A null model for Pearson coexpression networks

Andrea Gobbi\(^1,2\) and Giuseppe Jurman\(^*\)\(^1\)

\(^1\) Fondazione Bruno Kessler, Trento, Italy
\(^2\) University of Trento, Italy
\(\{\text{agobbi, jurman}\}@fbk.eu\)

Abstract

Gene coexpression networks inferred by correlation from high-throughput profiling such as microarray data represent a simple but effective technique for discovering and interpreting linear gene relationships. In the last years several approaches have been proposed to tackle the problem of deciding when the resulting correlation values are statistically significant. This is mostly crucial when the number of samples is small, yielding a non negligible chance that even high correlation values are due to random effects. Here we introduce a novel hard thresholding solution based on the assumption that a coexpression network inferred by randomly generated data is expected to be empty. The theoretical derivation of the new bound by geometrical methods is shown together with applications in onco- and neurogenomics.

1 Introduction

Universally acknowledged by the scientific community as the basic task of the systems biology, the network inference is the prototypal procedure for moving from the classical reductionist approach to the novel paradigm of data-driven complex systems in the interpretation of biological processes [1]. The core of all the network inference (or network reconstruction) procedures is the detection of the topology of a graph, \textit{i.e.}, its wiring diagram, whose nodes are a given set of biological entities, starting from measurements of the entities themselves. In the last 15 years, the reconstruction of the regulation mechanism of a gene network and of the interactions among proteins from high-throughput data such as expression microarray of, more recently, from Next Generation Sequencing data has become a major line of research for laboratories worldwide. The proposed solutions rely on techniques ranging from deterministic to stochastic, and their number is constantly growing in the literature. Nonetheless, network inference is still considered an open, unsolved problem [2]. In fact, in many practical cases, the performances of the reconstruction algorithms are poor, due to several factors limiting the inference accuracy [3, 4] to the point of making it no better than coin tossing in some situations [5]. The major problem is the underdeterminacy of the task [6], due to the overwhelming number of interactions to predict starting from a usually small number of available measurements. In general, size and quality of available data are critical factors for all inference algorithms.

In what follows the impact of data size is discussed for one of the simplest inference techniques, \textit{i.e.}, the gene coexpression network, where interaction strength between two genes is a function of the correlation between the corresponding expression levels across the available tissue samples. The biological underlying hypothesis is that functionally related genes have similar expression patterns [7], and thus that coexpression is correlated with functional relationships, although this does not imply causality. In particular, as highlighted in [8], correlation can help unveiling the underlying cellular processes, since coordinated coexpression of genes encode interacting proteins, and Pearson correlation coefficient can be used as the standard measure. However, as noted in [9], correlation between genes may sometimes be due to unobserved factors affecting expression levels. Coexpression analysis has been intensively used as an effective algorithm to explore the system-level functionality of genes, sometimes outperforming much more refined approaches [10, 11]. The observation that simpler approaches such as correlation can be superior even on synthetic data has been explained by some authors [12, 13] with the difficulties of complex algorithm in detecting the subtleties of the combinatorial regulation. Moreover, coexpression network can capture more important features that the conventional differential expression approach [14], and its use has been extended to other tasks, for instance the investigation of complex biological traits [15]. Finally, these network can be crucial for understanding regulatory mechanisms [16], for the development of personalized medicine [17] or, more recently, in metagenomics [18].

Despite its success, a major issue affects coexpression networks: deciding when a given correlation value between two nodes can be deemed statistically significant and thus worthwhile assigning a link connecting them. This translates mathematically into choosing (a function of) a suitable threshold, as in the case of mutual information and relevance networks [19]. As reported

\(^*\) Corresponding author
in [20], in literature statistical methods for testing the correlations are underdeveloped, and thresholding is often overlooked even in important studies [21]. The two main approaches known in literature can be classified as soft or hard thresholding. The soft thresholding is adopted in a well-known framework called Weighted Gene Coexpression Network Analysis (WGCNA) [22], recently used also for other network types [23, 24]. All genes are mutually connected, and the weight of the link is a positive power of the absolute value of the Pearson correlation, where the exponent is chosen as the best fit of the resulting network according to a scale-free model [25, 26]. This approach, without discarding any correlation, promotes high correlation values and penalizes low values. In the hard thresholding approach, instead, only correlation values larger than the threshold are taken into account, and an unweighted link is set for each of these values, so that a binary network is generated (see [27] for one of the earliest references). Clearly, an incorrectly chosen threshold value can jeopardize the discussed results with false negative links (for too strict threshold) or false positive links (for too loose threshold). Many different heuristics have been proposed for setting the threshold values, such as using the False Discovery Rate [28, 29, 30, 31], or the p-value of the correlation test [17], or employing partial correlation [32], or using rank-based techniques [33, 34, 35] or more complex randomization techniques [36]. Alternatively, correlation distribution has been studied, experimentally [37] or at level of single interaction, not as whole network [38]. However, in many studies in literature, the threshold is not chosen accordingly to a soundly bases procedure, but referring to standard choices [39, 40, 41, 42], or to heuristics not directly related to the correlation values, but rather with the resultining network topology [43, 44, 45, 46, 47, 48, 49, 50, 51]. In [52] a comparison of some coexpression thresholds is shown on a few microarray datasets.

Here we propose a new a priori and non-parametric model for the computation of an hard theshold based on the assumption that a random coexpression graph should not have any edge. The threshold is theoretically derived by means of a geometric approach based on the work of Bevington [53], and, as a deterministic independent null model, it depends only on the dimensions of the starting data matrix, with assumptions on the skewness of the data distribution compatible with the structure of gene expression levels data [54, 55]. By definition, this threshold is aimed at minimizing the possible false positive links, paying a price in terms of false negative detected edges.

To conclude with, we show four applications, in both the large and the small sample size settings. The first two are examples in a large sample size settings, with a synthetic dataset and with an ovarian epithelial carcinoma dataset on a large cohort of 285 cases [56, 57]. Two more applications in the opposite situations are demonstrated on two publicly available datasets, the former regarding a pancreatic cancer study [58] on a tiny cohort of six patients, and the latter on a Alzheimer dataset with 28 cases [59, 60, 61].

## 2 Distribution of Pearson correlation

Let $x, y \in \mathbb{R}^n$ with $n \geq 3$. The Pearson correlation coefficient $\rho$ between $x$ and $y$ is defined as:

$$
\rho(x, y) = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n} (x_i - \bar{x})^2} \sqrt{\sum_{i=1}^{n} (y_i - \bar{y})^2}} ,
$$

where $\bar{x}$ denotes the arithmetic mean $\frac{1}{n} \sum_{i=1}^{n} x_i$ of the $n$-dimensional vector $x$.

The first step towards the construction of a null model for random absolute Pearson coexpression network is the estimation, for $0 < p < 1$, of the function $F(n, p) = P(|\rho(x, y)| > p)$, where $x$ and $y$ are two independent normal vectors of length $n$. Define two new random variable $\tilde{x}$ and $\tilde{y}$ as follows:

$$
\tilde{x} = \frac{x - \bar{x}}{\sigma_x \sqrt{n-1}} , \quad \tilde{y} = \frac{y - \bar{y}}{\sigma_y \sqrt{n-1}} ,
$$

where $\sigma_x$ and $\sigma_y$ are the standard deviations of $x$ and $y$. From the definition, the following identities immediately descend:

$$
\sum_{i=1}^{n} \tilde{x}_i = 0 = \sum_{i=1}^{n} \tilde{y}_i ,
$$
$$
\sum_{i=1}^{n} \tilde{x}_i^2 = 1 = \sum_{i=1}^{n} \tilde{y}_i^2 ,
$$
$$
\rho(x, y) = \rho(\tilde{x}, \tilde{y}) = \sum_{i=1}^{n} \tilde{x}_i \tilde{y}_i .
$$

We can now state and prove two key results.

**Proposition 1.** Let $x, y, \tilde{x}, \tilde{y}$ as in Eq. 1. Then $\tilde{x}, \tilde{y} \in S_{n-1} \cap H \sim S_{n-2}$, where $H$ is the vectorial hyperplane defined as $\sum_{i=1}^{n} w_i = 0$ and $w_i$ are the coordinates of $\mathbb{R}^n$. 


Proof. Since $\|\tilde{x}\| = 1$, the following identity holds:

$$\sum_{i=1}^{n} \tilde{x}_i = \sum_{i=1}^{n} \frac{x_i - x}{\sigma \sqrt{n-1}} = \frac{1}{\sigma \sqrt{n-1}} \sum_{i=1}^{n} (x_i - x) = \frac{1}{\sigma \sqrt{n-1}} \left( \sum_{i=1}^{n} x_i - n x \right) = \frac{1}{\sigma \sqrt{n-1}} (n \tilde{x} - nx) = 0,$$

and the same holds for $\tilde{y}$, too. \qed

An example for $n = 3$ of the situation described in Prop. 1 is plotted in Fig. 2.

**Proposition 2.** Let $x, y$ as in Prop. 1 and $0 < p < 1$ be a real number. Then the function $F(n, p)$ has the following close form

$$F(n, p) = P(|\rho(x, y)| > p) = \frac{2}{\sqrt{n-1} \Gamma(n-\frac{1}{2})} \int_{0}^{\frac{\pi}{2}} \sin^{n-3}(\vartheta) d\vartheta,$$

where $\Gamma(x)$ is the gamma function $\Gamma(x) = \int_{0}^{+\infty} t^{x-1} e^{-t} dt$.

**Proof.** Using Eq. 1 and Eq. 2, we have that

$$\rho(x, y) = \rho(\tilde{x}, \tilde{y}) = \tilde{x} \tilde{y} = \cos \beta,$$

where $\beta$ is the angle between the two vectors $\tilde{x}$ and $\tilde{y}$. Eq. 4 and Prop. 1 yields that $P(|\rho(x, y)| > p)$ is the proportion between the area of the spherical cap in $n - 2$ dimensions included within an angle $\beta$ from $x$ and the whole surface of the $n - 2$-dimensional sphere [62]. A compact formula for the area $A_{n-1}^{\text{cap}}(r)$ of a $n - 2$-spherical cap is given in [63] as:

$$A_{n-1}^{\text{cap}}(r) = \frac{2\pi(n-2)/2}{\Gamma(n-2/2)} r^{n-2} \int_{0}^{\beta} \sin^{n-3}(\vartheta) d\vartheta,$$

and, since the area of the whole surface is

$$S_{n-2}(r) = \frac{2\pi(n-1)/2}{\Gamma(n-1/2)} r^{n-2},$$

the thesis follows from the setting $r = 1$. \qed

An alternative derivation of the same result can be found in [53].

In Prop. 1 the transformed vectors are assumed to be uniformly distributed on the spherical surface. This assumption holds in the case of a normal distribution, but it does not hold in general. However, in the following paragraph we show that is a good approximation, since $x$ and $y$ are independent. In fact, Prop. 3 can be generalised to other distributions [64, 65, 66, 67], when data skewness can be bounded [62].

Let $G^\delta(p, n)$ be an empirical distribution generated by $k$ couples of two vectors $x, y \in \mathbb{R}^n$ sampled according to a given distribution function $\delta$. Let then

$$E_k(F, G^\delta) = \left( \int_{0}^{1} |F(p, n) - G^\delta(p, n)|^t dp \right)^{\frac{1}{t}}$$

be the $t$-error function evaluating the difference between the theoretical distribution $F(p, n)$ and the empirical distribution $G^\delta(p, n)$. Hereafter we report the results of the simulations for $k = 50000$ and $n = 8, 20, 100$, where $\delta$ is one of the following three distribution functions:

- $U(0, 1)$, the uniform distribution in $[0, 1]$;
- $N(m, s)$, the normal distribution with mean $m$ and standard deviation $s$;
- $L(ml, sl)$, the lognormal distribution with mean-log $ml$ and standard deviation-log $sl$.

In particular, in Tab. 1 we list the values of $E_2(F, G^\delta)$ and in Fig. 1 we display the curves of the cumulative distribution functions (CDF) of $G^\delta(p, n)$ corresponding to the three functions $\delta$, separately for the different values of $n$.

Regardless of the value of $n$, the empirical distribution fits the exact formula Eq. 3 when $x$ and $y$ are uniformly sampled, while it does not fit the same equation when the two vectors come from extremely skewed distributions such as the lognormal. Note that non-Gaussian asymmetric distributions can occasionally being detected in some array studies [55]; however, techniques for reducing the skewness are routinely applied during preprocessing [54], and thus the aforementioned results can be safely used in the microarray framework.
Table 1: Error function $E_2(F, G^\delta)$, for $n = 8, 20, 100$ and different distributions $\delta$.

<table>
<thead>
<tr>
<th>$G^\delta(p, n)$</th>
<th>$x$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$G^\delta(p, 8)$</td>
<td></td>
</tr>
<tr>
<td>$U(0, 1)$</td>
<td>0.001832</td>
</tr>
<tr>
<td>$N(0, 1)$</td>
<td>0.001195</td>
</tr>
<tr>
<td>$L(2, 3)$</td>
<td>0.022961</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>$G^\delta(p, 20)$</th>
<th>$x$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$U(0, 1)$</td>
<td>0.0016851</td>
</tr>
<tr>
<td>$N(0, 1)$</td>
<td>0.0008008</td>
</tr>
<tr>
<td>$L(2, 3)$</td>
<td>0.022961</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>$G^\delta(p, 100)$</th>
<th>$x$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$U(0, 1)$</td>
<td>0.0006978</td>
</tr>
<tr>
<td>$N(0, 1)$</td>
<td>0.0009281</td>
</tr>
<tr>
<td>$L(2, 3)$</td>
<td>0.0238804</td>
</tr>
</tbody>
</table>

Finally, we conclude this paragraph deriving the mean and the variance of the function $|\rho|$. Starting from Eq. 3, the density function $f(p, n)$ can be computed as

$$f(p, n) = \frac{2}{\sqrt{\pi}} \frac{\Gamma\left(\frac{n-1}{2}\right)}{\Gamma\left(\frac{n-2}{2}\right)} (1 - p^2)^{\frac{n-4}{2}}.$$

Using the above expression for $f(p, n)$, the two moments follow straightforwardly:

$$\mathbb{E}(|\rho|, n) = \int_0^1 p f(p, n) dp = \frac{2}{\sqrt{\pi}} \frac{\Gamma\left(\frac{n-1}{2}\right)}{\Gamma\left(\frac{n-2}{2}\right)}.$$

$$\text{Var}(|\rho|, n) = \int_0^1 p^2 f(p, n) dp - \mathbb{E}^2(p, n) = \frac{1}{n-1} - \frac{4\Gamma^2\left(\frac{n-1}{2}\right)}{\pi(n-2)^2\Gamma^2\left(\frac{n-2}{2}\right)}.$$

Figure 1: CDFs relative to the different distributions $\delta = U$ and $\delta = L$ compared with the theoretical curve $F(n, p)$, for the three cases $n = 8$ (a), $n = 20$ (b) and $n = 100$ (c). In all cases, the red curve of $F(n, p)$ and the black curve for the double uniform distribution $U - U$ are almost coincident.
The proof follows immediately from the definition of $G$.

Proposition 3. The graph $G_p$ is an Erdős-Rényi model [68] with $m$ nodes and probability $p$ as in Eq. 3.

The first result characterizes the coexpression graphs in terms of null models:

Proposition 3. The graph $G_p$ is an Erdős-Rényi model [68] with $m$ nodes and probability $p$ as in Eq. 3.

The proof follows immediately from the definition of $G_p$ and Eq. 3.

Example. Consider a dataset $\mathcal{Y}$ consisting of $n = 3$ samples described by $m = 100$ genes. Then $\mathcal{Y}$ can be represented by 100 points in $[0,1]^3 \subset \mathbb{R}^3$ as shown in Fig. 2(a). The new variables $\tilde{x}$ are built through a two-stages procedure applied to each gene. First the mean is subtracted, so the transformed dataset lies on the hyperplane $H$ described in Prop. 1 as displayed in Fig. 2(b,c). Finally, each gene is normalized to unitary variance, and the resulting dataset lies on the circumference in Fig. 2(d). Using the results in the previous section, it is now possible to define, for $n$ genes measured on $m$ samples, the secure threshold $\tilde{p}$ as the minimum value of $p$ such that the corresponding random coexpression network $G_p$ is on average an empty graph, that is

$$\tilde{p} = \min_{p \in [0,1]} \left\{ \frac{F(n,p)m(m-1)}{2} < 1 \right\}.$$  \hspace{1cm} (5)

The underlying hypothesis for Eq. 5 is the assumption that in a random dataset we do not expect an kind of edge, i.e. any kind of relation. Due to its definition, the secure threshold $\tilde{p}$ is biased towards avoiding the false positive links, paying a price in terms of false negatives. In fact, all the links passing the filter are induced by correlation only due to the inference data, while all links whose correlation value can be generated either by relation between data or by random noise are discarded.

In Tab. 2 a collection of values of $\tilde{p}$ is listed for different $m$ and $n$, while in Fig. 3 the contourplot of the function $\tilde{p}(n,m)$ is shown first on a large range of values and then zooming on the small sample size area. In the Tab. 3 we show the comparison on a set of synthetic and array datasets of the secure threshold $\tilde{p}$ with another well known hard thresholding methods, the clustering coefficient-based threshold $C^*$ [48] and with the statistical thresholds based on the adjusted $p$-values of 0.01, 0.05 or 0.1. In almost all cases, the threshold $\tilde{p}$ is the strictest. As shown in the previous section, for not very skewed distribution,
the good approximation provided by the exact formula for $F(n, p)$ given in Eq. 3 guarantees the effectiveness of the secure threshold $\bar{p}$ in detecting actual links between nodes. Nonetheless, whenever a stricter threshold is needed, it is still possible to follow the construction proposed, with the following refinement. The edge-creation process in the Erdős-Rény model follows a binomial distribution, where $n$ is the number of trials and $p$ the probability associated to the success of a trial. The mean $np$ of this distribution is one of the contributing terms in the definition of secure threshold Eq. 5. To further restrict the number of falsely detected links, the variance term $(np(1-p))$ for the binomial distribution can be added to the formula through the Chebyshev’s inequality

$$P(|X - \mu| \geq k\sigma) \leq \frac{1}{k^2},$$

where $\mu$ and $\sigma$ are the mean and the standard deviation of $X$. Thus, the definition of secure threshold can be sharpened to $\tilde{p}_k$ as follows:

$$\tilde{p}_k = \min_{p \in [0,1]} \left\{ F(n, p) \frac{m(m-1)}{2} + k\sqrt{(1 - F(n, p))F(n, p) \frac{m(m-1)}{2}} < 1 \right\}.$$

For instance, the binomial distribution, for large values of $n$, can be approximated as a normal distribution for which the $95.45\%$ of the values lie in the interval $(\mu - 2\sigma, \mu + 2\sigma)$. In Tab 4 we show, for $\tilde{p}_2$, the analogous of Tab. 2 for $\bar{p}$. Finally, the Chebyshev’s inequality implies that at least the $96\%$ of the values lie in the interval $(\mu - 5\sigma, \mu + 5\sigma)$: the corresponding threshold values for $k = 5$ are listed in Tab. 5.

Table 3: Comparison of the secure threshold $\bar{p}$ with the clustering coefficient-based threshold $C^*$ [48] and the statistical thresholds based on the adjusted p-values B0.01, B0.05 or B0.1 on a collection of synthetic and array datasets.

<table>
<thead>
<tr>
<th>Dataset type</th>
<th>#samples</th>
<th>#nodes</th>
<th>$C^*$</th>
<th>B0.01</th>
<th>B0.05</th>
<th>B0.1</th>
<th>$\bar{p}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Simulated</td>
<td>50</td>
<td>1000</td>
<td>0.57</td>
<td>0.58</td>
<td>0.54</td>
<td>0.52</td>
<td>0.6152</td>
</tr>
<tr>
<td>Simulated</td>
<td>25</td>
<td>1000</td>
<td>0.69</td>
<td>0.76</td>
<td>0.72</td>
<td>0.70</td>
<td>0.7956</td>
</tr>
<tr>
<td>H-U133P</td>
<td>23</td>
<td>897</td>
<td>0.72</td>
<td>0.78</td>
<td>0.74</td>
<td>0.72</td>
<td>0.8125</td>
</tr>
<tr>
<td>H-U133P</td>
<td>10</td>
<td>897</td>
<td>0.78</td>
<td>0.96</td>
<td>0.94</td>
<td>0.93</td>
<td>0.9723</td>
</tr>
<tr>
<td>H-U133P</td>
<td>10</td>
<td>675</td>
<td>0.77</td>
<td>0.96</td>
<td>0.93</td>
<td>0.92</td>
<td>0.9681</td>
</tr>
<tr>
<td>H-U133P</td>
<td>9</td>
<td>897</td>
<td>0.79</td>
<td>0.97</td>
<td>0.96</td>
<td>0.95</td>
<td>0.9821</td>
</tr>
<tr>
<td>H-U133P</td>
<td>8</td>
<td>897</td>
<td>0.81</td>
<td>0.98</td>
<td>0.97</td>
<td>0.96</td>
<td>0.9899</td>
</tr>
<tr>
<td>H-U133P</td>
<td>7</td>
<td>897</td>
<td>0.81</td>
<td>0.99</td>
<td>0.99</td>
<td>0.98</td>
<td>0.9958</td>
</tr>
<tr>
<td>H-U133P</td>
<td>6</td>
<td>897</td>
<td>0.86</td>
<td>&gt;0.99</td>
<td>&gt;0.99</td>
<td>0.99</td>
<td>0.9987</td>
</tr>
<tr>
<td>H-U133P</td>
<td>5</td>
<td>897</td>
<td>0.92</td>
<td>&gt;0.99</td>
<td>&gt;0.99</td>
<td>&gt;0.99</td>
<td>0.9998</td>
</tr>
<tr>
<td>H-U133P</td>
<td>4</td>
<td>897</td>
<td>0.99</td>
<td>&gt;0.99</td>
<td>&gt;0.99</td>
<td>&gt;0.99</td>
<td>&gt; 0.999</td>
</tr>
<tr>
<td>H-U133A</td>
<td>4</td>
<td>675</td>
<td>0.99</td>
<td>&gt;0.99</td>
<td>&gt;0.99</td>
<td>&gt;0.99</td>
<td>&gt; 0.999</td>
</tr>
<tr>
<td>H-I6</td>
<td>4</td>
<td>675</td>
<td>0.99</td>
<td>&gt;0.99</td>
<td>&gt;0.99</td>
<td>&gt;0.99</td>
<td>&gt; 0.999</td>
</tr>
<tr>
<td>M-U74</td>
<td>4</td>
<td>401</td>
<td>0.97</td>
<td>&gt;0.99</td>
<td>&gt;0.99</td>
<td>&gt;0.99</td>
<td>0.9999</td>
</tr>
</tbody>
</table>
Table 4: A subset of values of the secure threshold $\hat{p}_2$ for different number of samples $m$ and genes $n$.

<table>
<thead>
<tr>
<th></th>
<th>100</th>
<th>500</th>
<th>1000</th>
<th>2000</th>
<th>10000</th>
<th>50000</th>
<th>100000</th>
</tr>
</thead>
<tbody>
<tr>
<td>8</td>
<td>0.97584</td>
<td>0.99179</td>
<td>0.99484</td>
<td>0.99675</td>
<td>0.99889</td>
<td>0.99962</td>
<td>0.99977</td>
</tr>
<tr>
<td>15</td>
<td>0.86282</td>
<td>0.91826</td>
<td>0.93437</td>
<td>0.94723</td>
<td>0.96810</td>
<td>0.98065</td>
<td>0.98439</td>
</tr>
<tr>
<td>20</td>
<td>0.78966</td>
<td>0.85726</td>
<td>0.87876</td>
<td>0.89686</td>
<td>0.92883</td>
<td>0.95068</td>
<td>0.95784</td>
</tr>
<tr>
<td>30</td>
<td>0.68082</td>
<td>0.75573</td>
<td>0.78151</td>
<td>0.80425</td>
<td>0.84759</td>
<td>0.88074</td>
<td>0.89256</td>
</tr>
<tr>
<td>50</td>
<td>0.55034</td>
<td>0.62269</td>
<td>0.64902</td>
<td>0.67302</td>
<td>0.72135</td>
<td>0.76137</td>
<td>0.77651</td>
</tr>
<tr>
<td>75</td>
<td>0.45887</td>
<td>0.52436</td>
<td>0.54881</td>
<td>0.57145</td>
<td>0.61820</td>
<td>0.65834</td>
<td>0.67394</td>
</tr>
<tr>
<td>100</td>
<td>0.40153</td>
<td>0.46116</td>
<td>0.48369</td>
<td>0.50471</td>
<td>0.54865</td>
<td>0.58703</td>
<td>0.60214</td>
</tr>
</tbody>
</table>

Table 5: A subset of values of the secure threshold $\hat{p}_5$ for different number of samples $m$ and genes $n$.

<table>
<thead>
<tr>
<th></th>
<th>100</th>
<th>500</th>
<th>1000</th>
<th>2000</th>
<th>10000</th>
<th>50000</th>
<th>100000</th>
</tr>
</thead>
<tbody>
<tr>
<td>8</td>
<td>0.98553</td>
<td>0.99508</td>
<td>0.99691</td>
<td>0.99805</td>
<td>0.99934</td>
<td>0.99978</td>
<td>0.99986</td>
</tr>
<tr>
<td>15</td>
<td>0.89287</td>
<td>0.93585</td>
<td>0.94842</td>
<td>0.95849</td>
<td>0.97486</td>
<td>0.98474</td>
<td>0.98768</td>
</tr>
<tr>
<td>20</td>
<td>0.82530</td>
<td>0.88080</td>
<td>0.89858</td>
<td>0.91361</td>
<td>0.94025</td>
<td>0.95853</td>
<td>0.96454</td>
</tr>
<tr>
<td>30</td>
<td>0.71934</td>
<td>0.78401</td>
<td>0.80647</td>
<td>0.82636</td>
<td>0.86445</td>
<td>0.89373</td>
<td>0.90420</td>
</tr>
<tr>
<td>50</td>
<td>0.58686</td>
<td>0.65162</td>
<td>0.67541</td>
<td>0.69720</td>
<td>0.74130</td>
<td>0.77803</td>
<td>0.79198</td>
</tr>
<tr>
<td>75</td>
<td>0.49164</td>
<td>0.55125</td>
<td>0.57373</td>
<td>0.59463</td>
<td>0.63803</td>
<td>0.67552</td>
<td>0.69015</td>
</tr>
<tr>
<td>100</td>
<td>0.43124</td>
<td>0.48595</td>
<td>0.50683</td>
<td>0.52640</td>
<td>0.56752</td>
<td>0.60368</td>
<td>0.61796</td>
</tr>
</tbody>
</table>

4 Applications in functional genomics

4.1 Large sample size

**Synthetic dataset** First a correlation matrix $M_G$ on 20 genes $G_1, \ldots, G_{20}$ is created, together with a dataset $G$ of the corresponding expression $G_i^{1000}$ across 1000 synthetic samples, so that $M_G(i, k) = \lvert \text{cor}(G_i^{1000}, G_j^{1000}) \rvert$ is the absolute Pearson correlation between the expression of the genes $G_i$ and $G_j$ from $G$.

In particular, $M_G$ has two $10 \times 10$ blocks highly correlated on the main diagonal, and two $10 \times 10$ poorly correlated blocks on the minor diagonal, as shown in Fig. 4. These blocks derived from the following generating rule, given uncorrelated starting element $G_1^{1000}$ and $G_{11}^{1000}$:

$$\lvert \text{cor}(G_k^{1000}, G_j^{1000}) \rvert \approx \begin{cases} 1 - 0.03j & \text{for } k = 1, 2 \leq j \leq 10 \\ 0.7 - 0.015j & \text{for } k = 11, 12 \leq j \leq 10 \end{cases}.$$  

Outside the two main blocks, all correlation values range between 0.002 and 0.074. In Fig. 4 we also show the heatmap of the gene expression dataset $G$. Then a subset of $n_s$ samples is selected from the starting 1000, and the corresponding coexpression networks is built, for the 100 hard threshold values $0.01j$, for $1 \leq j \leq 100$. The secure threshold for these cases are respectively 0.799, 0.596 and 0.389. These procedure is repeated 500 times for each value $n_s = 10, 20, 50$. The same experiment is then repeated adding a 20% and a 40% level of Gaussian noise to the original data. Using $M_G$ as the ground truth where all values outside the two main blocks are thresholded to 0, for each hard threshold $0.01j$ we evaluate the ratio of False Positive links, the ratio of False Negative links and the Hamming-Ipsen-Mikhailov (HIM) distance from the gold standard. The graphs summarizing the experiments, separately for sample size, are displayed in Fig. 5.

In all cases, the secure threshold $p$ corresponds to the strictest value yielding a coexpression network with no false positive links included, which its characterizing property. Moreover, in almost all displayed situations, thresholding at $p$ still guarantees an acceptable HIM distance from the ground truth, and a false negative ratio always smaller than 0.4.

**Ovarian cancer** The aforementioned results obtained in a synthetic case are then tested here in a large array study on 285 patients of ovarian cancer at different stages [57], recently used in a comparative study on conservation of coexpressed modules across different pathologies [56]. In details, a whole tumor gene expression profiling was conducted on 285 predominately high-grade and advanced stage serous cancers of the ovary, fallopian tube, and peritoneum; the samples were hybridized on the Affymetrix Human Genome HG-U133 Plus 2.0 Array, including 54621 probes. The goal of the original study was to identify novel molecular subtypes of ovarian cancer by gene expression profiling with linkage to clinical and pathologic

---

1The HIM distance [69, 70] is a metric between networks having the same nodes, ranging between 0 for identical networks and 1, attained comparing the clique with the empty graph.
features. As a major result, the authors presented two ranked gene lists supporting their claim that molecular subtypes show distinct survival characteristics. The two gene lists characterize the Progression Free Survival (PSF) and the poor Overall Survival (OS), respectively.

Following the procedure of the previous, synthetic example, first we individuate the sample subset corresponding to the homogeneous cohort of 161 grade three patients and a set $T$ of 20 genes, belonging to the top good OS and PFS genes (EDG7, LOC649242, SCGB1D2, CYP4B1, NQO1, MYCL1, PRSS21, MGC13057, PPP1R1B, KIAA1324, LOC646769) and to the top poor OS/PFS genes (THBS2, SFRP2, DPSG3, COL11A1, COL10A1, COL8A1, FAP, FABP4, POSTN), thus generating a dataset $O_T$ of dimension 161 samples and 20 features. The corresponding absolute Pearson correlation matrix $O_T$ is then used as the ground truth for the subsampling experiments: the levelplot of $O_T$ and the heatmap of $O_T$ is shown in Fig. 6. In these experiments, a random subdataset of $n_s$ samples is extracted from $O_T$, and the corresponding absolute Pearson coexpression network on the nodes $T$ is built, for increasing threshold values. In Fig. 7 we report the HIM and the ratio of False Positive and False Negative links for 500 runs of the experiments, separately for $n_s = 5, 10, 20$ and 50.

Again, the secure threshold $\bar{p}$ corresponds to the smallest Pearson value warranting no false positive links included. Moreover, in almost all displayed situations, the threshold $\bar{p}$ is approximately the value where the HIM distance starts growing quicker, while the false positive rate remains under 0.8.

### 4.2 Small sample size

When the sample size is very small, the novel hard thresholding introduced here can severely limit the conclusions than can be drawn without incurring in the risk of discussing false positive links. This problem can be particularly evident in differential network analysis tasks [71, 72, 73, 31, 74], where loosening the threshold may lead to consider unsupported variations between networks in different conditions. In what follows we show two cases of (almost) negative results, where the experimental conditions tightly bound the possible differential coexpression network analysis.

**Pancreatic Cancer** The first example is based on a pancreatic cancer dataset, publicly available at GEO [http://www.ncbi.nlm.nih.gov/geo/], at the accession number GDS4329 and originally analysed in [58]. The dataset consists of 24 samples from 6 patients suffering from pancreatic ductal adenocarcinoma, divided in 4 subgroup samples, i.e., circulating tumor cells (C), haematological cells (G), original tumour (T), and non-tumoural pancreatic control tissue (P). The aim of the original study was to develop a circulating tumor cells gene signature and to assess its prognostic relevance after surgery, while here we concentrate on the feasibility of a differential coexpression network analysis. Namely, we explore the Pearson correlation networks build separately on the four classes of samples on a specific set of genes $S$, defined by the differential expression analysis. In particular, the set $S$ include as nodes the genes resulting upregulated in the C subgroup and associated with both the p38 mitogen-activated protein kinase (MAPK) signaling pathway and the cell motility pathway, which were ranked as the pathways with the highest expression ratio. In details, the nine genes are Talin-1 (TLN1), signal transducer and activator of transcription 3 (STAT3), Vinculin (VCL), CCL5, autocrine motility factor receptor (AMFR), Tropomyosin alpha-4 chain (TMP4), arachidonate 12-lipoxygenase (ALOX12), Rho-guanine nucleotide exchange factor 2 (ARHGEF2), and engulfment and cell motility protein (ELMO1), respectively denoted by 1, . . . , 9 in the plots.

Following the formula in Definition 5, the secure threshold for nine genes and six samples is 0.8508; hard thresholding the four coexpression networks results in the graphs collected in Fig. 8. As shown by the plots, the number of edges that result statistically significant over the secure threshold 0.8508 is small: namely 6 for the class G, 4 for the classes C and P and none

![Figure 4](image.png)

Figure 4: Levelplot of the structure of the correlation matrix $M_G$ (left) and heatmap of the dataset $G$. The generating gene expression vectors $G_1^{1000}$ and $G_1^{1100}$ are marked with *. 

---

Figure 5: Coexpression inference of the $M_G$ network from random subsampling of the $G$ dataset, without noise (a,b,c), with 20% Gaussian noise (d,e,f) and with 40% Gaussian noise (g,h,i), on 10 (a,d,g), 20 (b,e,h) and 50 (c,f,i) samples. Solid lines indicate mean over 500 replicates of HIM distance (black), ratio of False Positive (blue) and ratio of False Negative (red); dotted lines of the same color indicate $\pm 1 \sigma$, while grey vertical dashed lines correspond to the secure threshold $\bar{p}$. 

CC-BY-ND 4.0 International license

peer-reviewed) is the author/funder. It is made available under a CC-BY-ND 4.0 International license.
for the primary tumoral cells T. In particular, the classes C and G share the links VCL–CCL5 and VCL–ALOX12, while P and G share the link TPM4–ALOX12 and P and C have no common links. Clearly, the paucity of statistically significant links prevents any further quantitative comparison: in Fig 9 we show, for each networks, the number of links at a given correlation.

Alzheimer data  A similar situation occurs with the Alzheimer dataset studied in [59, 60, 61] and available at GEO http://www.ncbi.nlm.nih.gov/geo/, at the accession number GSE4226. The dataset collect the expression of peripheral blood mononuclear cells from normal elderly control (NEC) and Alzheimer disease (AD) subjects. The NEC and AD subjects were matched for age and education; the Mini-Mental State Examination (MMSE) [75] was administered to all subjects, and the mean MMSE score of the AD group was significantly lower than that of the NEC subjects. Targets from biological replicates of female (F) and male (M) NEC and female and male AD were generated and the expression profiles were determined using the NIA Human MGC custom cDNA microarray. Each combinations of the sex and disease phenotypes has a cohort size of seven samples.

The original aim of the studies was the comparison between NEC and AD and the identification of genes with disease and gender expression patterns. In what follows, we show that, given the small sample size, very little can be assessed by a differential coexpression network analysis (see [76] for a recent larger miRNA coexpression study on a cohort of 363 individuals). In particular, from the KEGG Database http://www.genome.jp/kegg/ [77, 78] we extracted the Alzheimer’s disease pathway in Homo sapiens (KEGG accession hsa05010) and we extracted, from the original 32 genes included in the pathway, the 10 genes spotted on the platform with no missing value across the 28 total samples. The ten resulting genes are apolipoprotein E (APOE), amyloid beta (A4) precursor protein (APP), glycogen synthase kinase 3 beta (GSK3B), cyclin-dependent kinase 5 (CDK5), microtubule-associated protein tau (MAPT), presenilin 2 (Alzheimer disease 4) (PSEN2), amyloid beta A4 precursor protein-binding, family B, member 1 Fe65 (APBB1), lipoprotein lipase (LPL), synuclein alpha non A4 component of amyloid precursor (SNCA) and anterior pharynx defective 1 homolog A (APH1A), numbered from 1 to 10 in the above order in what follows. The resulting heatmap is shown in Fig. 10. The coexpression networks for the four combinations of sex (M/F) and disease (NEC/AD) are shown in Fig. 11, where the secure threshold is \( \bar{p} = 0.8166 \). Again, the number of links whose correlation value is above the secure threshold is very small: however, all the retrieved links are well known in literature [79] and in dedicated webservers such as GeneMANIA http://www.genemania.org [80]. Clearly, if we consider the two main classes AD and NEC, the number of samples grows to 14 for each class, and the threshold \( \bar{p} \) can be relaxed down to 0.5943. The two resulting networks are displayed in Fig. 12, together with the trend of the HIM distance between AD and NEC as a function of the threshold, both globally and separately for gender, where we can see that the selected threshold, in all cases, falls after the maximal distance between disease and control group. As a major effect emerging when comparing the coexpression network of the AD patients versus the NEC individuals we note that the connections between CDK5 and PSEN2, APBB1, LPL, SNCA are lost in the disease networks, while connections appear between APP and APOE, CDK5, MAPT, PSEN2; changing of regulation of CDK5 and APP in AD patients are well known in literature: see for instance [81, 82, 83].

5 Conclusion

A simple a priori, theoretical and non-parametric method is proposed for the selection of an hard threshold for the construction of correlation networks. This model is based on the requirements of filtering random data due to noise and reducing the number of false positive, and it is implemented by means of geometric properties of the Pearson correlation coefficient. This new approach can be especially useful in small sample size case, probably the most common situation in profiling studies in functional genomics. Finally, when the number of samples increase, coupling this method with soft thresholding approaches, can help recovering false negative links neglected by too strict thresholds.
Figure 7: Coexpression inference of the coexpression network from subsampling of the OT dataset, on 5 (a), 10 (b), 20 (c) and 50 (d) samples. Solid lines indicate mean over 500 replicates of HIM distance (black), ratio of False Positive (blue) and ratio of False Negative (red); dotted lines of the same color indicate +/- 1σ, while grey vertical dashed lines correspond to the secure threshold $\bar{p}$. 
Figure 8: Correlation networks on the set $S$ for the four classes T, G, C and P, thresholded at Pearson correlation coefficient 0.8508.

References


Figure 9: Number of links with correlation values larger than a given threshold for the coexpression networks C, P, T, and G; the vertical gray dashed line corresponds to Pearson correlation 0.8508, the secure threshold for 9 nodes and 6 samples.
Figure 10: Heatmap of the expression of the ten genes of the Alzheimer pathway on the 28 samples of the Alzheimer dataset. Vertical lines separate samples groups.


Figure 11: Correlation networks on the Alzheimer dataset $S$ for the four classes M-AD, M-NEC, F-AD, F-NEC, thresholded at Pearson correlation coefficient $0.8166$. 
Figure 12: Alzheimer dataset: correlation networks for the two classes AD and NEC, thresholded at Pearson correlation coefficient 0.5943 (AD, NEC) and HIM curve of the distance between AD and NEC network versus the Pearson correlation threshold, globally (black) and separately for Male (red) and Female (blue) patients and controls. Grey dashed vertical line indicates the secure threshold \( \bar{p} = 0.5943 \) for the global case, while the green line corresponds to the secure threshold \( \bar{p} = 0.8166 \) for the sex disaggregated case.


