Relating balance and conditional independence in graphical models

Alberto Zenere, Erik G. Larsson, and Claudio Altafini [*]
Department of Electrical Engineering, Linköping University, SE-58183 Linköping, Sweden

(Received 29 March 2022; accepted 13 September 2022; published 28 October 2022)

When data are available for all nodes of a Gaussian graphical model, then, it is possible to use sample correlations and partial correlations to test to what extent the conditional independencies that encode the structure of the model are indeed verified by the data. In this paper, we give a heuristic rule useful in such a validation process: When the correlation subgraph involved in a conditional independence is balanced (i.e., all its cycles have an even number of negative edges), then a partial correlation is usually a contraction of the corresponding correlation, which often leads to conditional independence. In particular, the contraction rule can be made rigorous if we look at concentration subgraphs rather than correlation subgraphs. The rule is applied to real data for elementary gene regulatory motifs.

DOI: 10.1103/PhysRevE.106.044309

I. INTRODUCTION

Graphical models are a popular tool for representing relationships among variables in a wide variety of contexts from Medicine to Biology from Socio-Economical Sciences to Psychology [1–6].

When in a graphical model the joint probability distribution is Gaussian, the dependency structure among the variables is completely represented by the so-called concentration matrix K, i.e., the inverse of the covariance matrix Σ. If we look at the graph whose adjacency matrix is K, then absence of an edge between the ith and the jth nodes ($K_{ij} = 0$) can be encoded as an independence relationship between the ith and the jth variables, marginal or conditioned on some other variables of the graph. In particular, the whole topology determined by K can be mapped into a set of conditional independencies, often referred to as Markov conditions [7]. This mapping can be achieved in both undirected and directed graphical models with only minor differences, the most important being the causality interpretation which can be attached to the directed case, and which leads to graphs having the structure of Directed Acyclic Graphs (DAGs).

When sample data are available for the random variables that form the graphical model, then K can be estimated by inverting the sample covariance matrix. This is known to be an optimal estimator when the sample size goes to infinity [8]. When instead the sample size is small, then inversion of the sample covariance is an ill-posed operation, and many alternative methods have been developed in the literature for topology inference or validation [8–14].

The setting we deal with in this paper is this latter one: We assume that we know the underlying topology of the graphical model and that we have a small number of observations available for all nodes of the graph. Our aim is, therefore, to use these data to verify how much of the conditional independencies that correspond to the topology is "visible" in the data, and to provide criteria for discriminating which conditional independencies are more likely to be validated by the data, and which are less likely and hence can be labeled as false positives. As we assume to be dealing with Gaussian multivariates, measures based on sample correlations are suitable for this task. In particular, it is well known that independence conditioned on certain variables is equivalent to vanishing of the corresponding partial correlation, i.e., vanishing of the correlation among the residuals obtained after projecting away the contribution of the conditioning variables. For sample data, conditional independencies can, therefore, be associated with sample partial correlations that are "small enough" [15].

The contribution of this paper is to relate partial correlation with the properties of the associated correlation graph. In particular, a correlation graph is a signed graph, and the property of signed graphs which we want to highlight here is the so-called *structural balance* (henceforth, for brevity, *balance*): a signed graph is balanced if all its cycles are positive, i.e., have an even number of negative edges, see Refs. [16–18]. The property is also called *frustration free* in the Statistical Physics literature [19–21]. It is closely related to the notion of positive association among random variables [22]. In particular, we provide a couple of heuristic rules, valid for small sample size:

(1) Partial correlations associated with balanced correlation graphs tend to be smaller in absolute value than those associated with unbalanced correlation graphs;

(2) partial correlations associated with balanced correlation graphs tend to be contractions of the corresponding correlations (contraction is intended in an elementwise sense:

[*] Corresponding author: claudio.altafini@liu.se

A partial correlation coefficient is less in absolute value than the corresponding correlation coefficient).

Both rules point to the fact that balanced correlation graphs tend to lead to validation of conditional independencies more often than unbalanced ones. It is straightforward to show on examples that both rules are only heuristic and not obeyed strictly. The second, however, can be made into a rigorous law if instead of correlation graphs we look at concentration graphs. The equivalent of balance for concentration graphs is an analogous property which we call *inverse balance* but which in Ref. [23] is called *signed* MTP_2 (multivariate totally positive of order 2). MTP_2 is a stronger form of positivity, which corresponds to the concentration matrix K being an M-matrix ([24]; see below for details of this characterization). Indeed we show in the paper that

(2′) partial covariances associated with inverse balanced correlation graphs are always contractions of the corresponding covariances.

Hence, for inverse balance, indeed, conditioning leads to contractions, which points even more to the balance property as being a predictor of the validity of a conditional independence.

As an application of the rules above to empirical data, we consider length-2 causal chains in high-resolution models of gene transcription and protein synthesis. Here the task is to discriminate which chains are true positives, in the sense that the associated conditional independence is validated by our omics data. It is shown that there is a systematic difference between the validation rate obtained in the balanced and inverse balanced case and that obtained in the unbalanced and inverse unbalanced case. In particular, balance and inverse balance appear to be good predictors of conditional independence, in accordance with the above-mentioned rules. These results are qualitatively similar to other studies we recently carried out on different datasets and/or different network motifs, see Refs. [25,26]. They are here integrated by a more thorough theoretical analysis of the problem.

II. GAUSSIAN GRAPHICAL MODELS AND CONDITIONAL INDEPENDENCE

Let us consider a Gaussian Bayesian network, i.e., a probabilistic graphical model on a DAG in which to each node is associated a Gaussian random variable X_i. Denote $\mathcal{X} = \{X_1, \dots, X_n\} \sim \mathcal{N}(0, \Sigma)$, with Σ an $n \times n$ positive definite (p.d.) covariance matrix. We assume that the joint probability distribution that represents the DAG factorizes according to the structure of the DAG, i.e., according to the topology determined by the concentration matrix $K = \Sigma^{-1}$ with the directionality of the edges prescribed by the DAG. Such factorization determines a set of conditional independencies (also called Markov conditions),

$$X_i \perp X_j | \mathcal{S}, \tag{1}$$

where the symbol "⊥" means independence, "| " means conditioning and $\mathcal{S}$ is a separating set for X_i, see Ref. [7]. We also assume that the nodes of the DAG are "well ordered" [7] so that local and global Markov conditions coincide.

For instance, for the length-2 chain $X_1 \to X_2 \to X_3$ shown in Fig. 1(a), the joint probability distribution factorizes as

$$p(X_1, X_2, X_3) = p(X_1)p(X_2|X_1)p(X_3|X_2).$$

Conditioning on X_2 and using Bayes' rule, we get

$$p(X_1, X_3|X_2) = \frac{p(X_1, X_2, X_3)}{p(X_2)} = p(X_1|X_2)p(X_3|X_2),$$

which shows that X_1 and X_3 are conditionally independent given X_2: $X_1 \perp X_3 \,|\, X_2$, see Ref. [27].

III. CORRELATION, PARTIAL CORRELATION, AND CONDITIONAL INDEPENDENCE

From $K = \Sigma^{-1}$, consider the normalized versions of both Σ and K, i.e., the correlation matrix $R = \Delta_1^{-1}\Sigma\Delta_1^{-1}$, where Δ_1 is the diagonal matrix $\Delta_1 = \mathrm{diag}(\sqrt{\Sigma_{11}}, \dots, \sqrt{\Sigma_{nn}})$, and $H = \Delta_2^{-1}K\Delta_2^{-1}$, where, analogously, $\Delta_2 = \mathrm{diag}(\sqrt{K_{11}}, \dots, \sqrt{K_{nn}})$. R and H are then related by $H = \Delta_3^{-1}R^{-1}\Delta_3^{-1}$ where $\Delta_3 = \Delta_1\Delta_2$. Both R and H are obviously p.d. when Σ, respectively, K are, and have the same sign pattern as Σ, respectively, K. Note that even if K (and H) is sparse, Σ (and R) is, in general, a dense matrix. The matrix H (and, hence, K) is related to the partial correlation matrix P, whose $(i,\ j)$th entry, denoted $R_{X_iX_j.\mathcal{S}}$, is the partial correlation between variables X_i and X_j given the remaining $n-2$ variables $\mathcal{S} = \mathcal{X} \setminus \{X_i, X_j\}$ and is defined as

$$R_{X_iX_j.\mathcal{S}} = -\frac{(R^{-1})_{ij}}{\sqrt{(R^{-1})_{ii}(R^{-1})_{jj}}}. \tag{2}$$

In matrix form

$$P = [R_{X_iX_j.\mathcal{S}}] = 2I - H. \tag{3}$$

Formula (3) shows that the normalized concentration matrix H and the partial correlation matrix P have the same nonzero pattern but with opposite signs in the off-diagonal part. P is, in general, not p.d. Conditional independence between X_i and X_j corresponds to absence of the $(i,\ j)$ edge in the concentration matrix,

$$X_i \perp X_j | \mathcal{S} \iff K_{ij} = H_{ij} = 0, \tag{4}$$

which, in turn, corresponds also to $R_{X_iX_j.\mathcal{S}} = 0$.

On sample data, the exact test (4) is normally replaced by a statistical test, such as

$$X_i \perp X_j | \mathcal{S} \iff |R_{X_iX_j.\mathcal{S}}| < \theta, \tag{5}$$

where θ is a significance threshold obtained, e.g., through a Fisher test, see Ref. [15]. This is especially needed when the sample size m is small (comparable to the number of variables n) because the matrix inversion operation in (2) becomes ill conditioned even when R is p.d. This ill conditioning calls for alternative tests to be developed in order to check conditional independence on sample data, topic which is discussed next.

IV. BALANCE AND INVERSE BALANCE ON GRAPHICAL MODELS

For Gaussian variables X_i and X_j, one says that X_i and X_j are *positively associated* if for all monotone functions f and g, $R_{f(X_i)g(X_j)} \geqslant 0$, which, in particular, implies that $R_{X_iX_j} \geqslant 0$ [22]. Consequently, if the entire vector of Gaussian random

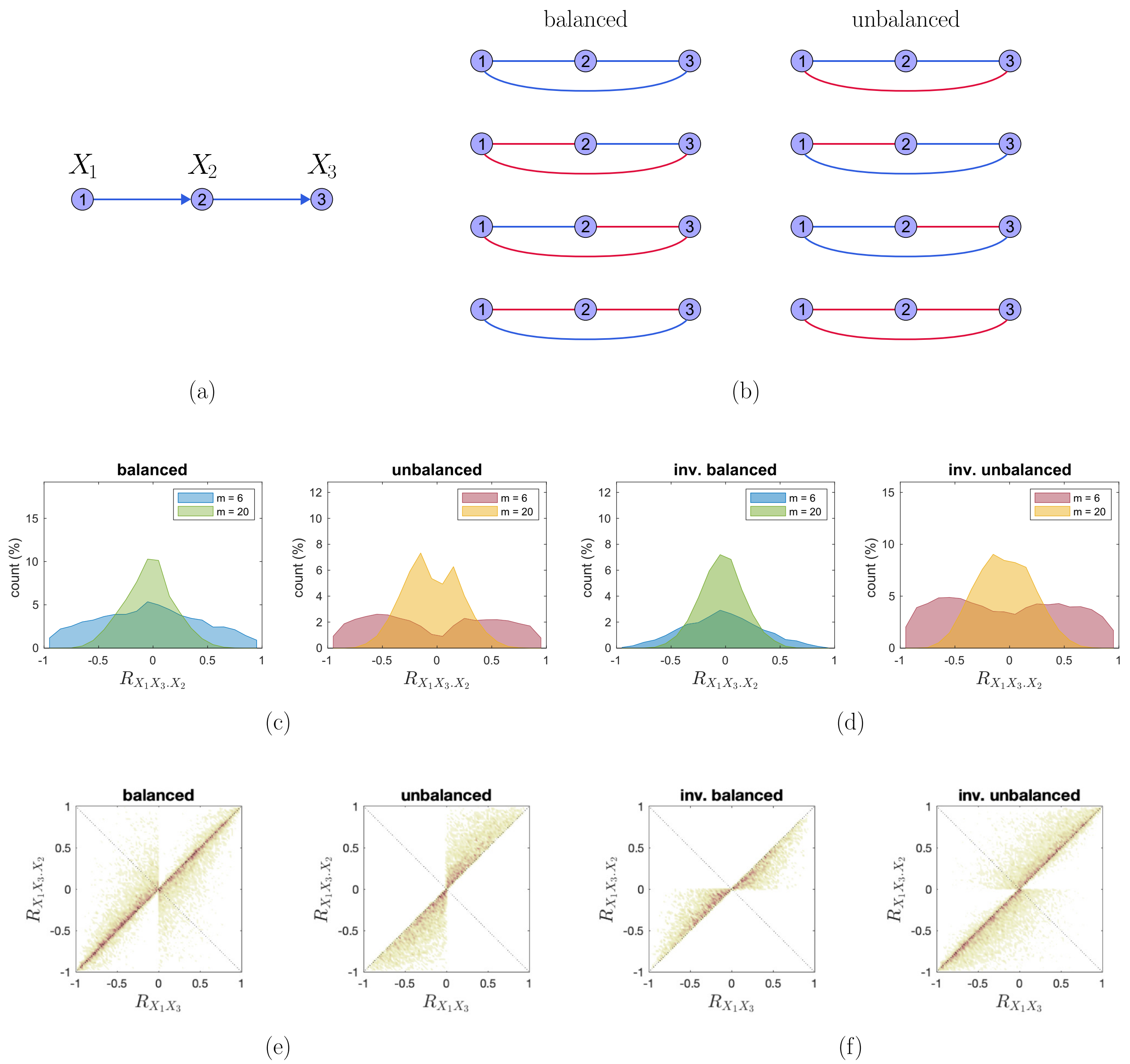


FIG. 1. Conditional independence on a length-2 chain motif and graphical tests based on balance and unbalance. (a) The length-2 chain motif. (b) Possible correlation graphs $\mathcal{G}(R)$ associated with the chain motif. Four of the $\mathcal{G}(R)$ are balanced and four unbalanced. (c) Distribution of the partial correlations $R_{X_1X_3.X_2}$ according to balance (left) or unbalance (right) for two different sample sizes $m = 6$ and $m = 20$. (d) Distribution of the partial correlations $R_{X_1X_3.X_2}$ according to inverse balance (left) or inverse unbalance (right) for $m = 6$ and $m = 20$. (e) Scatter plot of $R_{X_1X_3}$ vs $R_{X_1X_3.X_2}$ in the balanced (left) and unbalanced (right) case. (f) Scatter plot of $R_{X_1X_3}$ vs $R_{X_1X_3.X_2}$ in the inverse balanced (left) and inverse unbalanced (right) case.

variables $\mathbf{X} = [X_1 \cdots X_n] \in \mathcal{N}(0, \Sigma)$ is positively associated, the graph $\mathcal{G}(R)$ formed by taking the correlation matrix R as the adjacency matrix has all nonnegative entries. In general, however, R contains both positive and negative off-diagonal entries, meaning that $\mathcal{G}(R)$ is a signed graph. Among all signed graphs, a special class stands out because it has properties that are similar to those of a nonnegative graph: It is the class of balanced graphs (also called frustration-free graphs). A graph is said *balanced* if all its cycles are positive, i.e., they contain an even number of negative edges. An equivalent condition is that there exists a diagonal signature matrix $D = \text{diag}(d_1, \dots, d_n)$ with $d_i = \pm 1$, such that after the change of basis with $D = D^{-1}$ (sometimes called a "gauge transformation"), $\bar{R} = DRD$ is non-negative, and, therefore, $\mathcal{G}(\bar{R})$ has all non-negative edges [18]. We then say that a vector of Gaussian random variables $\mathbf{X}$ is *balanced* if it has a balanced correlation graph $\mathcal{G}(R)$, i.e., if the variables $D\mathbf{X}$ are positively associated (they have correlation $\bar{R}$).

In Gaussian graphical models, a concept stronger than positive association is often used, that of MTP$_2$ [24]. In terms of the concentration matrix K, $\mathbf{X} \in \mathcal{N}(0, \Sigma)$ is MTP$_2$ if and only if K is an M-matrix. Recall that a matrix K is said to be

TABLE I. Summary of properties of random Gaussian variables, their equivalent graphical characterization, and their dependencies.

Property	Graphical test
X positively associated	$\mathcal{G}(R)$ has all non-negative edges
X balanced	$\mathcal{G}(R)$ has all positive cycles
X MTP$_2$	$\mathcal{G}(P)$ has all non-negative edges
X inverse balanced (i.e., signed MTP$_2$)	$\mathcal{G}(P)$ has all positive cycles
Dependencies	
X MTP$_2$ $\substack{\Longrightarrow \\ \not\Longleftarrow}$ **X** positively associated	
X inverse balanced $\substack{\Longrightarrow \\ \not\Longleftarrow}$ **X** balanced	

an M-matrix if it can be written as $K = sI - B$ with $B \geqslant 0$ and $s > \rho(B)$, where $B \geqslant 0$ means elementwise non-negative, and $\rho(B)$ is the spectral radius of B. **X** being MTP$_2$ is a sufficient but not necessary condition for **X** being positively associated (meaning that K M-matrix implies R (and Σ) non-negative but not vice versa, see Ref. [24]). In practice, K being an M-matrix implies that K (and H) have all off-diagonal entries which are nonpositive. Hence, from (3), P has all non-negative off-diagonal entries. Summarizing: **X** MTP$_2$ is equivalent to say that the partial correlation graph $\mathcal{G}(P)$ has all non-negative edges. A generalization from non-negative $\mathcal{G}(P)$ to signed $\mathcal{G}(P)$ can be obtained in the same way as we discussed above for $\mathcal{G}(R)$. In particular, in this paper we call *inverse balanced* a Gaussian probability distribution $\mathbf{X} \in \mathcal{N}(0, \Sigma)$ such that K becomes a M-matrix after a gauge transformation with a diagonal signature matrix D, i.e., $\bar{K} = DKD$ is a M-matrix. It is straightforward to show that if D renders $\bar{K}$ a M-matrix, then it also renders DRD (and $D\Sigma D$) non-negative, meaning that the same gauge transformation D that renders **X** MTP$_2$ also renders **X** positively associated. The same D is also such that all partial correlations $d_i R_{X_i X_j.\mathcal{S}} d_j$ become non-negative for any conditioning subset $\mathcal{S} \subset \mathcal{X} \setminus \{X_i, X_j\}$. The notion of inverse balance is referred to as *signed* MTP$_2$ in Ref. [23].

Given a p.d. covariance matrix Σ and its concentration matrix K, then checking inverse balance is a purely graphical condition: **X** is inverse balanced (i.e., signed MTP$_2$) if and only if $\mathcal{G}(P)$ is balanced. Since **X** which is MTP$_2$ is a sufficient but not necessary condition for **X** to be positively associated, then also **X** which is inverse balanced (i.e., signed MTP$_2$) is a sufficient but not necessary condition for **X** to be balanced. The various properties with their graphical characterizations are summarized in Table I.

V. CONTRACTION PROPERTY OF INVERSE BALANCED PARTIAL COVARIANCES

It is known that in MTP$_2$ distributions, partial correlations and covariances over all conditioning sets are nonnegative, see Ref. [22]. On the other hand, conditioning corresponds to projecting the data on the orthogonal complement of the variables being conditioned upon, hence, intuitively, in reducing the variance and covariance of the data. Combining this two concepts, we have the following elementwise contraction property.

Theorem 1. Consider n Gaussian random variables $X_i \in \mathcal{X} \sim \mathcal{N}(0, \Sigma)$ such that their joint probability distribution is inverse balanced (i.e., signed MTP$_2$). Then, for any two sets $\mathcal{S}_1,\ \mathcal{S}_2 \subset \mathcal{X}$, $\mathcal{S}_1 \subset \mathcal{S}_2$, and $X_i,\ X_j \notin \mathcal{S}_2$, it is

$$|\Sigma_{X_i X_j.\mathcal{S}_2}| \leqslant |\Sigma_{X_i X_j.\mathcal{S}_1}|, \tag{6}$$

where $\Sigma_{X_i X_j.\mathcal{S}_k}$ is the covariance of X_i and X_j conditioned on $\mathcal{S}_k$ and $|\cdot|$ is the absolute value.

Proof. We can use the recursive formula for conditioning a covariance of k variables, see Ref. [28]. By reordering the indices, let $i, j = 1, \ldots, q$, and the conditioning be on the indices $q+1, q+2, \ldots, p$. Then, if $\Sigma_{ij.q+1\cdots p}$ is the short-hand notation for $\Sigma_{X_i X_j.X_{q+1}\cdots X_p}$,

$$\Sigma_{ij.q+1\cdots p} = \Sigma_{ij.q+2\cdots p} - \frac{\Sigma_{iq+1.q+2\cdots p}\Sigma_{jq+1.q+2\cdots p}}{\Sigma_{q+1q+1.q+2\cdots p}}. \tag{7}$$

If the probability distribution of $\mathcal{X}$ is inverse balanced, then $\exists$ a diagonal signature matrix $D = \text{diag}\{\pm 1\}$ such that $\bar{\Sigma} = [\bar{\Sigma}_{ij}] = D\Sigma D \geqslant 0$. Since $\bar{\mathcal{X}} = D\mathcal{X}$ is MTP$_2$, all its partial covariances $\bar{\Sigma}_{ij.\mathcal{S}}$ are non-negative for any $\mathcal{S}$, hence, in particular, all three terms appearing in (7) are non-negative, which implies $0 \leqslant \bar{\Sigma}_{ij.q+1\cdots p} \leqslant \bar{\Sigma}_{ij.q+2\cdots p}$. Iterating over successive conditioning leads to $0 \leqslant \bar{\Sigma}_{ij.\mathcal{S}_2} \leqslant \bar{\Sigma}_{ij.\mathcal{S}_1}$. Consequently, a similar expression holds for $\mathcal{X}$ with the absolute values as in (6). ■

Since $\Sigma_{X_i X_j.\mathcal{S}}$ is a scalar, the contraction in (7) is elementwise. A similar result is likely valid for the corresponding correlations, although a formal proof is still missing.

Remark 1. If a distribution is balanced but not inverse balanced, then the contraction property (6) can be violated when $\Sigma_{X_i X_j.\mathcal{S}}$ changes sign with respect to $\Sigma_{X_i X_j}$. Consider, for example,

$$\Sigma = \begin{bmatrix} 1.14 & 0.67 & 0.08 \\ 0.67 & 0.87 & 0.29 \\ 0.08 & 0.29 & 0.89 \end{bmatrix}.$$

For instance, $\Sigma_{13.2} = -0.14$ is such that $|\Sigma_{13.2}| > \Sigma_{13} = 0.08$. Similar violations can occur for the corresponding correlations.

VI. USING BALANCE AND INVERSE BALANCE OF THE SAMPLES FOR CHECKING CONDITIONAL INDEPENDENCE

If all nodes of the graphical model are measured, then we can form the sample correlation matrix R (which we can assume p.d.). By construction, correlation matrices are symmetric $R_{ij} = R_{ji}$, meaning that $\mathcal{G}(R)$ is undirected. The sample correlation obtained from the data is also typically full, hence $\mathcal{G}(R)$ is typically a complete graph. This implies that all variables pairs are connected by an undirected edge in $\mathcal{G}(R)$, even those that are not so in the underlying graphical model. It is this passage from the “true” topology of the graphical model (a DAG) to the fully connected topology of $\mathcal{G}(R)$ that provides an opportunity for checking consistency of the data, as the balance of the induced cycles on $\mathcal{G}(R)$ helps us discriminate the reliable data from the unreliable ones. We stress that $\mathcal{G}(R)$ is now a *sample* correlation graph, meaning that we always rely exclusively on the data to determine balance and inverse balance.

As an example, in Fig. 1 we consider the chain $X_1 \to X_2 \to X_3$. In this case $\mathcal{G}(R)$ is a triangle of nodes 1, 2 and 3. On this triangle, looking at the signs of the edges, eight different $\mathcal{G}(R)$ are possible, four of which are balanced and four unbalanced, see Fig. 1(b). If we associate an empirical regulatory action to each edge of $\mathcal{G}(R)$ by taking the corresponding value of R, then it is easy to realize that if $\mathcal{G}(R)$ is balanced the three regulatory actions are compatible with each other, whereas if $\mathcal{G}(R)$ is unbalanced, then at least one of these empirical regulations is incompatible with the others. For instance, if the data for X_1 and X_2 are positively correlated (meaning, e.g., X_1 activates X_2), but X_2 and X_3 are negatively correlated (X_2 inhibits X_3), then one expects that X_1 and X_3 should also be negatively correlated (X_1 indirectly inhibits X_3; balanced case), whereas if instead X_1 and X_3 are positively correlated (X_1 indirectly activates X_3; unbalanced case) then the three data series of X_1, X_2, and X_3 are incompatible with each other and should not belong to the same "true" regulatory graph.

To check what happens on our length-2 chain, we generated 10^4 sample realizations using the "mvnrnd" function of MATLAB with sample size $m = 6$ and $m = 20$. For the conditional independence $X_1 \perp X_3 \,|\, X_2$, the correlation $R_{X_1X_3}$, and the partial correlation $R_{X_1X_3.X_2}$ were computed, as well as the balance and the inverse balance properties of the resulting graphs. The histograms of the resulting partial correlations, classified according to balance and unbalance and inverse balance and inverse unbalance, are shown in Figs. 1(c) and 1(d) for both values of m, and the scatter plots of correlations vs partial correlations for the case of $m = 6$ in Figs. 1(e) and 1(f). The case of $m = 6$ is the most interesting: the partial correlations for the balanced cases tend to be more concentrated around the origin, whereas those associated with unbalanced cases tend to stay away from the origin, see Fig. 1(c). Something similar is visible also for inverse balance, although in a less pronounced way, see Fig. 1(d). When m grows, the effect tends to disappear in both unbalanced and inverse unbalanced cases because numerically the sample correlations and sample partial correlations become more precise. Note in Fig. 1(e) how in the balanced case the contraction rule $|R_{X_1X_3.X_2}| < |R_{X_1X_3}|$ is tendentially satisfied. The same relationship appears to be always satisfied for inverse balanced graphs (even though our proof of Theorem 1 is only for covariances and not for correlations). The unbalanced cases instead do not obey to any contraction rule, on the contrary. To summarize, the "sign consistency" encoded in the balance of $\mathcal{G}(R)$ reflects in the amplitude of the partial correlations, and, in particular, leads to $|R_{X_1X_3.X_2}| < |R_{X_1X_3}|$ much more often than unbalance, which, in turn, leads to verification of the conditional independence $X_1 \perp X_3 \,|\, X_2$ more often than unbalance. These are heuristic properties and, indeed, exceptions are visible in the data. A consequence is that balance can be used as a (heuristic) test of the validity of a conditional independence among the variables of the graphical model. Since inverse balance is computed from samples in an analogous way, also inverse balance of a graphical model can be used to set up a heuristic test of independence. What can be made rigorous in this case is the property stated in Theorem 1 that for inverse balance conditioning always leads to an elementwise contraction towards 0 of the "residual" partial covariance.

These observations for chains $X_1 \to X_2 \to X_3$ and artificial data are confirmed in the next section on experimental data.

Similar arguments can be set up for any graphical model, or subset of a graphical model, involving, at least, two adjacent edges (not forming a collider motif). It gives us a way to label certain interaction patterns as "more plausible" (candidates to be true positives) or as "less plausible" (candidates to be false positives) based on the sample correlations, without relying on the computation of partial correlations. This sign consistency test on $\mathcal{G}(R)$ is akin to a parity check in the error correcting code [29,30], and, in fact, the notion of balance can be mapped into solvability of an XOR-SAT problem, i.e., a linear system over a binary field [20]. The test is purely on the sample correlations, and, even when applied to Bayesian networks, it does not lead to any causality violation on the corresponding DAG.

VII. AN EXPERIMENT: ELEMENTARY CHAIN MOTIFS IN GENE TRANSCRIPTION AND TRANSLATION

We consider high-resolution models of transcription and translation processes for around 4,800 human genes and proteins, see Fig. 2. In particular, for each gene and protein the variables we consider are:

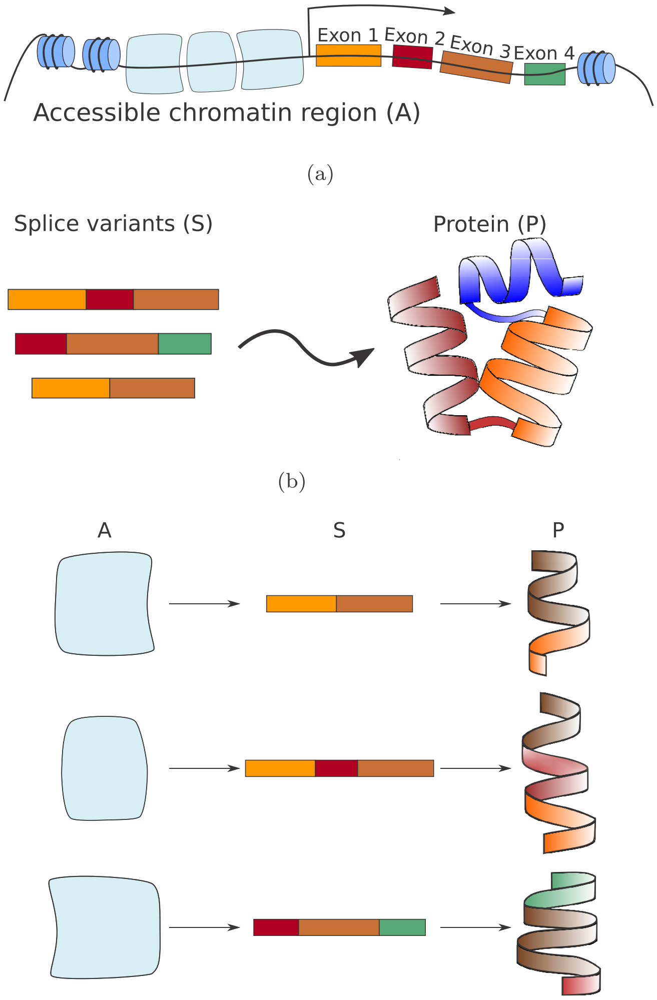
(1) The chromatin accessibility state in sites (called "peaks") in the promoter region of the gene;

(2) the RNA expression of the alternative splice variants that characterize a gene.

(3) Protein abundance.

Omics measurements of these quantities are available through ATAC-seq, RNA-seq, and mass spectrometry, see Refs. [25,26] for details. We use the subindex "A" to denote ATAC sites, "S" to denote splice variant, and "P" for protein. An "open" chromatin around the promoter region of a gene favors the binding of the transcription factors and hence enhances the regulatory effects on the gene (both positive and negative regulations). ATAC-seq (assay for transposase-accessible chromatin using sequencing) essays probe the chromatin state by estimating the accessible sites ("peaks") on the DNA in the promoter region of a gene, in particular in proximity of the transcription start site of each splice variant of a gene. The measurements it produces are assumed to be in direct proportionality with this accessibility. Gene expression itself is composed by the abundance of an ensemble of transcripts representing alternative splice variants of the gene, all measurable in modern deep RNA-seq experiments.

The DNA regions targeted by ATAC-seq reads can be of relevance for some splice variants but not for others, hence different measured ATAC peaks can be associated with different splice variants of the same gene. By knowing the position on the DNA of both the ATAC peaks and the transcriptional starting sites of each splice variant of a gene, it is possible to create a bioinformatic map of putative $A \to S$ interactions for each gene. The splice variants of a gene, in principle, can lead to different protein isoforms, which, however, are not distinguishable by current mass-spec proteomics. Hence, all splice variants of a gene are necessarily associated to the same protein variable, leading to a set of causal interactions $S \to P$, some of which are "real" whereas some other are false positives. To summarize, for each protein a bioinformatic analysis

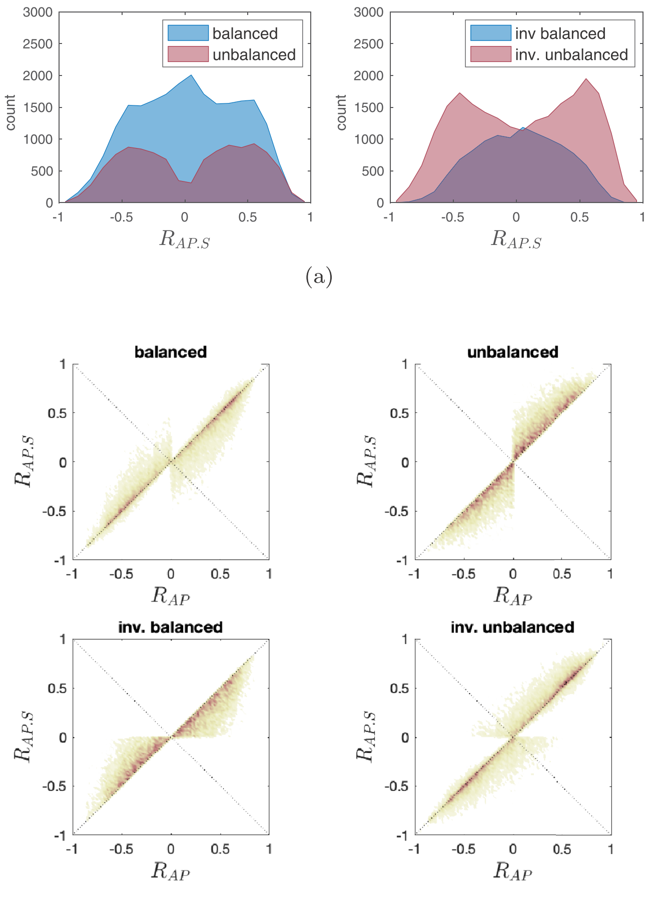


(a)

(b)

(c)

FIG. 2. (a) Sketch of the gene transcription/translation process. The open or close status of the chromatin in the sites in proximity of the promoter region of a gene influences gene transcription. A gene is itself composed of various exons which, during transcription, combine in different ways yielding alternative splice variants of a gene. (b) All of the splice variants of the gene, in principle, can contribute to protein synthesis. (c) To each gene and protein are associated multiple elementary chain motifs $A \rightarrow S \rightarrow P$ corresponding to the three variables chosen ($A =$ chromatin accessibility state, $S =$ splice variants of a gene, and $P =$ protein).



(a)

(b)

FIG. 3. Balance and unbalance for elementary chain motifs $A \rightarrow S \rightarrow P$ in gene transcription and translation. (a) $R_{AP.S}$ classified according to balance and unbalance (left panel) and according to inverse balance and unbalance (right panel). (b) Scatter plots of R_{AP} vs $R_{AP.S}$ classified according to balance and unbalance (upper panels) and to inverse balance and inverse unbalance (lower panels).

provides us with a set of putative interactions among A, S, and P. Since causality flows unidirectionally from chromatin accessibility to RNA transcription and then to translation into proteins, both sets of interactions $A \rightarrow S$ and $S \rightarrow P$ have an unambiguous directionality. The motif we are interested in is the length-2 chain $A \rightarrow S \rightarrow P$. The number of such chains in our data is 34,026. The set of putative chains obtained in this way contains, however, a large number of false positives. Our task is to use the notion of balance and inverse balance to prune some of these false positive interactions.

More specifically, only some of the ATAC sites, in practice, influence transcription, and, similarly, only some of the splice variants S are of relevance for protein synthesis. Our task is to find out which ones by checking for which triples $\{A, S, P\}$ involved in a chain it is indeed $A \perp P|S$. For each triplet, the sample correlations R_{AS} and R_{SP} can be used to estimate the direct regulatory effects and R_{AP} for the indirect one. For each $\{A, S, P\}$, the three correlations form a triangle $\mathcal{G}(R)$ whose balance can be computed, as well as the associated inverse balance on $\mathcal{G}(P)$. Because of the edge directionality, the balance can be interpreted straightforwardly as coherence between the regulatory signs inferred for the "physical path" $A \rightarrow S \rightarrow P$ and that associated with the "nonphysical" indirect path $A \rightarrow P$. These motifs have some similarity with the so-called feedforward loops used in biology, where balance is, indeed, referred to as "coherence," see Ref. [31]. The conditional independence $A \perp P|S$ can be checked also via partial correlations using (5). The distribution of such partial correlations for our 34,026 chains, classified according to balance and unbalance (and inverse balance and unbalance) is shown in Fig. 3(a), and the scatter plots of R_{AP} vs $R_{AP.S}$ are given in Fig. 3(b). These plots show the following two behaviors: (i) partial correlations $R_{AP.S}$ associated with balanced $\mathcal{G}(R)$ tend to concentrate around 0, whereas those associated with unbalanced $\mathcal{G}(R)$ tend to stay away from 0, and (ii)

$|R_{AP.S}| < |R_{AP}|$ occurs much more often in the balanced cases than in the unbalanced ones. Both properties lead to the same conclusion: balance is much more a proxy for conditional independence than unbalance. Both properties keep holding for inverse balanced graphs where, in particular, the contraction property $|R_{AP.S}| < |R_{AP}|$ is always verified, see lower left panel in Fig. 3(b).

The main contribution of this paper is to introduce the notion of balance and inverse balance as easy-to-check proxies for conditional independence in graphical models. If in this paper we have used them as a validation test for graphs of known topology, a more challenging task ahead is to try to use them as a structure discovery tool, in the case in which the topology of the graphical model is not known and its conditional independences (and concentration matrix) must be found relying only on the sample data.

ACKNOWLEDGMENTS

This work was supported, in part, by grants from SSF (Grant No. SB16-0011 to C.A.) and from the ELLIIT research framework at LiU.

[1] R. G. Cowell, P. Dawid, S. L. Lauritzen, and D. J. Spiegelhalter, *Probabilistic Networks and Expert Systems: Exact Computational Methods for Bayesian Networks*, Information Science and Statistics (Springer, New York, 2003).

[2] S. Epskamp, D. Borsboom, and E. I. Fried, Estimating psychological networks and their accuracy: A tutorial paper, Behav. Res. Methods **50**, 195 (2018).

[3] N. Friedman, Inferring cellular networks using probabilistic graphical models, Science **303**, 799 (2004).

[4] D. Koller and N. Friedman, *Probabilistic Graphical Models: Principles and Techniques*, Adaptive computation and machine learning (MIT Press, Cambridge, MA, 2009).

[5] O. Pourret, P. Naïm, and B. Marcot, *Bayesian Networks: A Practical Guide to Applications*, Statistics in Practice (Wiley, Hoboken, NJ, 2008).

[6] D. R. Williams, Bayesian estimation for gaussian graphical models: Structure learning, predictability, and network comparisons, Multivar. Behav. Res. **56**, 336 (2021).

[7] S. L. Lauritzen, *Graphical Models*, Oxford Statistical Science Series (Clarendon, Oxford, 1996).

[8] M. Slawski and M. Hein, Estimation of positive definite M-matrices and structure learning for attractive Gaussian markov random fields, Linear Algebra Appl. **473**, 145 (2015).

[9] O. Banerjee, L. El Ghaoui, and A. d'Aspremont, Model selection through sparse maximum likelihood estimation for multivariate gaussian or binary data, J. Mach. Learn. Res. **9**, 485 (2008).

[10] J. Engel, L. Buydens, and L. Blanchet, An overview of large-dimensional covariance and precision matrix estimators with applications in chemometrics, J. Chemom. **31**, e2880 (2017).

[11] J. Friedman, T. Hastie, and R. Tibshirani, Sparse inverse covariance estimation with the graphical lasso, Biostatistics **9**, 432 (2008).

[12] N. Meinshausen and P. Bühlmann, High-dimensional graphs and variable selection with the Lasso, Ann. Stat. **34**, 1436 (2006).

[13] J. Peng, P. Wang, N. Zhou, and J. Zhu, Partial correlation estimation by joint sparse regression models, J. Am. Stat. Assoc. **104**, 735 (2009).

[14] M. Scanagatta, A. Salmeron, and F. Stella, A survey on bayesian network structure learning from data, Prog. Artif. Intell. **8**, 425 (2019).

[15] M. Kalisch and P. Bühlmann, Estimating high-dimensional directed acyclic graphs with the PC-algorithm, J. Mach. Learn. Res. **8**, 613 (2007).

[16] F. Harary, On the measurement of structural balance, Behav. Sci. **4**, 316 (1959).

[17] J. A. Davis, Clustering and structural balance in graphs, Hum. Relat. **20**, 181 (1967).

[18] G. Facchetti, G. Iacono, and C. Altafini, Computing global structural balance in large-scale signed social networks, Proc. Natl. Acad. Sci. USA **108**, 20953 (2011).

[19] M. Mezard, G. Parisi, and M. Virasoro, *Spin Glass Theory and Beyond* (World Scientific, Singapore, 1986).

[20] M. Mezard and A. Montanari, *Information, Physics, and Computation* (Oxford University Press, New York, 2009).

[21] D. M. Malioutov, J. K. Johnson, and A. S. Willsky, Walk-sums and belief propagation in gaussian graphical models, J. Mach. Learn. Res. **7**, 2031 (2006).

[22] S. Karlin and Y. Rinott, Classes of orderings of measures and related correlation inequalities. I. multivariate totally positive distributions, J. Multivariate Anal. **10**, 467 (1980).

[23] S. Lauritzen, C. Uhler, and P. Zwiernik, Maximum likelihood estimation in Gaussian models under total positivity, Ann. Statist. **47**, 1835 (2019).

[24] S. Karlin and Y. Rinott, M-matrices as covariance matrices of multinormal distributions, Linear Algebra Appl. **52-53**, 419 (1983).

[25] A. Zenere, O. Rundquist, M. Gustafsson, and C. Altafini, Using high-throughput multi-omics data to investigate structural balance in elementary gene regulatory network motifs, Bioinformatics **38**, 173 (2022).

[26] A. Zenere, O. Rundquist, M. Gustafsson, and C. Altafini, Multi-omics protein-coding units as massively parallel Bayesian networks, iScience **25**, 104048 (2022).

[27] C. M. Bishop, *Pattern recognition and machine learning*, Information science and statistics (Springer, New York, NY, 2006).

[28] T. W. Anderson, *An Introduction to Multivariate Statistical Analysis*, Wiley Series in Probability and Statistics (Wiley, Hoboken, NJ, 2003).

[29] T. M. Cover and J. A. Thomas, *Elements of Information Theory (Wiley Series in Telecommunications and Signal Processing)* (Wiley-Interscience, Hoboken, NJ, 2006).

[30] R. G. Gallager, *Low-Density Parity-Check Codes* (MIT Press, Cambridge, MA, 1963).

[31] S. Mangan and U. Alon, Structure and function of the feed-forward loop network motif, Proc. Natl. Acad. Sci. USA **100**, 11980 (2003).