

# LARGE SCALE CORRELATION MINING FOR BIOMOLECULAR NETWORK DISCOVERY

By

Alfred Hero  
Bala Rajaratnam

Technical Report No. 2015-02  
January 2015

Department of Statistics  
STANFORD UNIVERSITY  
Stanford, California 94305-4065



LARGE SCALE CORRELATION MINING FOR  
BIOMOLECULAR NETWORK DISCOVERY

By

Alfred Hero  
University of Michigan, Ann Arbor

Bala Rajaratnam  
Stanford University

Technical Report No. 2015-02  
January 2015

**This research was supported in part by  
Air Force Office of Scientific Research  
grant FA9550-13-1-0043 and by  
National Science Foundation grants  
DMS 1106642 and CMG 1025465.**

Department of Statistics  
STANFORD UNIVERSITY  
Stanford, California 94305-4065

<http://statistics.stanford.edu>

# 1 Large scale correlation mining for biomolecular network discovery

---

Alfred Hero\* and Bala Rajaratnam†

\* University of Michigan, Ann Arbor MI

† Stanford University, Stanford CA

## Abstract

Continuing advances in high-throughput mRNA probing, gene sequencing and microscopic imaging technology is producing a wealth of biomarker data on many different living organisms and conditions. Scientists hope that increasing amounts of relevant data will eventually lead to better understanding of the network of interactions between the thousands of molecules that regulate these organisms. Thus progress in understanding the biological science has become increasingly dependent on progress in understanding the data science. Data mining tools have been of particular relevance since they can sometimes be used to effectively separate the “wheat” from the “chaff”, winnowing the massive amount of data down to a few important data dimensions. Correlation mining is a data mining tool that is particularly useful for probing statistical correlations between biomarkers and recovering properties of their correlation networks. However, since the number of correlations between biomarkers is quadratically larger than the number biomarkers, the scalability of correlation mining in the big data setting becomes an issue. Furthermore, there are phase transitions that govern the correlation mining discoveries that must be understood in order for these discoveries to be reliable and of high confidence. This is especially important to understand at big data scales where the number of samples is fixed and the number of biomarkers becomes unbounded, a sampling regime referred to as the “purely-high dimensional setting.” In this chapter, we will discuss some of the main advances and challenges in correlation mining in the context of large scale biomolecular networks with a focus on medicine. A new correlation mining application will be introduced: discovery of correlation sign flips between edges in a pair of correlation or partial correlation networks. The pair of networks could respectively correspond to a disease (or treatment) group and a control group.

## 1.1 Introduction

Data mining at a large scale has matured over the past 50 years to a point where every hour millions of searches over billions of data dimensions are routinely handled by search engines at Google, Yahoo, LinkedIn, Facebook, Twitter, and other media. Similarly, large ontological databases like GO [3] and DAVID [4] have enabled large scale text data mining for researchers in the life sciences [3]. Curated repositories in the NCBI databases [5], such as the NCI pathway interaction database [6], or aggregated repository search engines, such as Pathway Commons [7], can be used to search over the network of interactions between genes and proteins as reported in scientific publications. These reported interactions are sometimes based on causal analysis, e.g., the result of gene-knockout studies that identify causal interactions between a knocked-out gene and some other set of downstream genes. However, with increasing frequency life-science researchers have been reporting networks of associations between thousands of variables measured in a high-throughput assay like gene chip, RNAseq, or chromosomal conformational capture (3C, 4C, HiC). These networks can yield information on direct and indirect interactions in very large dimensions. Indeed there exist many algorithms for reconstructing gene interaction networks over several thousand gene microarray probes. Several of these algorithms have been compared [8].

Data mining is based on computing a set of scores, indexed over the variables in the database, which are ranked in decreasing order of magnitude to yield a rank ordered list. Variables at the top of the list are considered to be the best match to the data mining criterion. The matching criterion will depend on the objective of the experimenter and the nature of the data. For “big data” applications there are two major issues: computational challenges and false positives, where the latter are defined as the occurrence of spurious scores near the top of the list that should be lower down on the list. Correlation mining\* is a kind of data mining that searches for patterns of statistical correlation, or partial correlation, in an interaction network or across several interaction networks. These correlations can usually only be empirically estimated from multiple samples of the variables, as in gene microarray or RNAseq data collected over a population of subjects. From the estimated correlations one obtains an empirical correlation network, or partial correlation network, that can be mined for interconnectivity patterns such as edges, connected nodes, hub nodes, or sub-graphs. Correlation mining was introduced by the authors in [1] for mining connected nodes from correlation networks and in [2] for mining hub nodes from correlation or partial correlation networks.

In this chapter we focus on the related problem of mining edges from correlation or partial correlation networks. In particular, we will emphasize significance

\*Our definition of correlation mining is not to be confused with “correlated graph pattern mining” that seeks to find co-occurring subgraphs in graph databases [9].

---

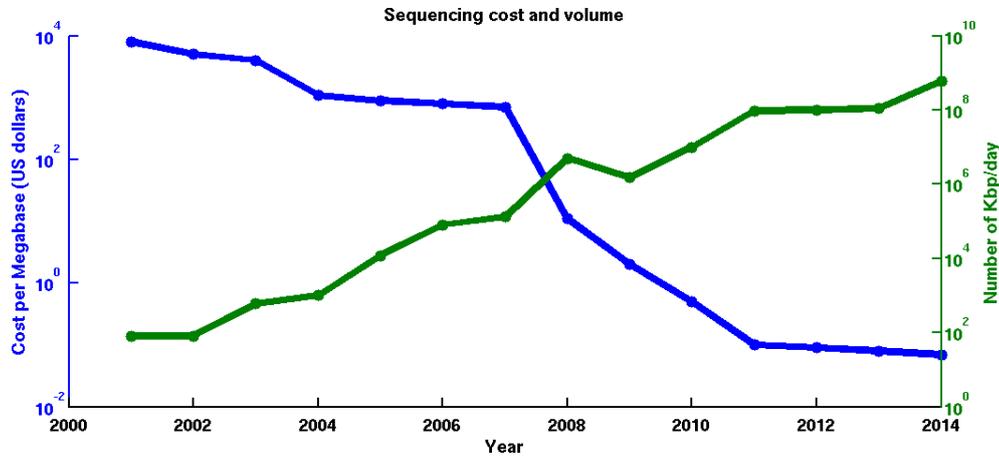
testing for network edges and hubs in the emerging “big data” setting where there are an exceedingly large number of variables. We will provide theory for reliable recovery of edges that applies to arbitrarily large numbers of variables and we will illustrate this theory by implementing correlation mining on gene-gene correlation network of over 12,000 variables (gene probes), which can in principle have on the order of 7 million edges. We will also provide perspectives on future challenges in correlation mining biological data with a special focus on health and medical applications.

Accurate estimation of correlation requires acquiring multiple samples, e.g., technical replication of a gene chip assay  $n$  times or biological replication of the assay on  $n$  different members of the population. With the emergence of increasingly low cost and higher throughput technology, e.g., oligonucleotide gene microchips and RNAseq assays, the expression of increasing numbers of biomarkers can be determined from a single biological sample at increasingly lower cost. On the other hand, the cost of acquiring additional reliable samples does not appear to be decreasing at the same rate as the cost of including additional biomarkers in the assay. For example, running a controlled experiment or challenge study with animals or human volunteers is labor-intensive and very costly, and will likely remain so for the foreseeable future.

Therefore, at least for experimental biology applications, correlation mining practitioners face a deluge of biomarkers (variables) with a relative paucity of biological replicates (samples). This situation poses great difficulty in correlation mining due to the multiple comparisons problem: with so few samples one is bound to find spurious correlations between some pairs of the many variables. Controlling, or at least predicting, the number of spurious correlations must be the principal scientific objective in such situations. Achieving this objective lies in the realm of high dimensional statistical inference in the “large  $p$  small  $n$  regime.” Statisticians have studied several sub-regimes, e.g., those characterized as “high dimensional,” “very-high dimensional” or “ultra-high dimensional” settings [10]. These settings, however, still require that both the dimension  $p$  and the sample size  $n$  go to infinity. Consequently they may not be very useful for biological experimentalists who lack the budget to collect an increasing number of samples, especially given how large  $p$  is. A more useful and relevant regime, which will be the focus of this chapter, is the “purely-high dimensional” setting where the number of samples  $n$  remains fixed while the number of variables  $p$  is allowed to grow without bound. This high dimensional regime is in fact the highest possible dimensional regime, short of having no samples at all. Thus it is appropriate to call this purely-high dimensional regime the “ultimately-high dimensional regime,” and we shall use these two terms interchangeably in the rest of this chapter. A table illustrating our asymptotic framework is given below (Table 1.1), and also serves to compare and contrast our framework to previously proposed asymptotic regimes.

Asymptotic framework	Terminology	Sample size	Dimension	Application setting	References
		$n$	$p$		
Classical (or sample increasing)	small dimensional	$\rightarrow \infty$	fixed	"small data"	Fisher [11, 12], Rao [13, 14], Neyman and Pearson [15], Wilks [16], Wald [17, 18, 19, 20], Cramér [21, 22], Le Cam [23, 24], Chernoff [25], Kiefer and Wolfowitz [26], Bahadur [27], Efron [28]
	high dimensional	$\rightarrow \infty$	$\rightarrow \infty$	"medium sized" data	Donoho [29], Zhao and Yu [30], Meinshausen and Bühlmann [31], Candès and Tao [32], Bickel, Ritov, and Tsybakov [33], Peng, Wang, Zhou, and Zhu [34], Wainwright [35, 36], Khare, Oh, and Rajaratnam, [37]
Mixed asymptotics	very high dimensional	$\rightarrow \infty$	$\rightarrow \infty$	"medium sized" data (mega or giga scales)	Candès and Tao [32], Bickel, Ritov, and Tsybakov [33], Peng, Wang, Zhou, and Zhu [34], Wainwright [35, 36], Khare, Oh, and Rajaratnam, [37]
	ultra high dimensional	$\rightarrow \infty$	$\rightarrow \infty$		
		$\rightarrow \infty$	$\rightarrow \infty$		
Purely high dimensional	purely high dimensional	fixed	$\rightarrow \infty$	"Big Data" (tera, peta and exascales)	Hero and Rajaratnam [1] Hero and Rajaratnam [2] Firouzi, Hero and Rajaratnam [38]

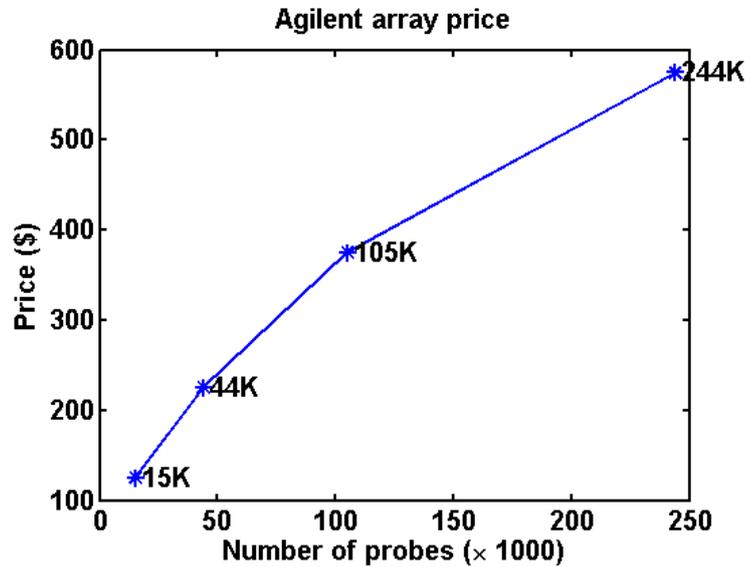
**Table 1.1** Overview of different asymptotic regimes of statistical data analysis. These regimes are determined by the relation between the number  $n$  of samples drawn from the population and the number  $p$ , called the dimension, of variables or biomarker probes. In the classical asymptotic regime the number  $p$  is fixed while  $n$  goes to infinity. This is the regime where most of the well known classical statistical testing procedures, such as student  $t$  tests of the mean, Fisher  $F$  tests of the variance, and Spearman tests of the correlation, can be applied reliably. Mixed asymptotic regimes where  $n$  and  $p$  both go to infinity have received much attention in modern statistics. However, in this era of big data where  $p$  is exceedingly large, the mixed asymptotic regime is inadequate since it still requires that  $n$  go to infinity. The recently introduced "purely high dimensional regime" [1], [2], which is the regime addressed in this paper, is more suitable to big data problems where  $n$  is finite.



**Figure 1.1** The cost in US dollars per Megabase (blue curve) and the number of Kilobases (red curve) of DNA sequenced per day over the past 14 years. The cost of determining one megabase of DNA has fallen at a very rapid rate since the transition in early 2008 from the Sanger-based technology to second-generation technology. At the same time the total volume of DNA sequenced has risen at an even more rapid rate as the price has fallen, demand for sequencing has grown, and sequencing centers have proliferated (Note: volume is a product of sequencing depth and number of DNA samples sequenced). Source: NCBI.

The purely-high dimensional regime of large  $p$  and fixed  $n$  poses several challenges to those seeking to perform correlation mining on biological data. These include both computational challenges and the challenge of error control and performance prediction. Yet this regime also holds some rather pleasant surprises. Remarkably there are modest benefits to having few samples in terms of computation and scaling laws. There is a scalable computational complexity advantage relative to other high dimensional regimes where both  $n$  and  $p$  are large. In particular, correlation mining algorithms can take advantage of numerical linear algebra shortcuts and approximate  $k$  nearest neighbor search to compute large sparse correlation or partial correlation networks. Another benefit of purely-high dimensionality is an advantageous scaling law of the false positive rates as a function of  $n$  and  $p$ . Even small increases in sample size can provide significant gains in this regime. For example, when the dimension is  $p = 10,000$  and the number of samples is  $n = 100$ , the experimenter only needs to double the number of samples in order to accommodate an increase in dimension by six orders of magnitude ( $p = 10,000,000,000$ ) without increasing the false positive rate [2].

The relevance of the purely-high dimensional regime in biology can be understood more concretely in the context of developments in the technology of DNA sequencing and gene microarrays. The cost of sequencing has dropped very rapidly to the point where, as of this writing, the sequencing of a individual's full



**Figure 1.2** The price in US dollars of an Agilent array as a function of the number of gene probes per array. The price increases sublinearly in the number of probes in the array. Data is from the Agilent Custom Gene Expression Microrarrays G2309F, G2513F, G4503A, G4502A (Feb 2013). Source: BMC RNA Profiling Core .

genome (3000 megabases) is approaching \$1000 US. Recent RNAseq technology allows DNA sequencing to be used to measure levels of gene expression. Figure 1.1 shows how the cost of sequencing (in dollars per megabase) vs. the volume of DNA sequenced (in kilobases per day) as a function of time (Source: [39], [40]). While there has been a flattening of the cost curve over the past two years, the cost is still decreasing, albeit more slowly than before. Despite the recent flattening out of cost, the volume sequenced has been increasing at a rapid rate. This suggests that the usage of high throughput sequencing technology will expand, creating a demand for efficient correlation mining methods like the ones discussed in this chapter.

The trends are similar in gene microarray technology, a competitor to sequencing based gene expression technologies such as RNAseq. One such technology is the 60mer oligonucleotide gene expression technology by Agilent, which is similar to the shorter 25mer oligonucleotide technology sold by Affymetrix as the "genechip." In Fig. 1.2 is shown the price per slide of the Agilent Custom Microarray as a function of the number of biomarkers (probes) included on the slide. The price increases sublinearly in the number of probes. As the probe density increases the price will come down further at the high density end of the scale. Thus it is becoming far less costly to collect more biomarker probes ( $p$  variables) than it is to collect more biological replicates ( $n$  samples). One can

therefore expect an accelerated demand for high dimensional analysis techniques such as the correlation mining methods discussed in this chapter.

The outline of the chapter is as follows. In Sec. 1.2 we give an illustrative example to motivate the utility of correlation measures for mining biological data. In Sec. 1.2.2 we formally introduce correlation and partial correlation in the context of recovering networks of biomarker interactions. In Sec. 1.3, we briefly describe rules of thumb for setting correlation screening thresholds to protect against these errors. In Sec. 1.4 we describe some future challenges in correlation mining of biological data, In Sec. 1.5 we provide some concluding remarks.

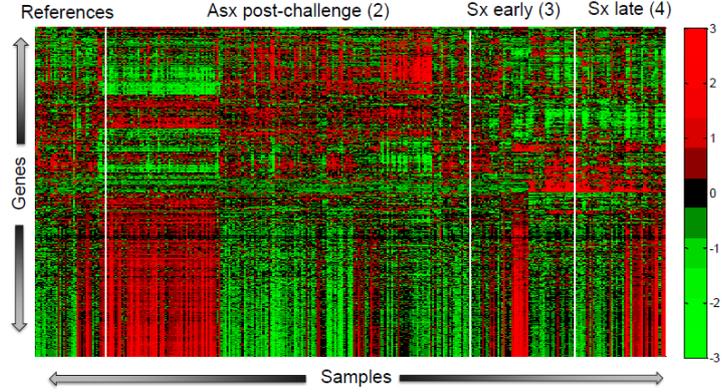
## 1.2 Illustrative example

We illustrate the utility of our correlation mining framework in the context of gene chip data collected from an acute respiratory infection (ARI) flu study [41],[42]. In this ARI flu study 17 human male and female volunteers were infected with a virus and their peripheral blood gene expression was assayed at 16 time points pre- and post-infection. The study resulted in a matrix of 12,023 gene probes by 267 samples was formed from the 267 available Affymetrix HU133 chips.

Roughly half of the volunteers became symptomatically ill (Sx) while the rest (Asx) did not become ill despite exposure to the virus. Each volunteer had two reference samples collected before viral exposure. The samples of the Sx volunteers were subdivided into early and late stages of infection using an unsupervised Bayesian factor analysis method [42]. Figure 1.3 shows the  $12,023 \times 267$  matrix with columns arranged over the 4 categories: *reference*, *Asx post-challenge*, *Sx early*, *Sx late*. From this raw data one can see some patterns but they do not clearly differentiate the classes of individuals. This lack of definition is due to the fact that these classes are not homogeneous. The immune response of the volunteers evolves over time, the temporal evolution is not synchronous across the population, and some genes exhibit different responses between the men and women in the study.

To better discriminate the genes that differentiate between the classes, a score function can be designed to rank and select the genes with highest scores. Since the number of samples is limited, the score function and the cutoff threshold should be carefully selected according to statistical principles to avoid false positives. Two classes of score functions will be discussed here: first order and second order.

Let's say that an experimenter wishes to find genes that have very different mean expression values  $\mu$  when averaged over two sub-populations  $A$  and  $B$ . Let  $\{X_k^A(g)\}_{k=1}^{n_A}$  be the measured expression levels of gene  $g$  over a population of  $n_A$



**Figure 1.3** A heatmap of the  $12,023 \times 267$  matrix (6.5 mega pixels) of gene expression from a viral flu challenge study involving samples taken before viral exposure (reference), subjects that did not become ill after exposure (Asx post-challenge), subjects that became symptomatic (Sx early and Sx late). These are 4 different sub-population classes, called phenotypes, that can be mined for genes whose expression patterns change between pairs of classes. There is no easily discernible discriminating gene pattern that one can see just by looking at the raw data heatmap. The objective of correlation mining is to extract genes whose pattern of correlation changes over different sub-populations.

samples in population  $A$ . The sample mean  $\mu_A(g)$  is

$$\mu_A(g) = n_A^{-1} \sum_{k=1}^{n_A} X_k^A(g).$$

Similarly define the sample mean  $\mu_B$  over sub-population  $B$ . The difference  $\mu_B - \mu_A$  between the sample means over the sub-populations is an example of a first order score function. Thresholding this score function will produce a list of genes that have the highest contrast in their means. More generally, a first order score function is any function of the data that is designed to contrast the sample means across sub-populations. The student-t test statistic and the Welch test statistic [43] are also first order score functions that are variance normalized sample mean differences between two populations.

### 1.2.1 Pairwise correlation

A different objective for the experimenter, and the motivation for this chapter, is to find pairs of genes that have very different correlation coefficients  $\rho$  over two sub-populations. Let  $\{X_k^A(g), X_k^A(\gamma)\}_{k=1}^{n_A}$  and  $\{X_k^B(g), X_k^B(\gamma)\}_{k=1}^{n_B}$  be the measured expression levels of two genes  $g$  and  $\gamma$  over two sub-populations  $A$  and  $B$ . The standard Pearson product moment correlation coefficient between these

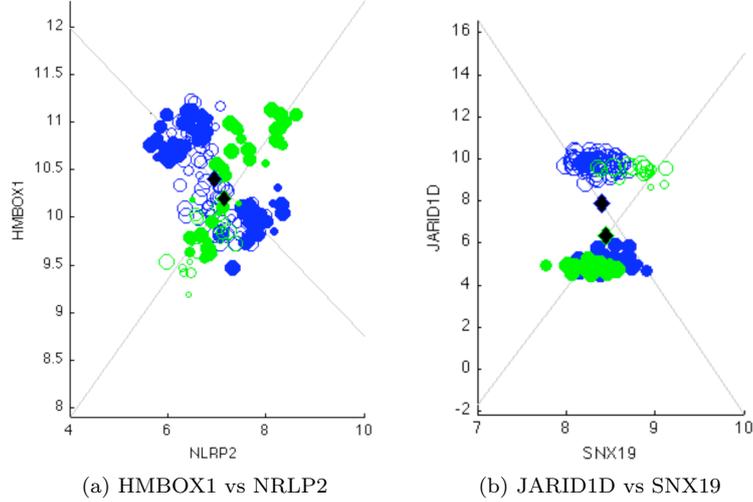
two genes for sub-population  $A$  is

$$\rho_A(g, \gamma) = \frac{\sum_{k=1}^{n_A} (X_k^A(g) - \mu_A(g))(X_k^A(\gamma) - \mu_A(\gamma))}{\sqrt{\sum_{k=1}^{n_A} (X_k^A(g) - \mu_A(g))^2} \sqrt{\sum_{k=1}^{n_A} (X_k^A(\gamma) - \mu_A(\gamma))^2}} \in [-1.1],$$

and similarly for  $\rho_B(g, \gamma)$ . The difference between sample correlations  $\Delta\rho_{A,B} = \rho_A(g, \gamma) - \rho_B(g, \gamma)$  is a contrast that is an example of a higher order score function. The magnitude of  $\Delta\rho_{A,B}$  takes a maximum value of 2, which it approaches when the correlation is high but of opposite sign in both sub-populations. Genes that have high contrast  $\Delta\rho_{A,B}$  in their correlations may not have high contrast in their means. Furthermore, unlike the mean, the correlation is directly related to the predictability of one gene from another gene, since the mean squared error of the optimal linear predictor is proportional to  $(1 - \rho^2)$ : a variable that is highly correlated to another variable is easy to predict from that other variable. Finally, it has been observed by many researchers that, while the mean often fluctuates over sub-population, the correlation between biomarkers is preserved. Thus the sample correlation can be more stable than the sample mean, especially when used to compare multiple populations or multiple species [44], [45], [46], [47] which often makes it a better discovery tool.

For an illustration of the differences between mining with the mean vs. mining with the correlation the reader is referred to Fig. 1.4. In this figure, two pairs of genes are shown from the ARI challenge study whose sample correlation coefficients flip from a highly positive coefficient in the Asx sub-population to a highly negative coefficient in the Sx sub-population. The HMBOX1 and NRLP2 genes shown in the left panel are known transcription factors that are involved in immune response. HMBOX1, a homeobox transcription factor, negatively regulates natural killer cells (NK) in the immune system defense mechanism [48]. NRLP2, an intracellular nod-like receptor, regulates the level of immune response to the level of threat [49]. The flip from positive correlation (Asx) to negative correlation (Sx) between NRLP2 and HMBOX1 is an interesting and discriminating biomarker pattern between Sx and Asx phenotypes. Note that neither of these genes exhibit a significant change in mean over these sub-populations. Hence, while a second order contrast function assigns these genes very high scores, revealing them as highly differentiated over Sx and Asx, a first order contrast function would miss them entirely.

The right panel of Fig. 1.4 shows a correlation flip between JARID1D and SNX19 that is capturing differences between the immune response of women (solid circles) and men (hollow circles). These differences would not be revealed by a first order analysis without stratifying the population into male and female volunteers. Without such stratification, the first order analysis would detect no significant change in mean over the Sx and Asx sub-populations.



**Figure 1.4** Two pairs of genes (rows of matrix in Fig. 1.3) discovered by a second order analysis (polarity-flip correlation screening) between Asx and Sx sub-populations in a acute respiratory infection (ARI) dataset (expression levels of 12,023 genes measured over 267 samples). Solid circles denote women, hollow circles denote men. Green symbols denotes Asx subjects and blue symbols denotes Sx subjects. Size of symbol encodes time index of sample (smaller size and larger size symbols encode early and late post-exposure times, respectively). In both left and right panels there is a flip from high positive correlation to high negative correlation (indicated by the diagonal and antidiagonal lines in each scatterplot). The sample mean would not reveal these genes as differentiating the Sx and Asx subjects.

### 1.2.2 From pairwise correlation to networks of correlations

Assume that there are  $p$  genes that are assayed in a population and let  $\rho(g, \gamma)$  be the sample correlation between the pair  $g, \gamma$ . Define the  $p \times p$  matrix of sample correlation coefficients  $\mathbf{R} = [[\rho(g, \gamma)]]_{g, \gamma=1}^p$ . By thresholding the magnitude of  $\mathbf{R}$  one obtains an adjacency matrix whose non-zero entries specify the edges between pairs of genes in a correlation graph. The higher the correlation threshold the sparser will be the adjacency matrix and the fewer edges will be contained in the graph. The correlation graph, also called a correlation network, has  $p$  nodes and up to  $\binom{p}{2}$  edges. To avoid confusion between the correlation network and the partial correlation network, discussed below, the former is often called the *marginal correlation network*.

Another network associated with the interactions between genes is the *partial correlation network*. The partial correlation network is computed by thresholding the diagonally normalized Moore-Penrose generalized inverse of the sample correlation, which we denote  $\mathbf{P}$ , where

$$\mathbf{P} = \text{diag}(\mathbf{R}^\dagger)^{-1/2} \mathbf{R}^\dagger \text{diag}(\mathbf{R}^\dagger)^{-1/2} \quad (1.1)$$

and  $\mathbf{R}^\dagger$  is the Moore-Penrose generalized inverse and  $\text{diag}(\mathbf{R}^\dagger)$  is the matrix

obtained by zeroing out the off diagonal entries of  $\mathbf{R}^\dagger$ . Thus edges in the partial correlation graph depend on the correlation matrix only through its inverse.

Correlation graphs and partial correlation graphs have different properties, a fact that is perhaps easily understood in the context of sparse Gaussian graphical models [50], also known as Gauss-Markov random fields. For a centered Gaussian graphical model, the joint distribution  $f(\mathbf{X})$  of the variables  $\mathbf{X} = [X(1), \dots, X(p)]^T$  is multivariate Gaussian with mean zero mean and covariance parameter  $\Sigma$ . Consider samples  $\{\mathbf{X}_k\}_{k=1}^n$  which are assumed to be independent and identically distributed (i.i.d.). For this Gaussian model the sparsity of  $(i, j)^{th}$  element of the inverse correlation matrix, and hence sparsity of the partial correlation network, implies that components  $i$  and  $j$  are independent given the remaining variables. This conditional independence property is referred to as the “pairwise” Markov property. Another kind of Markov property, the so-called local Markov property, states that a specified variable given its nearest neighboring variables (in the partial correlation network) is conditionally independent of all the remaining variables. Yet another Markov property is the global Markov property which states that two blocks of variables A and B in a graph are conditionally independent given another set of variables C when the third block C separates in the graph the original two blocks of variables (see [50] for more details). Here “separate” means that every path between a pair of nodes, one in block A and another in block B, has to traverse C. It can be shown that the pairwise, local and global Markov properties are equivalent under very mild conditions [50]. In other words, a Markov network constructed from the pairwise Markov property allows one to read off complex multivariate relationships at the level of groups of variables by simply using the partial correlation graph. Thus the sparsity of the inverse correlation captures Markov properties between variables in the dataset. Since the marginal correlation graph encodes pairwise or bivariate relationships, the partial correlation network is often at least as sparse, and usually much sparser, than the marginal correlation network. This key property makes partial correlation networks more useful for obtaining a parsimonious description of high dimensional data.

One of the principal challenges faced by practitioners of data mining is the problem of phase transitions. We illustrate this problem in the context of correlation mining as a prequel to the theory presented in Sec. 1.3. As explained in Sec. 1.3, as one reduces the threshold used to recover a correlation or partial correlation network, one eventually encounters an abrupt phase transition point where we start to see an increasingly large number of false positive edges or false positive nodes connected in the recovered graph. This phase transition threshold can be mathematically approximated and, using the theory below, a threshold can be selected that guarantees a prescribed false positive rate under an assumption of sparsity.

We illustrate the behavior of partial correlation and marginal correlation networks on the reference samples of the ARI challenge study dataset. In this data set there were 34 reference samples taken at two time points before viral-exposure

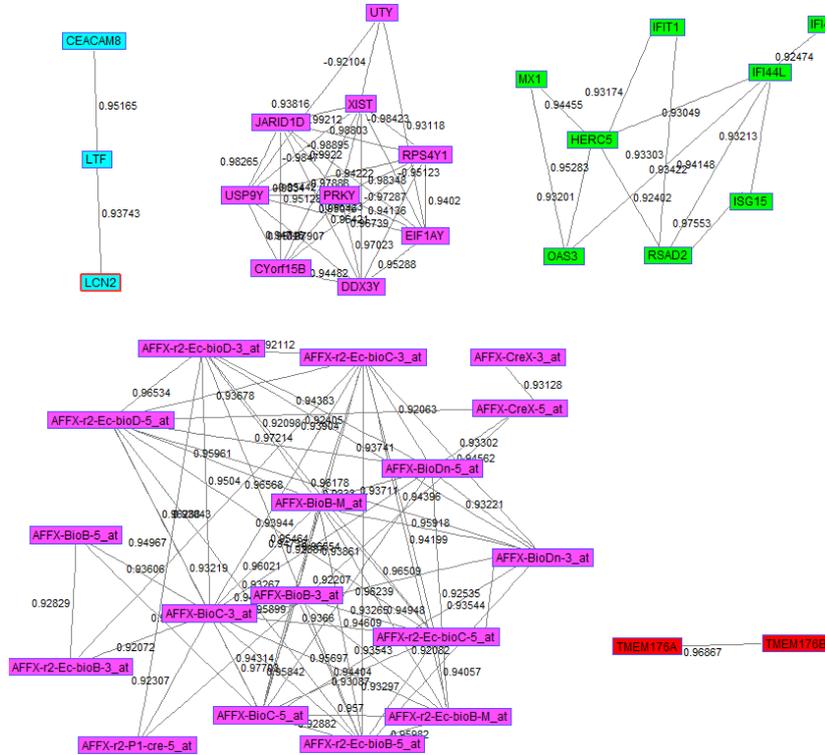
from the 17 volunteers enrolled in the study. Using theory discussed in Sec. 1.3 we selected and applied a threshold (0.92) to the correlation and partial correlation matrices. This threshold was determined using Thm 1 to approximate the false positive rate  $P_e$  with the expression (1.2) and using this expression to select the threshold  $\rho$  that gives  $P_e = 10^{-6}$  when there are  $p = 12,023$  variables and  $n = 34$  samples. This false positive rate constraint is equivalent to a constraint that on the average there are fewer than 0.22 nodes that are mistakenly connected in the graph.

The application of the threshold  $\rho = 0.92$  to the sample correlation resulted in a correlation network having 8718 edges connecting 1658 nodes (genes). On the other hand, using the same threshold on the sample partial correlation, which according to Thm 1 also guarantees  $10^{-6}$  false positive rate, the recovered partial correlation network had only 111 edges connecting 39 genes. This later network is shown in Fig. 1.5. The fact that the partial correlation network is significantly sparser strongly suggests that the genes behave in a Markovian manner with Markov structure specified by the graph shown in Fig. 1.5. This graph reveals four connected components that are conditionally independent of each other since there are no edges between them.

By investigating the genes in the partial correlation network's connected components in Fig. 1.5, the four modules can be putatively associated with different biological functions. In particular, the connected component colored in green at top right corresponds to sentinel genes like HERC5, MX1, RSAD2, and OAS3 in the immune system. The purple colored component in the top middle is composed of genes like UTY and PRKY that are involved in protein-protein interactions and kinase production. The large connected component in purple at bottom center is a module of housekeeping biomarkers, e.g., AFX-BioDn-T\_at, that are used by the gene chip manufacturer (Affymetrix) for calibration purposes. While all of these genes are also present in the connected components of the much less sparse marginal correlation network, the picture is not nearly as simple and clear.

### 1.3 Principles of correlation mining for big data

In big data collection regimes one is often stuck with an exceedingly large number of variables (large  $p$ ) and a relatively small and fixed number of samples (small  $n$ ). This regime is especially relevant for bio-molecular applications where size  $p$  of the genome, proteome or interactome can range from tens of thousands to millions of variables, while the number  $n$  of samples in a given sub-population is fixed and only on the order of tens or hundreds. Recently an asymptotic theory has been developed expressly for correlation mining in this purely-high dimensional regime [1], [2]. Unlike other high dimensional regimes where both  $p$  and  $n$  go to infinity, the theory of [1], [2] only requires  $p$  to go to infinity while  $n$  remains fixed. This property makes this purely-high dimensional regime more



**Figure 1.5** Recovered partial correlation gene network over the 12,023 genes and 34 reference samples assayed in the ARI challenge study data. The applied partial correlation threshold used to obtain this network was determined using the theory in Sec. 1.3 in order to guarantee a  $10^{-6}$  false positive rate or less. There are only 39 nodes (genes) connected by 111 edges in the displayed partial correlation network as compared to over 1658 nodes and 8718 edges recovered in the (marginal) correlation network (not shown).

relevant to limited sampling applications, which characterize big data. Before discussing the theory of correlation mining we contrast it to other correlation recovery methods.

We define correlation mining in a very specific manner in order to differentiate it from the many other methods that have been proposed to recover properties of correlation matrices. Most of these methods are concerned with the so called covariance selection problem [51]. The objective of covariance selection is to find non-zero entries in the inverse covariance matrix for which many different approaches have been proposed [31], [52], [53],[54], [55], [56], [34], [57], [58], [59], [60], [61], [37], [62], [63],[64], [65] and some have been applied to bioinformatics applications [66], [67],[68]. Covariance selection adopts an estimation framework where one attempts to fit a sparse covariance (or inverse covariance) model to the data using a goodness of fit criterion. For example, one can minimize residual

fitting error, such as penalized Frobenius norm squared, or maximize likelihood. The resulting optimization problem is usually solved by iterative algorithms with a stopping criterion. These methods are not immediately scalable to large numbers of variables except in very sparse situations [58].

As contrasted to covariance selection methods, which seek to recover the entire covariance or inverse covariance including the zero values, correlation mining methods use thresholding to identify variables and edges connected in the correlation or partial correlation network that have the the strongest correlations. Thus correlation mining does not penalize for estimation error nor does it recover zero correlations. Correlation mining is fundamentally a testing problem as contrasted to an approximation problem - a characteristic that clearly differentiates correlation mining from covariance selection. Correlation mining can be looked at as complementary to covariance selection and has several advantages in terms of computation and error control. Since it only involves simple thresholding operations correlation mining is highly scalable. Since correlation mining filters out all but the highest correlations, a certain kind of extreme value theory applies. This theory specifies the asymptotic distribution of the false positive rate in the large  $p$  small  $n$  big data regime. This can then be used to accurately predict the onset of phase transitions and to select the applied threshold to ensure a prescribed level of error control. Inverse covariance estimation problems do not in general admit this property.

In [1], [2] three different categories of correlation mining problems were defined. These correlation mining problems aimed to identify variables having various degrees of connectivity in the correlation or partial correlation network. These were: correlation screening over a single population, correlation screening over multiple populations, and hub screening over a single population. In each of these problems theory was developed for the purely-high dimensional regime of large  $p$  and fixed  $n$  under the assumption of block sparse covariance. A block sparse  $p \times p$  covariance matrix is a positive definite symmetric matrix  $\Sigma$  for which there exists a row-column permutation matrix  $\Pi$  such that  $\Pi\Sigma\Pi^T$  is block diagonal with a single block of size  $m \times m$  with  $m = o(p)$ . For more details the reader is referred to the original papers [1], [2]. While not specifically emphasized in these papers, the asymptotic theory also directly applies to screening edges in the network. Indeed the proofs of the asymptotic limits in [1], [2] use an obvious equivalence between the incidence of false edges and false connected nodes in the recovered network. In particular, we have the following theorem that is proved as an immediate corollary to [2, Prop. 2].

**THEOREM 1.1** *Assume that the  $n$  samples  $\{\mathbf{X}_k\}_{k=1}^n$  are i.i.d. random vectors in  $\mathbb{R}^p$  with bounded elliptically contoured density and block sparse  $p \times p$  covariance matrix. Let  $\rho$  be the threshold applied to either the sample correlation matrix or the sample partial correlation matrix. Assume that  $p$  goes to infinity and  $\rho = \rho_p$  goes to one according to the relation  $\lim_{p \rightarrow \infty} p(p-1)(1-\rho_p^2)^{(n-2)/2} = e_n$ . Then the probability  $P_e$  that there exists at least one false edge in the correlation or*

partial correlation network satisfies

$$\lim_{p \rightarrow \infty} P_e = 1 - \exp(-\kappa_n/2), \quad (1.2)$$

where

$$\kappa_n = e_n a_n (n-2)^{-1}$$

where  $a_n$  is the volume of the  $n-2$  dimensional unit sphere in  $\mathbb{R}^{n-1}$  and is given by  $a_n = \frac{\Gamma((n-1)/2)}{\sqrt{\pi}\Gamma((n-2)/2)} = 2B((n-2)/2, 1/2)$ .

We now provide an intuitive way to understand the flavor of our results. Consider a null model where the covariance matrix is diagonal, i.e., when there are no true non-zero correlations, and so any edge that is detected as non-zero will be a false edge. Under this setting one can generate the distribution of the sample correlation coefficients for various dimensions. Fig. 1.6 illustrates these distributions in the setting when  $p = 100$  for various values of  $n$ . It is clear that when the sample size  $n$  is low relative to the dimension  $p$ , the probability of obtaining false edges are higher, and consequently a higher threshold level  $\rho_p$  is required in order to avoid detecting false edges. In such situations consider once more the quantity  $P_e$ , the probability of obtaining at least one false edge. It is clear that for a fixed sample size  $n$  and fixed threshold  $\rho$ , as the dimension  $p \rightarrow \infty$  the probability of detecting a false edge  $P_e$  tends to 1. So it makes sense to let the threshold  $\rho$  tend to 1 in such settings so that the probability of detecting a false edge is small. However, for a fixed sample size  $n$  and fixed dimension  $p$ , if the threshold  $\rho$  tends to 1, then the probability of detecting a false edge goes to zero. This is because as  $\rho$  gets larger and tends to 1, it will eventually surpass the largest sample correlation. Thus increasing  $\rho$  to 1 will eventually threshold all sample correlations to zero, and will result in the probability of detecting a false edge tending to 0. The two scenarios described above therefore lead to degenerate or trivial limits. The theory above resolves this degeneracy by letting the threshold  $\rho_p$  tend to 1 at the correct rate as  $p \rightarrow \infty$  to obtain a non-degenerate limit. This in turn leads to a very useful expression for the probability of detecting a false edge  $P_e$  in the “purely high dimensional” setting, when only the dimension  $p \rightarrow \infty$ , while the sample size  $n$  is fixed.

Theorem 1.1 gives a limit for the probability  $P_e$  that is universal in the following senses: 1) it applies equally to correlation and partial correlation networks; 2) the limit does not depend on the true covariance.

The quantity  $e_n$  in Theorem 1.2 does not depend on  $p$ . However, we can remove the limit from the definition of  $e_n$  and substitute it into the expression for  $\kappa$  to obtain a useful large  $p$  approximation to  $P_e$

$$P_e = 1 - \exp(-\lambda(\rho, n)/2), \quad (1.3)$$

with

$$\lambda_{\rho, n} = p(p-1)P_0(\rho, n),$$

and,  $P_0$  is the normalized incomplete Beta function

$$P_0(\rho, n) = 2B((n-2)/2, 1/2) \int_{\rho}^1 (1-u^2)^{\frac{n-4}{2}} du, \quad (1.4)$$

with  $B(a, b)$  the Beta function. A bound on the fidelity of the approximation  $1 - \exp(-\lambda_{\rho, n}/2)$  to  $P_e$  was obtained in [2, Sec. 3.2] for hub screening and it also applies to the case of edge screening treated here.

The large  $p$  approximation (1.3) to  $P_e$  resembles the probability that a Poisson random variable with rate function  $\lambda_{\rho, n}/2$  exceeds zero. It can be shown that  $\lambda_{\rho, n}/2$  is asymptotically equal to the expected number of false edges. Hence, in plain words, Theorem 1 says that the incidence of false edges in the thresholded sample correlation or sample partial correlation network asymptotically behave as if the edges were Poisson distributed.

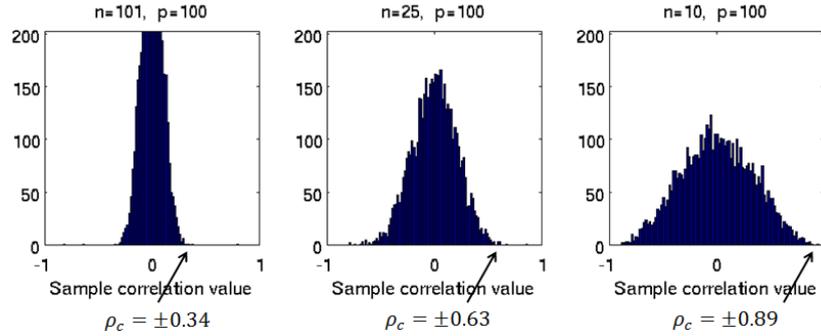
Remarkably, the Poisson rate  $\lambda_{\rho, n}/2$  in the large  $p$  approximation (1.3) does not depend on the true covariance matrix. This is a consequence of the block sparse assumption on the covariance. When the covariance is only row sparse, i.e., the number of non-zeros in each row increases only as order  $o(p)$ , then the same theorem holds except that now  $\lambda_{\rho, n}$  will generally depend on the true covariance matrix.

The behavior of  $\lambda_{\rho, n}$  as a function of  $\rho$ ,  $n$ , and  $p$  specifies the behavior of the approximation (1.3) to  $P_e$ . In particular for fixed  $n$  and  $p$  there is an abrupt phase transition in  $P_e$  as the applied correlation threshold  $\rho$  decreases from 1 to 0 (see Fig. 1.6). An asymptotic analysis of the incomplete Beta function yields a sharp approximation to the phase transition threshold  $\rho_c$  [1]. This threshold, defined as the knee in the curve  $\lambda(\rho, n)$  defining the asymptotic mean number of edges surviving the threshold  $\rho$ , has the form [2, Eq. (10)]:

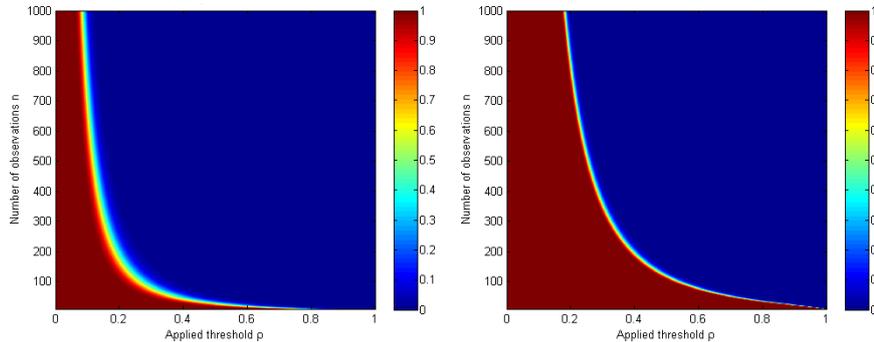
$$\rho_c = \sqrt{1 - (c_n(p-1))^{-2/(n-4)}}, \quad (1.5)$$

where  $c_n = 2B((n-2)/2, 1/2)$ .

In Fig. 1.7 the large  $p$  approximation (1.3) to the false discovery probability  $P_e$  is rendered as a heatmap over  $n$  and  $\rho$  for the cases  $p = 10$  and  $p = 10,000$ . The phase transition clearly delineates the region of low  $P_e$  from the region of high  $P_e$  and becomes more abrupt for the larger value of  $p$ . The critical phase transition point increases much less slowly in  $n$  than it does in  $p$ . For example, from the heatmap we can see that if the biologist wanted to reliably detect edges of magnitude greater than 0.5 with  $p = 10$  variables she would need more than  $n = 60$  samples. However, if the number of variables were to increase by a factor of 1000 to  $p = 10,000$  she would only need increase the number of samples by a factor of 2 to  $n = 120$  samples. This represents a very high return on investment into additional samples. However, as can be seen from the heatmaps, there are rapidly diminishing returns as the number of samples increase for fixed  $p$ . Though we do not undertake a detailed demonstration of Theorem 1.1 and further development thereof for use in graphical model selection, we note that this topic is the subject of ongoing and future work.



**Figure 1.6** Histograms of the number of edge discoveries in a sample correlation graph over  $p = 100$  biomarkers as a function of the threshold  $\rho$  applied to the magnitude of each entry of the correlation matrix. In this simulation the true covariance matrix is diagonal so every edge discovery is a false discovery. In the three panels from left to right the number of samples is reduced from  $n = 101$  to  $n = 10$ . The theoretically determined critical phase transition threshold  $\rho_c$ , given by expression (1.5), closely tracks the observed phase transition threshold in all cases (Figure adapted from [1, Fig. 1]).



**Figure 1.7** Heatmap of the large  $p$  approximation to the false edge discovery probability  $P_e$  in (1.2) as a function of the number of samples  $n$  and the applied threshold  $\rho$  for  $p = 10$  (left) and for  $p = 10,000$  (right). The phase transition clearly delineates the region of low  $P_e$  (blue) from the region of high  $P_e$  (red) and becomes increasingly abrupt as  $p$  increases to  $p = 10,000$ .

### 1.3.1 Correlation mining for correlation flips between two populations

In Fig. 1.4 of Sec. 1.2.1 was shown a pair of genes whose correlation flipped from a positive value in one population to a negative value in another population. The correlation mining theory of the previous subsection is easily extended to this case using the results of [1, Prop 3] in persistency correlation screening of multiple independent populations. For two populations of samples  $\{\mathbf{X}_k\}_{k=1}^{n_A}$  and  $\{\mathbf{Y}_k\}_{k=1}^{n_B}$  on the same domain  $\mathbb{R}^p$  let  $\rho_A$  and  $\rho_B$  be two thresholds in  $[0, 1]$ .

When each of these thresholds is applied to the respective correlation matrices one obtains correlation networks  $G_A$  and  $G_B$ , respectively. We define the  $p$  node *correlation flip network* as the  $p$  node network obtained by placing an edge between two nodes if there exists a corresponding edge in  $G_A$  and  $G_B$  and the correlations associated with these two edges are of opposite sign. The number of false positive edges will depend on the two thresholds and the number of samples in each population. Specifically, we have the following theorem that formalizes this statement.

**THEOREM 1.2** *Assume that the samples  $\{\mathbf{X}_k\}_{k=1}^{n_A}$  and  $\{\mathbf{Y}_k\}_{k=1}^{n_B}$  are both i.i.d. and mutually independent random vectors in  $\mathbb{R}^p$  with bounded elliptically contoured densities and block sparse  $p \times p$  covariance matrices. Let  $\rho_A$  and  $\rho_B$  be the associated correlation thresholds. Assume that  $p$  goes to  $\infty$  and  $\rho_A$  and  $\rho_B$  go to one according to the relations  $\lim_{p \rightarrow \infty} p^{1/2}(p-1)(1-\rho_A^2)^{(n_A-2)/2} = e_{n_A}$  and  $\lim_{p \rightarrow \infty} p^{1/2}(p-1)(1-\rho_B^2)^{(n_B-2)/2} = e_{n_B}$ . Then the probability  $P_e$  that there exists at least one false edge in the correlation flip network satisfies*

$$\lim_{p \rightarrow \infty} P_e = 1 - \exp(-\kappa_{n_A, n_B}), \quad (1.6)$$

where

$$\kappa_{n_A, n_B} = e_{n_A} e_{n_B} a_{n_A} a_{n_B} (n_A - 2)^{-1} (n_B - 2)^{-1} / 2$$

As in Thm. 1, Thm. 2 can be used to obtain a large  $p$  approximation to  $P_e$

$$P_e = 1 - \exp(-\lambda_{\rho_A, \rho_B, n_A, n_B}), \quad (1.7)$$

with

$$\lambda_{\rho_A, \rho_B, n_A, n_B} = e_{n_A} e_{n_B} p(p-1)^2 P_0(\rho_A, n_A) P_0(\rho_B, n_B) / 2,$$

and  $P_0(\cdot, \cdot)$  as given in (1.4).

Again the form of the limit (1.7) is the probability that Poisson random variable is not equal to zero, where the rate of the Poisson random variable is  $\lambda_{\rho_A, \rho_B, n_A, n_B}$ . This expression for the Poisson rate differs by a factor of 1/2 from the expression for the Poisson rate in the persistency-screening case considered in [1, Prop 3]. This is simply due to the fact that the version of persistency screening in [1, Prop 3] does not carry the restriction that the correlations be of opposite sign in the two populations. It can be shown, using the results of [2], that the theorem equally applies to mining edges in a partial correlation flip network.

To apply the theorem to correlation flip mining, the two thresholds  $\rho_A$  and  $\rho_B$  will not be chosen independently. Rather, as in persistency screening [1], they should be selected in a coupled manner according to the relation [1, Sec. 3.3] so as to equalize the mean number of false positive edges in the marginal networks  $G_A$  and  $G_B$ .

For illustration, we apply the large  $p$  approximation (1.7) to the ARI challenge study example shown in 1.4. There were  $n_A = 170$  Sx samples and  $n_B = 152$  Asx

samples collected in the study and the number of genes is  $p = 12023$ . The large  $p$  approximation to  $P_e$  specifies the two (coupled) thresholds that guarantee at most  $10^{-6}$  false edges in the correlation flip network:

$$\rho_A = 0.44, \quad \rho_B = 0.47.$$

This can be compared to the thresholds that guarantee the same error rate on false edges if the individual correlation networks were screening independently:

$$\rho_A = 0.54, \quad \rho_B = 0.57,$$

which simply reflects the fact that we can reliably detect lower value correlations in the correlation flip network since, for equal thresholds, false edges are rarer than in the individual correlation networks.

In the correlation flip example above we have deliberately not compared the results of our proposed threshold-based approach to other more complex optimization-based methods in the literature such as glasso, elastic net, or SPACE. These methods are simply not scalable to the dimension ( $p = 12023$ ) that is considered in the above example, at least not without making additional restrictive assumptions.

### 1.3.2 Large scale implementation of correlation mining

Only a subset of the correlations are required to construct a correlation network whose edges correspond to correlations exceeding the applied correlation threshold  $\rho$ . This fact makes large scale implementation of correlation mining scalable to large dimension  $p$ . In particular, it is not necessary to either store or compute the full correlation matrix  $\mathbf{R}$  in order to find the correlation graph. This is due to the following two reasons:

1. **Z-score representation of  $\mathbf{R}$  and  $\mathbf{P}$ :** The sample correlation matrix  $\mathbf{R}$  and the partial correlation matrix  $\mathbf{P}$ , as defined in (1.1), have Gram product representations [2, Sec. 2.3]

$$\mathbf{R} = \mathbb{T}^T \mathbb{T}, \quad \mathbf{P} = \mathbb{Y}^T \mathbb{Y}$$

where  $\mathbb{T} = [\mathbf{T}_1, \dots, \mathbf{T}_p]$  is an  $n \times p$  matrix and  $\mathbb{Y} = [\mathbf{Y}_1, \dots, \mathbf{Y}_p]$  is an  $(n-1) \times p$  matrix. The columns of these matrices are Z-scores that are defined in [2, Eq. (4)] and [2, Eq. (9)]. Therefore, all that needs to be stored are the smaller matrices  $\mathbb{T}$  and  $\mathbb{Y}$  (recall that  $n \ll p$ ). Furthermore, computation of the Z-score matrix  $\mathbb{Y}$  only requires a single  $(n-1) \times (n-1)$  matrix inversion. This property is especially useful in typical cases where  $n \ll p$ .

2. **Ball graph representation of correlation network:** Due to the Z-score representations above, the correlation and partial correlation networks are equivalent to ball graphs, also called a Euclidean proximity graphs, for which fast and scalable algorithms exist [69]. We explain this equivalence for the

correlation graph that thresholds  $\mathbf{R}$ ; the partial correlation graph equivalence is similarly explained.

A Euclidean proximity graph with ball radius  $r$  places an edge between vectors  $\mathbf{T}_i$  and  $\mathbf{T}_j$  in  $\mathbb{R}^n$  iff the Euclidean distance  $\|\mathbf{T}_i - \mathbf{T}_j\|$  does not exceed  $r$ . The ball graph is easily converted to a correlation network since two vectors  $\mathbf{T}_i, \mathbf{T}_j$  in  $\mathbb{R}^n$  with sample correlation  $\rho_{ij}$  are at mutual Euclidean distance  $\|\mathbf{T}_i - \mathbf{T}_j\| = \sqrt{2(1 - \rho_{ij})}$ . Hence if a ball graph with ball radius  $r = \sqrt{2(1 - \rho)}$  is constructed on the Z-score sample  $\{\mathbf{T}_i\}_{i=1}^p$  the resulting graph will be a one-sided positive correlation network, i.e. a network whose edges represent positive correlations exceeding  $\rho$ . A one-sided negative correlation network is obtained by applying the same ball graph construction with pairwise distances  $\|(-\mathbf{T}_i) - \mathbf{T}_j\| = \|\mathbf{T}_i + \mathbf{T}_j\|$ , resulting in a network whose edges represent negative correlations less than  $-\rho$ . Merging the two one sided correlation networks yields the correlation network.

By exploiting the Z-score and ball graph representations of the sample correlation and sample partial correlation, correlation mining algorithms can be implemented on very high dimensional biomarker spaces as long as the number of samples is small. Matlab and R code for implementing correlation mining algorithms for edge recovery and hub recovery is currently being developed for online access.

## 1.4 Perspectives and future challenges

As discussed above, one of the basic challenges in gleaning information from big data is the large  $p$  small  $n$  problem: a deluge of variables and a paucity of samples of these variables to adequately represent the population. This is of critical importance when one searches for correlations among the variables as correlations requires multiple samples. That this challenge is especially relevant to molecular biology applications is due to advances in high throughput sequencing technology: the cost of collecting additional variables ( $p$ ) is dropping much faster than the cost of obtaining additional independent samples ( $n$ ). Thus the theory and practice of correlation mining is well poised to tackle big data problems in the medical and biological sciences. However, not surprisingly, there exist challenges that will require future work. Before discussing these challenges, we summarize the current state of the art of knowledge in correlation mining.

### 1.4.1 State-of-the-art in correlation mining

As discussed in Sec. 1.1 the objective and formulation of correlation mining are different from those of covariance selection. Correlation mining seeks to locate a few nodes, edges or hubs associated with high (partial) correlation correlation while covariance selection seeks to estimate a sparse (inverse) covariance

matrix, thereby localizing the zero valued (partial) correlations. Put more succinctly, correlation mining is founded on testing a few high (partial) correlations while covariance selection is focused on minimization of approximation error for the entire sample (inverse) covariance. This difference leads to sharp finite sample theory and simplified implementation of correlation mining as compared to covariance selection.

To illustrate we consider covariance selection for recovering the inverse covariance matrix. As mentioned above, partial correlation networks arise naturally when the experimenter assumes a Gaussian graphical model for the data [10]. Under a sparse inverse covariance hypothesis, maximum likelihood and other parametric estimation methods can then be used to recover an approximation to the true network. There are many methods for partial correlation network recovery that are based on the GGM. These include among others 1) GeneNet [70] which uses a regularized generalized-inverse of the covariance and uses the bootstrap to estimate the most suitable regularization parameter, 2) the work of [57] which proposes two algorithms: a block coordinate descent method and another based on Nesterov's first order method to maximize the  $\ell_1$ -regularized Gaussian log-likelihood, and 3) the block coordinatewise method g-lasso of [52]. In recent years, a fast proximal Newton-like method (QUIC) was proposed by [58]. This work was followed up by even faster proximal gradients methods, G-ISTA and G-AMA, in [59] and [60] respectively. Currently, the state-of-the-art method in the field, G-AMA, is able to handle problem sizes as high as 5,000, though handling problem sizes of 10,000 or 20,000 is more challenging. Some of these methods are compared in [8].

There have also been efforts to go beyond the  $\ell_1$ -regularized Gaussian likelihood approach to  $\ell_1$ -regularized pseudo-likelihood and regression based approaches. These include methods such as "neighborhood selection" (NS) [31], SPACE [34], SPLICE [61] and SYMLASSO [71] (see also [72]). In a recent major contribution to this area, a convex pseudo-likelihood framework for high dimensional partial correlation estimation with convergence guarantees was recently proposed in [37]. This method, CONCORD, overcomes the many hurdles faced by existing pseudo-likelihood based methods but at the same time retains all their attractive properties. In particular, it yields an approach which is provably convergent, respects symmetry, enjoys asymptotic statistical guarantees, is computationally tractable, and has excellent finite sample performance. The CONCORD pseudo-likelihood in [37] is solved by coordinatewise minimization. In even more recent work, ISTA and FISTA type proximal gradient methods have been proposed in [62] to scale up the implementation of CONCORD. In particular, the authors in [62] demonstrate convincingly that pseudo-likelihood methods can also benefit from leveraging recent advances in convex optimization theory.

Despite the tremendous advances described above,  $\ell_1$ -regularized and other proposed methods are not currently as easily scalable due to the iterative nature of the corresponding algorithms. Unlike correlation mining methods described in

this manuscript, the methods above also do not have the tight theory provided by Theorem 1 to control false positive, and neither do they have the results obtained in [2].

The following summarizes some of the main features of the proposed correlation mining framework for biology applications.

- Correlation mining has its niche in the purely-high dimensional regime characterizing big biological datasets where  $n$  is fixed while  $p$  is large and increasing. This is where the asymptotic theory provides tight bounds on edge, node and hub recovery performance.
- In this purely-high dimensional regime correlation mining algorithms are scalable to very large  $p$ . This is because they are non-iterative, only involve inversion of  $n \times n$  matrices, and can be implemented using approximate nearest neighbor search algorithms.
- Correlation mining algorithms and theory apply equally to correlation mining in both (marginal) correlation networks and partial correlation networks.
- The theory provides a mathematical expression for the critical phase transition  $\rho_c$  in the mean number of false positive edges in the recovered correlation network as a function of the applied threshold  $\rho$ .
- In the purely-high dimensional regime, the theory indicates that the curse of high dimensionality in the variables can be counterbalanced by a certain *benefit of high dimensionality* in the samples: to maintain a given phase transition value only a small sample size increase is necessary to accommodate a big increase in biomarker dimension. This benefit diminishes as the sample size approaches the biomarker dimension.
- The theory can be used to determine the right sample size for the experimentalist to be able to reliably detect correlations that are above the phase transition threshold and are therefore both biologically significant and statistically significant.
- This sample sizing can be performed without knowledge of the underlying distribution or true covariance matrix in cases where the distribution is elliptically contoured, the covariance is block sparse, and the collected data are i.i.d.
- If the true data covariance cannot be assumed to be block sparse, this assumption can be replaced by the weaker assumption of row sparse covariance to obtain the same type of asymptotic limit as in Thm. 1.2 (see [1], [2]). However, without block sparsity the Poisson-type rate constant  $\kappa_n$  in the theorem may now be a function of the true covariance and the expression (1.7) for the false positive rate may no longer be accurate. An empirical estimate of the true rate constant may be constructed by estimating a certain “J functional” of the density (see [73]), which may be used to determine proper sample size. This would be practical when sufficient amounts of baseline data exist on which to train, e.g. obtained by pooling publicly available data from prior studies, or when a model for baseline is known.

- If the collected data cannot be assumed to be i.i.d. and has either some temporal or spatial dependency, our correlation mining framework is still flexible enough to deal with such settings. First, correlation screening thresholds from the i.i.d. setup can still be used as lower bounds for the correct threshold when dependency is present. This follows from the fact that screening thresholds are a decreasing function of sample size. Second, if the dependency structure is known such as in an AR(1) model, the effective sample size can be used to adjust the screening threshold appropriately. Third, if the data is a Gaussian temporally stationary multivariate time series then the spectral correlation screening framework of [74] can be applied.

#### 1.4.2 Future challenges in correlation mining biomolecular networks

There remain several challenges and open problems in large scale correlation mining of molecular networks in biology and biomedicine. Many of these align with currently recognized grand challenges in engineering and the life sciences [75]. We discuss several opportunities and challenges below, with a focus on health and medicine applications.

Correlation mining can play an important role in emerging translational medicine applications such as: personalized medicine [76]; genomic medicine [77]; and network medicine [78]. In personalized medicine, correlation mining might be used to identify personalized molecular signatures and their nominal variations in order to specify a baseline of health for an individual. A principal objective of genomic medicine is early detection of disease by testing for activation of the gene expression pathways in the host immune system that presage inflammatory response and acute symptoms. Network medicine recognizes the dynamic nature of molecular networks regulating pathogenesis and immune response and proposes precisely timed drug treatments that target certain transitions in the network. Due to the potentially large number of biomarker variables and limited samples, correlation mining can be a key player in determining the complex molecular interactions for these translational areas.

There has been much recent interest in biochronicity and other time varying phenomena underlying gene regulation occurring in healthy hosts. Biomarkers that manifest as periodically changing over time are plentiful, e.g., clock genes that regulate 24 hour circadian and 12 hour hemicircadian rhythms of monocyte cell-cycles [79]. Quantifying and understanding the role of biochronicity in gene regulation is of independent interest but could also lead to periodic detrending of other biomarkers. For example, this could result in greatly improving sensitivity to subtle pathogen induced changes and lead to improved early detection performance. Fourier-based correlation mining extensions, such as developed in [80], [74], could possibly be applied to detecting the biochronicity genes as they manifest their periodic components over time. In this extension, correlation mining is performed independently on each complex-valued DFT coefficient of the

time varying variables. A challenge that would need to be overcome is the non-sinusoidal nature of many biochronicity signals. The

There are continuing efforts to establish a biomarker-based baseline of health for preventative medicine. Molecular biomarker signatures are highly variable across time for a single healthy individual and across individuals for individuals in a healthy population. Some of this variability can be explained by fixed factors such as age and gender or deterministic factors. e.g., trait-related gene expression variation [81], [82], or biochronicity, discussed above. Such deterministic factors might be removed using simple detrending or regression approaches. However, much of the variability of a healthy persons molecular patterns remains unexplained and might be better modeled as random over time but correlated over different biomarkers. Under this model, correlation mining could be used to identify the principal hubs and edges in the correlation network that characterize a persons healthy baseline. If some non-healthy samples were available, these correlations could be incorporated into a logistic regression to detect abnormal deviations from baseline. Alternatively, the predictive correlation screening framework of [83] could be used to tune correlation mining specifically to the task of prediction.

Multiplatform assays probe complementary manifestations of biological system states and behaviors. For example, challenge study protocols may simultaneously assay a tissue sample for gene expression, biased and unbiased protein expression, metabolite levels, and antibodies. Correlation mining and other statistical methods of correlation analysis have been principally developed for a single platform. Development of a theory and practice of multiplatform correlation mining would be worthwhile. Principal hurdles would be the need to account for cross-platform calibration, multiplatform normalization, and different levels of technical noise inherent to each platform. Several approaches to correlation networks for multiplatform assays include: multiattribute networks [84], which create separate nodes for each platform; multivariate normal networks [85] and Kronecker network [65], which accommodate networks with vector valued nodes. While the models used in these approaches are worth considering, the methods are not immediately extendible to correlation mining as they all focus on covariance estimation. Exploring extensions of these models to the testing framework of correlation mining would be worthwhile pursuing further.

Integration of ontological information relevant to biological function is commonly used to place an investigator's empirical findings in the context of the global knowledge base. For example, a gene list consisting of a discovered hub gene and its neighboring genes might be analyzed in the context of previously reported molecular interactions and pathways using PID [6]. More generally, ontologies can reveal transcription, methylation and phosphorylation correspondences; gene-protein intra-actions and inter-actions; molecular-disease associations; epigenetics; and other types of phenotype specific data. Such data is often in the form of text, trees and graphs that can be integrated into the correlation network discovered via correlation mining. However, a major challenge is deter-

---

mining provenance and reliability of the ontological data. Another challenge is the direct use of ontological data as side information that can directly affect the correlation mining outcomes. Possible approaches to integration of such information into correlation mining include: correlation weighting to de-emphasize connections not supported by ontology [86], non-uniform thresholding of the correlation entries, and direct incorporation of ontologically determined interactions and non-interactions as topological constraints [87],[88].

## 1.5 Conclusion

The biomedical sciences are one of the major generators of big data of our era. An aspect of such data is that with the emergence of next-generation sequencing technology, it is becoming far less costly to collect more biomarker probes ( $p$  variables) than it is to collect more biological replicates ( $n$  samples). This high dimensional regime will be much better characterized by the “purely-high dimensional” setting (large  $p$  and fixed finite  $n$ ) than the standard “ultra-high dimensional setting” (large  $p$  and large  $n$ ) common in current high dimensional statistics. It is in this purely-high dimensional regime that our scalable correlation mining frameworks will play a significant role in extracting information and making sense of the next generation of biological data.

### Acknowledgements

The authors gratefully acknowledge several people with whom interactions have been relevant to the work described in this chapter. In particular we thank Rob Brown at UCLA. We also thank Geoff Ginsburg, Tim Veldman, Chris Woods, and Aimee Zaas, at the Duke University School of Medicine, and Euan Ashley, Phil Tsao, Frederick Dewey, at the Stanford University School of Medicine, Division of Cardiovascular Medicine and the Cardiovascular Institute (CVI). The authors acknowledge Michael Tsiang for LaTeX assistance. The authors also thank Joseph Romano for assistance with the naming convention for the different asymptotic regimes. This work was partially supported by the Air Force Office of Scientific Research under grant FA9550-13-1-0043. B.R. was also supported in part by the National Science Foundation under Grant Nos. DMS-1106642, DMS-CMG-1025465 and DMS CAREER-1352656.

## References

- [1] A. Hero and B. Rajaratnam, “Large-scale correlation screening,” *Journal of the American Statistical Association*, vol. 106, no. 496, pp. 1540–1552, 2011.
- [2] —, “Hub discovery in partial correlation models,” *IEEE Trans. on Inform. Theory*, vol. 58, no. 9, pp. 6064–6078, 2012, available as Arxiv preprint arXiv:1109.6846.
- [3] M. Ashburner, C. A. Ball, J. A. Blake, D. Botstein, H. Butler, J. M. Cherry, A. P. Davis, K. Dolinski, S. S. Dwight, J. T. Eppig, *et al.*, “Gene ontology: tool for the unification of biology,” *Nature genetics*, vol. 25, no. 1, pp. 25–29, 2000.
- [4] G. Dennis Jr, B. T. Sherman, D. A. Hosack, J. Yang, W. Gao, H. C. Lane, R. A. Lempicki, *et al.*, “David: database for annotation, visualization, and integrated discovery,” *Genome biol*, vol. 4, no. 5, p. P3, 2003.
- [5] E. W. Sayers, T. Barrett, D. A. Benson, E. Bolton, S. H. Bryant, K. Canese, V. Chetvernin, D. M. Church, M. DiCuccio, S. Federhen, *et al.*, “Database resources of the national center for biotechnology information,” *Nucleic acids research*, vol. 39, no. suppl 1, pp. D38–D51, 2011.
- [6] C. F. Schaefer, K. Anthony, S. Krupa, J. Buchoff, M. Day, T. Hannay, and K. H. Buetow, “Pid: the pathway interaction database,” *Nucleic acids research*, vol. 37, no. suppl 1, pp. D674–D679, 2009.
- [7] E. G. Cerami, B. E. Gross, E. Demir, I. Rodchenkov, Ö. Babur, N. Anwar, N. Schultz, G. D. Bader, and C. Sander, “Pathway commons, a web resource for biological pathway data,” *Nucleic acids research*, vol. 39, no. suppl 1, pp. D685–D690, 2011. [Online]. Available: <http://www.pathwaycommons.org/about/>
- [8] J. D. Allen, Y. Xie, M. Chen, L. Girard, and G. Xiao, “Comparing statistical methods for constructing large scale gene networks,” *PloS one*, vol. 7, no. 1, p. e29348, 2012.
- [9] C. Jiang, F. Coenen, and M. Zito, “A survey of frequent subgraph mining algorithms,” *The Knowledge Engineering Review*, vol. 28, no. 01, pp. 75–105, 2013.
- [10] P. Bühlmann and S. van de Geer, *Statistics for High-Dimensional Data: Methods, Theory and Applications*. Springer, 2011.
- [11] R. Fisher, “On the mathematical foundations of theoretical statistics,” *Philosophical Transactions of the Royal Society of London, Series A*, vol. 222, pp. 309–368, 1922.
- [12] —, “Theory of statistical estimation,” *Proceedings of the Cambridge Philosophical Society*, vol. 22, pp. 700–725, 1925.

- 
- [13] C. Rao, "Large sample tests of statistical hypotheses concerning several parameters with applications to problems of estimation," *Mathematical Proceedings of the Cambridge Philosophical Society*, vol. 44, pp. 50–57, 1947.
- [14] —, "Criteria of estimation in large samples," *Sankhyā: The Indian Journal of Statistics, Series A*, vol. 25, pp. 189–206, 1963.
- [15] J. Neyman and E. Pearson, "On the problem of the most efficient tests of statistical hypotheses," *Philosophical Transactions of the Royal Society of London, Series A*, vol. 231, pp. 289–337, 1933.
- [16] S. Wilks, "The large-sample distribution of the likelihood ratio for testing composite hypotheses," *Annals of Mathematical Statistics*, vol. 9, pp. 60–62, 1938.
- [17] A. Wald, "Asymptotically most powerful tests of statistical hypotheses," *Annals of Mathematical Statistics*, vol. 12, pp. 1–19, 1941.
- [18] —, "Some examples of asymptotically most powerful tests," *Annals of Mathematical Statistics*, vol. 12, pp. 396–408, 1941.
- [19] —, "Tests of statistical hypotheses concerning several parameters when the number of observations is large," *Transactions of the American Mathematical Society*, vol. 54, pp. 426–482, 1943.
- [20] —, "Note on the consistency of the maximum likelihood estimate," *Annals of Mathematical Statistics*, vol. 20, pp. 595–601, 1949.
- [21] H. Cramér, *Mathematical Methods of Statistics*. Princeton, NJ: Princeton University Press, 1946.
- [22] —, "A contribution to the theory of statistical estimation," *Scandinavian Actuarial Journal*, vol. 29, pp. 85–94, 1946.
- [23] L. Le Cam, "On some asymptotic properties of maximum likelihood estimates and related Bayes' estimates," *University of California publications in statistics*, vol. 1, pp. 277–330, 1953.
- [24] —, *Asymptotic Methods in Statistical Decision Theory*. New York: Springer-Verlag, 1986.
- [25] H. Chernoff, "Large-sample theory: Parametric case," *Annals of Mathematical Statistics*, vol. 27, pp. 1–22, 1956.
- [26] J. Kiefer and J. Wolfowitz, "Consistency of the maximum likelihood estimator in the presence of infinitely many incidental parameters," *Annals of Mathematical Statistics*, vol. 27, pp. 887–906, 1956.
- [27] R. Bahadur, "Rates of convergence of estimates and test statistics," *Annals of Mathematical Statistics*, vol. 38, pp. 303–324, 1967.
- [28] B. Efron, "Maximum likelihood and decision theory," *Annals of Statistics*, vol. 10, pp. 340–356, 1982.
- [29] D. Donoho, "For most large underdetermined systems of linear equations the minimal  $\ell_1$ -norm solution is also the sparsest solution," *Communications on Pure and Applied Mathematics*, vol. 59, pp. 797–829, 2006.
- [30] P. Zhao and B. Yu, "On model selection consistency of lasso," *Journal of Machine Learning Research*, vol. 7, pp. 2541–2563, 2006.
- [31] N. Meinshausen and P. Bühlmann, "High-dimensional graphs and variable selection with the lasso," *Annals of Statistics*, vol. 34, no. 3, pp. 1436–1462, June 2006.
- [32] E. Candès and T. Tao, "The Dantzig selector: statistical estimation when  $p$  is much larger than  $n$ ," *Annals of Statistics*, vol. 35, pp. 2313–2351, 2007.

- 
- [33] P. Bickel, Y. Ritov, and A. Tsybakov, "Simultaneous analysis of Lasso and Dantzig selector," *Annals of Statistics*, vol. 37, pp. 1705–1732, 2009.
- [34] J. Peng, P. Wang, N. Zhou, and J. Zhu, "Partial correlation estimation by joint sparse regression models," *Journal of the American Statistical Association*, vol. 104, no. 486, 2009.
- [35] M. Wainwright, "Information-theoretic limitations on sparsity recovery in the high-dimensional and noisy setting," *IEEE Transactions on Information Theory*, vol. 55, pp. 5728–5741, 2009.
- [36] —, "Sharp thresholds for high-dimensional and noisy sparsity recovery using  $\ell_1$ -constrained quadratic programming (Lasso)," *IEEE Transactions on Information Theory*, vol. 55, pp. 2183–2202, 2009.
- [37] K. Khare, S. Oh, and B. Rajaratnam, "A convex pseudo-likelihood framework for high dimensional partial correlation estimation with convergence guarantees," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, to appear, 2014. [Online]. Available: <http://arxiv.org/abs/1307.5381>
- [38] H. Firouzi, A. Hero, and B. Rajaratnam, "Variable Selection for ultra high dimensional regression," *Technical Report, University of Michigan and Stanford University*, 2014.
- [39] B. Mole, "The gene sequencing future is here," *Science News*, Feb 6, 2014. [Online]. Available: <https://www.sciencenews.org/article/gene-sequencing-future-here>
- [40] W. KA, "Dna sequencing costs: Data from the nhgri genome sequencing program (gsp)," Aug. 22, 2014. [Online]. Available: <https://www.sciencenews.org/article/gene-sequencing-future-here>
- [41] A. Zaas, M. Chen, J. Varkey, T. Veldman, A. Hero, J. Lucas, Y. Huang, R. Turner, A. Gilbert, R. Lambkin-Williams, *et al.*, "Gene Expression Signatures Diagnose Influenza and Other Symptomatic Respiratory Viral Infections in Humans," *Cell Host & Microbe*, vol. 6, no. 3, pp. 207–217, 2009.
- [42] Y. Huang, A. Zaas, A. Rao, N. Dobigeon, P. Woolf, T. Veldman, N. Øien, M. McClain, J. Varkey, B. Nicholson, *et al.*, "Temporal dynamics of host molecular responses differentiate symptomatic and asymptomatic influenza a infection," *PLoS Genet*, vol. 7, no. 8, p. e1002234, 2011.
- [43] P. J. Bickel and K. A. Doksum, *Mathematical Statistics: Basic Ideas and Selected Topics*. Holden-Day, San Francisco, 1977.
- [44] H. Jeong, S. P. Mason, A.-L. Barabasi, and Z. N. Oltvai, "Lethality and centrality in protein networks," *Nature*, vol. 411, no. 6833, pp. 41–42, May 2001. [Online]. Available: <http://dx.doi.org/10.1038/35075138>  
<http://www.nature.com/nature/journal/v411/n6833/abs/411041a0.html>
- [45] M. C. Oldham, S. Horvath, and D. H. Geschwind, "Conservation and evolution of gene coexpression networks in human and chimpanzee brains," *Proceedings of the National Academy of Sciences*, vol. 103, no. 47, pp. 17 973 –17 978, Nov. 2006. [Online]. Available: <http://www.pnas.org/content/103/47/17973.abstract>
- [46] P. Langfelder and S. Horvath, "WGCNA: an R package for weighted correlation network analysis." *BMC bioinformatics*, vol. 9, no. 1, p. 559, Jan. 2008. [Online]. Available: <http://www.biomedcentral.com/1471-2105/9/559>
- [47] A. Li and S. Horvath, "Network neighborhood analysis with the multi-node topological overlap measure." *Bioinformatics (Oxford, Eng-*

- land), vol. 23, no. 2, pp. 222–31, Jan. 2007. [Online]. Available: <http://bioinformatics.oxfordjournals.org/cgi/content/abstract/23/2/222>
- [48] L. Wu, C. Zhang, and J. Zhang, “Hmbox1 negatively regulates nk cell functions by suppressing the nkg2d/dap10 signaling pathway,” *Cellular & molecular immunology*, vol. 8, no. 5, pp. 433–440, 2011.
- [49] A. Y. Istomin and A. Godzik, “Understanding diversity of human innate immunity receptors: analysis of surface features of leucine-rich repeat domains in nlrs and tlrs,” *BMC immunology*, vol. 10, no. 1, p. 48, 2009.
- [50] S. L. Lauritzen, *Graphical Models*. Oxford University Press, 1996.
- [51] A. Dempster, “Covariance selection,” *Biometrics*, vol. 28, no. 1, pp. 157–175, 1972.
- [52] J. Friedman, T. Hastie, and R. Tibshirani, “Sparse inverse covariance estimation with the graphical lasso,” *Biostatistics*, vol. 9, no. 3, pp. 432–441, 2008.
- [53] B. Rajaratnam, H. Massam, and C. Carvalho, “Flexible covariance estimation in graphical gaussian models,” *Annals of Statistics*, vol. 36, pp. 2818–2849, 2008.
- [54] K. Khare and B. Rajaratnam, “Wishart distributions for decomposable covariance graph models,” *The Annals of Statistics*, vol. 39, no. 1, pp. 514–555, Mar. 2011. [Online]. Available: <http://projecteuclid.org/euclid.aos/1297779855>
- [55] A. J. Rothman, P. Bickel, E. Levina, and J. Zhu, “Sparse permutation invariant covariance estimation,” *Electronic Journal of Statistics*, vol. 2, pp. 494–515, 2008.
- [56] P. Bickel and E. Levina, “Covariance regularization via thresholding,” *Annals of Statistics*, vol. 34, no. 6, pp. 2577–2604, 2008.
- [57] O. Banerjee, L. E. Ghaoui, and A. d’Aspremont, “Model selection through sparse maximum likelihood estimation for multivariate Gaussian or binary data,” *Journal of Machine Learning Research*, vol. 9, pp. 485–516, March 2008.
- [58] C.-J. Hsieh, M. A. Sustik, I. Dhillon, P. Ravikumar, and R. Poldrack, “Big & quic: Sparse inverse covariance estimation for a million variables,” in *Advances in Neural Information Processing Systems*, 2013, pp. 3165–3173.
- [59] D. Guillot, B. Rajaratnam, B. T. Rolfs, A. Maleki, and I. Wong, “Iterative Thresholding Algorithm for Sparse Inverse Covariance Estimation,” in *Advances in Neural Information Processing Systems 25*, 2012. [Online]. Available: <http://arxiv.org/abs/1211.2532>
- [60] O. Dalal and B. Rajaratnam, “G-AMA: Sparse Gaussian Graphical Model Estimation via Alternating Minimization,” *Technical Report, Department of Statistics, Stanford University (in revision)*, 2014. [Online]. Available: <http://arxiv.org/abs/1405.3034>
- [61] G. Rocha, P. Zhao, and B. Yu, “A path following algorithm for Sparse Pseudo-Likelihood Inverse Covariance Estimation (SPLICE),” Statistics Department, UC Berkeley, Berkeley, CA, Tech. Rep., 2008. [Online]. Available: <http://www.stat.berkeley.edu/~binyu/ps/rocha.pseudo.pdf>
- [62] S. Oh, O. Dalal, K. Khare, and B. Rajaratnam, “Optimization methods for sparse pseudo-likelihood graphical model selection,” in *Advances in Neural Information Processing Systems 27*, 2014.
- [63] G. Marjanovic and A. O. Hero III, “On lq estimation of sparse inverse covariance,” in *Proc. of IEEE Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, Florence, May 2014.
- [64] —, “l<sub>0</sub> sparse inverse covariance estimation,” *arXiv preprint arXiv:1408.0850*, 2014.

- [65] T. Tsiligkaridis, A. Hero, and S. Zhou, "Convergence properties of Kronecker Graphical Lasso algorithms," *IEEE Trans on Signal Processing (also available as arXiv:1204.0585)*, vol. 61, no. 7, pp. 1743–1755, 2013.
- [66] R. Gill, S. Datta, and S. Datta, "A statistical framework for differential network analysis from microarray data," *BMC bioinformatics*, vol. 11, no. 1, p. 95, 2010.
- [67] N. Kramer, J. Schafer, and A.-L. Boulesteix, "Regularized estimation of large-scale gene association networks using graphical gaussian models," *BMC Bioinformatics*, vol. 10, no. 384, pp. 1–24, 2009.
- [68] V. Pihur, S. Datta, and S. Datta, "Reconstruction of genetic association networks from microarray data: a partial least squares approach," *Bioinformatics*, vol. 24, no. 4, p. 561, 2008.
- [69] D. Mount and S. Arya, "Approximate Nearest Neighbor Code," <http://www.cs.umd.edu/~mount/ANN>.
- [70] J. Schäfer and K. Strimmer, "An empirical bayes approach to inferring large-scale gene association networks," *Bioinformatics*, vol. 21, no. 6, pp. 754–764, 2005.
- [71] J. Friedman, T. Hastie, and R. Tibshirani, "Applications of the lasso and grouped lasso to the estimation of sparse graphical models," 2010. [Online]. Available: <http://www-stat.stanford.edu/~tibs/research.html>
- [72] J. Lee and T. Hastie, "Learning the structure of mixed graphical models," *Journal of Computational and Graphical Statistics, to appear*, 2014.
- [73] K. Sricharan, A. Hero, and B. Rajaratnam, "A local dependence measure and its application to screening for high correlations in large data sets," in *Information Fusion (FUSION), 2011 Proceedings of the 14th International Conference on*. IEEE, 2011, pp. 1–8.
- [74] H. Firouzi, D. Wei, and A. Hero, "Spectral correlation hub screening of multivariate time series," in *Excursions in Harmonic Analysis: The February Fourier Talks at the Norbert Wiener Center*, R. Balan, M. Begué, J. J. Benedetto, W. Czaja, and K. Okoudjou, Eds. Springer, 2014.
- [75] B. He, R. Baird, R. Butera, A. Datta, S. George, B. Hecht, A. Hero, G. Lazzi, R. C. Lee, J. Liang, *et al.*, "Grand challenges in interfacing engineering with life sciences and medicine." *IEEE Transactions on bio-medical engineering (BME)*, vol. 4, no. 4, 2013.
- [76] R. Chen, G. I. Mias, J. Li-Pook-Than, L. Jiang, H. Y. Lam, R. Chen, E. Miriami, K. J. Karczewski, M. Hariharan, F. E. Dewey, *et al.*, "Personal omics profiling reveals dynamic molecular and medical phenotypes," *Cell*, vol. 148, no. 6, pp. 1293–1307, 2012.
- [77] J. J. McCarthy, H. L. McLeod, and G. S. Ginsburg, "Genomic medicine: a decade of successes, challenges, and opportunities," *Science translational medicine*, vol. 5, no. 189, pp. 189sr4–189sr4, 2013.
- [78] J. T. Erler and R. Linding, "Network medicine strikes a blow against breast cancer," *Cell*, vol. 149, no. 4, pp. 731–733, 2012.
- [79] D. B. Boivin, F. O. James, A. Wu, P. F. Cho-Park, H. Xiong, and Z. S. Sun, "Circadian clock genes oscillate in human peripheral blood mononuclear cells," *Blood*, vol. 102, no. 12, pp. 4143–4145, 2003.
- [80] H. Firouzi, D. Wei, and A. Hero, "Spatio-temporal analysis of gaussian wss processes via complex correlation and partial correlation screening," in *Proceedings of IEEE GlobalSIP Conference. Also available as arxiv:1303.2378*, 2013.

- 
- [81] J. J. Eady, G. M. Wortley, Y. M. Wormstone, J. C. Hughes, S. B. Astley, R. J. Foxall, J. F. Doleman, and R. M. Elliott, "Variation in gene expression profiles of peripheral blood mononuclear cells from healthy volunteers," *Physiological Genomics*, vol. 22, no. 3, pp. 402–411, 2005.
  - [82] A. R. Whitney, M. Diehn, S. J. Popper, A. A. Alizadeh, J. C. Boldrick, D. A. Relman, and P. O. Brown, "Individuality and variation in gene expression patterns in human blood," *Proceedings of the National Academy of Sciences*, vol. 100, no. 4, pp. 1896–1901, 2003.
  - [83] H. Firouzi, A. Hero, and B. Rajaratnam, "Predictive correlation screening: Application to two-stage predictor design in high dimension," in *Proceedings of AIS-TATS*. Also available as *arxiv:1303.2378*, 2013.
  - [84] N. Katenka, E. D. Kolaczyk, *et al.*, "Inference and characterization of multi-attribute networks with application to computational biology," *The Annals of Applied Statistics*, vol. 6, no. 3, pp. 1068–1094, 2012.
  - [85] S. Zhou, "Gemini: Graph estimation with matrix variate normal instances," *The Annals of Statistics*, vol. 42, no. 2, pp. 532–562, 2014.
  - [86] P. Langfelder and S. Horvath, "Wgcna: an r package for weighted correlation network analysis," *BMC bioinformatics*, vol. 9, no. 1, p. 559, 2008.
  - [87] D. Zhu, A. Hero, H. Cheng, R. Kanna, and A. Swaroop, "Network constrained clustering for gene microarray data," *Bioinformatics*, vol. 21, no. 21, pp. 4014–4021, 2005.
  - [88] A. Rao and A. O. Hero, "Biological pathway inference using manifold embedding," in *Acoustics, Speech and Signal Processing (ICASSP), 2011 IEEE International Conference on*. IEEE, 2011, pp. 5992–5995.