects of alcohol and grape juice on the genes are similar. Specifically, we test the null hypothesis $H_0 : J_1 = J_2$ where $J_1$ and $J_2$ are the $J$-values for alcohol and grape juice respectively. The p-value for the above test was found to be 0.4192. Because the p-value is fairly high, we do not reject our null hypothesis $H_0$.

We now detect dependent genes in the alcohol data set using correlation screening at a false alarm rate of 0.01. We
again use two thresholds: the first threshold is determined under the alternative hypothesis $H_1^*: J_j = 1$ and the second threshold is determined under the null hypothesis $H_0$ by using $J_j = J_u$, where $J_u$ is the upper 95% confidence interval on $J_2$. Determining the dependency adjusted higher screening threshold using $J_j = J_u$ reduces the number of false positive gene pairs discovered at a desired false alarm rate as compared to using the naive threshold determined by approximating $J_1$ as 1. Using the value $J_j = J_u$ to determine the screening threshold in place of the approximation $J_j = 1$ reduced the number of discovered gene pairs from 671 to 468. This is further illustrated via dependency graphs in Fig. 8(a) and Fig. 8(b).

VI. CONCLUSION

Estimating the dependence measure $J$ is vital to choosing appropriate thresholds in correlation screening problems. In this paper, we introduced a simple kernel estimator and developed asymptotic analysis of the bias and variance for the estimator. We also established that the estimator converges weakly to a normal distribution. We used our analysis of the MSE to optimally choose free parameters in the estimator. We subsequently used the central limit theorem to obtain confidence intervals on $J$ and to test for independence of the variates. We used the confidence intervals to test for dependence between variables in simulated data sets and to analyze interactions between email spam harvesters. Finally, we used the estimate of $J$ to determine screening thresholds in correlation screening problems for achieving desired false alarm rates. We applied our results on a gene expression data set to detect dependencies between genes at a desired false alarm rate.

Figure 7. 3-dimensional projections of the Z-scores for the experimental beverage data under each of the 4 treatments: alcohol, grape juice, water and wine (clockwise from top left). For visualization the 22,283 variables (gene probes) were down sampled by a factor of 8 and a randomly selected set of four samples in each treatment were used to produce these figures. These projections show that the Z-scores are not uniformly distributed on the sphere $S_2$.

Figure 8. Dependency graph between 50 of the 22,283 genes in the data set under the alcohol treatment. The genes are represented by the red dots. An edge between two genes indicates that the corresponding gene pair under the alcohol treatment was discovered in the correlation screening experiment at a false alarm rate of 0.01. Spurious discoveries due to a naive lower screening threshold obtained by using the approximation $J_j = 1$ are eliminated by using a dependency adjusted lower screening threshold computed using the upper 95% confidence envelope $J_j = J_u$. The upper confidence level $J_u$ was determined using an independent set of gene expression data under grape juice treatment. Using the dependency adjusted threshold in place of the naive threshold therefore helps reduce the number of false positives that are discovered at any given false alarm rate.

ACKNOWLEDGMENT

This work is partially funded by the Air Force Office of Scientific Research, grant number FA9550-09-1-0471. The authors would like to thank Kevin Xu for collecting the email spam data.

REFERENCES