Astronomy and Computing 56 (2026) 101129



Contents lists available at ScienceDirect

Astronomy and Computing

journal homepage: www.elsevier.com/locate/ascom



Full length paper



Probabilistic graphical models in astronomy Application to exoplanetary systems

Abigail Sheerin, Giuseppe Vinci *

Department of Applied and Computational Mathematics and Statistics, University of Notre Dame, Notre Dame, 46556, IN, USA

ARTICLE INFO

Dataset link: https://github.com/gvincistat/astroggm, https://github.com/gvincistat/astroggm, https://exoplanetarchive.ipac.caltech.edu/

Keywords:
Astrostatistics
Exoplanets
Gaussian graphical model
Multiple hypothesis testing
Regularization
Stars

ABSTRACT

The field of astronomy has been experiencing a data explosion driven by significant advances in observational instrumentation, and classical methods often fall short of addressing the complexity of modern astronomical datasets. Probabilistic graphical models offer powerful tools for uncovering the dependence structures and data-generating processes underlying a wide array of cosmic variables. By representing variables as nodes in a network, these models allow for the visualization and analysis of the intricate relationships that underpin theories of hierarchical structure formation within the universe. We highlight the value that graphical models bring to astronomical research by demonstrating their practical application to the study of exoplanets and host stars.

1. Introduction

Astrophysical theory has long been central to understanding the cosmos, from planetary formation to the evolution of the universe. However, astrophysics requires statistical models for the proper interpretation of astronomical data to test cosmological theories (Feigelson et al., 2021). The study of our universe through statistical analysis, *astrostatistics*, relies on analyzing diverse types of data about galaxies, stars, planets, and other celestial objects. Key among these data types are photometry (Eyer et al., 2019), spectrophotometry (Christian and Janes, 1977), spectroscopy (Lasue et al., 2012), galactic properties (Corbin et al., 2001), luminosity (Fabbiano and Shapley, 2002), age and metallicity (Carraro et al., 1998), rotation and velocity (Das et al., 2015), element abundance (Nissen and Gustafsson, 2018), and geological mapping (Bhatt et al., 2019).

The field of astrostatistics relies on diverse statistical and machine learning methods for astrophysical data analysis. These methods can be broadly categorized into unsupervised and supervised learning. Unsupervised learning techniques allow for the discovery of patterns or structures in data without prespecified labels. Various unsupervised learning methods for dimensionality reduction, clustering, and data completion have been applied in astrostatistics. These methods include: principal component analysis (Efstathiou and Fall, 1984; Fiorentin et al., 2007; Bailey, 2012; Eyer et al., 2019); independent component analysis (Maino et al., 2002; Lu et al., 2006; Allen et al., 2013; Das et al., 2015; Akutsu et al., 2020; Pati et al., 2021); sparse dictionary learning (Vinci et al., 2014; Constantina et al., 2017); and k-means clustering (Chattopadhyay et al., 2010; Das et al., 2015; Fraix-Burnet et al., 2015; Turner et al., 2018; Nascimento, 2022; Ikotun et al., 2023).

Supervised learning methods, including regression and classification, are widely used in astrostatistics for prediction. While they help extract meaningful patterns from complex datasets like unsupervised methods, supervised learning methods require a response variable and predictors. These methods provide powerful tools for analyzing astronomical data, enabling discoveries in areas such as galaxy classification, stellar dynamics, and elemental mapping (Feigelson and Babu, 2012). Various supervised learning methods have been applied in astrostatistics. These methods include: linear regression (Feigelson and Babu, 1992; Magorrian et al., 1998; Kelly, 2007; Riess et al., 2016; Bhatt et al., 2019); generalized linear models (Andreon and Hurn, 2010; de Souza et al., 2015a,b; Elliott et al., 2015); random forests, k-Nearest Neighbors, support vector machines, and deep learning for automated classification of galaxy morphologies and stars (Freeman et al., 2013; Dieleman et al., 2015; Kim and Brunner, 2017; Baumstark and Vinci, 2024; Nemade et al., 2024); and graph neural networks to infer matter density (Villanueva-Domingo and Villaescusa-Navarro, 2022), and resolve weak lensing alignments (Jagvaral et al., 2022).

New observational instruments now allow for more detailed and frequent observations, resulting in an unprecedented and exponentially growing volume of data, with sample sizes ranging from dozens to

* Corresponding author.
E-mail address: gvinci@nd.edu (G. Vinci).

https://doi.org/10.1016/j.ascom.2026.101129
Received 24 January 2026; Accepted 8 May 2026
Available online 14 May 2026
2213-1337/© 2026 The Authors. Published by Elsevier B.V. This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/).

billions of objects (Cisewski-Kehe and Schafer, 2019; Feigelson et al., 2021). This data surge has propelled astronomy into the "big data" era (Zhang and Zhao, 2015), exceeding the capabilities of traditional statistical frameworks. This massive influx of information presents considerable challenges across the entire data lifecycle. From cleaning and integration to processing, indexing, mining, visualization, and analysis, each stage is impacted. Consequently, the statistical methods traditionally applied in astronomy require continuous improvements to effectively handle the scale and complexity of this new data landscape (Feigelson and Babu, 2012).

Probabilistic graphical models (Lauritzen, 1996) are powerful tools for describing the dependencies among many random variables. In a graphical model, the dependence structure of d random variables $X_1, \dots, X_d$ is encoded by a graph $G = (V, E)$, where the vertices or nodes $V = \{1, \dots, d\}$ represent the variables, and the edges in E connect nodes to reflect conditional dependence relationships. Given the data, we are interested in estimating E. Graphical models have been applied in numerous fields, including neuroscience (Ortiz et al., 2015; Vinci et al., 2018a,b; Zhu and Cribben, 2018), genomics (Dobra et al., 2004; Yin and Li, 2011; Chun et al., 2013; Gan et al., 2022), proteomics (Wang et al., 2016), metabolomics (Krumsiek et al., 2011), forensic science (Dawid et al., 2007; Xu and Vinci, 2024), and finance (Carvalho and West, 2007). Despite their potential, graphical models remain underutilized in astrostatistics.

Graphical models are well suited to analyze astronomical data and present numerous advantages compared to other statistical learning approaches. First, graphical models allow us to disentangle confounding effects in multivariate astrophysical data, while traditional marginal correlation neglects the effects of other variables on the dependence between the others. Second, unlike traditional regression modeling, which describes relationships asymmetrically (response vs. predictor), graphical models capture mutual interactions by mapping conditional dependencies across an entire system simultaneously. Third, unlike dimension reduction techniques, which mathematically compress datasets and destroy the physical meaning of original astronomical variables, graphical models handle high-dimensional feature spaces by maintaining the original dimensions, thereby allowing astrophysicists to analyze exact physical properties rather than uninterpretable mathematical components. Finally, clustering is useful for categorizing objects (e.g., galaxies by shape), but it does not let us explain the underlying physical processes linking individual variables.

Probabilistic graphical models are related to graph neural networks, which have been applied in astronomy as discussed above. Yet, the two approaches are quite different. Probabilistic graphical modeling discovers hidden graph structures from raw data by modeling statistical conditional dependencies between variables, while graph neural networks utilize predefined graph topologies to perform deep representation learning via message passing, optimizing node features for complex predictive tasks (Kipf and Welling, 2017). However, together, the two methodologies might inherently serve entirely complementary purposes in modern network analysis, where an estimated probabilistic graphical model may inform on the graph structure utilized in a graph neural network. We leave this investigation for future research.

This article demonstrates the utility of probabilistic graphical models in astronomy through the analysis of exoplanetary systems. In Section 2, we review undirected and directed graphical models, detailing their estimation via hypothesis testing and penalized maximum likelihood, making these methodologies accessible to astronomers that are not familiar with the topic. In Section 3, we apply these models to exoplanetary systems — a compelling domain characterized by highly interrelated planetary and stellar parameters. We demonstrate that these models successfully recover most relationships consistent with established astrophysical laws. These results suggest that graphical models are not only capable of validating known physics but also may serve as powerful exploratory tools for gaining insight into astronomical variables that are not yet well understood. Finally, in Section 4, we discuss our results and outlines directions for future research.

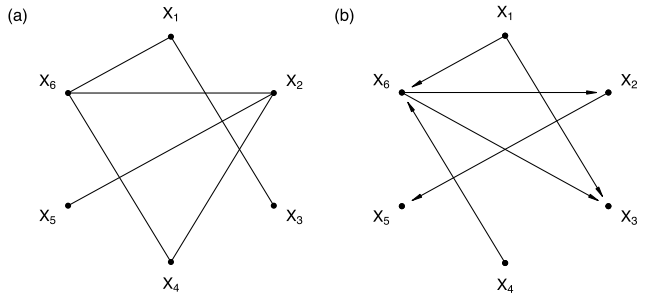


Fig. 1. (a): An undirected graph where any two variables are not connected by an edge if and only if they are independent conditionally on all other variables. For example, X_1 and X_5 are independent conditionally on all other variables. (b): A directed acyclic graph (DAG) where a variable X_i is independent of a variable X_j conditionally on the parents of X_i if and only if X_j is neither a parent nor a descendant of X_i. For example, X_6 is the only parent of X_2 so, conditionally on X_6, X_2 is independent of (X_1, X_3, X_4), but it is dependent on its descendant X_5.

2. Probabilistic graphical models

Probabilistic graphical models are multivariate probability distributions of random variables $X_1, \dots, X_d$ with a dependence structure represented by a *graph* $G = (V, E)$, where $V = \{1, \dots, d\}$ is a set of *nodes* (or vertices) representing d random variables, and $E \subset V \times V$ is a set of *edges* encoding dependence relationships among the variables. Two important classes of graphical models are *undirected* and *directed* graphical models.

2.1. Undirected graphs

A graph $G = (V, E)$ is *undirected* if the edge set E is symmetric, i.e., $(i, j) \in E \Leftrightarrow (j, i) \in E$ (Lauritzen, 1996). Fig. 1(a) shows an example of an undirected graph. We say that a d-variate random vector $X = (X_1, \dots, X_d)^T$ forms a Markov Random Field (MRF) with respect to an undirected graph $G = (V, E)$ if the *pairwise Markov property* holds: X_i and X_j are independent conditionally on all other variables if and only if $(i, j) \notin E$. If the distribution of X is positive, then the *local* and *global* Markov properties are also satisfied: any variable X_i is conditionally independent of all other variables given $X_{\mathcal{N}_E(i)} = (X_k)_{k \in \mathcal{N}_E(i)}$, where $\mathcal{N}_E(i) = \{j : (i, j) \in E\}$ are the neighbors of i; and any two distinct subsets of variables are independent conditionally on a separating set, i.e., the set of variables that, if removed from the graph, would disconnect the two variable sets. For example, in Fig. 1(a), X_1 and X_2 are independent conditionally on all other variables (pairwise Markov property); X_1 is independent of (X_2, X_4, X_5) conditionally on its neighbors (X_3, X_6) (local Markov property); and (X_2, X_5) are independent of (X_1, X_3) conditionally on (X_4, X_6) (global Markov property).

A well established MRF is the Gaussian Graphical Model (GGM), where $X \sim N(\mu, \Theta^{-1})$, $\mu \in \mathbb{R}^d$ is the mean vector, $\Theta \succ 0$ is a $d \times d$ symmetric positive definite *precision matrix*, and the edge set is $E = \{(i, j) \in V \times V : i \neq j, \Theta_{ij} \neq 0\}$. Equivalently, the edge set in a GGM is

$$E = \left\{(i, j) \in V \times V : i \neq j,\ \rho_{ij} \neq 0\right\} \tag{1}$$

where

$$\rho_{ij} = -\Theta_{ij} \left(\Theta_{ii}\Theta_{jj}\right)^{-1/2} \tag{2}$$

is the *partial correlation* of (X_i, X_j), i.e., the correlation between (X_i, X_j) conditionally on all the other $d - 2$ variables. Note that $-1 < \rho_{ij} < 1$, just like the *marginal Pearson's correlation*

$$r_{ij} = \Sigma_{ij}(\Sigma_{ii}\Sigma_{jj})^{-1/2}, \tag{3}$$

where $\Sigma = \Theta^{-1}$ is the *covariance matrix* of X. However, r_{ij} neglects the effects of other variables on the dependence between (X_i, X_j), so it is

possible to have $r_{ij} \neq 0$ while $\rho_{ij} = 0$, or even $r_{ij} < 0$ and $\rho_{ij} > 0$, and vice versa.

In real data analysis, the primary goal is to use data to estimate $G = (V, E)$, which, in a GGM, reduces to estimating the support of Θ or, equivalently, identifying all pairs (i, j) such that $\rho_{ij} \neq 0$.

2.1.1. Estimating a GGM via multiple hypothesis testing

Suppose we observe n independent and identically distributed (i.i.d.) data vectors $X^{(1)}, \dots, X^{(n)} \overset{\text{i.i.d.}}{\sim} N(\mu, \Theta^{-1})$, where $X^{(r)} = (X_1^{(r)}, \dots, X_d^{(r)})^T$. An approach to estimating E requires first obtaining the *maximum likelihood estimate* (MLE) of (μ, Θ),

$$(\hat{\mu}, \hat{\Theta}) = \underset{\mu \in \mathbb{R}^d, \Theta > 0}{\arg\max}\ L(\mu, \Theta; X^{(1)}, \dots, X^{(n)}), \tag{4}$$

where

$$L(\mu, \Theta; X^{(1)}, \dots, X^{(n)}) = \prod_{i=1}^{n} p(X^{(i)}; \mu, \Theta) \tag{5}$$

is the likelihood function, and

$$p(x; \mu, \Theta) = \sqrt{\frac{\det \Theta}{(2\pi)^d}} \exp\left(-\tfrac{1}{2}(x-\mu)^T \Theta (x-\mu)\right) \tag{6}$$

is the probability density function (p.d.f.) of $N(\mu, \Theta^{-1})$. If $n > d$, we have $\hat{\mu} = \frac{1}{n}\sum_{i=1}^n X^{(i)}$ and $\hat{\Theta} = \hat{\Sigma}^{-1}$, where

$$\hat{\Sigma} = \frac{1}{n}\sum_{i=1}^{n} (X^{(i)} - \hat{\mu})(X^{(i)} - \hat{\mu})^T \tag{7}$$

is the $d \times d$ sample covariance matrix.

The matrix $\hat{\Theta}$ is dense almost surely, so it does not readily provide us with an estimate of the graph structure of Θ. Thus, we use hypothesis testing to assess whether $\rho_{ij} \neq 0$ based on the partial correlation MLE (or sample partial correlation)

$$\hat{\rho}_{ij} = -\hat{\Theta}_{ij}(\hat{\Theta}_{ii}\hat{\Theta}_{jj})^{-1/2}, \tag{8}$$

for $1 \leq i < j \leq j$. For large n, we have

$$g\left(\hat{\rho}_{ij}\right) \approx N\left(g(\rho_{ij}),\ s_n^2\right), \tag{9}$$

where $s_n^2 = (n - d - 1)^{-1}$, and $g(\omega) = \frac{1}{2}\log((1+\omega)/(1-\omega))$ is the Fisher transformation, a monotone increasing function that maps the interval $(-1, 1)$ onto the real line $\mathbb{R} = (-\infty, \infty)$, which is the support of the Gaussian distribution (Fisher, 1924). Thus, to test the hypothesis $H_{0,ij} : \rho_{ij} = 0$ versus $H_{1,ij} : \rho_{ij} \neq 0$, we compute the p-value

$$P_{ij} = 2\left(1 - \Phi\left(|g(\hat{\rho}_{ij})| s_n^{-1}\right)\right), \tag{10}$$

where

$$\Phi(z) = \int_{-\infty}^{z} (2\pi)^{-1/2} e^{-\omega^2/2} d\omega \tag{11}$$

is the cumulative distribution function (c.d.f.) of the standard normal (Gaussian) distribution $N(0, 1)$. Finally, we apply methods for multiple hypothesis testing to obtain estimates of the edge set E.

Two possible approaches for multiple hypothesis testing are family-wise type I error rate control and false discovery rate control. The *family-wise type I error rate* (FWER) is the probability of (erroneously) discovering at least one edge while assuming that the true graph contains no edges. By applying the Bonferroni correction (Bonferroni, 1936; Dunn, 1961), we can define the edge set estimator

$$\hat{E}(\alpha) = \left\{(i, j) \in V \times V : i \neq j, P_{ij} \leq \alpha/m\right\}, \tag{12}$$

where $\alpha \in (0, 1)$ is the desired FWER level, and $m = \binom{d}{2}$ is the number of hypotheses. Equivalently, based on Eq. (9), an approximate $1 - \alpha$ confidence interval for ρ_{ij} is

$$C_{ij}(\alpha) = \left[g^{-1}\left(g(\hat{\rho}_{ij}) - z_{\frac{\alpha}{2}} s_n\right), g^{-1}\left(g(\hat{\rho}_{ij}) + z_{\frac{\alpha}{2}} s_n\right)\right], \tag{13}$$

where $g^{-1}(z) = (\exp(2z) - 1)/(\exp(2z) + 1)$ is the inverse Fisher transformation, $z_{\alpha/2} = \Phi^{-1}(1 - \alpha/2)$ is the $1 - \alpha/2$ quantile of $N(0, 1)$, and $\Phi^{-1}()$ is the inverse of the c.d.f. of $N(0, 1)$ in Eq. (11). Thus, the edge set in Eq. (12) can be rewritten as

$$\hat{E}(\alpha) = \left\{(i, j) \in V \times V : i \neq j, 0 \notin C_{ij}\left(\alpha/m\right)\right\} \tag{14}$$

The *false discovery rate* (FDR) is the expected proportion of incorrect edges among all discovered edges. By applying the FDR control of Benjamini and Yekutieli (2001) for dependent tests, we can define the edge set estimator

$$\hat{E}(\beta) = \left\{(i, j) \in V \times V : i \neq j, P_{ij} \leq P_\beta^*\right\}, \tag{15}$$

where $P_\beta^* = P^{(k^*)}$ is the k^*th smallest p-value among the $m = \binom{d}{2}$ p-values $\{P_{ij}\}_{1 \leq i < j \leq d}$,

$$k^* = \max\left\{k : P^{(k)} \leq \beta k \left(m \sum_{j=1}^{m} j^{-1}\right)^{-1}\right\}, \tag{16}$$

and $\beta \in (0, 1)$ is the desired FDR level.

In the data analysis in Section 3, we will also estimate *marginal correlation graphs* based on MLE and multiple hypothesis testing, i.e., graphs analogous to Eqs. (12) and (15) but based on the marginal correlation MLEs (or sample marginal correlations),

$$\hat{r}_{ij} = \hat{\Sigma}_{ij}(\hat{\Sigma}_{ii}\hat{\Sigma}_{jj})^{-1/2}, \tag{17}$$

rather than the partial correlation estimates $\hat{\rho}_{ij}$, and with $s_n^2 = (n-3)^{-1}$. Since the marginal correlation r_{ij} neglects the effects of other variables on the dependence between (X_i, X_j), a correlation graph is typically much denser (if not fully connected) than a conditional dependence graph.

2.1.2. Estimating a GGM via penalized MLE

If the sample size n is not large enough compared with the number of nodes d, the MLE $\hat{\Theta}$ may have undesirable statistical properties or may not even have a unique solution (for example, if $n < d$, $\hat{\Sigma}$ is not invertible). To circumvent the issues of the MLE, several *penalized MLE* methodologies have been proposed. A notable example is the *Graphical LASSO* (GLASSO) (Yuan and Lin, 2007), which is the solution of

$$\hat{\Theta}(\lambda) = \underset{\Theta > 0}{\arg\max}\ \ell_n(\Theta) - \lambda \|\Theta\|_{1,\text{off}}, \tag{18}$$

where

$$\ell_n(\Theta) = \ln L(\hat{\mu}, \Theta; X^{(1)}, \dots, X^{(n)}) \tag{19}$$

is the log likelihood function maximized with respect to μ (maximizer $\hat{\mu} = \frac{1}{n}\sum_{i=1}^n X^{(i)}$), and λ is a tuning parameter controlling the sparsity effect of the L_1 norm $\|\Theta\|_{1,\text{off}} := \sum_{i \neq j} |\Theta_{ij}|$ on $\hat{\Theta}(\lambda)$. For a given value of λ, the GLASSO estimate of E is

$$\hat{E}(\lambda) = \left\{(i, j) \in V \times V : i \neq j, \hat{\Theta}_{ij}(\lambda) \neq 0\right\}, \tag{20}$$

which is generally sparser for larger values of $\lambda > 0$. Note that, for $n > d$, $\hat{\Theta}(0) = \hat{\Theta}$, the MLE in Eq. (4).

Thus, in GLASSO, graphical model selection reduces to selecting the tuning parameter λ. A well established approach to select λ is the Extended Bayesian Information Criterion (EBIC; Foygel and Drton (2010)), which picks λ by minimizing the risk function

$$\text{EBIC}(\lambda) = -2\ell_n(\tilde{\Theta}(\lambda)) + \kappa(\lambda)(\log n + 4\gamma \log d), \tag{21}$$

where

$$\tilde{\Theta}(\lambda) = \underset{\Theta > 0, \Theta_{\hat{E}(\lambda)^c} = 0}{\arg\max}\ \ell_n(\Theta) \tag{22}$$

is the MLE of Θ constrained to have no edges beyond the GLASSO edge set $\hat{E}(\lambda)$ (Eq. (20)), $\kappa(\lambda) = |\hat{E}(\lambda)|/2$ is the number of edges in $\hat{E}(\lambda)$, and γ is a penalty term to restrict model complexity and prevent false positives in networks. Foygel and Drton (2010) recommend setting $\gamma = 0.5$ because it perfectly balances sensitivity and

specificity in high-dimensional datasets. Mathematically, it neutralizes the quadratic explosion of potential network edges, ensuring a conservative, sparse, and highly reliable model that effectively prevents false-positive connections.

GLASSO is guaranteed to have a unique solution if $\lambda > 0$ (Ravikumar et al., 2011), and various algorithms have been developed to estimate graphs with thousands of nodes efficiently (Yuan and Lin, 2007; Friedman et al., 2008; Hsieh et al., 2011; Mazumder and Hastie, 2012; Hsieh et al., 2013). Furthermore, the excellent statistical performance of GLASSO has been demonstrated under numerous settings (Rothman et al., 2008; Yuan, 2010; Ravikumar et al., 2011).

2.1.3. Estimating a GGM from non-Gaussian data

Multivariate data often do not follow a Gaussian distribution, and other graphical models may be more appropriate. However, GGM estimation is computationally convenient, and various approaches exist to align empirical data with GGM assumptions. Some approaches include applying suitable monotonic transformations, such as the logarithm or the square root (Vinci et al., 2018b), or embedding the GGM within a hierarchical model (Vinci et al., 2018a). Alternatively, we can transform the data *nonparametrically* (Liu et al., 2012) as follows. First, we define the empirical c.d.f. of X_i

$$\hat{F}_i(x) = \frac{1}{n}\sum_{r=1}^{n} I(X_i^{(r)} \le x), \tag{23}$$

where $I()$ is the indicator function, and its Winsorized version

$$\tilde{F}_i(x) = \begin{cases} \delta_n, & \text{if } \hat{F}_i(x) < \delta_n \\ 1-\delta_n, & \text{if } \hat{F}_i(x) > 1-\delta_n \\ \hat{F}_i(x), & \text{otherwise} \end{cases} \tag{24}$$

where $\delta_n = (n+1)^{-1}$. Then, we compute the *normal scores* $\tilde{X}_i^{(1)}, \dots, \tilde{X}_i^{(n)}$, where

$$\tilde{X}_i^{(r)} = \Phi^{-1}\left(\tilde{F}_i(X_i^{(r)})\right) \tag{25}$$

and $\Phi^{-1}(\cdot)$ is the inverse of the c.d.f. of the standard normal (Gaussian) distribution $N(0,1)$ in Eq. (11). This transformation makes $\tilde{X}_i^{(r)}$ approximately follow $N(0,1)$. Finally, we estimate a GGM based on the normal scores of the data. We apply this nonparametric approach in our data analysis in Section 3.

2.2. Directed graphs

A graph $G = (V, E)$ is *directed* if the edge set E consists of ordered pairs of nodes, where $(i,j) \in E$ indicates $i \to j$, i.e., there is an *edge-arrow* pointing from the *parent* node i to the *child* node j (Lauritzen, 1996). Fig. 1(b) shows an example of a directed graph. A *directed path* is a set of arrows all pointing in the same direction and linking an *ancestor* node i to a *descendant* node j, for example, $i \to k \to s \to j$. We say that the triple (i,j,k) forms a *collider* at j if $i \to j \leftarrow k$; this collider is *unshielded* if i and k are not connected by any edge; otherwise, it is *shielded*. For example, in Fig. 1(b), the triple $(1,6,4)$ forms an unshielded collider at 6, while the triple $(1,3,6)$ forms a shielded collider at 3.

2.2.1. DAGs

A *directed acyclic graph* (DAG) is a directed graph that contains no cycles, i.e., directed paths that start and end with the same node. We say that a DAG $G = (V, E)$ represents the probability distribution of a random vector $X = (X_1, \dots, X_d)^T$ if the joint probability density function (p.d.f.) or probability mass function (p.m.f.) of X can be expressed as

$$p_X(x_1, \dots, x_d) = \prod_{i=1}^{d} p_{X_i|X_{\pi(i)}=x_{\pi(i)}}(x_i), \tag{26}$$

where

$$p_{X_i|X_{\pi(i)}=x_{\pi(i)}}(x_i) = \frac{p_{X_i,X_{\pi(i)}}(x_i, x_{\pi(i)})}{p_{X_{\pi(i)}}(x_{\pi(i)})} \tag{27}$$

is the conditional p.d.f. or p.m.f. of X_i given $X_{\pi(i)} = x_{\pi(i)}$, and $\pi(i) = \{j : (j,i) \in E\}$ is the set of parents of node i. It can be shown that X_i is independent of X_j conditionally on $X_{\pi(i)}$ if and only if j is neither a parent nor a descendant of i.

2.2.2. Estimating DAGs

A probability distribution may be represented by multiple DAGs that imply the same conditional independencies. The collection of all such DAGs is called a *Markov equivalence class*. All DAGs in an equivalence class have the *same skeleton* (undirected version of the DAG) and the *same unshielded colliders* (Verma and Pearl, 1990), while they may differ in the orientation of other edges in the skeleton. An equivalence class can be represented by a completed partial DAG (CPDAG), a hybrid graph consisting of all directed edges common to all DAGs in the equivalence class and undirected edges wherever at least two DAGs have arrows in opposite directions.

Thus, in general, it is only possible to estimate the equivalence class rather than a unique DAG. A widely used approach to estimating the equivalence class is the PC algorithm (Spirtes and Glymour, 1991), which we outline in Algorithm 1. The PC algorithm allows us to estimate the skeleton and the set of all unshielded colliders, which, together with the no-cycle constraint, is all we need to estimate the equivalence class of DAGs representing the multivariate distribution of interest.

Algorithm 1 PC algorithm

Input: Data $X^{(1)}, \dots, X^{(n)}$; adjacency set $C = \{(i,j) : i \neq j\}$; collection of empty separator sets $\{\mathcal{A}_{ij}\}_{i \neq j}$; significance level α.

1. Skeleton
   For $k = 0, 1, \dots, d$:
   For every pair (i,j) such that $|\mathcal{N}_C(i) \setminus \{j\}| \ge k$:
   For every $A \subseteq \mathcal{N}_C(i) \setminus \{j\}$ such that $|A| = k$:
   Test H_0: $(X_i, X_j)|X_A$ indep. VS H_1: $(X_i, X_j)|X_A$ dep.
   H_0 not rejected $\Rightarrow C \equiv C \setminus \{(i,j),(j,i)\}$ & $\mathcal{A}_{ij} \equiv \mathcal{A}_{ij} \cup A$.
2. Unshielded Colliders
   For every $(i,j) \notin C$:
   For every $k \in \mathcal{N}_C(i) \cap \mathcal{N}_C(j)$:
   If $k \notin \mathcal{A}_{ij}$, update $C \equiv C \setminus \{(k,i),(k,j)\} \cup \{(i,k),(j,k)\}$.
3. CPDAG
   Assign arrows to undirected edges in C without creating any new unshielded colliders or cycles.

Output: CPDAG $= C$.

Algorithm 1 takes as input the data $X^{(1)}, \dots, X^{(n)}$, a starting adjacency set $C = \{(i,j) : i \neq j\}$, a starting collection of separator sets $\mathcal{A} = \{\mathcal{A}_{ij}\}_{i \neq j}$, where $\mathcal{A}_{ij} = \emptyset$, for all $i \neq j$, and a test size α. In step 1, we find the skeleton by implementing a sequence of hypothesis tests on the independence of variable pairs (X_i, X_j) conditionally on other variables $(X_A = (X_k)_{k \in A})$ that are neighbors with i and/or j in C. Each hypothesis test may be implemented in various ways. For example, if the data are multivariate Gaussian, as per Eq. (9), we can use the test statistic $T_{ij,A} = |g(\hat{\rho}_{ij|A})|\sqrt{n - |A| - 3}$, where $\hat{\rho}_{ij|A}$ is the sample partial correlation between (X_i, X_j) given X_A (i.e., $\hat{\rho}_{ij|A}$ is computed as in Eq. (8) but with $\hat{\Theta}$ being the inverse of the sample covariance matrix $\hat{\Sigma}_{(i,j,A)\times(i,j,A)}$ relative to nodes i, j, A), and $g(\omega)$ is the Fisher transformation; we reject H_0 if $T_{ij} > z_{\alpha/2}$, where $z_{\alpha/2} = \Phi^{-1}(1-\alpha/2)$ is the $1-\alpha/2$ quantile of $N(0,1)$, and $\Phi^{-1}(\cdot)$ is the inverse of the c.d.f. of $N(0,1)$ in Eq. (11). We use this test in the data analysis of Section 3, where we apply Algorithm 1 to the normal scores (Eq. (25)) of the data.

Implementing this test for all (i, j) and all suitable potential separators A, the skeleton estimate is finally given by what remains in the set C. During this process, the collection of separator sets $\{\mathcal{A}_{ij}\}_{i\neq j}$ is also updated.

In step 2, we identify all unshielded colliders. We look through all triples (i, k, j) where i, j are not connected in the skeleton, while k is adjacent to both i and j and is not a separator of (i, j) (i.e., $k \notin \mathcal{A}_{ij}$). Any such triple (i, k, j) forms an unshielded collider at k. During this process, the adjacency set C is updated to reflect the orientation (i.e., $i \to k \leftarrow j$) of the edges in the unshielded colliders.

Finally, in step 3, we attempt to assign orientations to all remaining undirected edges in C by simply applying the constraints that (i) no cycles and (ii) no additional unshielded colliders may be created. The output of the Algorithm is the CPDAG C, which may still contain undirected edges for all those pairs (i, j) where a unique orientation could not be determined.

Remarks. Although the PC algorithm effectively estimates equivalence classes, its directed edges should be interpreted causally with caution, particularly in observational studies. The algorithm's reliance on multiple hypothesis tests makes it vulnerable to incorrect edge identification. Moreover, unmeasured latent variables can easily confound the results, removing any guarantee of retrieving a true causal graph. Consequently, the PC algorithm is best treated as a tool for generating causal hypotheses that require external validation. Accordingly, in Section 3, we contextualize our estimated directed edges using known astrophysical laws.

A DAG estimated via the PC algorithm typically appears much sparser than an estimated undirected Gaussian graphical model, even when both use the same significance level for their respective conditional independence tests. This sparsity difference arises primarily from how the two models handle conditioning sets. In a undirected graphical model, an edge is dropped if and only if two variables are independent conditionally on all other variables in the network. Conversely, the PC algorithm drops an edge if it finds any separating subset of variables. Consequently, for any unshielded collider $i \to j \leftarrow k$ identified by the PC algorithm, the nodes i and k are likely to be connected in the undirected graph. This occurs because the undirected graph tests the independence of i and k by conditioning on all other variables, which inevitably includes the collider j; conditioning on a collider induces a statistical dependence between i and k, resulting in a "moralized" edge in the undirected graph.

3. Analysis of exoplanets and host stars data

In this section, we apply graphical models to estimate the dependence structure among multiple astrophysical variables characterizing exoplanets and their host stars. All computations were implemented with the R package `astroggm`, which we developed and made available on GitHub.

Astronomers analyze interconnected planetary and stellar parameters to study exoplanetary systems. A planet's mass indicates its bulk composition — whether rocky, ice, or gas (Chen and Kipping, 2017). Combined with mass, a planet' orbital period determines the star–planet distance via Kepler's Third Law, which helps infer whether the planet lies within the habitable zone (Carroll and Ostlie, 2017), while the surface temperature of the host star defines the radiation environment and habitable zone boundaries (Kopparapu et al., 2013). Furthermore, orbital eccentricity dictates climate stability; highly elliptical paths cause extreme seasonal variations and potential tidal heating (Barnes et al., 2013). Stellar mass governs system dynamics and lifespan, while stellar radius is mathematically required to calculate a transiting planet's true physical size (Winn, 2010). Additionally, high stellar metallicity strongly correlates with gas giant formation (Fischer and Valenti, 2005), and surface gravity confirms the star's evolutionary stage. A star's distance from Earth and apparent magnitude in the Gaia G-band determine its observable brightness (Prusti et al., 2016). Given the highly intricate web of relationships between all of these factors, conditional dependence graphical models are exceptionally useful in exoplanetary science.

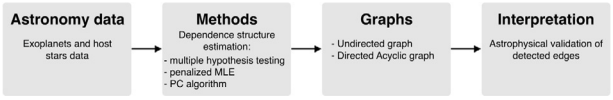


Fig. 2. Full analysis pipeline.

Table 1
Variables in the exoplanet and host star data.

P.mass	Planet mass relative to Earth.
P.orb.per	Planet orbital period.
P.orb.ecc	Planet orbital eccentricity.
S.temp	Star surface temperature.
S.radius	Star radius relative to the Sun.
S.mass	Star mass relative to the Sun.
S.metal	Star abundance of elements heavier than *H* and *He*.
S.gravity	Star surface gravitational pull.
S.dist	Star distance from Earth to the star.
S.gaia	Star apparent magnitude in the Gaia G-band.

The exoplanet and host star data we analyze were taken from the NASA Exoplanet Archive and consist of astrophysical measurements from confirmed exoplanets identified through a variety of detection methods. We focus our analysis on the $d = 10$ quantitative variables listed in Table 1. To apply the methods presented in Section 2, all data vectors must be complete. Thus, all cases containing any missing values in at least one of the 10 variables were removed (total 3508 incomplete cases), yielding a total number of $n = 1067$ exoplanets. Moreover, we apply the nonparametric transformation in Eq. (25) to all variables before pursuing any analysis. In Fig. 2, we summarize the full analysis pipeline: starting from astronomical data, we apply probabilistic graphical model estimation methods to obtain graphical representations of the dependence structure governing the ten variables characterizing the exoplanetary systems, and finally we contextualize the connections using known astrophysical laws.

In Fig. 3(a), we show the histograms of the ten variables standardized via Eq. (25) (diagonal plots), scatterplots of each variable versus every other (upper off-diagonals), and contour plots of estimated bivariate Gaussian distributions. In Fig. 3(b), we show the heatmap of the sample marginal correlations (upper diagonals; Eq. (17)) and sample partial correlations (lower diagonals; Eq. (8)) of the ten variables, where blue colors denote negative correlations, red colors denote positive correlations, and the gray color denotes correlations that are not significantly different from zero (FWER $\leq$ 1%, Bonferroni correction), as per Table 2. In this table, we report the sample marginal correlations (upper diagonals) and the sample partial correlations (lower diagonals) as percentages, with 99% Bonferroni corrected confidence intervals (Eq. (13), with $\alpha = 0.01/\binom{10}{2}$).

In Fig. 4(a)–(b), we show the estimated marginal correlation (marginal dependence) graphs based on 1% FWER control via Bonferroni correction (Table 2, upper diagonals) and 1% FDR control, respectively. In Fig. 4(c)–(e), we show the partial correlation (conditional dependence) graphs based on 1% FWER control via Bonferroni correction (Eq. (12); Table 2, lower diagonals), 1% FDR control (Eq. (15)), and GLASSO (Eq. (18)) selected via EBIC (Eq. (21)), respectively. In all these graphs, blue edges denote negative correlations, red edges denote positive correlations, the thickness of each edge is proportional to the magnitude of the correlation, and the radius of each node is proportional to the sum of the magnitudes of the correlations relative to the node. Finally, in Fig. 4(f), we show the CPDAG representing the Markov equivalence class estimated via the PC algorithm (Algorithm 1, $\alpha = 0.01$). Since no undirected edge is present, the estimated

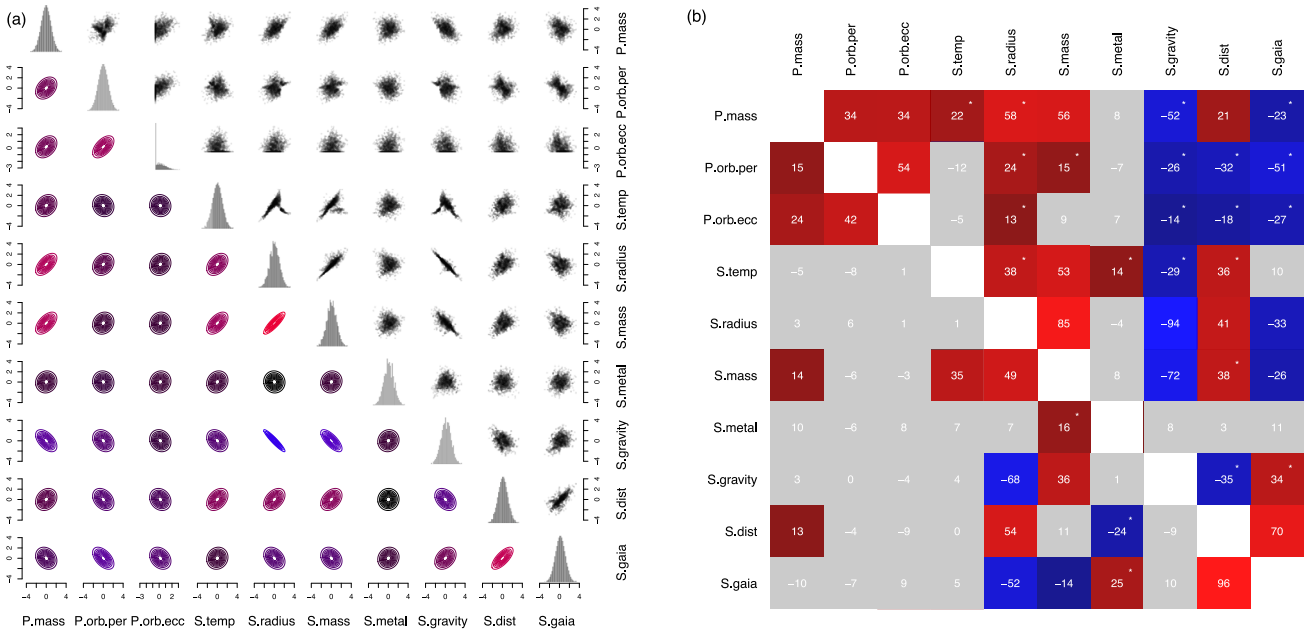


Fig. 3. (a): Histograms of the ten variables listed in Table 1 (diagonal plots) standardized via Eq. (25); scatterplots of each variable versus every other (upper off-diagonals); and contour plots of estimated bivariate Gaussian distributions (lower diagonals). (b): Heatmap of the sample marginal correlations $\hat{r}_{ij}$ (upper diagonals; Eq. (17)) and sample partial correlations $\hat{\rho}_{ij}$ (lower diagonals; Eq. (8)) of the ten variables, where blue colors denote negative correlations, red colors denote positive correlations, and gray color denotes correlations not significantly different from zero (FWER $\leq$ 1%, Bonferroni correction). Values of the correlations and partial correlations are shown as rounded signed percentages. The symbol $*$ denotes cases where only the marginal or the partial correlation is significant.

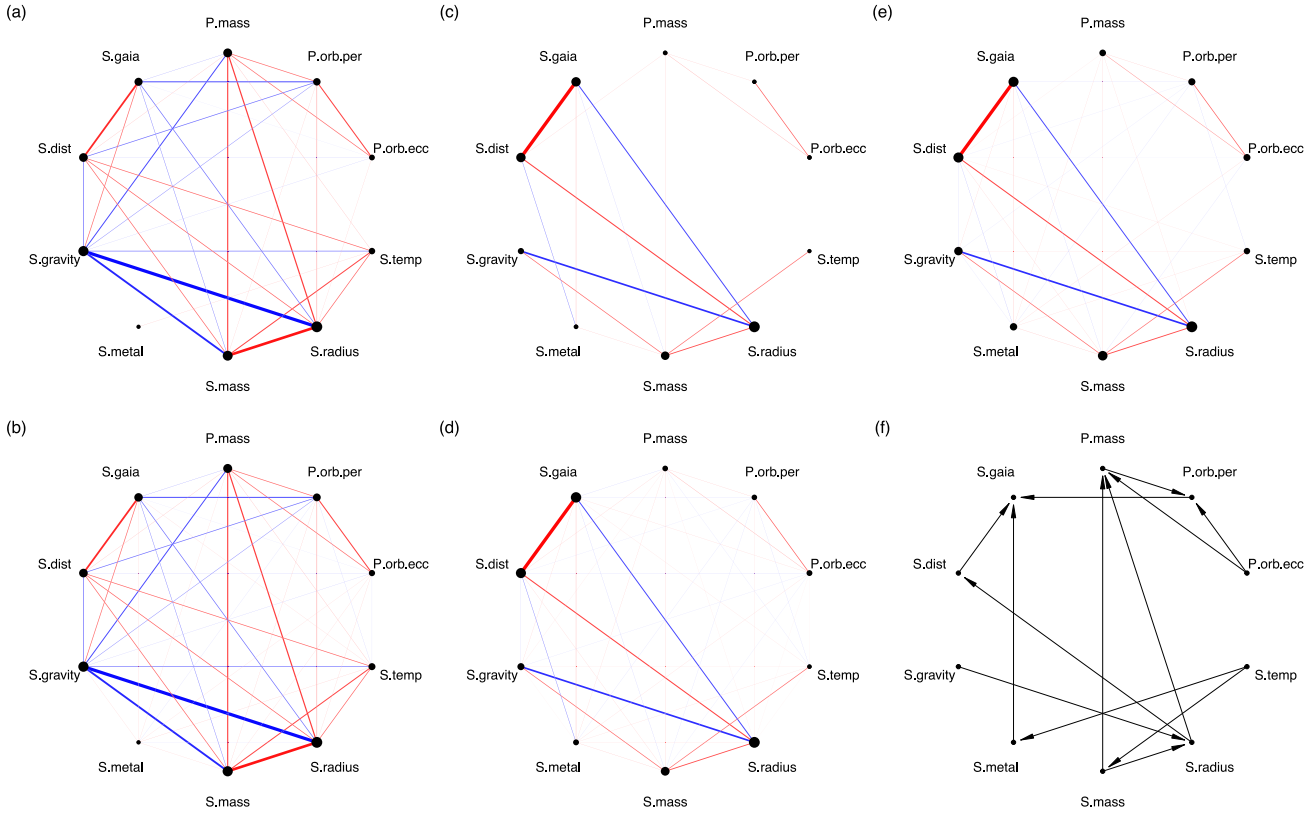


Fig. 4. (a): Marginal correlation graph based on 1% FWER control via Bonferroni correction (Table 2, upper diagonals). (b): Marginal correlation graph based on 1% FDR control. (c): Partial correlation (conditional dependence) graphs based on 1% FWER control via Bonferroni correction (Eq. (12); Table 2, lower diagonals). (d): Partial correlation graph based on 1% FDR control (Eq. (15)). (e): Partial correlation graph based on GLASSO (Eq. (20)) with EBIC tuning parameter selection (Eq. (21)). In all graphs in (a)–(e), blue edges denote negative correlations, red edges denote positive correlations, the thickness of each edge is proportional to the magnitude of the correlation, and the radius of each node is proportional to the sum of the magnitudes of the correlations relative to the node. (f): CPDAG obtained via PC algorithm (Algorithm 1, $\alpha = 0.01$).

Table 2
Sample marginal correlations $\hat{r}_{ij}$ (upper diagonals; Eq. (17)) and sample partial correlations $\hat{\rho}_{ij}$ (lower diagonals; Eq. (8)) as rounded signed percentages with 99% Bonferroni corrected confidence intervals (Eq. (13), $\alpha = 0.01/\binom{10}{2}$). Values that are significantly different from zero are in bold face.

	P.mass	P.orb.per	P.orb.ecc	S.temp	S.radius	S.mass	S.metal	S.gravity	S.dist	S.gaia
P.mass	⋆	(24,**34**,44)	(24,**34**,44)	(11,**22**,32)	(49,**58**,65)	(47,**56**,63)	(−4,8,19)	(−60,−**52**,−43)	(10,**21**,31)	(−33,−**23**,−12)
P.orb.per	(3,**15**,26)	⋆	(45,**54**,61)	(−23,−12,0)	(13,**24**,34)	(4,**15**,26)	(−18,−7,5)	(−36,−**26**,−15)	(−42,−**32**,−22)	(−59,−**51**,−43)
P.orb.ecc	(13,**24**,35)	(32,**42**,51)	⋆	(−16,−5,6)	(2,**13**,24)	(−3,9,20)	(−4,7,19)	(−25,−**14**,−3)	(−29,−**18**,−7)	(−37,−**27**,−16)
S.temp	(−16,−**5**,7)	(−19,−8,3)	(−10,1,12)	⋆	(28,**38**,48)	(44,**53**,60)	(3,**14**,25)	(−39,−**29**,−18)	(26,**36**,46)	(−2,10,21)
S.radius	(−9,3,14)	(−6,6,17)	(−10,1,13)	(−11,1,12)	⋆	(82,**85**,88)	(−15,−4,7)	(−95,−**94**,−92)	(31,**41**,50)	(−42,−**33**,−22)
S.mass	(3,**14**,25)	(−17,−6,5)	(−14,−3,8)	(24,**35**,44)	(40,**49**,57)	⋆	(−3,8,19)	(−77,−**72**,−66)	(28,**38**,47)	(−36,−**26**,−15)
S.metal	(−1,10,21)	(−17,−6,5)	(−3,8,20)	(−5,7,18)	(−4,7,18)	(5,**16**,27)	⋆	(−3,8,19)	(−9,3,14)	(0,11,22)
S.gravity	(−8,3,14)	(−11,0,12)	(−16,−4,7)	(−7,4,15)	(−74,−**68**,−62)	(26,**36**,46)	(−11,1,12)	⋆	(−45,−**35**,−25)	(24,**34**,43)
S.dist	(2,**13**,24)	(−15,−4,8)	(−21,−9,2)	(−11,0,11)	(45,**54**,61)	(−1,11,22)	(−34,−**24**,−13)	(−20,−9,2)	⋆	(64,**70**,76)
S.gaia	(−21,−10,1)	(−19,−**7**,4)	(−3,9,20)	(−6,5,16)	(−60,−**52**,−43)	(−24,**14**,−2)	(14,**25**,35)	(−1,10,21)	(95,**96**,97)	⋆

Markov equivalence class contains only one DAG. Even though this DAG contains fewer links than the undirected conditional dependence graphs, it does not necessarily encode a much sparser and/or very different dependence structure. As explained in Section 2.2.2, undirected and directed graphs handle conditioning sets in different ways. For example, S.gravity and S.mass are connected in the undirected graphs but are disconnected in the DAG. Yet, S.gravity → S.radius ← S.mass is an unshielded collider, so S.gravity and S.mass are actually dependent conditionally on all other variables (S.radius included) also in the DAG. Similar observation can be made about the unshielded collider S.dist → S.gaia ← S.metal.

A considerable number of dependencies can be observed, both positive and negative, spanning stellar, planetary, and orbital parameters. The partial correlation graphs in Fig. 4(c)–(e) are sparser than the marginal correlation graphs in Fig. 4(a)–(b), with (c) being the sparsest of all as Bonferroni multiple hypothesis testing correction is known to be very conservative. Thus, numerous apparent pairwise relationships weaken or disappear once the effects of other variables are accounted for in the graphical model. Indeed, conditional dependencies provide a more complete and reliable explanation of the dependence structure among variables. Let us focus on the Bonferroni graphs in Fig. 4(a) and (c), whose hypothesis tests are summarized in Fig. 3(b), where the symbol ∗ denotes cases where only the marginal or the partial correlation is significant. A total of twenty marginal correlations that were found to be significantly different from zero have corresponding partial correlations that are not statistically significant. For example, the marginal correlation between S.temp and S.dist, likely reflects observational selection biases since exoplanet surveys preferentially detect planets around bright, nearby stars, which tend to be hotter. However, this relationship disappears in the conditional dependence graphs, demonstrating how graphical models can help disentangle true physical dependencies from observational selection effects. Finally, note that the conditional dependence graph in Fig. 4c is stable across other common values of FWER $\in [0.001, 0.05]$, with only the edge S.dist–P.mass being dropped for FWER = 0.001.

We now focus on interpreting the conditional dependence graphs from an astrophysical perspective. Most of the identified conditional dependencies in Fig. 4(c)–(e), especially the strongest ones that are present in all graphs, correctly reproduce several established astrophysical principles. The negative partial correlation between stellar surface gravity (S.gravity) and stellar radius (S.radius), and the positive partial correlation between stellar surface gravity and stellar mass (S.mass) are consistent with the *surface gravity formula* S.gravity $\propto$ S.mass/S.radius^2 (Carroll and Ostlie, 2017). The link between planet mass (P.mass) and orbit period (P.orb.per) reflects *Newton's version of Kepler's Third Law*, although P.mass is expected to play a small role in the determination of P.orb.period. The link between planet orbit period (P.orb.per) and eccentricity (P.orb.ecc) may reflect *tidal circularization*, for which planets with short periods orbit close to stars, whereby strong tidal forces cause friction and circularizing orbits over time (Penev et al., 2018; Mahmud et al., 2023). The link between planet mass (P.mass) and orbit eccentricity (P.orb.ecc) is consistent with the findings of Jurić and Tremaine (2008), Chatterjee et al. (2008) and Van Eylen and Albrecht (2015), for which higher mass planets tend to have higher orbital eccentricity. The positive partial correlations connecting stellar mass (S.mass) to stellar radius (S.radius) and stellar effective temperature (S.temp) correspond to well-established *main-sequence scaling* relations: more massive stars are larger (S.radius $\approx$ S.mass^a, $0.6 \leq a \leq 0.8$) and hotter (S.temp $\approx$ S.mass$^{0.5}$) (Kippenhahn et al., 1990).

The positive partial correlations connecting planet mass (P.mass) to stellar mass (S.mass) and stellar metallicity (S.metal) align with the *core accretion model* (Laughlin et al., 2004): more massive stars with higher metallicity possess heavier protoplanetary disks, providing a larger accumulation of solids to build massive planets. Our results are also consistent with Lozovsky et al. (2021), who found that more massive stars host planets that are systematically larger, and with Owen and Murray-Clay (2018), who found that metal-rich stars favor giant planet formation.

The positive partial correlation between apparent magnitude (S.gaia) and stellar distance (S.distance) reflects the *distance–modulus relation* (Carroll and Ostlie, 2017), which explains why more distant stars appear fainter. Finally, the positive partial correlation between apparent magnitude (S.gaia) and stellar metallicity (S.metal) is consistent with *stellar opacity theory* (Kippenhahn et al., 2012), whereby a star's chemical metals alter its atmospheric absorption, thereby shifting its perceived brightness in the Gaia photometric filters (Prusti et al., 2016).

Also the dependence structure represented by the CPDAG in Fig. 4(f) is mostly consistent with the aforementioned astrophysical principles, as we summarize in Table 3. The link S.mass → P.mass aligns with the core accretion model (Laughlin et al., 2004) and the results of Lozovsky et al. (2021). The unshielded collider S.mass → S.radius ← S.gravity is consistent with the law of gravitational acceleration at a surface (Carroll and Ostlie, 2017). The link S.temp → S.mass is consistent with the main sequence relation S.temp $\approx$ S.mass$^{0.5}$ (Kippenhahn et al., 1990), although the direction appears to be reversed. Apparent magnitude (S.gaia) is a descendant of all other nodes. In particular, the unshielded collider S.metal → S.gaia ← S.distance reflects the distance–modulus relation (Binney and Merrifield, 1998) and stellar opacity theory (Kippenhahn et al., 2012). The link S.temp → S.metal may reflect stellar opacity effects governed by the *Vogt–Russell Theorem* and *Kramers' Opacity Law* (Kippenhahn et al., 1990; Carroll and Ostlie, 2017), although the direction is most likely reversed. Moreover, the link P.orb.per → S.gaia is reasonable since a planet's orbital period drives the frequency of its star's Gaia G magnitude variations. Short-period planets produce frequent, detectable brightness changes via three main mechanisms: transits (blocking light), tidal distortions (ellipsoidal variations), and relativistic Doppler beaming (Shporer, 2017). Finally, the full connectivity between planet mass (P.mass), planet orbital period (P.orb.per), and planet orbit eccentricity (P.orb.ecc), with planet mass being the one directly affected by

Table 3
Astrophysical validation of the CPDAG in Fig. 4f.

S.mass → P.mass	*Core accretion model*
S.mass → S.radius ← S.gravity	*Law of gravitational acceleration at a surface*
S.metal → S.gaia ← S.dist	*Distance–modulus relation*
P.orb.per → S.gaia	*Transits, tidal distortions, relativistic Doppler beaming*
P.mass → P.orb.per	*Newton's version of Kepler's Third Law*
S.temp → S.mass	*Main sequence relation* (reversed)
P.orb.ecc → P.orb.per	*Tidal circularization* (reversed)
P.orb.ecc → P.mass	*Empirically observed* (reversed)
S.temp → S.metal	*Vogt–Russell Theorem* and *Kramers' Opacity Law* (reversed)
S.radius → P.mass	Indirect relationship (through S.mass)
S.radius → S.dist	Possible observational bias

stellar parameters, reflects Newton's version of Kepler's Third Law, tidal circularization (Penev et al., 2018; Mahmud et al., 2023), and empirical findings (Jurić and Tremaine, 2008; Chatterjee et al., 2008; Van Eylen and Albrecht, 2015), as discussed above, although some arrow directions appear to be reversed.

As discussed at the end of Section 2.2.2, a DAG estimated from observational data should not be interpreted strictly as a causal graph. The PC algorithm effectively estimates equivalence classes, but the algorithm's reliance on multiple hypothesis tests makes it vulnerable to incorrect edge identification. Moreover, unmeasured latent variables and selection bias can confound the results. Indeed, the directions of some of the identified connections appear to be reversed (S.temp → S.mass, P.orb.ecc → P.orb.per, P.orb.ecc → P.mass, S.temp → S.metal). These reversals may also be due to the way these parameters are collected. For example, the incorrect direction S.temp → S.mass may have arisen because, in observational practice, S.temp is measured and S.mass is inferred. The link S.radius → P.mass correctly reflect more of an indirect relationship through S.mass (core accretion model; Laughlin et al. (2004)) rather than a direct causality. Finally, undoubtedly, a star's physical radius (S.radius) does not causally affect its distance (S.dist) from Earth, so the link S.radius → S.dist appears to be an observational artifact.

While constraints based on known astrophysics could be imposed on the graph to avoid these situations, the beauty of probabilistic graphical models lies in their potential to shed light on the automated discovery of relationships among astronomical variables that are not yet well studied, where constraints cannot yet be imposed. Nonetheless, we intend to address constrained graph estimation in our future research.

4. Discussion

In this article, we explored the application of probabilistic graphical models in astronomy. Graphical models allow us to efficiently study the complex multivariate relationships among numerous variables. In an undirected graph, two nodes are disconnected if and only if the two variables they represent are independent, conditionally on all other variables. In a directed acyclic graph, any variable, conditionally on its parents, depends only on its parents and descendants.

After providing an overview of undirected and directed probabilistic graphical models, we applied these frameworks to detect dependencies among variables concerning exoplanets and host stars. Most of the detected dependencies reflected established planetary and stellar astrophysical principles, such as main-sequence scaling relations (Kippenhahn et al., 1990), the surface gravity formula (Carroll and Ostlie, 2017), the core accretion model (Laughlin et al., 2004), the distance–modulus relation (Carroll and Ostlie, 2017), and stellar opacity theory (Kippenhahn et al., 2012). Conditional dependence graphs capture the dependence structure of variables more precisely than marginal correlations, which can often be misleading. For example, in our data analysis, stellar surface gravity (S.gravity) and stellar mass (S.mass) exhibited a strong negative marginal correlation and a positive partial correlation, with the latter being the one consistent with astrophysical theory.

Our results suggest that graphical models not only can accurately retrieve known relations but also may serve as powerful exploratory tools for gaining insight into astronomical variables that are not yet well understood. Therefore, while still not widely applied in astronomy, we expect probabilistic graphical models to become a common statistical tool in future astronomical research.

The presented approaches and data analysis results are certainly not free from limitations. Just as marginal correlation can be misleading because it neglects the effects of other variables, partial correlation can also be misleading if it is not computed conditionally on all variables that may affect the dependence between the two variables. Thus, if there are unobserved variables that are not part of the graph, spurious connections may be created. Approaches like Chandrasekaran et al. (2012) could be used to address this issue. Moreover, missing data corrections (e.g., removing incomplete observations) can also induce sample selection bias. Methods for dependence estimation from structurally incomplete data (Vinci et al., 2019; Vinci, 2024; Steneman and Vinci, 2024) could help. Finally, while the graphs were selected using specific criteria, they should be regarded as point estimates, and methods of uncertainty quantification or edge stability could be implemented based on bootstrapping or Bayesian modeling (Wang, 2012; Vinci et al., 2018a,b). The latter may also be used to incorporate prior astrophysical knowledge that may improve undirected graph estimation and prevent or resolve ambiguities in the direction of arrows in DAGs. The investigation of all these other approaches, as well as other forms of regularization (Meinshausen and Bühlmann, 2006; Cai et al., 2011), will be the focus of our future research.

CRediT authorship contribution statement

Abigail Sheerin: Writing – review & editing, Writing – original draft, Validation, Investigation, Funding acquisition, Data curation. **Giuseppe Vinci:** Writing – review & editing, Writing – original draft, Visualization, Validation, Supervision, Software, Resources, Project administration, Methodology, Investigation, Funding acquisition, Formal analysis, Data curation, Conceptualization.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This research was funded by the Beutter Family Endowment for Excellence in Applied Mathematics and Computer Science and the Sheila Flynn Endowment for Excellence at the University of Notre Dame.

Data and code availability

The R package `astroggm` used to implement the analyses is available on GitHub at https://github.com/gvincistat/astroggm. The data set used in our analyses can also be found at https://github.com/gvincistat/astroggm, and was originally retrieved from the NASA Exoplanet Archive https://exoplanetarchive.ipac.caltech.edu/.

References

Akutsu, T., Ando, M., Arai, K., Arai, Y., Araki, S., Araya, A., et al., 2020. Application of independent component analysis to the iKAGRA data. Prog. Theor. Exp. Phys. 2020 (5), 053F01. doi:10.1093/ptep/ptaa056.

Allen, J.T., Hewett, P.C., Richardson, C.T., Ferland, G.J., Baldwin, J.A., 2013. Classification and analysis of emission-line galaxies using mean field independent component analysis. Mon. Not. R. Astron. Soc. 430 (4), 3510–3532. doi:10.1093/mnras/stt151.

Andreon, S., Hurn, M.A., 2010. The Poisson regression in astronomy, with an application to the luminosity function of galaxy clusters. Mon. Not. R. Astron. Soc. 404 (4), 1922–1937. doi:10.1111/j.1365-2966.2010.16410.x.

Bailey, S., 2012. Principal component analysis with noisy and/or missing data. Publ. Astron. Soc. Pac. 124 (919), 1015.

Barnes, R., Mullins, K., Goldblatt, C., Meadows, V.S., Kasting, J.F., Heller, R., 2013. Tidal venuses: Triggering a climate catastrophe via tidal heating. Astrobiology 13 (3), 225–250.

Baumstark, M.J., Vinci, G., 2024. Spiral-elliptical automated galaxy morphology classification from telescope images. Astron. Comput. 46, 100770.

Benjamini, Y., Yekutieli, D., 2001. The control of the false discovery rate in multiple testing under dependency. Ann. Statist. 29 (4), 1165–1188.

Bhatt, M., Wöhler, C., Grumpe, A., Hasebe, N., Naito, M., 2019. Global mapping of lunar refractory elements: Multivariate regression vs. machine learning. Astron. Astrophys. 627, A155.

Binney, J., Merrifield, M., 1998. Galactic Astronomy. Princeton University Press, Princeton, NJ.

Bonferroni, C., 1936. Teoria statistica delle classi e calcolo delle probabilita. In: Pubblicazioni Del R Istituto Superiore Di Scienze Economiche E Commericiali Di Firenze. vol. 8, pp. 3–62.

Cai, T., Liu, W., Luo, X., 2011. A constrained ℓ_1 minimization approach to sparse precision matrix estimation. J. Amer. Statist. Assoc. 106 (494), 594–607.

Carraro, G., Ng, Y.K., Portinari, L., 1998. On the galactic disc age–metallicity relation. Mon. Not. R. Astron. Soc. 296 (4), 1045–1056.

Carroll, B.W., Ostlie, D.A., 2017. An Introduction to Modern Astrophysics. Cambridge University Press.

Carvalho, C.M., West, M., 2007. Dynamic matrix-variate graphical models. Bayesian Anal. 2 (1), 69–97.

Chandrasekaran, V., Parrilo, P.A., Willsky, A.S., 2012. Latent variable graphical model selection via convex optimization. Ann. Statist. 40 (4), 1935–1967. doi:10.1214/11-AOS949.

Chatterjee, S., Ford, E.B., Matsumura, S., Rasio, F.A., 2008. Dynamical outcomes of planet-planet scattering. Astrophys. J. 686 (1), 580–602.

Chattopadhyay, T., Karmakar, S., Chattopadhyay, A.K., 2010. Exploring the BATSE Gamma-Ray burst catalog through cluster analysis. Astrophys. J. 711 (1), 258–267. doi:10.1088/0004-637X/711/1/258.

Chen, J., Kipping, D., 2017. Probabilistic forecasting of the masses and radii of other worlds. Astrophys. J. 834 (1), 17.

Christian, C., Janes, K., 1977. Multivariate analysis of spectrophotometry.. Publ. Astron. Soc. Pac. 89 (529), 415.

Chun, H., Chen, M., Li, B., Zhao, H., 2013. Joint conditional Gaussian graphical models with multiple sources of genomic data. Front. Genet. 4, 294.

Cisewski-Kehe, J., Schafer, C., 2019. Special issue on astrostatistics. CHANCE 32 (3), 4–5.

Constantina, K., Tsakalides, P., et al., 2017. Denoising galaxy spectra with coupled dictionary learning. In: 25th European Signal Processing Conference. EUSIPCO, pp. 161–165. doi:10.23919/EUSIPCO.2017.8081190.

Corbin, M.R., Urban, A., Stobie, E., Thompson, R.I., Schneider, G., 2001. A multivariate analysis of galaxies in the hubble deep field-north. Astrophys. J. 551 (1), 23.

Das, S., Chattopadhayay, T., Davoust, E., 2015. Multivariate analysis of the globular clusters in M87. Publ. Astron. Soc. Aust. 32, e041.

Dawid, A.P., Mortera, J., Vicard, P., 2007. Object-oriented Bayesian networks for complex forensic DNA profiling problems. Forensic Sci. Int. 169 (2–3), 195–205.

de Souza, R.S., Cameron, E., Killedar, M., Hilbe, J., Vilalta, R., Maio, U., Biffi, V., Ciardi, B., Riggs, J.D., 2015a. The overlooked potential of generalized linear models in astronomy - I. Binomial regression. Astron. Comput. 12, 21–34. doi:10.1016/j.ascom.2015.04.002.

de Souza, R.S., Hilbe, J.M., Buelens, B., Riggs, J.D., Cameron, E., Ishida, E.E.O., Chies-Santos, A.L., Killedar, M., 2015b. The overlooked potential of generalized linear models in astronomy - III. Bayesian negative binomial regression and globular cluster populations. Mon. Not. R. Astron. Soc. 453 (2), 1928–1940. doi:10.1093/mnras/stv1825.

Dieleman, S., Willett, K.W., Dambre, J., 2015. Rotation-invariant convolutional neural networks for galaxy morphology prediction. Mon. Not. R. Astron. Soc. 450 (2), 1441–1459. doi:10.1093/mnras/stv632.

Dobra, A., Hans, C., Jones, B., Nevins, J.R., Yao, G., West, M., 2004. Sparse graphical models for exploring gene expression data. J. Multivariate Anal. 90 (1), 196–212.

Dunn, O.J., 1961. Multiple comparisons among means. J. Amer. Statist. Assoc. 56 (293), 52–64.

Efstathiou, G., Fall, S.M., 1984. Multivariate analysis of elliptical galaxies. Mon. Not. R. Astron. Soc. 206 (3), 453–464.

Elliott, J., de Souza, R.S., Krone-Martins, A., Cameron, E., Ishida, E.E.O., Hilbe, J., 2015. The overlooked potential of generalized linear models in astronomy - II. Gamma regression and photometric redshifts. Astron. Comput. 10, 61–72. doi:10.1016/j.ascom.2014.12.001.

Eyer, L., Süveges, M., De Ridder, J., Regibo, S., Mowlavi, N., Holl, B., Rimoldini, L., Bouchy, F., 2019. Multivariate time-series analysis of variable objects in the gaia mission. Publ. Astron. Soc. Pac. 131 (1002), 088001.

Fabbiano, G., Shapley, A., 2002. A multivariate statistical analysis of spiral galaxy luminosities. II. Morphology-dependent multiwavelength emission properties. Astrophys. J. 565 (2), 908.

Feigelson, E.D., Babu, G.J., 1992. Linear regression in astronomy. II. Astrophys. J. 397, 55. doi:10.1086/171765.

Feigelson, E.D., Babu, G.J., 2012. Modern Statistical Methods for Astronomy: With R Applications. Cambridge University Press.

Feigelson, E.D., De Souza, R.S., Ishida, E.E., Babu, G.J., 2021. Twenty-first-century statistical and computational challenges in astrophysics. Annu. Rev. Stat. Appl. 8 (1), 493–517.

Fiorentin, P.R., Bailer-Jones, C., Lee, Y.S., Beers, T.C., Sivarani, T., Wilhelm, R., Prieto, C.A., Norris, J., 2007. Estimation of stellar atmospheric parameters from SDSS/SEGUE spectra. Astron. Astrophys. 467 (3), 1373–1387.

Fischer, D.A., Valenti, J., 2005. The planet-metallicity correlation. Astrophys. J. 622 (2), 1102–1117.

Fisher, R.A., 1924. The distribution of the partial correlation coefficient. Metron 3, 329–332.

Foygel, R., Drton, M., 2010. Extended Bayesian information criteria for Gaussian graphical models. Adv. Neural Inf. Process. Syst. 23.

Fraix-Burnet, D., Thuillard, M., Chattopadhyay, A.K., 2015. Multivariate approaches to classification in extragalactic astronomy. Front. Astron. Space Sci. 2, 3.

Freeman, P., Izbicki, R., Lee, A., Newman, J., Conselice, C., Koekemoer, A., Lotz, J., Mozena, M., 2013. New image statistics for detecting disturbed galaxy morphologies at high redshift. Mon. Not. R. Astron. Soc. 434 (1), 282–295.

Friedman, J., Hastie, T., Tibshirani, R., 2008. Sparse inverse covariance estimation with the graphical lasso. Biostatistics 9 (3), 432–441.

Gan, L., Vinci, G., Allen, G.I., 2022. Correlation imputation for single-cell RNA-seq. J. Comput. Biol. 29 (5), 465–482.

Hsieh, C.-J., Dhillon, I.S., Ravikumar, P.K., Sustik, M.A., 2011. Sparse inverse covariance matrix estimation using quadratic approximation. In: Advances in Neural Information Processing Systems. pp. 2330–2338.

Hsieh, C.-J., Sustik, M.A., Dhillon, I.S., Ravikumar, P.K., Poldrack, R., 2013. BIG & QUIC: Sparse inverse covariance estimation for a million variables. Adv. Neural Inf. Process. Syst. 26.

Ikotun, A.M., Ezugwu, A.E., Abualigah, L., Abuhaija, B., Heming, J., 2023. K-means clustering algorithms: A comprehensive review, variants analysis, and advances in the era of big data. Inform. Sci. 622, 178–210.

Jagvaral, Y., Singh, S., Mandelbaum, R., 2022. Intrinsic alignments of bulges and discs. Mon. Not. R. Astron. Soc. 514 (1), 1021–1033.

Jurić, M., Tremaine, S., 2008. Dynamical origin of extrasolar planet eccentricity distribution. Astrophys. J. 686 (1), 603–620.

Kelly, B.C., 2007. Some aspects of measurement error iterative estimation in linear regression of astronomical data. Astrophys. J. 665 (2), 1489–1506. doi:10.1086/519947.

Kim, E.J., Brunner, R.J., 2017. Star-galaxy classification using deep convolutional neural networks. Mon. Not. R. Astron. Soc. 464 (4), 4463–4475. doi:10.1093/mnras/stw2672.

Kipf, T.N., Welling, M., 2017. Semi-supervised classification with graph convolutional networks. In: International Conference on Learning Representations. ICLR, URL https://openreview.net/forum?id=SJU4ayYgl.

Kippenhahn, R., Weigert, A., Weiss, A., 1990. Stellar Structure and Evolution, vol. 192, Springer.

Kippenhahn, R., Weigert, A., Weiss, A., 2012. Stellar Structure and Evolution, second ed. In: Astronomy and Astrophysics Library, Springer-Verlag, Berlin Heidelberg, doi:10.1007/978-3-642-30304-3.

Kopparapu, R.K., Ramirez, R., Kasting, J.F., Eymet, V., Robinson, T.D., Mahadevan, S., Terrien, R.C., Domagal-Goldman, S., Meadows, V., Deshpande, R., 2013. Habitable zones around main-sequence stars: New estimates. Astrophys. J. 765 (2), 131.

Krumsiek, J., Suhre, K., Illig, T., Adamski, J., Theis, F.J., 2011. Gaussian graphical modeling reconstructs pathway reactions from high-throughput metabolomics data. BMC Syst. Biol. 5 (1), 21.

Lasue, J., Wiens, R., Clegg, S., Vaniman, D., Joy, K., Humphries, S., Mezzacappa, A., Melikechi, N., McInroy, R., Bender, S., 2012. Remote laser-induced breakdown spectroscopy (LIBS) for lunar exploration. J. Geophys. Res.: Planets 117 (E1).

Laughlin, G., Bodenheimer, P., Adams, F.C., 2004. The core accretion model predicts few jovian-mass planets orbiting red dwarfs. Astrophys. J. 612 (1), L73.

Lauritzen, S.L., 1996. Graphical Models. Oxford University Press.

Liu, H., Han, F., Yuan, M., Lafferty, J., Wasserman, L., 2012. High-dimensional semiparametric Gaussian copula graphical models. Ann. Statist. 40 (4), 2293–2326.

Lozovsky, M., Helled, R., Pascucci, I., Dorn, C., Venturini, J., Feldmann, R., 2021. Why do more massive stars host larger planets? Astron. Astrophys. 652, A110.

Lu, H., Zhou, H., Wang, J., Wang, T., Dong, X., Zhuang, Z., Li, C., 2006. Ensemble learning independent component analysis of normal galaxy spectra. Astron. J. 131 (2), 790–805. doi:10.1086/499304.

Magorrian, J., Tremaine, S., Richstone, D., Bender, R., Bower, G., Dressler, A., Faber, S.M., Gebhardt, K., Grillmair, C., Kormendy, J., Lauer, T., 1998. The demography of massive dark objects in galaxy centres. Astron. J. 115 (6), 2285–2305. doi:10.1086/300351.

Mahmud, M.M., Penev, K.M., Schussler, J.A., 2023. Measuring tidal dissipation in giant planets from tidal circularization. Mon. Not. R. Astron. Soc. 525 (1), 876–897.

Maino, D., Farusi, A., Baccigalupi, C., Perrotta, F., Banday, A.J., Bedini, L., Burigana, C., De Zotti, G., Górski, K.M., Salerno, E., 2002. All-sky astrophysical component separation with fast independent component analysis (FastICA). Mon. Not. R. Astron. Soc. 334 (1), 53–68. doi:10.1046/j.1365-8711.2002.05425.x.

Mazumder, R., Hastie, T., 2012. The graphical lasso: New insights and alternatives. Electron. J. Stat. 6, 2125.

Meinshausen, N., Bühlmann, P., 2006. High-dimensional graphs and variable selection with the Lasso. Ann. Statist. 34 (3), 1436–1462. doi:10.1214/009053606000000281.

Nascimento, M.L.F., 2022. In search of star clusters: An introduction to the K-means algorithm. J. Humanist. Math. 12 (1), 243–255. doi:10.5642/jhummath.202201.19.

Nemade, R., Jha, S., Anand, H., Chandrasekaran, S., 2024. Morphological classification of galaxies using enhanced deep learning model. In: 2024 5th International Conference on Communication, Computing & Industry 6.0. C2I6, pp. 1–6. doi:10.1109/c2i663243.2024.10895611.

Nissen, P.E., Gustafsson, B., 2018. High-precision stellar abundances of the elements: Methods and applications. Astron. Astrophys. Rev. 26 (1), 6.

Ortiz, A., Munilla, J., Álvarez-Illán, I., Górriz, J.M., Ramírez, J., Alzheimer's Disease Neuroimaging Initiative, 2015. Exploratory graphical models of functional and structural connectivity patterns for Alzheimer's disease diagnosis. Front. Comput. Neurosci. 9, 132.

Owen, J.E., Murray-Clay, R., 2018. Metallicity-dependent signatures in the Kepler planets. Mon. Not. R. Astron. Soc. 480 (2), 2206–2216.

Pati, R., Pujari, A.K., Gahan, P., Kumar, V., 2021. Independent component analysis: A review with emphasis on commonly used algorithms and contrast function. Comput. Y Sist. 25 (1), 97–115. doi:10.13053/cys-25-1-3449.

Penev, K., Bouma, L., Winn, J.N., Hartman, J.D., 2018. Empirical tidal dissipation in exoplanet hosts from tidal spin-up. Astron. J. 155 (4), 165.

Prusti, T., De Bruijne, J., Brown, A.G., Vallenari, A., Babusiaux, C., Bailer-Jones, C., Bastian, U., Biermann, M., Evans, D.W., Eyer, L., et al., 2016. The gaia mission. Astron. Astrophys. 595, A1.

Ravikumar, P., Wainwright, M.J., Raskutti, G., Yu, B., 2011. High-dimensional covariance estimation by minimizing ℓ_1-penalized log-determinant divergence. Electron. J. Stat. 5, 935–980.

Riess, A.G., et al., 2016. A 2.4% determination of the local value of the hubble constant. Astrophys. J. 826 (1), 56. doi:10.3847/0004-637X/826/1/56.

Rothman, A.J., Bickel, P.J., Levina, E., Zhu, J., 2008. Sparse permutation invariant covariance estimation. Electron. J. Stat. 2, 494–515.

Shporer, A., 2017. The astrophysics of visible-light orbital phase curves in the space age. Publ. Astron. Soc. Pac. 129 (977), 072001.

Spirtes, P., Glymour, C., 1991. An algorithm for fast recovery of sparse causal graphs. Soc. Sci. Comput. Rev. 9 (1), 62–72.

Steneman, J., Vinci, G., 2024. Covariance matrix completion via auxiliary information. doi:10.48550/arXiv.2402.05767, arXiv preprint arXiv:2402.05767.

Turner, S., Kelvin, L.S., Baldry, I.K., Lisboa, P.J., Longmore, S.N., Collins, C.A., Holwerda, B.W., Hopkins, A.M., Liske, J., 2018. Reproducible k-means clustering in galaxy feature data from the GAMA survey. Mon. Not. R. Astron. Soc. 482 (1), 126–144. doi:10.1093/mnras/sty2690.

Van Eylen, V., Albrecht, S., 2015. Eccentricity from transit photometry: Small planets in Kepler multi-planet systems have low eccentricities. Astrophys. J. 808 (2), 126.

Verma, T., Pearl, J., 1990. Equivalence and synthesis of causal models, uncertainty in artificial intelligence, Vol. 6, P. Bonissone, M. Henrion, LN Kanal and JF Lemmer.

Villanueva-Domingo, P., Villaescusa-Navarro, F., 2022. Learning cosmology and clustering with cosmic graphs. Astrophys. J. 937 (2), 115.

Vinci, G., 2024. Linked factor analysis. doi:10.48550/arXiv.2401.02953, arXiv preprint arXiv:2401.02953v1.

Vinci, G., Dasarathy, G., Allen, G.I., 2019. Graph quilting: Graphical model selection from partially observed covariances. doi:10.48550/arXiv.1912.05573, arXiv preprint arXiv:1912.05573.

Vinci, G., Freeman, P., Newman, J., Wasserman, L., Genovese, C., 2014. Estimating the distribution of galaxy morphologies on a continuous space. Proc. Int. Astron. Union 10 (S306), 68–71.

Vinci, G., Ventura, V., Smith, M., Kass, R.E., 2018a. Adjusted regularization in latent graphical models: Application to multiple-neuron spike count data. Ann. Appl. Stat. 12 (2), 1068–1095.

Vinci, G., Ventura, V., Smith, M.A., Kass, R.E., 2018b. Adjusted regularization of cortical covariance. J. Comput. Neurosci. 45 (2), 83–101.

Wang, H., 2012. Bayesian graphical lasso models and efficient posterior computation. Bayesian Anal. 7 (4), 867–886.

Wang, T., Ren, Z., Ding, Y., Fang, Z., Sun, Z., MacDonald, M.L., Sweet, R.A., Wang, J., Chen, W., 2016. FastGGM: An efficient algorithm for the inference of gaussian graphical model in biological networks. PLoS Comput. Biol. 12 (2), e1004755.

Winn, J.N., 2010. Exoplanet Transits and Occultations, vol. 55, University of Arizona Press Tucson, AZ.

Xu, X., Vinci, G., 2024. Forensic science and how statistics can help it: Evidence, likelihood ratios, and graphical models. Wiley Interdiscip. Rev.: Comput. Stat. 16 (5), e70006.

Yin, J., Li, H., 2011. A sparse conditional gaussian graphical model for analysis of genetical genomics data. Ann. Appl. Stat. 5 (4), 2630.

Yuan, M., 2010. High dimensional inverse covariance matrix estimation via linear programming. J. Mach. Learn. Res. 11 (Aug), 2261–2286.

Yuan, M., Lin, Y., 2007. Model selection and estimation in the Gaussian graphical model. Biometrika 94 (1), 19–35.

Zhang, Y., Zhao, Y., 2015. Astronomy in the big data era. Data Sci. J. 14, 11–11.

Zhu, Y., Cribben, I., 2018. Sparse graphical models for functional connectivity networks: best methods and the autocorrelation issue. Brain Connect. 8 (3), 139–165.