Proceedings of the IEEE Intl. Symposium on Information Theory, July 2012.

Large scale correlation detection

Francesca Bassi[1,2]
[1]LSS – CNRS – SUPELEC – Univ Paris Sud
Gif-sur-Yvette, France

Alfred O. Hero III[2]
[2]University of Michigan, EECS Department
Ann Arbor, MI

***Abstract*—This work addresses the problem of correlation detection in a group of elliptically-contoured variables, when the number p of variates greatly exceeds the number n of observed samples. We exploit the properties inherent to the Z-score representation of the data set to devise two different decision tests, whose performances are assessed by upper bounding the Type I and Type II error probabilities. The results specifically apply to the asymptotic regime where the number of variates p is large, and the number of samples n is finite and fixed.**

I. Introduction

An increasing number of practical applications require the solution of inference problems on high dimensional data sets, where the amount of considered features is large (some examples are sensor networks, gene expressions arrays, multimedia databases, multivariate financial time series, and traffic in communication networks). Their unifying trait is the small number n of available samples, in comparison to the feature dimension p. In this context, especially challenging tasks are the correlation screening problem [1], *i.e.* the identification of significant levels of linear dependence, and the covariance structure testing problem [2]. In gene expression analysis, for example, a biologist might be interested to test for structure of the gene regulatory network, but only has a handful (n) of samples to construct the sample correlation matrix between tens of thousands (p) of gene probes. In this situation classical large n asymptotic techniques cannot be reliably applied.

This work addresses the simpler problem of deciding whether the high dimensional data set is completely uncorrelated. Building upon the U-score representation introduced in [1] we test properties of the empirical distribution to decide the diagonality of the covariance matrix. This work extends previous results that hold only in the limiting asymptotic regime $p \to \infty$, $n \to \infty$. For normal multivariate samples, for instance, [3] provides the asymptotic distribution of the maximum eigenvalue of the sample correlation matrix. This statistic, however, is only meaningful for $p < n$. In [4] a quadratic form of the sample covariance matrix is shown to be a consistent statistic even for $p > n$, if n and p have equal growth rate. Under the same conditions, [5] and [6] derive the asymptotic distribution of the maximum sample correlation coefficient, test statistic first considered in [7], [8].

Differently from this approach, the results presented here specifically apply to the asymptotic regime $p \to \infty$ with n fixed and finite, also considered in [1], [9]. Unlike in correlation screening for sparse correlation matrices [1], we are interested in global tests for diagonal covariance, here derived applying the theory of exchangeability and the method of types.

The paper is organized as follows. Section II reviews the U-score representation of the data set introduced in [1]. The decision problem is formally defined in Section III. In Section III-A and Section III-B two different tests are proposed, whose performances are characterized by means of the respective Type I and Type II error exponents. Section IV provides experimental results.

II. The geometry of correlation

Let the size p random vector $\boldsymbol{X}$ be distributed according to an elliptically-contoured density $f_{\boldsymbol{X}}(\boldsymbol{x})$ (*e.g.* the multivariate normal, or the multivariate t-distribution, see [10, Sec. 2.7] and references therein). The elliptically-contoured density is specified by the parameters $\boldsymbol{\mu}$, $\boldsymbol{\Sigma}$, and $g(\cdot)$, where $g(\cdot)$ is a non-negative monotonic function:

$$f_{\boldsymbol{X}}(\boldsymbol{x}) = |\boldsymbol{\Sigma}|^{-1/2}\, g\big((\boldsymbol{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\big).$$

The elements of $\boldsymbol{X}$ are assumed to have finite first and second moments. This section introduces different possible representations of the data set collected by making n independent drawings from the *a priori* distribution $f_{\boldsymbol{X}}(\boldsymbol{x})$.

A. Z-score and U-score representations of the data

Raw data set: The matrix $\mathbb{X}_{[n\times p]}$ is formed by (vertically) stacking n independent drawings from $f_{\boldsymbol{X}}(\boldsymbol{x})$.

Z-scores: The matrix $\mathbb{Z}_{[n\times p]}$ is obtained centering and normalizing the columns of $\mathbb{X}$ with respect to their sample mean and standard deviation. Let $m_i = (n)^{-1}\mathbf{1}^T\boldsymbol{x}^{(i)}$ denote the i-th sample mean, and $s_i^2 = (n-1)^{-1}(\boldsymbol{x}^{(i)} - m_i\mathbf{1})^T(\boldsymbol{x}^{(i)} - m_i\mathbf{1})$ the i-th sample variance. The columns of $\mathbb{Z}$ are obtained from the columns of $\mathbb{X}$ as

$$\boldsymbol{z}^{(i)} = \frac{\boldsymbol{x}^{(i)} - m_i\mathbf{1}}{\sqrt{(n-1)}\ s_i}, \qquad \forall i \in \{1, 2, \cdots, p\}. \tag{1}$$

Geometrically, the columns of $\mathbb{Z}$ are points living in the $\mathbb{R}^n$ space. Each $\boldsymbol{z}^{(i)}$ belongs to the intersection of the hyperplane $\mathcal{M}_{n-1} = \{\boldsymbol{z} \in \mathbb{R}^n : \mathbf{1}^T\boldsymbol{z} = 0\}$ with the unit hypersphere $\mathcal{S}_{n-1} = \{\boldsymbol{z} \in \mathbb{R}^n : \parallel \boldsymbol{z} \parallel = 1\}$.

U-scores [1]: The matrix $\mathbb{U}_{[(n-1)\times p]}$ is obtained from $\mathbb{Z}$ as a result of the rotation by $\mathbf{H}^*$ of its columns, and of their subsequent projection over $\mathcal{M}_{n-1}$. Define $\mathbf{H}^*$ as an orthonormal matrix[1] whose first row is parallel to $\mathbf{1}^T$,

[1] As explained in [1], the matrix $\mathbf{H}^*$ can be easily obtained via Gram-Schmidt orthogonalization of the matrix composed by $\mathbf{1}^T$ and any other arbitrarily chosen $n-1$ linearly independent row vectors.

the vector orthogonal to $\mathcal{M}_{n-1}$. The first row in $\mathbf{H}^*\mathbb{Z}$ is identically null, and is removed by the projection onto $\mathcal{M}_{n-1}$. The U-score matrix is given by $\mathbb{U} = \mathbf{H}\mathbb{Z}$, where $\mathbf{H}$ is $\mathbf{H}^*$ deprived of its first row. Geometrically, the columns of $\mathbb{U}$ are points in the $\mathbb{R}^{n-1}$ space. The transformation $\mathbf{H}\boldsymbol{z}^{(i)}$ is norm preserving, so that the columns of $\mathbb{U}$ belong to $\mathcal{S}_{n-2}$.

The geometry of correlation [1]*:* In [1, Sec. 2] the analysis of the properties of the U-scores for elliptically-contoured $f_{\boldsymbol{X}}(\boldsymbol{x})$ is presented. When the elements of $\boldsymbol{X}$ are uncorrelated, and $\boldsymbol{\Sigma}$ is diagonal, the distribution of the columns of $\mathbb{U}$ is uniform on $\mathcal{S}_{n-2}$. For non-diagonal $\boldsymbol{\Sigma}$ the distribution is, in general, far from uniform. This property is the key to establishing the correlation detection procedure proposed here.

B. Statistical properties

In this subsection some relevant statistical properties of the matrices $\mathbb{X}$, $\mathbb{Z}$ and $\mathbb{U}$ are derived. The matrix $\mathbb{X}$, composed of independent drawings from $f_{\boldsymbol{X}}(\boldsymbol{x})$, is row–exchangeable [11], *i.e.* its probability law is unchanged upon row permutation. The conversion in Z-scores, detailed in (1), preserves exchangeability (though not independence) of the rows of $\mathbb{Z}$. These considerations justify the following proposition.

Proposition 1. *The matrix* $\boldsymbol{\Gamma}^{(ij)} = \mathrm{cov}\big(\boldsymbol{Z}^{(i)}, \boldsymbol{Z}^{(j)}\big)$ *has the form*

$$\boldsymbol{\Gamma}^{(ij)} = \beta^{(ij)}\ \mathbf{1}\mathbf{1}^T + (\alpha^{(ij)} - \beta^{(ij)})\ \mathbf{I}. \tag{2}$$

If X_i and X_j are uncorrelated, then $\alpha^{(ij)} = \beta^{(ij)} = 0$.

Proposition 2. *The matrix* $\boldsymbol{\Upsilon}^{(ij)} = \mathrm{cov}\big(\boldsymbol{U}^{(i)}, \boldsymbol{U}^{(j)}\big)$ *has the form*

$$\boldsymbol{\Upsilon}^{(ij)} = (\alpha^{(ij)} - \beta^{(ij)})\ \mathbf{I}. \tag{3}$$

If X_i and X_j are uncorrelated, then $\boldsymbol{\Upsilon}^{(ij)} = 0\ \mathbf{1}\mathbf{1}^T$.

Proof: The U-scores covariance matrix can be represented as $\boldsymbol{\Upsilon}^{(ij)} = \mathbf{H}\boldsymbol{\Gamma}^{(ij)}\mathbf{H}^T$. Combining this representation with (2) yields $\boldsymbol{\Upsilon}^{(ij)} = \beta^{(ij)}\mathbf{H}\mathbf{1}\mathbf{1}^T\mathbf{H}^T + (\alpha^{(ij)} - \beta^{(ij)})\mathbf{H}\mathbf{H}^T$. Recall that the rows of the matrix $\mathbf{H}$ belong, by construction, to the hyperplane $\mathcal{M}_{n-1}$. This implies $\mathbf{H}\mathbf{1} = \mathbf{0}$, and hence (3). ■

III. CORRELATION DETECTION

Making use of the representation of the data set exposed in Section II we formulate the hypothesis test in the $\mathbb{R}^{(n-1)}$ space associated with the U-scores. From now on we denote $\mathbb{U} = \big(\boldsymbol{U}^{(1)}, \cdots, \boldsymbol{U}^{(p)}\big)$ the vector of random points belonging to $\mathcal{S}_{n-2}$. The proposed hypotheses to be tested are

$$\begin{cases} \mathcal{H}_0\ :\ f_{\mathbb{U}}(\boldsymbol{u}^{(1)}, \cdots, \boldsymbol{u}^{(p)}) = \mathcal{U}(\mathcal{S}_{n-2}) \\ \mathcal{H}_1\ :\ f_{\mathbb{U}}(\boldsymbol{u}^{(1)}, \cdots, \boldsymbol{u}^{(p)}) \neq \mathcal{U}(\mathcal{S}_{n-2}) \end{cases}, \tag{4}$$

where $\mathcal{U}(\mathcal{S}_{n-2})$ is the uniform distribution on $\mathcal{S}_{n-2}$. As a consequence of the properties of the elliptically-contoured density $f_{\boldsymbol{X}}(\boldsymbol{x})$, $\mathcal{H}_0$ is verified only if $\boldsymbol{\Sigma}$ is diagonal. The testing of the hypotheses (4) is difficult, since the observed matrix $\mathbb{U}$ represents a single drawing from $f_{\mathbb{U}}(\boldsymbol{u}^{(1)}, \cdots, \boldsymbol{u}^{(p)})$. We propose a simple testing strategy, based on the following observation: since the test aims to *detect* the presence of correlated variables in the vector $\boldsymbol{X}$, but not to *identify* the correlated pairs, the information about the identities (*i.e.* the labeling) of the random variables is redundant, and can be discarded. This conceptual operation is formalized as the resampling (without replacement) of the data set $\mathbb{U}$. The resampled data set $\mathbb{V}$ is defined by $\mathbb{V} = \mathbb{U}\boldsymbol{P}$, where the matrix $\boldsymbol{P}$ is random and uniformly distributed over the set $\mathcal{P} \subset \mathbb{B}^{p\times p}$ of permutation matrices. Therefore, for $p = 2$ the density of $\mathbb{V}$ for the arguments $(\boldsymbol{v}, \boldsymbol{w})$ is equal to

$$f_{\boldsymbol{V}^{(1)}\boldsymbol{V}^{(2)}}(\boldsymbol{v}, \boldsymbol{w}) = \frac{1}{2} f_{\boldsymbol{U}^{(1)}\boldsymbol{U}^{(2)}}(\boldsymbol{v}, \boldsymbol{w}) + \frac{1}{2} f_{\boldsymbol{U}^{(2)}\boldsymbol{U}^{(1)}}(\boldsymbol{v}, \boldsymbol{w}),$$

where $1/2$ is the probability of choosing any of the two permutation matrices. Extending this argument to general values of p yields the following proposition.

Proposition 3. *The probability density function of the resampled data set $f_{\mathbb{V}}(\boldsymbol{v}^{(1)}, \cdots, \boldsymbol{v}^{(p)})$ has form*

$$f_{\mathbb{V}}(\boldsymbol{v}^{(1)}, \cdots, \boldsymbol{v}^{(p)}) = \frac{1}{p!} \sum_{a=1}^{p!} f_{\mathbb{U}\boldsymbol{P}|\boldsymbol{P}=\boldsymbol{P}_a}\big(\boldsymbol{v}^{(1)}, \cdots, \boldsymbol{v}^{(p)}\big). \tag{5}$$

As it is evident from (5), the distribution $f_{\mathbb{V}}(\boldsymbol{v}^{(1)}, \cdots, \boldsymbol{v}^{(p)})$ is invariant upon permutation of the arguments. The resampled data set $\mathbb{V}$ is hence an exchangeable sequence [11] of random points in the $\mathbb{R}^{n-1}$ space. This implies, in particular, that each column in $\mathbb{V}$ is distributed according to the same marginal $f_{\boldsymbol{V}}(\boldsymbol{v})$. The covariance between any pair of columns of $\mathbb{V}$ is given by

$$\mathrm{cov}\,(\boldsymbol{V}, \boldsymbol{W}) = \sum_{(ij)\in\mathcal{Q}} \frac{\boldsymbol{\Upsilon}^{(ij)}}{p(p-1)} = \mathbf{I} \sum_{(ij)\in\mathcal{Q}} \frac{\big(\alpha^{(ij)} - \beta^{(ij)}\big)}{p(p-1)}. \tag{6}$$

The testing procedure will be performed on $\mathbb{V}$. It is straightforward to verify that, when $\mathcal{H}_0$ is in force, $f_{\mathbb{V}}(\boldsymbol{v}^{(1)}, \cdots, \boldsymbol{v}^{(p)}) = f_{\mathbb{U}}(\boldsymbol{v}^{(1)}, \cdots, \boldsymbol{v}^{(p)})$. The hypotesis testing problem on $f_{\mathbb{V}}(\boldsymbol{v}^{(1)}, \cdots, \boldsymbol{v}^{(p)})$ can now be solved easily, as shown below. Since exchangeability allows to consider each column of $\mathbb{V}$ as a drawing from $f_{\boldsymbol{V}}(\boldsymbol{v})$, the high dimensionality p of the problem can be exploited to test, without requiring large sample size n.

A. The empirical squared distance test

Let $\mathcal{Q}^* = \{(i,j)\ :\ i,j \in \{1, \cdots, p\}, i < j\}$. The vector $\boldsymbol{D}(\mathbb{V}) = (D_1, \cdots, D_{p(p-1)/2})$ is constructed by stacking in sequence, $\forall (i,j) \in \mathcal{Q}^*$, the pairwise squared Euclidean distances $D_k = \parallel \boldsymbol{V}^{(i)} - \boldsymbol{V}^{(j)} \parallel^2$. As consequence of the exchangeability of $\mathbb{V}$, the distribution $f_{\boldsymbol{D}}(d_1, \cdots, d_{p(p-1)/2})$ is invariant with respect to the permutation of the arguments, and $\boldsymbol{D}$ is exchangeable as well. The explicit expression of $\mathbb{E}[D]$, *i.e.* the expectation of the marginal distribution of D, is given in the following proposition.

Proposition 4. *The expectation of the random variable D is given by*

$$\mathbb{E}[D] = 2\big(1 - \mathrm{tr}\big(\mathrm{cov}\,(\boldsymbol{V}, \boldsymbol{W})\big) - \parallel \mathbb{E}\,[\boldsymbol{V}] \parallel^2\big) = 2(1 - \gamma), \tag{7}$$

where $\gamma = \mathrm{tr}\big(\mathrm{cov}\,(\boldsymbol{V}, \boldsymbol{W})\big) + \|\mathbb{E}[\boldsymbol{V}]\|^2$.

Proof: Because of the exchangeability of the elements in $\mathbb{V}$, $\mathbb{E}\left[D\right] = \mathbb{E}\big[\|\boldsymbol{V} - \boldsymbol{W}\|^2\big]$. Consider the following relation

$$\mathrm{tr}\big(\mathrm{cov}\left(\boldsymbol{V}, \boldsymbol{W}\right)\big) = \mathbb{E}\big[\boldsymbol{V}^T\boldsymbol{W}\big] - \|\,\mathbb{E}[\boldsymbol{V}]\,\|^2 \tag{8}$$

$$= 1 - \frac{1}{2}\,\mathbb{E}\big[\|\,\boldsymbol{V} - \boldsymbol{W}\,\|^2\big] - \|\,\mathbb{E}[\boldsymbol{V}]\,\|^2, \tag{9}$$

where (8) is a consequence of the linearity of the trace and of the expectation, and (9) is due to the fact that $\|\boldsymbol{V}\|^2 = \|\boldsymbol{W}\|^2 = 1$. Rearranging (9) yields (7). ■

The value $\mathbb{E}\left[D\right]$ belongs to the interval $[0, 2]$. When $\mathcal{H}_0$ is true, $\|\,\mathbb{E}[\boldsymbol{V}]\,\|^2 = 0$, and $\boldsymbol{\Upsilon}^{(ij)} = 0\,\mathbf{I}$, $\forall (i,j) \in \mathcal{Q}$, so that, after (6), $\mathrm{tr}\big(\mathrm{cov}\left(\boldsymbol{V}, \boldsymbol{W}\right)\big) = 0$. As a consequence $\mathbb{E}\big[D\big|\mathcal{H}_0\big] = 2$, *i.e.* the expectation $\mathbb{E}\left[D\right]$ is maximized when $\mathcal{H}_0$ is true. This motivates using the average of the vector $\boldsymbol{D}(\mathbb{V})$ as the test statistic T

$$T = \frac{2}{p(p-1)} \sum_{k=1}^{p(p-1)/2} D_k. \tag{10}$$

By thresholding T we obtain the hypothesis test

$$\begin{cases} \big|T - 2\big| \leq \tau & : \quad \mathcal{H}_0 \\ \big|T - 2\big| > \tau & : \quad \mathcal{H}_1 \end{cases}. \tag{11}$$

The threshold value τ will satisfy $\tau \geq 2/(p-1)$.

The performance of the test is characterized evaluating the Type I and Type II error probabilities.

Proposition 5. *The Type I error probability relative to the threshold test* (11) *is bounded by*

$$\mathbb{P}_{\text{I}} \;\leq\; 2\; e^{-\frac{1}{4}\;p(p-1)\;\tau^2}. \tag{12}$$

The Type II error probability is bounded by

$$\mathbb{P}_{\text{II}} \leq \begin{cases} e^{-\frac{1}{4}p(p-1)(\gamma-\tau)^2} - e^{-\frac{1}{4}p(p-1)(\gamma+\tau)^2}, & 2\gamma \geq \tau \\ 1 - e^{-\frac{1}{4}p(p-1)(\gamma+\tau)^2} - e^{-\frac{1}{4}p(p-1)(\gamma-\tau)^2}, & 2\gamma < \tau \end{cases}. \tag{13}$$

Proof: The proof relies on Lemma 8 and Lemma 9 in Appendix A. Lemma 8 proves that the variables in $\boldsymbol{D}$ have negative association [12]. This property is intuitively clear: since the surface $\mathcal{S}_{n-2}$ is bounded, the increase of the distance of one point on $\mathcal{S}_{n-2}$ with another induces a decrease in distance with respect to a third point. The important consequence of negative association of the variables in $\boldsymbol{D}$ is that it permits the application of Chernoff-type large deviation bounds, proven in Lemma 9.

The Type I error probability $\mathbb{P}_{\text{I}}$ is, by definition,

$$\mathbb{P}_{\text{I}} = \mathbb{P}\left(\big|T - 2\big| > \tau \;\middle|\; \mathcal{H}_0\right) = \mathbb{P}\left(\big|T - \mathbb{E}\left[D\right]\big| > \tau\right), \tag{14}$$

where the last equality follows from $\mathbb{E}\left[D|\mathcal{H}_0\right] = 2$. Using (10) and (26) with (14) yields (12). The Type II probability of error $\mathbb{P}_{\text{II}}$ is given by

$$\mathbb{P}_{\text{II}} = 1 - \mathbb{P}\big(|T - 2| \geq \tau \mid \mathcal{H}_1\big) \tag{15}$$

$$= 1 - \mathbb{P}(T - \mathbb{E}\left[D\right] \geq 2\gamma + \tau) - \mathbb{P}(T - \mathbb{E}\left[D\right] \leq 2\gamma - \tau) \tag{16}$$

$$= 1 - e^{-\frac{1}{4}p(p-1)(2\gamma+\tau)^2} - \mathbb{P}(T - \mathbb{E}\left[D\right] \leq 2\gamma - \tau), \tag{17}$$

where (16) is deduced using the relation $2 = \mathbb{E}\left[D\right] + 2\gamma$ introduced in (7). Lemma 7 in Appendix A allows to establish the relation $2\gamma \geq -2/(p-1)$. As a consequence of the threshold choice $\tau \geq 2/(p-1)$ it is concluded $2\gamma + \tau \geq 0$. This justifies (17), derived making use of the upper tail bound (30) in Lemma 9 in Appendix A.

For $2\gamma < \tau$, the last term in (17) is evaluated using the lower tail bound (31) derived in Lemma 9, and this yields the second part of (13). For $2\gamma \geq \tau$, (17) can be developed as

$$\mathbb{P}_{\text{II}} = 1 - e^{-\frac{1}{4}p(p-1)(2\gamma+\tau)^2} - 1 + \mathbb{P}(T - \mathbb{E}\left[D\right] > 2\gamma - \tau),$$

from which the first part of (13) is finally obtained by substitution of (31). ■

As it is clear from (12) and (13), for $2\gamma \geq \tau$ the test is characterized by a fast decrease of the error probability as the dimension p increases, forcing it to zero for $p \to \infty$. The constant γ is null when $\mathcal{H}_0$ is in force, and, for high dimensional problems (*i.e.* when the modelization $p \to \infty$ is allowed), is always positive (Lemma 7 in Appendix A). It increases whenever there is correlation among the elements of $\mathbb{V}$, and/or the marginal $f_{\boldsymbol{V}}(\boldsymbol{v}^{(1)}, \cdots, \boldsymbol{v}^{(p)})$ over $\mathcal{S}_{n-2}$ deviates from symmetry about the origin of $\mathbb{R}^{n-1}$. Thus γ can be understood as a measure of divergence of the empirical $f_{\mathbb{V}}(\boldsymbol{v}^{(1)}, \cdots, \boldsymbol{v}^{(p)})$ from the uniform distribution on $\mathcal{S}_{n-2}$, as a function of its first and second order statistics. If $2\gamma < \tau$, this will almost surely induce a Type II error, for $p \to \infty$.

B. The empirical entropy test

As discussed above, testing the average of the squared distance between the columns of $\mathbb{V}$ allows to detect the deviation, expressed by γ, of the empirical distribution from the uniform distribution on the sphere. For some *a priori* densities $f_{\boldsymbol{X}}(\boldsymbol{x})$, however, the covariance term $\mathrm{tr}\big(\mathrm{cov}\left(\boldsymbol{V}, \boldsymbol{W}\right)\big)$ contributing to γ may be small. This happens, for example, when the random vector $\boldsymbol{X}$ is composed of elements that are both positively and negatively correlated[2], or for the sparse correlation regime, when only $\kappa \ll p$ elements are correlated. Under these circumstances the test on the empirical squared distances will have reduced power of rejecting the null hypothesis for symmetric, but not uniform, marginal distributions. This section outlines an alternative test, based on the method of types [13]. As it will be shown, the Type I error exponent increases less rapidly in p, but allows a Type II error exponent that is better behaved for symmetric marginals under the alternative hypothesis.

Define the quantizer $\mathbb{Q} : \mathbb{R}^{n-1} \to \{1, \cdots, m\}$, given by a tessellation of $\mathcal{S}_{n-2}$ in m Voronoi cells of equal volume. The (column by column) quantization $\mathbb{Q}(\mathbb{V})$ produces a p-dimensional vector $\boldsymbol{\nu}$ of quantization indexes. Counting how many instances of each quantization index appear in $\boldsymbol{\nu}$, and normalizing for $1/p$, gives the m-dimensional vector $\boldsymbol{\mu}$, describing a probability mass function on the support $\{1, \cdots, m\}$. Under the high dimensionality assumption $p \to \infty$, by effect of the law of large numbers, the empirical

[2] In (6) $\mathrm{cov}\left(\boldsymbol{V}, \boldsymbol{W}\right)$ is defined as the average of the covariances between the U-scores, which may cancel each other out. It can in fact be proven that the sign of the term $(\alpha^{(ij)} - \beta^{(ij)})$ in (3) is equal to the sign of $\mathrm{cov}\left(X_i, X_j\right)$.

distribution $\boldsymbol{\mu}$ (the type) almost surely converges to the *a priori* distribution $\overline{\boldsymbol{\mu}}$ of $\mathbb{Q}(\mathbb{V})$, where $\overline{\mu}_k = \int_{\mathcal{V}_k} f_{\boldsymbol{V}}(\boldsymbol{v})\ \mathrm{d}\boldsymbol{v}, \quad \forall k \in \{1, \cdots, m\}$. The statistic chosen to perform the hypothesis test is the entropy of the empirical distribution $H(\boldsymbol{\mu})$. Observe that, when hypothesis $\mathcal{H}_0$ is true, $H(\boldsymbol{\mu} \mid \mathcal{H}_0) \xrightarrow{\text{a.s.}} \boldsymbol{H}(\overline{\boldsymbol{\mu}}_0) = \log(m)$. Here $\overline{\boldsymbol{\mu}}_0$ is the uniform probability mass function over the index set $\{1, \cdots, m\}$. The thresholding test takes the form

$$\begin{cases} H(\boldsymbol{\mu}) \geq \tau : & \mathcal{H}_0 \\ H(\boldsymbol{\mu}) < \tau : & \mathcal{H}_1 \end{cases}. \tag{18}$$

The performance of the test is determined by the Type I and Type II probabilities of error.

Proposition 6. *The Type I probability of error for the threshold test* (18) *is bounded by*

$$\mathbb{P}_{\text{I}} \leq e^{-p\ (\log m - \tau)}, \tag{19}$$

while the Type II probability of error is bounded by

$$\mathbb{P}_{\text{II}} \leq e^{-pD(\boldsymbol{\mu}^* || \overline{\boldsymbol{\mu}})}, \tag{20}$$

where $\boldsymbol{\mu}^*$ *is defined as* $\boldsymbol{\mu}^* = \arg\min_{\boldsymbol{\mu} \in F} D(\boldsymbol{\mu} || \overline{\boldsymbol{\mu}})$, *given* $F = \{\boldsymbol{\mu} : H(\boldsymbol{\mu}) \geq \tau\} = \{\boldsymbol{\mu} : D(\boldsymbol{\mu} || \overline{\boldsymbol{\mu}}_0) \leq \log m - \tau\}$.

Proof: Consider the set of probability mass functions defined on the support $\{1, \cdots, m\}$, and define its partition (F, F^c). We establish convexity of F and F^c. By convexity of the Kullback-Lieber divergence [14, Thm. 2.7.2], for $\boldsymbol{\mu}_3 = \alpha \boldsymbol{\mu}_1 + (1 - \alpha) \boldsymbol{\mu}_2$, with $\boldsymbol{\mu}_1, \boldsymbol{\mu}_2 \in F$ and $\alpha \in [0, 1]$

$$D(\boldsymbol{\mu}_3 || \boldsymbol{\mu}_0) \leq \alpha D(\boldsymbol{\mu}_1 || \boldsymbol{\mu}_0) + (1 - \alpha) D(\boldsymbol{\mu}_2 || \boldsymbol{\mu}_0) \leq \log m - \tau$$

so that $\boldsymbol{\mu}_3 \in F$. Similarly, by concavity of the entropy [14, Thm. 2.7.3], for $\boldsymbol{\mu}_1, \boldsymbol{\mu}_2 \in F^c$

$$H(\boldsymbol{\mu}_3) \geq \alpha H(\boldsymbol{\mu}_1) + (1 - \alpha) H(\boldsymbol{\mu}_2) \geq \tau,$$

so that $\boldsymbol{\mu}_3 \in F^c$ is established.

Finally the Type I probability of error (19) can be evaluated using Sanov's theorem [14, Thm. 12.4.1] as

$$\mathbb{P}_{\text{I}} = \mathbb{P}(F^c \mid \mathcal{H}_0) \leq e^{-p \min_{\boldsymbol{\mu} \in F^c} D(\boldsymbol{\mu} || \overline{\boldsymbol{\mu}}_0)} = e^{-p\ (\log m - \tau)}.$$

Similarly, the Type II probability of error (20) is given by

$$\mathbb{P}_{\text{II}} = \mathbb{P}(F \mid \mathcal{H}_1) \leq e^{-p \min_{\boldsymbol{\mu} \in F} D(\boldsymbol{\mu} || \overline{\boldsymbol{\mu}})} = e^{-pD(\boldsymbol{\mu}^* || \overline{\boldsymbol{\mu}})},$$

with $\boldsymbol{\mu}^*$ as defined above. ■

As pointed out above, the test will almost surely induce a Type II error whenever the empirical distribution $\boldsymbol{\mu}$ falls in F. The Type II probability of error can be decreased by increasing the cardinality m of the quantizer.

IV. Experimental results

This section presents an experimental assessment of the performance of the empirical squared distance test for finite dimension. The ROC curves are obtained generating 10^4 data matrices, each given by n i.i.d. drawings from a Gaussian p-variate distribution with random covariance matrix and mean vector. In Figure 1 the performance of the empirical squared distance (SqDist) test proposed in Section III-A is compared with the ROC curves for the maximum Pearson's correlation coefficient (MaxCor) and maximum Spearman's rank correlation coefficient (MaxRCor) tests [7], and for the S biggest correlation coefficients test (MaxCorS) [8], where $S = 5$. All the tests have comparable complexity.

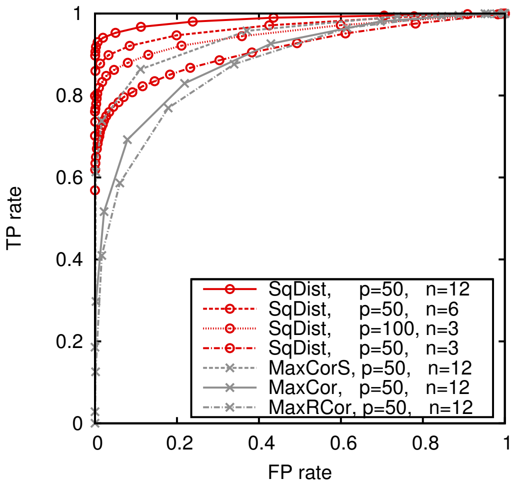

Figure 1. ROC curves for the proposed empirical squared distance(SqDist) test, compared to the maximum correlation (MaxCor), maximum ranked correlation (MaxRCor), and S-biggest correlation (MaxCorS) tests.

For $p = 50$ and $n = 12$ the SqDist test outperforms the alternatives. Figure 1 presents also the SqDist ROC curves for $n = 6$ and $n = 3$: it can be observed that the SqDist test ($n = 3$) achieves a performance comparable to the MaxCor and MaxCorS tests ($n = 12$), but requiring only $1/4$ the number of samples. For comparison, the SqDist curve for $p = 100$ and $n = 3$ is depicted as well, showing that, as expected, increasing p for fixed n improves the performance.

V. Conclusion

In this work two tests for correlation detection in large data sets of elliptically-contoured variables are presented. Their performance is characterized for finite values of n, using Chernoff and Sanov bounds. The properties of the U-scores allow to take advantage of the high dimension of the problem without requiring n to go to infinity. Using vector quantization to discretize the empirical distribution of the U-scores leads to a simple test statistic whose Type I and Type II error exponents can be computed using the method of types.

References

[1] A. O. Hero and B. Rajaratnam, "Large scale correlation screening," *J. Amer. Statistical Assoc.*, vol. 106, no. 496, pp. 1540–1552, Dec 2011.

[2] T. T. Cai and T. Jiang, "Limiting laws of coherence of random matrices with applications to testing covariance structure and construction of compressed sensing matrices," *Ann. Stat.*, vol. 39, pp. 1496–1525, 2011.

[3] I. M. Johnstone, "On the distribution of the largest eigenvalue in principal component analysis," *Ann. Stat.*, vol. 29, pp. 295–327, 2001.

[4] O. Ledoit and M. Wolf, "Some hypothesis tests for the covariance matrix when the dimension is large compared to the sample size," *Ann. Stat.*, vol. 30, no. 4, pp. 1081–1102, 2002.

[5] T. Jiang, "The asymptotic distributions of the entries of sample correlation matrices," *Ann. Appl. Prob.*, vol. 14, no. 2, pp. 865–880, 2004.

[6] W. Zhou, "Asymptotic distribution of the largest off-diagonal entry of correlation matrices," *Trans. Am. Math. Soc.*, vol. 359, no. 11, pp. 5345–5363, 2007.

[7] G. K. Eagleson, “A robust test for multiple comparisons of correlation coefficients,” *Aust. J. Stat.*, vol. 25, no. 2, pp. 256–263, 1983.

[8] M. A. Cameron and G. K. Eagleson, “A new procedure for assessing large sets of correlations,” *Aust. J. Stat.*, vol. 27, no. 1, pp. 84–95, 1985.

[9] A. O. Hero and B. Rajaratnam, “Hub discovery in partial correlation graphs,” *IEEE Trans. Inf. Th.*, 2012, to appear.

[10] T. W. Anderson, *An introduction to multivariate statistical analysis*. Wiley, 2007.

[11] J. F. C. Kingman, “Uses of exchangeability,” *Ann. Prob.*, vol. 6, no. 2, pp. 183–197, 1978.

[12] K. Joag-Dev and F. Proschan, “Negative association of random variables, with applications,” *Ann. Stat.*, vol. 11, no. 1, pp. 286–295, 1983.

[13] I. Csiszar, “The method of types,” *IEEE Trans. Inf. Th.*, vol. 44, no. 6, pp. 2505–2523, 1998.

[14] T. Cover and J. Thomas, *Elements of information theory*. Wiley, 2006.

[15] A. Panconesi and A. Srinivasan, “Randomized distributed edge coloring via an extension of the Chernoff–Hoeffding bounds,” *SIAM J. on Computing*, vol. 26, no. 2, pp. 350–368, 1997.

APPENDIX

Lemma 7. *The trace of the covariance* $\mathrm{cov}\,(\boldsymbol{V}, \boldsymbol{W})$ *is lower bounded by* $-1/(p-1) \leq \mathrm{tr}\big(\mathrm{cov}\,(\boldsymbol{V}, \boldsymbol{W})\,\big)$.

Proof: In order to derive the lower bound, some preliminary results are needed. Define $\overline{\boldsymbol{U}} = \sum_{i=1}^{p} \boldsymbol{U}^{(i)}$. Positiveness of the quantity $\mathrm{tr}\big(\mathrm{cov}\,(\overline{\boldsymbol{U}}, \overline{\boldsymbol{U}})\,\big)$ is established by

$$\mathrm{tr}\big(\mathrm{cov}\,(\overline{\boldsymbol{U}}, \overline{\boldsymbol{U}})\,\big) = \mathbb{E}\big[\mathrm{tr}\big(\overline{\boldsymbol{U}}^T \overline{\boldsymbol{U}}\big)\big] - \mathrm{tr}\big(\mathbb{E}[\overline{\boldsymbol{U}}]^T \mathbb{E}[\overline{\boldsymbol{U}}]\big) \quad (21)$$

$$= \mathbb{E}\big[\parallel \overline{\boldsymbol{U}} \parallel^2\big] - \parallel \mathbb{E}[\overline{\boldsymbol{U}}] \parallel^2 \geq 0, \quad (22)$$

where (21) is obtained by linearity of the trace and of the expectation. The inequality in (22) follows applying Jensen's inequality to the convex function $\parallel \cdot \parallel^2$. Now consider the following expression:

$$\mathrm{tr}\big(\mathrm{cov}\,(\overline{\boldsymbol{U}}, \overline{\boldsymbol{U}})\,\big) = \mathrm{tr}\Big(\textstyle\sum_{i=1}^{p} \boldsymbol{\Upsilon}^{(ii)} + \sum_{(ij)\in\mathcal{Q}} \boldsymbol{\Upsilon}^{(ij)}\Big)$$

$$= \textstyle\sum_i (1 - \parallel \mathbb{E}[\boldsymbol{U}^{(i)}] \parallel^2) + p(p-1)\mathrm{tr}\big(\mathrm{cov}\,(\boldsymbol{V}, \boldsymbol{W})\,\big). \quad (23)$$

The first term in (23) is obtained by linearity of the trace and of the expectation, and using the relation $\mathbb{E}\left[\parallel \boldsymbol{U}^{(i)} \parallel^2\right] = 1$; the second term is obtained applying (6). Rearranging (23) yields

$$\frac{\mathrm{tr}\big(\mathrm{cov}\,(\overline{\boldsymbol{U}}, \overline{\boldsymbol{U}})\big) + \sum \parallel \mathbb{E}[\boldsymbol{U}^{(i)}] \parallel^2}{p} = 1 + (p-1)\mathrm{tr}\big(\mathrm{cov}\,(\boldsymbol{V}, \boldsymbol{W})\big).$$

Noticing that the first term is positive, because of (22), yields the lower bound. ■

Lemma 8. *The elements of the random vector* $\boldsymbol{D}$ *are negatively associated,* i.e. *for every pair* $\mathcal{A}_1, \mathcal{A}_2$ *of disjoint subsets of* $\{1, 2, \cdots, p(p-1)/2\}$*, and for any pair of non-decreasing functions* $f_1(\cdot), f_2(\cdot)$*, the following holds:*

$$\mathrm{cov}\,(f_1(D_i, i \in \mathcal{A}_1), f_2(D_j, j \in \mathcal{A}_2)) \leq 0. \quad (24)$$

Proof: The proof is obtained via a slight modification of the proof of [12, Thm. 2.11]. Let $\boldsymbol{D}_1, \boldsymbol{D}_2$ denote an arbitrary partition of the vector $\boldsymbol{D}$, and let $f_1(\cdot), f_2(\cdot)$ denote a pair of non-decreasing, permutation invariant functions. Using [12, (1.1)] it is possible to write

$$\begin{aligned}\mathrm{cov}\,(f_1(\boldsymbol{D}_1), f_2(\boldsymbol{D}_2)) &= \mathbb{E}\left[\mathrm{cov}\,(f_1(\boldsymbol{D}_1), f_2(\boldsymbol{D}_2)|\mathbb{U}, I)\right] \\ &\quad + \mathrm{cov}\,(\mathbb{E}\left[f_1(\boldsymbol{D}_1)|\mathbb{U}, I\right], \mathbb{E}\left[f_2(\boldsymbol{D}_2)|\mathbb{U}, I\right]),\end{aligned} \quad (25)$$

where I is a random variable identifying the minimum valued component in the vector $\boldsymbol{D}$. The distribution $f_{\boldsymbol{D}|\mathbb{U},I}$ is a permutation distribution in the sense defined in [12, Def. 2.10], and hence is negatively associated [12, Thm. 2.11]. This implies that the first term in the right hand of (25) is negative. The negativity of the second term follows by the same argument in the proof to [12, Thm. 2.11]. ■

Lemma 9. *The average of the random vector* $\boldsymbol{D}$ *obeys Chernoff-type large deviation bounds. In particular, for* $\epsilon > 0$

$$\mathbb{P}\left(\left|\frac{2\sum_{i=1}^{p(p-1)/2} D_i}{p(p-1)} - \mathbb{E}\left[D\right]\right| > \epsilon\right) \leq 2\; e^{-\frac{1}{4}\; p(p-1)\; \epsilon^2}. \quad (26)$$

Proof: The proof relies on [15, Thm. 3.2]. Recall that the elements of $\boldsymbol{D}$ are positive, bounded in the interval $[0, 2]$. In order to apply [15, Thm. 3.2] we need to prove that the elements in $\boldsymbol{D}$ are λ-correlated, as defined in [15, Def. 3.1]. Define a vector $\boldsymbol{X}$ of $p(p-1)/2$ independent random variables on the support $[0, 2]$, and such that $\mathbb{E}\left[X_i\right] = \mathbb{E}\left[D\right],\ \forall i$. This implies $\sum_{i=1}^{p(p-1)/2} \mathbb{E}\big[D_i\big] = \sum_{i=1}^{p(p-1)/2} \mathbb{E}\big[X_i\big]$. Using linearity of the expectation it is easy to see that condition (i) in [15, Def. 3.1] is satisfied. Now consider a non-negative function $f(\cdot)$. Since the variables in $\boldsymbol{D}$ are negatively associated, invoking [12, Property 2] gives

$$\mathbb{E}\left[\prod_{i=1}^{p(p-1)/2} f(D_i)\right] \leq \prod_{i=1}^{p(p-1)/2} \mathbb{E}\big[f(D_i)\big] = \prod_{i=1}^{p(p-1)/2} \mathbb{E}\big[f(X_i)\big], \quad (27)$$

where the last equality is obtained because $\mathbb{E}\left[D_i\right] = \mathbb{E}\left[X_i\right]$. Inspection of (27) confirms that condition (ii) in [15, Def. 3.1] is verified as well for $\lambda = 1$.

Now that the elements of $\boldsymbol{D}$ have been established to be λ-correlated, [15, Thm. 3.2] can be used to evaluate the upper tail bound as follows:

$$\mathbb{P}\Big(\sum_i D_i > (1+\varepsilon)\sum_i \mathbb{E}\left[D\right]\Big) \leq e^{-\frac{1}{p(p-1)}(\sum_i \mathbb{E}[D])^2 \varepsilon^2} \quad (28)$$

$$\mathbb{P}\left(\frac{2\sum_i D_i}{p(p-1)} - \mathbb{E}\left[D\right] > \varepsilon\; \mathbb{E}\left[D\right]\right) \leq e^{-\frac{1}{4}\; p(p-1)\; \mathbb{E}[D]^2 \varepsilon^2} \quad (29)$$

$$\mathbb{P}\left(\frac{2\sum_i D_i}{p(p-1)} - \mathbb{E}\left[D\right] > \epsilon\right) \leq e^{-\frac{1}{4}\; p(p-1)\; \epsilon^2}, \quad (30)$$

where (28) follows directly from [15, Thm. 3.2], obtained for $\mathbb{E}\left[X_i\right] = \mathbb{E}\left[D_i\right] = \mathbb{E}\left[D\right]$, $\lambda = 1$, $D_i \in [0, 2]$. Algebraic manipulation yields (29), and the substitution $\epsilon = \varepsilon\; \mathbb{E}\left[D\right]$ yields (30).

The lower tail bound is obtained as follows. Define the random variables $C_i = \mathbb{E}\left[D\right] + 1 - D_i$ and $Y_i = \mathbb{E}\left[X\right] + 1 - X_i$. It is straightforward to verify that the elements of the vector $\boldsymbol{C}$ are λ-correlated for $\lambda = 1$. Hence, apply [15, Thm. 3.2] to obtain, in a similar manner to (29),

$$\mathbb{P}\left(\frac{2\sum_i C_i}{p(p-1)} - \mathbb{E}\left[C\right] > \epsilon\; \mathbb{E}\left[C\right]\right) \leq e^{-\frac{1}{4}\; p(p-1)\; \mathbb{E}[C]^2 \epsilon^2}. \quad (31)$$

Substitution of the expression C_i in (31) shows that the lower tail bound is equal to the upper tail bound (30). This establishes (26). ■