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 variables which are strongly correlated to each other. Indeed, in high dimensional settings when the number of samples n is very small compared to the number of variables p, screening out all but the highest sample correlations may be the only practical approach to discovering such dependencies.

Hero and Rajaratnam [4] showed that the threshold used to screen the sample correlation matrix must be carefully chosen due to an abrupt phase transitions phenomenon: when the threshold falls below a certain critical value, the number of discoveries increases rapidly. They further established that for large p the number of discoveries follows an asymptotic Poisson-type distribution, with a mean parameter which depends on an information theoretic measure J. If J were known these asymptotic results could be used to select the screening threshold to control Type I error and specify p-values. However, 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

$$P_o(\rho, n) = a_n \int_\rho^1 (1-u^2)^{\frac{n-4}{2}} du.$$

where a_n is

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

The quantity $P_o(\rho, n)/2$ 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 $\mathbf{U}$ 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

$$P_o(\rho, n) = \frac{a_n}{n-2}(1-\rho^2)^{(n-2)/2}(1+O(1-\rho^2)).$$

For random variables $\mathbf{U}$ and $\mathbf{V}$ with joint density $f_{\mathbf{U},\mathbf{V}}$ on $S_{n-2} \times S_{n-2}$ with marginals $f_{\mathbf{U}}$ and $f_{\mathbf{V}}$ define

$$J_\rho(f_{\mathbf{U},\mathbf{V}}) = \frac{Pr\left(||\mathbf{U}-\mathbf{V}|| \leq \sqrt{2(1-\rho)}\right)}{P_o(\rho,n)},$$

and the limit

$$J(f_{\mathbf{U},\mathbf{V}}) = \lim_{\rho \to 1} J_\rho(f_{\mathbf{U},\mathbf{V}}) = |S_{n-2}| \int_{S_{n-2}} f_{\mathbf{U},\mathbf{V}}(u,u)du.$$

The authors of [4] have shown that the number of discoveries in correlation screening are related to the measure J. The limit is equal to 1 when U and V are independent and uniformly distributed on S_{n-2}. Thus $J(f_{\mathbf{U},\mathbf{V}}) - 1$ is a measure of the deviation of the joint density from uniform $f_{\mathbf{U},\mathbf{V}}(u,v) = |S_{n-2}|^2$.

We give an intuitive interpretation of $J(f_{\mathbf{U},\mathbf{V}})$ as a measure of dependence between $\mathbf{U}$, $\mathbf{V}$. It is equal to the Bhattacharyya affinity between the product of the marginal distributions $m_{\mathbf{U},\mathbf{V}}(w) = f_{\mathbf{U}}(w)f_{\mathbf{V}}(w)$ and the product of the conditional distributions $c_{\mathbf{U},\mathbf{V}}(w) = f_{\mathbf{U}|\mathbf{V}}(w|w)f_{\mathbf{V}|\mathbf{U}}(w|w)$:

$$J(f_{\mathbf{U},\mathbf{V}}) = |S_{n-2}| \int \sqrt{m_{\mathbf{U},\mathbf{V}}(w)c_{\mathbf{U},\mathbf{V}}(w)}dw,$$
$$\leq |S_{n-2}| \left(\int m_{\mathbf{U},\mathbf{V}}(w)dw\right)^{1/2} \left(\int c_{\mathbf{U},\mathbf{V}}(w)dw\right)^{1/2},$$

where equality occurs iff $f_{\mathbf{U},\mathbf{V}}(u,u) = f_{\mathbf{U}}(u)f_{\mathbf{U}}(v)$.

The measure J is maximized when $f_{\mathbf{U},\mathbf{V}}(u,u) = f_{\mathbf{U}}(u)f_{\mathbf{V}}(u)$ 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 $\mathbf{X} = [X_1, ..., X_p]^T$ be a vector of random variables with mean μ and $p \times p$ covariance matrix $\mathbf{\Sigma}$. Denote the correlation matrix $\mathbf{\Gamma} = \mathbf{D_\Sigma}^{-1/2}\mathbf{\Sigma}\mathbf{D_\Sigma}^{-1/2}$ where $\mathbf{D_\Sigma} = diag_i(\mathbf{\Sigma}_{ii})$ is the diagonal matrix of variances of components of $\mathbf{X}$. Assume that n independent identically distributed (i.i.d.) samples of $\mathbf{X}$ are available and arrange these samples in a $n \times p$ data matrix

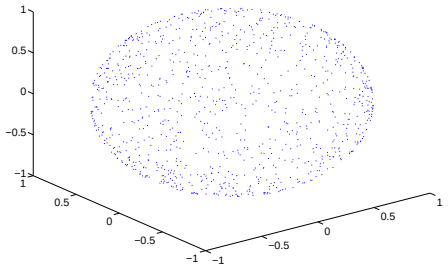
$$\mathcal{X} = [\mathbf{X}_1, \ldots, \mathbf{X}_p] = [\mathbf{X}_{(1)}^T, \ldots, \mathbf{X}_{(n)}^T]^T,$$

where $\mathbf{X}_i = [X_{1i}, \ldots, X_{ni}]^T$ and $\mathbf{X}_{(i)} = [X_{i1}, ..., X_{ip}]$ denote the i-th column and row, respectively, of $\mathcal{X}$.

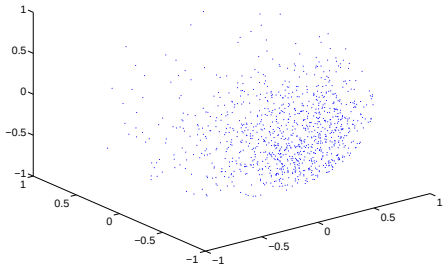
Define the sample mean of the i-th column $\overline{X}_i = n^{-1}\sum_{j=1}^n X_{ji}$, the vector of sample means $\overline{\mathbf{X}} = [\overline{X}_1, \ldots, \overline{X}_p]$, the $p \times p$ sample covariance matrix $\mathbf{S} = (n-1)^{-1}(\mathbf{X}_{(i)} - \underline{\mathbf{X}})^T(\mathbf{X}_{(i)} - \underline{\mathbf{X}})$, and the $p \times p$ sample correlation matrix

$$\mathbf{R} = \mathbf{D_S}^{-1/2}\mathbf{S}\mathbf{D_S}^{-1/2},$$

where $\mathbf{D_S} = diag_i(\mathbf{S}_{ii})$ is the diagonal matrix of component sample variances. Let the ij-th entry of the ensemble covariance $\mathbf{I}$ be denoted γ_{ij} and the ij-th entry of the sample covariance $\mathbf{R}$ be r_{ij}.



(a) Z-scores for multivariate sample with diagonal covariance matrix. The Z-scores are uniformly distributed over the unit sphere S_{n-2}.



(b) Z-scores for multivariate sample with non-diagonal covariance matrix. The Z-scores are concentrated in clumps over the unit sphere S_{n-2}.

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.

We now define the Z-scores, $\mathcal{U} = [\mathbf{U}_1, \ldots, \mathbf{U}_p]$, $\mathbf{U}_i \in \mathcal{R}^{n-1}$ as follows. The Z-scores are constructed to lie on the $(n-2)$-sphere $S_{n-2} \in \mathcal{R}^{n-1}$ and are determined by projecting away the components of $\mathbf{X}_i$ orthogonal to the $n-1$ dimensional hyperplane $\mathbf{u} \in \mathcal{R}^n : \mathbf{1}^T\mathbf{u} = 0, i = 1, \ldots, p$. The Z-scores can be computed from $\mathcal{X}$ using a Gramm-Schmidt procedure. The Z-scores satisfy the relation

$$\mathbf{R} = \mathcal{U}^T\mathcal{U}.$$

Furthermore, the sample correlation between $\mathbf{X}_i$ and $\mathbf{X}_j$ can be computed using the inner product or the Euclidean distance between associated Z-scores

$$r_{ij} = \mathbf{U}_i^T\mathbf{U}_j = 1 - \frac{||\mathbf{U}_i - \mathbf{U}_j||^2}{2}.$$

Pairs of Z-scores that are close to each other, as measured by Euclidean distance, therefore have high associated sample correlation. The Z-score lives in a geometry, the $(n-2)$-sphere of co-dimension 1 shown in Fig. 1. It is well known that when the rows of the data matrix $\mathcal{X}$ follow a diagonal elliptical distribution the Z-scores are uniformly distributed on S_{n-2} [1].

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 $\mathbf{X}$. This experiment produces the data matrix $\mathcal{X}$, which is a $n \times p$ matrix. From this data matrix, extract the Z-score matrix $\mathcal{U}$. We then construct the sample correlation matrix $\mathbf{R} = \mathcal{U}^T\mathcal{U}$.

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 ρ. 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. Assume that the correlation threshold ρ depends on the number of variables p as

$$\lim_{p\to\infty} p(p-1)(1-\rho^2)^{(n-2)/2} = d_n$$

for some finite constant d_n. Then the expected number of discoveries satisfy

$$\lim_{p\to\infty} \mathbf{E}[N] = \kappa_n J(f_{\mathbf{U}_{\cdot},\mathbf{U}_{*-\cdot}}), \tag{1}$$

where $\kappa_n = a_n d_n/(n-2)$, $f^s_{\mathbf{U},\mathbf{V}}(u,v) = (1/2)(f_{\mathbf{U},\mathbf{V}}(u,v) + f_{\mathbf{U},\mathbf{V}}(u,-v))$ and

$$f_{\mathbf{U}_{\cdot},\mathbf{U}_{*-\cdot}}(u,v) \quad = \quad \frac{1}{p(p-1)} \sum_{i=1}^{p} \sum_{j\neq i}^{p} \left(f^s_{\mathbf{U}_i,\mathbf{U}_j}(u,v) \right),$$

is the average Z-score. Furthermore, under a suitable weak dependency condition [4], the variable N also satisfies the following Poisson-type limit

$$\lim_{p\to\infty} Pr(N > 0) = e^{-\lambda/2}, \tag{2}$$

where $\lambda = \kappa_n J(f_{\mathbf{U}_{\cdot},\mathbf{U}_{*-\cdot}})$ is the limiting mean.

Contribution: We note that when the variables are independent, the value of J is identically equal to 1. For variables which are weakly dependent, Hero et.al. [4] use this fact to determine screening thresholds for desired false alarm rates using Eq. 2 by approximating the value J as 1. However, if the dependency is stronger, the value of J can significantly differ from 1 and it therefore becomes vital to estimate J to determine screening thresholds. The rest of this paper is concerned with the estimation of J.

IV. Estimation of J

Consider a p dimensional joint density $f_{\mathbf{U}_1,\ldots,\mathbf{U}_p}$ defined on $S_{n-2} \times \ldots \times S_{n-2}$. Now define the average pairwise density

$$f_{\mathbf{U}_{\cdot},\mathbf{U}_{*-\cdot}}(\mathbf{u},\mathbf{v}) = \frac{1}{p(p-1)} \sum_{i=1}^{p} \sum_{j\neq i}^{p} \left(f^s_{\mathbf{U}_i,\mathbf{U}_j}(u,v) \right).$$

We are interested in estimating $J(f_{\mathbf{U}_{\cdot},\mathbf{U}_{*-\cdot}})$ from a single sample $[\mathbf{U}_1, \ldots, \mathbf{U}_p]$ which is drawn from the joint density $f_{\mathbf{U}_1,\ldots,\mathbf{U}_p}$. Note that

$$J(f_{\mathbf{U}_{\cdot},\mathbf{U}_{*-\cdot}}) = \frac{1}{p(p-1)} \sum_{i=1}^{p} \sum_{j\neq i}^{p} J(f_{\mathbf{U}_i,\mathbf{U}_j}).$$

In order to estimate $J(f_{\mathbf{U}_{\cdot},\mathbf{U}_{*-\cdot}})$, we build an uniform kernel estimator $\hat{J}$, 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 ρ, let ρ_s be the estimation threshold which satisfies (i) $\lim_{p\to\infty} \rho_s = 1$, and (ii) $\lim_{p\to\infty}(1-\rho)/(1-\rho_s) = 0$.

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 $\mathbf{U}_j \in A(r,\mathbf{U}_i)$. Note that there are $p(p-1)/2$ distinct combinations of $\{i,j\}$. Therefore the cardinality of the set $\mathcal{B} = \{b_{ij}, i<j\}$ is $p(p-1)/2$.

We sample from the set $\mathcal{B}$ without replacement to obtain a finite sequence of random variables $\mathcal{V} = \{\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_m\}$, where $m = p(p-1)/2$. We now define our estimator of $J_\rho(f_{\mathbf{U}_{\cdot},\mathbf{U}_{*-\cdot}})$ to be

$$\hat{J} = \frac{1}{P_o(\rho_s,n)} \left(\frac{1}{m} \sum_{i=1}^{m} \mathbf{v}_i \right). \tag{3}$$

B. Intuition

Before stating the theoretical results, we provide some intuition behind the requirement $(1-\rho)/(1-\rho_s) \to 0$ which translates to the condition that the estimation threshold ρ_s approaches 1 slower than the screening threshold ρ.

The condition $\rho_s \to 1$ guarantees that the radius $r = \sqrt{2(1-\rho_s)}$ of the region $A(r,v)$ decays to 0, which in turn implies that the bias of $\hat{J}$ should decay to 0. On the other hand, the condition $\rho_s \to 1$ slower than $\rho \to 1$ guarantees that the average number of points $\mathbf{U}_j$ which fall in $A(r,\mathbf{U}_i)$ grows to ∞ as $p \to \infty$, which implies that the variance should decay to 0.

For notational convenience, henceforth denote $P_o(\rho_s,n)$ by P_o. Note that $P_o = \frac{a_n}{n-2}(1-\rho^2)^{(n-2)/2}(1+O(|1-\rho|^2))$, which implies that $p(p-1)P_o \to a_n d_n/(n-2)$, as $p \to \infty$. Also let p_s correspond to ρ_s according to the relation $\lim_{p\to\infty} p_s(p_s-1)(1-\rho_s^2)^{(n-2)/2} = d_n$. Then p_s satisfies the following conditions: $p_s/p \to 0$, $1/p_s \to 0$ as $p \to \infty$. We will now state results on the bias, variance and asymptotic

distribution of the estimator $\hat{J}$. The proofs for these results can be found in our technical report [9].

C. Bias

Define the maximal gradient of the average pairwise density to be

$$\nabla M = \sup_{u,v \in S_{n-2}} ||\nabla_v f_{\mathbf{U}_{\cdot},\mathbf{U}_{*-\cdot}}(u,v)|_{u=v}||.$$

The bias of $\hat{J}$ is bounded by

$$\begin{aligned} &|\,\mathbf{E}[\hat{J}] - J(f_{\mathbf{U}_{\cdot},\mathbf{U}_{*-\cdot}})| \leq 2a_n^2 \nabla M \sqrt{2(1-\rho_s)} \\ &= 2\sqrt{2} a_n^2 d_n^{(2-n)} \nabla M \left(\frac{1}{p_s}\right)^{\frac{2}{n-2}} (1+o(1)), \end{aligned} \quad (4)$$

which implies that the estimator $\hat{J}$ is asymptotically unbiased.

D. Variance

Define

$$M = \left[\underset{i \neq j \neq k}{avg} \sqrt{\mathbf{V}[f_{\mathbf{U}_j|\mathbf{U}_i}(\mathbf{U}_i|\mathbf{U}_i)]\,\mathbf{V}[f_{\mathbf{U}_k|\mathbf{U}_i}(\mathbf{U}_i|\mathbf{U}_i)]} \right].$$

Also define,

$$\begin{aligned} \delta_{ijkl} &= f_{\mathbf{U}_i,\mathbf{U}_j,\mathbf{U}_k,\mathbf{U}_l}(u_i,u_j,u_k,u_l) \\ &\quad - f_{\mathbf{U}_i,\mathbf{U}_j}(u_i,u_j) f_{\mathbf{U}_k,\mathbf{U}_l}(u_k,u_l), \end{aligned}$$

and

$$\delta_p = \underset{i \neq j \neq k \neq l}{avg} \left(\sup_{u_i,u_j,u_k,u_l} |\delta_{ijkl}| \right),$$

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

$$\begin{aligned} \mathbf{V}[\hat{J}] &= (P_o(\rho_s,n))^{-2}\,\mathbf{V}\left[(1/m)\sum_{i=1}^m \mathbf{v}_i\right] \\ &= ((P_o(\rho_s,n))^{-2}/m)\,\mathbf{V}[\mathbf{v}_1] \\ &\quad + (P_o(\rho_s,n))^{-2}(1-1/m)\,\mathbf{C}[\mathbf{v}_1,\mathbf{v}_2] \\ &\leq \frac{2J(f_{\mathbf{U}_{\cdot},\mathbf{U}_{*-\cdot}})}{\kappa_n}\left(\frac{p_s}{p}\right)^2 (1+o(1)) \\ &\quad + \left(\frac{2M}{3}\left(\frac{1}{p}\right) + a_n^2 \delta_p\right)(1+o(1)), \end{aligned} \quad (5)$$

and the variance of our estimator tends to 0 as $p \to \infty$.

E. Optimization of MSE

Note that our intuition behind choosing $(1-\rho)/(1-\rho_s) \to 0$ is verified by our expressions for the MSE. ρ_s approaches 1 slower than ρ, 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 $O(p_s{}^{-4/(n-2)} p^{-1})$ 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 $\mathcal{V} = \{\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_m\}$ are exchangeable [6] for every p, we can assume that $\mathcal{V} = \{\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_m\}$ are a finite segment of an infinite length exchangeable sequence. Consider the normalized version of our estimator $\hat{J}$ given by

$$\tilde{J} = \frac{\hat{J} - J(f_{\mathbf{U}_{\cdot},\mathbf{U}_{*-\cdot}})}{(p_s/p)\sqrt{2J(f_{\mathbf{U}_{\cdot},\mathbf{U}_{*-\cdot}})/\kappa_n}}. \quad (6)$$

We will show that $\tilde{J}$ is asymptotically normal in distribution. Set $X_i = P_o(\rho_s,n)^{-1}\mathbf{v}_i$ and consider the sum

$$S_p = \sqrt{m}\sum_{i=1}^m \frac{X_i - \mathbf{E}[X_i]}{\sqrt{\mathbf{V}[X_i]}} = \sqrt{m}\sum_{i=1}^m Y_i,$$

where Y_i are the normalized exchangeable random sequence $(X_i - \mathbf{E}[X_i])/\sqrt{\mathbf{V}[X_i]}$. Because $|J(f_{\mathbf{U}_{\cdot},\mathbf{U}_{*-\cdot}}) - J_{\rho_s}(f_{\mathbf{U}_{\cdot},\mathbf{U}_{*-\cdot}})| \to 0$ and $|\,\mathbf{V}[X_i]/m - (p_s/p)^2 2J(f_{\mathbf{U}_{\cdot},\mathbf{U}_{*-\cdot}})/\kappa_n| \to 0$, it follows that $\tilde{J}$ 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

$$\mathbf{C}[Y_1,Y_2] = ((P_o^{-2}\kappa_n)/(p_s^2 J(f_{\mathbf{U}_{\cdot},\mathbf{U}_{*-\cdot}})))\,\mathbf{C}[\mathbf{v}_1,\mathbf{v}_2] \to 0,$$

as $p \to \infty$. It is similarly possible to show that $\mathbf{C}[{Y_1}^2, {Y_2}^2] \to 0$ as $p \to \infty$. Then, using our central limit theorem for asymptotically uncorrelated interchangeable processes (Theorem 3. [10]), it follows that

$$\lim_{p\to\infty} Pr\{\tilde{J} \leq \alpha\} = \lim_{p\to\infty} Pr\{S_p \leq \alpha\} = \phi(\alpha), \quad (7)$$

where $\phi(.)$ is the distribution function of a Gaussian random variable with mean 0 and variance 1. Bounds on the Wassertein distance between the distribution of $\tilde{J}$ 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

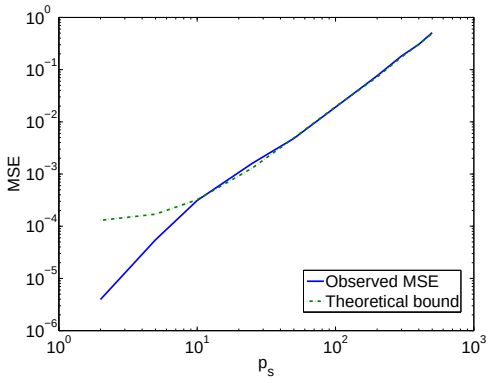


Figure 2. Variation of experimentally observed MSE of $\hat{J}$ and theoretical bound with varying p_s for a $n = 10$, $p = 1000$, $r = 5$ multivariate normal sample. The MSE indeed decreases with decreasing p_s as predicted by our theory in Sec. IV-E2. Furthermore, there is good agreement between the predicted bound and the observed MSE for $p_s > 10$ (see Sec. V-A for explanation.)

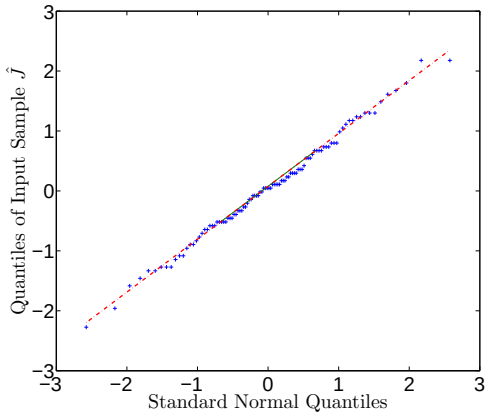


Figure 3. q-q plot of normalized quantiles of $\hat{J}$ and standard normal quantiles. The linear agreement between the quantiles corroborates that the asymptotic distribution of $\tilde{J}$ is indeed standard normal.

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 $r = 5$ variables 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 $(2J(f_{\mathbf{U}_{\cdot,}\mathbf{U}_{*-\cdot}})/\kappa_n)\,(p_s/p)^2$ 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 J: $\check{J} = 2N/a_n d_n$, where N is the number of discoveries at threshold ρ. We note that the estimator $\hat{J}$ for $p_s = p$ is identical to $\check{J}$. From our theory in Sec. IV-E2 and the corresponding simulation results in Fig. 2, we see that the MSE performance of $\check{J}$ 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 $\tilde{J}$. The q-q plot in Fig. 3 shows that the asymptotic distribution of $\tilde{J}$ is indeed standard normal.

B. Test for independence

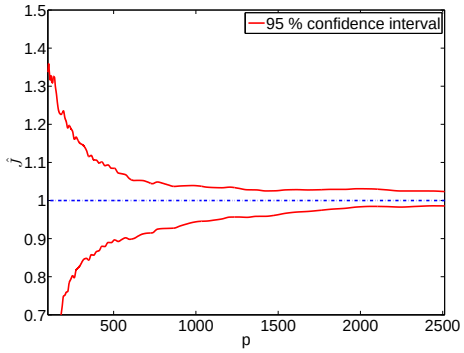
In this set of experiments, we are interested in verifying if the variates in a given data set are independent or not, i.e. to test if $J = 1$. We first consider two training data sets: (i) a multivariate normal sample where all the variates are independent, and (b) a multivariate normal sample where 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 $\tilde{J}$ 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.

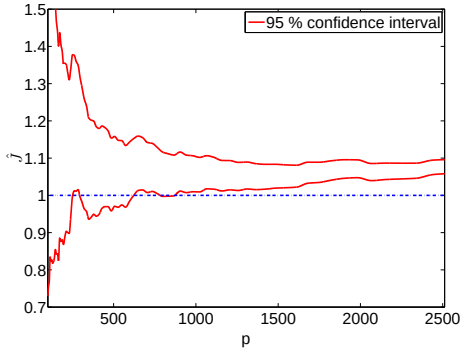
C. Correlations between spam harvesters

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 recipientâs email address.

We are specifically interested in studying the correlation between harvesters in a given month. Using data from Project

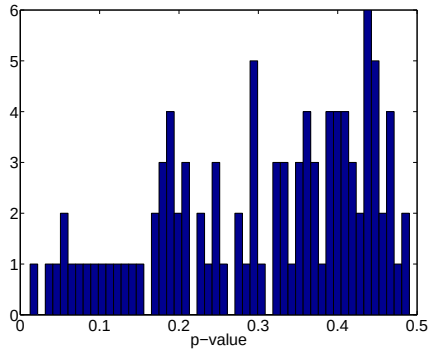


(a) 95% confidence intervals on J for independent variates. The confidence intervals sandwich the value $J = 1$.

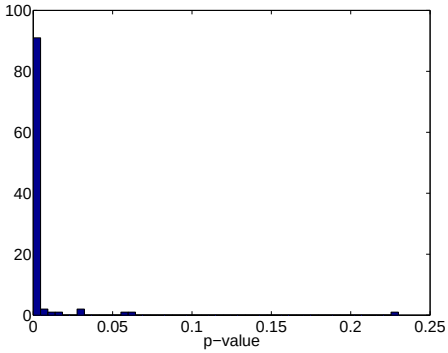


(b) 95% confidence intervals on J for dependent variates. The confidence intervals sandwich a value $J \neq 1$.

Figure 4. Confidence intervals on J with increasing dimension p. The confidence intervals for independent and dependent variates sandwich around values $J = 1$ and $J \neq 1$ respectively. For sufficiently large dimension p, the estimated confidence intervals can therefore be used to tightly bound J with high probability.



(a) Histogram of p-values for independent variates. A fraction of 1% of the p-values are below 0.01.



(b) Histogram of p-values for dependent variates. 90% of the p-values are below 0.01.

Figure 5. Histograms of p-values for the hypothesis testing problem $H_0 : J = 1$. A very small fraction of p-values are smaller than 0.01 in the independent variates case as opposed to a fraction of 90% in the dependent variates case. The statistic $\hat{J}$ can therefore be used to test for independence of the variates.

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 the 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 a 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.

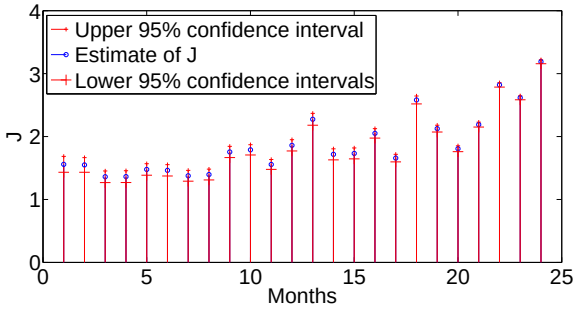


Figure 6. Dependence measure J between harvesters for each of the 24 months between January 2006 and December 2007 with confidence intervals. The dependence measure indicates increasing co-ordinated harvester activity with time.

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

False alarm rates for independent variates; $J = 1$					
Desired	.20	.10	.05	.02	.01
Observed	.18	.111	.043	.016	.010

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.

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

False alarm rates for r-spare variates; $J \neq 1$					
Desired	.20	.10	.05	.02	.01
Observed ($J = 1$)	.300	.180	.063	.033	.020
Observed ($J = J_u$)	.216	.123	.043	.026	.013

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.

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_l, 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_i-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}.

95% confidence intervals on J for gene expression data			
Treatments	MST length	Lower level (J_l)	Upper level (J_u)
Alcohol	0.81	1.87	1.88
Grape Juice	0.76	1.81	1.83
Water	0.73	2.01	2.03
Wine	0.69	2.50	2.52

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 $H_1 : J_1 = 1$ and the second threshold is determined under the null hypothesis H_0 by using

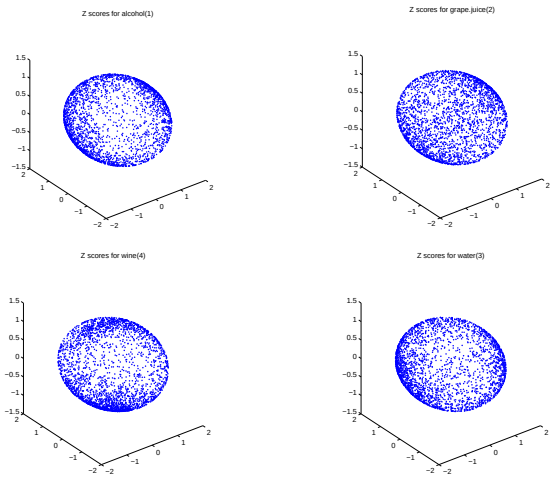
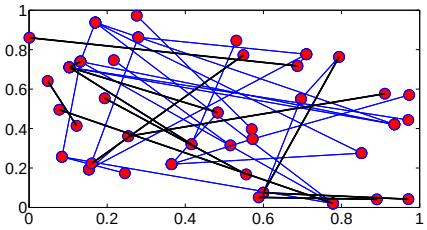
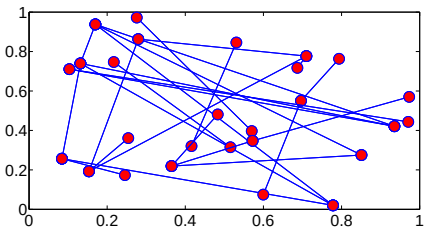


Figure 7. 3-dimensional projections of the Z-scores for the experimental beverage data under each of the 4 treatments: alcohol, grape juice, water and wine (clockwise from top left). For visualization the 22,283 variables (gene probes) were down sampled by a factor of 8 and a randomly selected set of 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.



(a) Dependency graph at false alarm rate 0.01 using the approximation $J_1 = 1$ to determine screening threshold. Spurious discoveries using this naive screening threshold are shown in black.



(b) Dependency graph at false alarm rate 0.01 using the value $J_1 = J_u$ to determine screening threshold. Spurious false positives are eliminated by using this lower screening threshold.

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 $J_1 = 1$ are eliminated by using a dependency adjusted lower screening threshold computed using the upper 95% confidence envelope $J_1 = J_u$. 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.

$J_1 = J_u$, where J_u is the upper 95% confidence interval on J_2. Determining the dependency adjusted higher screening threshold using $J_1 = J_u$ 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 $J_1 = J_u$ to determine the screening threshold in place of the approximation $J_1 = 1$ 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.

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

[1] T. W. Anderson. *An Introduction to Multivariate Statistical Analysis, 2nd Edition*. Wiley-Interscience, 2 edition, September 1984.

[2] M. Austin. Spam at epic levels. ITPro. [Online]. Available: http://www.itpro.co.uk/97589/spam-at-epic levels, Novemeber 2006.

[3] F. Baty, M. Facompre, J. Wiegand, J. Schwager, and M. Brutsche. Blood response to various beverages: time course. NCBI GEO, record number GDS2767. Available: http://www.ncbi.nlm.nih.gov/sites/entrez. 2006.

[4] A. O. Hero III and B. Rajaratnam. Large Scale Correlation Screening. *ArXiv e-prints*, February 2011.

[5] R. Hoffman and A. K. Jain. A test of randomness based on the minimal spanning tree. *Pattern Recognition Letters*, 1(3):175 – 180, 1983.

[6] J. F. C. Kingman. Uses of exchangeability. *Ann. Probab.*, 6:183–197, 1978.

[7] Project Honey Pot. Available online at http://www.projecthoneypot.org.

[8] H. Ruben. Probability content of regions under spherical normal distributions, iii: The bivariate normal integral. *Annals of Mathematical Statistics*, 32:171–186, March 1961.

[9] K. Sricharan, A. O. Hero III, and B. Rajaratnam. Local information measures in large scale correlation screening. *Technical Report, Communications and Signal Processing Laboratory, The University of Michigan*, April 2011. (To appear).

[10] K. Sricharan, R. Raich, and A. O. Hero III. Empirical estimation of entropy functionals with confidence. *ArXiv e-prints*, December 2010.