Bayesian graphical modeling for heterogeneous causal effects

Federico Castelletti | Guido Consonni

Department of Statistical Sciences, Università Cattolica del Sacro Cuore, Milan, Italy

Correspondence
Federico Castelletti, Università Cattolica del Sacro Cuore, Department of Statistical Sciences, 20123 Milan, Italy.
Email: federico.castelletti@unicatt.it

There is a growing interest in current medical research to develop personalized treatments using a molecular-based approach. The broad goal is to implement a more precise and targeted decision-making process, relative to traditional treatments based primarily on clinical diagnoses. Specifically, we consider patients affected by Acute Myeloid Leukemia (AML), an hematological cancer characterized by uncontrolled proliferation of hematopoietic stem cells in the bone marrow. Because AML responds poorly to chemotherapeutic treatments, the development of targeted therapies is essential to improve patients’ prospects. In particular, the dataset we analyze contains the levels of proteins involved in cell cycle regulation and linked to the progression of the disease. We evaluate treatment effects within a causal framework represented by a Directed Acyclic Graph (DAG) model, whose vertices are the protein levels in the network. A major obstacle in implementing the above program is represented by individual heterogeneity. We address this issue through a Dirichlet Process (DP) mixture of Gaussian DAG-models where both the graphical structure as well as the allied model parameters are regarded as uncertain. Our procedure determines a clustering structure of the units reflecting the underlying heterogeneity, and produces subject-specific estimates of causal effects based on Bayesian Model Averaging (BMA). With reference to the AML dataset, we identify different effects of protein regulation among individuals; moreover, our method clusters patients into groups that exhibit only mild similarities with traditional categories based on morphological features.

KEYWORDS
directed acyclic graph, Dirichlet process mixture, personalized treatment, subject-specific graph, tumor heterogeneity

1 | INTRODUCTION

1.1 | Background and motivation

Heterogeneity of individual responses to treatment is a pervasive aspect in a variety of clinical domains. An example is cancer which is not a single disease as it involves subtypes characterized by distinct sets of molecules; this implies that patients will react differently to the same treatment. Currently the best treatment is identified based on clinical diagnoses as opposed to a molecular-based approach which would allow a more precise and targeted approach to decision-making,
leading to better treatments and eventually more favorable prognoses. At present there exists a variety of methods to identify cancer subtypes,\textsuperscript{1,2} however, they cannot establish whether subtypes show heterogeneous responses toward a treatment. Identifying subtypes with heterogeneous treatment effects is a causal problem that need be addressed for a more informed decision strategy.\textsuperscript{3}

In this paper we consider a dataset concerning patients affected by Acute Myeloid Leukemia (AML). AML represents an aggressive hematological cancer characterized by uncontrolled proliferation of hematopoietic stem cells in the bone marrow. AML responds very poorly to chemotherapeutic treatments, with a 5-year overall survival rate of about 25%; the development of new targeted treatments therefore represents a key strategy to improve patients’ prospects. In addition, AML is widely heterogeneous with numerous genetic aberrations. Consequently, knowledge facilitating individualization of targeted therapies under development in AML is sorely needed. Several subtypes of AML have been identified mainly on the basis of morphologic features. However, interest has been recently focused on developing improved classification schemes that more accurately explain the AML heterogeneity and in particular its response to therapies.

The dataset analyzed here includes protein levels for 256 newly diagnosed AML patients and is provided as a supplement to Kornblau et al.\textsuperscript{4} Because protein function regulates the phenotypic characteristics of cancer, a proteomic-based classification can provide relevant information for pathogenesis and prognosis of patients. In this direction tumor profiling accounting for patients heterogeneity can provide insights on the effect of targeted therapies on individualized basis. We apply our methodology to the AML dataset to investigate the existence of clustering characterized by differential graph structures as well as model parameter configurations. In addition, we evaluate a battery of causal effects on selected proteins in the network whose regulation has been established to play a crucial role in AML progression and response to therapy. We then pair clusters with heterogeneity of causal effects, and identify broad patterns of potential use for targeted therapy.

1.2 Overview of our model

In this subsection we provide the broad picture underlying our model, while leaving technical specifications to later sections. In our setup we have a collection of variables \(\{X_1, \ldots, X_q\}\), exemplified by protein expressions in the AML dataset. When we are interested in isolating one of the variables as a response we label it as \(Y\). We consider a sample of size \(n\) from the \(q\)-dimensional population. To account for potential heterogeneity of the subjects we assume that the statistical model is a mixture of Directed Acyclic Graph (DAG) models. This means that, if we knew the mixture component of each sample case, its distributional family would factorize according to the corresponding DAG, equivalently it would be Markov with respect to that DAG. Graphical models\textsuperscript{5} are very effective for encoding conditional independencies. In our setting however we view a DAG as a causal model;\textsuperscript{6} in particular this means that the joint \textit{interventional} distribution of the variables, following a hypothetical intervention on a variable in the system, would still be Markov relative to the DAG, save for the component subjected to intervention, which will be specified separately. This stability assumption is crucial to define the notion of total causal effect on \(Y\) following an intervention on \(X_i\).\textsuperscript{7}

From a modeling perspective, we start by assuming that the joint observational distribution belongs to a Dirichlet Process (DP) mixture of Gaussian DAG models; see Müller and Mitra\textsuperscript{8} for an overview of DP mixtures. The mixing measure is over the space of priors on \((\mu, \mathbf{\Omega}, D)\) where \(D\) is a DAG and \((\mu, \mathbf{\Omega})\) the parameters of a \(q\)-variate Normal distribution having precision \(\mathbf{\Omega} \in P_D\), where \(P_D\) is the space of symmetric and positive-definite matrices constrained by being Markov relative to \(D\). We take the baseline measure of the DP as \(p(\mu, \mathbf{\Omega} | D)p(D)\), and specify \(p(\mu, \mathbf{\Omega} | D)\) as a Normal-DAG-Wishart prior after a suitable reparameterization involving a Cholesky-type decomposition of \(\mathbf{\Omega}\).\textsuperscript{9} We employ a constructive procedure which produces a prior under \textit{any} DAG model based on the elicitation of a \textit{single} standard Normal-Wishart prior on \((\mu, \mathbf{\Omega})\), with \(\mathbf{\Omega}\) unconstrained. In this way, not only is the elicitation procedure drastically simplified, but the marginal likelihood of DAGs belonging to the same Markov class can be shown to be constant (score equivalence), besides being available in closed form. Computations for our model are performed through a Markov Chain Monte Carlo (MCMC) strategy based on a slice sampler\textsuperscript{10} and the Partial Analytic Structure algorithm of Godsill.\textsuperscript{11}

Our contribution can be summarized as follows. We provide a Bayesian modeling framework to evaluate heterogeneous causal effects based exclusively on \textit{observational}, i.e. non-experimental data. Specifically, our model: (i) accounts for individual heterogeneity through an infinite mixture of DAGs models; (ii) allows for structure (DAG) and parameter uncertainty; (iii) determines a clustering structure of the units; (iv) produces subject-specific causal effects incorporating
uncertainty on various aspects of the model through BMA. When applied to the AML data it highlights that protein regulation produces heterogeneous effects, and identifies clusters of patients potentially benefiting from selective interventions.

1.3 Related work

The literature on the analysis of heterogeneous causal effects has been growing lately, especially in the machine learning community where tree-based methods, in particular random forests, have found natural applications. Works in this area have been carried out mostly within the potential outcome framework; see for instance Athey and Imbens, Wager and Athey, Lee et al. and Hahn et al. for a Bayesian viewpoint.

A more decision theoretic approach is presented in Shpitser and Sherman who discuss identification of personalized effects in causal pathways and explore settings where the goal is to map a unit’s characteristics to a treatment tailored to maximize the expected outcome for that unit. This line of approach is also connected to literature spanning from dynamic treatment regime to mediation analysis.

Moving to a Bayesian perspective, inference on causal effects within the causal graph framework has been traditionally carried out assuming a homogeneous population, that is the observations are (conditionally) independent and identically distributed from a zero-mean Gaussian DAG-model; see for instance Castelletti and Consonni.

So far heterogeneity has been mostly linked to differential Bayesian structural learning as in multiple graphical modeling, where each model is associated to a specific group which is known in advance; see Peterson et al. and Castelletti et al. for directed graphs. Also, Ni et al. provide an extension to multi-dimensional graphs. An important contribution for learning subject-specific graphical structures is contained in Ni et al. who incorporate heterogeneity in the model through covariates that must be available at subject-level. A previous important attempt to deal explicitly with heterogeneity using a DP mixture of Gaussian graphical models is Rodriguez et al. There are however notable differences with respect to our work. Firstly they only consider structural and parameter learning for undirected graphs; secondly there is no discussion of causal inference. Similarly to Rodriguez et al., Talluri et al. consider Bayesian inference of both finite and infinite mixtures of undirected graphs. Moving to directed graphs, an early paper dealing with mixtures of DAGs is Ickstadt et al. There are substantial differences in comparison with our framework. First, the authors consider a unique graph structure which is common to all clusters. Accordingly, differences between clusters are allowed only in terms of DAG parameters (ie, regression coefficients and conditional variances of the allied Structural Equation Model, SEM). In addition, since the interest is in learning the (common) DAG structure and clustering of the units only rather than in estimating model parameters, the authors implement a marginal (ie, “collapsed” w.r.t. DAG parameters) algorithm. As a consequence, their method does not provide inference on cluster-specific parameters as well as causal effects, which is instead given by our method. Also, when networks are suitable modeled through a Reciprocal Graphical Model, then a relevant paper for heterogeneity is Ni et al. Finally, Bayesian nonparametric techniques have been used in graphical modeling also for robustness purposes: see Finegold and Drton and Cremaschi et al. for an extension to more general measures.

Bayesian nonparametric methods for causal inference have been also employed in a potential outcome framework; see for instance Rubin. In particular, Roy et al. consider marginal structural models to evaluate the causal effect of a treatment on a survival outcome and allow for heterogeneity by implementing a dependent DP for the response given a set of confounders. Moreover, Oganisian et al. implement a DP mixture of zero inflated regression models for pathological data exhibiting excesses of zeros; their method allows for prediction, causal effect estimation and clustering of patients into homogeneous groups sharing the same propensity score distribution.

We close this subsection by pointing to a more foundational line of research that connects invariance, causality and robustness where an important role is played by heterogeneity; the latter however relates to different “environments,” seen and unseen, and thus has a broader scope than the notion of heterogeneity employed in this paper.

1.4 Structure of the paper

The organization of the rest of this paper is as follows. In Section 2 we provide some background results on causal effect estimation based on DAGs. We then introduce in Section 3 our DP mixture of Gaussian DAG models with particular
emphasis on the specification of the baseline prior distribution for model parameters. Posterior inference is discussed in Section 4. In Section 5 we conduct extensive simulation experiments to evaluate the proposed method in terms of structural learning, clustering and causal effect estimation. Section 6 is entirely devoted to the analysis of the AML data, highlighting the heterogeneity of causal effects as well as the clustering structure. Finally, Section 7 offers a few points for discussion. Some theoretical results on parameter prior distributions and computational details are reported in Appendix S1.

2  |  BACKGROUND

2.1  |  DAGs and causal effects

Let \(X_1, \ldots, X_q\) be a collection of real-valued continuous random variables with joint p.d.f. \(f(x_1, \ldots, x_q)\). Let also \(D = (V, E)\) be a DAG, where \(V = \{1, \ldots, q\}\) is a set of nodes associated to variables \(X_1, \ldots, X_q\) and \(E \subseteq V \times V\) is a set of edges. If \((u, v) \in E\), then \((v, u) \notin E\) and we say that \(u\) is a parent of node \(v\); by converse we say that \(v\) is a child of \(u\). The set of all parents of \(v\) in \(D\) is then denoted by \(\text{pa}_D(v)\). Under DAG \(D\) the joint density factorizes as

\[
f(x_1, \ldots, x_q | D) = \prod_{j=1}^{q} f(x_j | \text{pa}_D(j)).
\]  

Factorization (1) is often called the Markov property and determines conditional independence relations among (set of) nodes which can be read-off from the DAG using graphical criteria. We also assume faithfulness of \(f(\cdot)\) to \(D\) which prescribes that the conditional independencies implied by (1) are exactly those graphically encoded by \(D\). We remark that faithfulness holds, up to sets of Lebesgue measure zero, under many common families of distributions such as the Gaussian model (Section 2.2).

Consider now a (deterministic) intervention on variable \(X_s\) which consists in setting \(X_s\) to the value \(\tilde{x}\) and is denoted as \(\text{do}(X_s = \tilde{x})\). The postintervention density is\(^6\)

\[
f(x_1, \ldots, x_q | \text{do}(X_s = \tilde{x})) = \begin{cases} \prod_{j=1,j\neq s}^{q} f(x_j | \text{pa}_D(j)) |_{x_j = \tilde{x}} & \text{if } x_s = \tilde{x}, \\ 0 & \text{otherwise}, \end{cases}
\]

where, importantly, each term \(f(x_j | \text{pa}_D(j))\) in (2) is the corresponding (pre-intervention) conditional density of Equation (1); this implies that the data generating mechanism is stable under intervention, because the latter only affects the local component distribution \(f(x_s | \text{pa}_D(s))\) which is reduced to a point mass on \(\tilde{x}\).

In particular, we are interested in evaluating the causal effect of an intervention \(\text{do}(X_s = \tilde{x})\) on a response variable \(Y\); by convention we set \(X_1 = Y\). The postintervention distribution of \(Y\) is then obtained by integrating (2) w.r.t. \(x_2, \ldots, x_q\) which simplifies to

\[
f(y | \text{do}(X_s = \tilde{x})) = \int f(y | \tilde{x}, \text{pa}_D(s)) f(x_{\text{pa}_D(s)}) \, dx_{\text{pa}_D(s)};
\]

if \(Y \notin \text{pa}_D(s)\); on the other hand if \(Y \in \text{pa}_D(s)\) then \(f(y | \text{do}(X_s = \tilde{x})) = f(y)\); see Pearl\(^6\) (theorem 3.2.2). Equation (3) uses the most common adjustment set, namely the set of parents of \(X_s\); other valid adjustment sets are however possible; see Witte et al.\(^{15}\). Also, it is common to summarize the causal effect on \(Y\) of an intervention on \(X_s\) through the derivative of the expected value of (3)

\[
\gamma_s := \frac{\partial}{\partial \tilde{x}} \mathbb{E}(Y | \text{do}(X_s = \tilde{x})) |_{x_s=\tilde{x}}.
\]

Clearly if \(f(\cdot)\) belongs to some parametric family indexed by \(\theta \in \Theta_D\) (a parameter specific to the underlying DAG), the causal effect \(\gamma_s\) will be a function of \(\theta\); accordingly, inference on \(\theta\) will drive inference on \(\gamma_s\).
2.2 Gaussian DAG models

In the following we focus on Gaussian DAG models and assume

$$X_1, \ldots, X_q \mid \mu, \Omega \sim \mathcal{N}_q(\mu, \Omega^{-1}),$$  \tag{5}$$

where $\mu = (\mu_1, \ldots, \mu_q)^T \in \mathbb{R}^q$ and $\Omega \in \mathcal{P},$ the set of all symmetric positive definite (s.p.d.) precision matrices Markov w.r.t. $D.$

Equation (5) can be alternatively written as a SEM. Given $\Sigma = \Omega^{-1},$ consider the reparameterization

$$L_{-j} = \sum_{k=1}^{j-1} \Sigma_{e_k e_j}, \quad D_j = \Sigma_{j \mid \text{pa}_D(j)}, \quad \eta_j = \mu_j + L_{-j}^T \mu_{\text{pa}_D(j)},$$  \tag{6}$$

for $j = 1, \ldots, q,$ where $\Sigma_{j \mid \text{pa}_D(j)} = \Sigma_j - \Sigma_{j >} \Sigma_{-j}, < j ] = \text{pa}_D(j) \times j, [ j >] = \text{pa}_D(j) \times j > \text{pa}_D(j).$ Parameters $L_{-j}$ correspond to the nonzero elements of a $(q, q)$ matrix $L$ with all diagonal entries equal to one. Moreover, if we let $D$ be a $(q, q)$ diagonal matrix with $(j,j)$-element $D_j$ and $\eta = (\eta_1, \ldots, \eta_q)^T,$ the SEM representation of (5) is given by

$$\eta + L^T (X_1, \ldots, X_q)^T = \epsilon,$$

where $\epsilon \sim \mathcal{N}_q(0, D).$ Equivalently, we can write

$$f(x_1, \ldots, x_q \mid \mu, \Omega, D) = \prod_{j=1}^{q} d\mathcal{N}(x_j \mid \eta_j - L_{-j}^T \mu_{\text{pa}_D(j)}, D_j).$$  \tag{7}$$

where $d\mathcal{N}(x \mid \mu, \sigma^2)$ denotes the Normal density of $\mathcal{N}(\mu, \sigma^2).$ Equation (7) represents the density of a Gaussian DAG model after the reparameterization $(\mu, \Omega) \leftrightarrow (\eta, L, D);$ compare also Equation (1). Finally we note that $\Omega = LD^{-1}L^T.$

Consider now the causal effect of an intervention $\text{do}(X_s = \bar{x})$ on $Y = X_1$ as defined in Equation (4). Under the Gaussian model (5), the postintervention distribution of $Y$ can be written as

$$f(y \mid \text{do}(X_s = \bar{x}), \mu, \Omega, D) = \int f(y \mid \bar{x}, x_{\text{pa}_D(s)}) f(x_{\text{pa}_D(s)} \mid \mu, \Omega) \, dx_{\text{pa}_D(s)},$$  \tag{8}$$

where each density under the integral sign is a suitable Normal. Taking the expectation of the postintervention distribution on the left-hand side of (8), and interchanging the order of integration in the right-hand side, one obtains

$$\mathbb{E}(Y \mid \text{do}(X_s = \bar{x}), \mu, \Omega, D) = y_0 + y_0 \bar{x} + y^T \mu_{\text{pa}_D(s)}$$  \tag{9}$$

so that, using (4), the causal effect of $\text{do}(X_s = \bar{x})$ on $Y$ is $\gamma_s,$ the coefficient associated to $X_s$ in the conditional expectation of $Y$ given $x_{\text{pa}_D(s)}$ with $\text{fa}_D(s) = s \cup \text{pa}_D(s).$ Therefore, the causal effect $\gamma_s$ can be retrieved from the covariance matrix $\Sigma = \Omega^{-1}$ as

$$\gamma_s = \left[ \Sigma_{Y \mid \text{fa}_D(s)} \left( \Sigma_{\text{fa}_D(s) \mid \text{fa}_D(s)} \right)^{-1} \right]_1,$$  \tag{10}$$

where subscript 1 refers to the first element of the vector, having implicitly assumed that variable “s” appears first in the set $\text{fa}_D(s).$

3 DP MIXTURE OF GAUSSIAN DAG MODELS

We consider a DP mixture of Gaussian DAG models so that

$$X_1, \ldots, X_q \mid H \sim \int f(x_1, \ldots, x_q \mid \mu, \Omega, D) \, H(d\mu, d\Omega, dD) \quad H \sim \text{DP}(a_0, M),$$  \tag{11}$$
where \( f(x_1, \ldots, x_q | \mu, \Omega, D) \) denotes the density of a Gaussian DAG model defined in (7), and \( H \) follows a DP with parameters \( a_0 \) (precision) and \( M \) (baseline), written \( H(\cdot) \sim \text{DP}(a_0, M) \). With regard to the baseline measure we set

\[
M(d\mu, d\Omega, dD) = p(\mu, \Omega | D)p(D) d\mu d\Omega dD,
\]

where priors \( p(\mu, \Omega | D) \) and \( p(D) \) will be shortly defined in Section 3.1.

Let now \( x_i = (x_{i1}, \ldots, x_{iq})^T, i = 1, \ldots, n \), be \( n \) independent draws from (11). Recall that in a DP mixture each sample \( x_i, i = 1, \ldots, n \), has potentially a distinct parameter \( \theta_i = (\mu_i, \Omega_i, D_i) \). Let \( K \leq n \) be the number of unique values among \( \theta_1, \ldots, \theta_n \) and \( \xi_1, \ldots, \xi_n \) a sequence of indicator variables, with \( \xi_i \in \{1, \ldots, K\} \), such that \( \theta_1 = \theta_{\xi_1}^* \). Denote now with \( X \) the \((n, q)\) data matrix obtained by row-binding the individual observations \( x_i^T \)'s. It is instructive to write the DP mixture models in terms of the random partition induced by the \{\( \xi_i \)\}'s,

\[
f(X | \xi_1, \ldots, \xi_n, K) = \prod_{k=1}^{K} \left\{ \int \prod_{j: \xi_j = k} f(x_i | \mu^*_k, \Omega^*_k, D^*_k) M(d\mu^*_k, d\Omega^*_k, dD^*_k) \right\}.
\]

Representation (13) is easily interpretable. The model groups observations into homogeneous classes, with samples within each class generated from a standard Gaussian DAG model. In practice, for a given partition, \( X \) is split into \( K \) submatrices \( X^{(k)}, k = 1, \ldots, K \), each \( X^{(k)} \) collecting all observations \( x_i \) such that \( \xi_i = k \).

### 3.1 Prior on DAG parameters

We now detail our choice of prior distributions \( p(\mu, \Omega | D) \) and \( p(D) \).

For a given DAG \( D \), let \( (\mu, \Omega) \) be the corresponding parameters, where \( \mu \in \mathbb{R}^q, \Omega \in \mathcal{P}_P \). We first consider the reparameterization \( (\mu, \Omega) \mapsto (\eta, L, D) \) introduced in Section 2.2. Our elicitation procedure relies on the method of Geiger and Heckerman. A main feature of this approach is that we only need to specify a prior for the parameters of a complete DAG model, \( \mathcal{N}_q(\mu, \Omega^{-1}) \), with \( \Omega \in \mathcal{P} \) unconstrained; the prior for any other (incomplete) DAG is then derived automatically, as we detail in Appendix S1. Additionally, and importantly, this procedure guarantees compatibility of priors in the sense that Markov equivalent DAGs are scored with the same marginal likelihood. Specifically, we show that a proper Normal-Wishart prior, \( (\mu, \Omega) \sim \mathcal{N}_q(\alpha_{\mu}, \mathbf{m}, a_{\Omega}, \mathbf{U}) \), leads to the compatible prior

\[
p(\eta, D, L | D) = \prod_{j=1}^q p(\eta_j, L_{<j|}, D_{<j})
\]

\[
= \prod_{j=1}^q p(\eta_j | L_{<j|}, D_{<j})p(L_{<j|} | D_{<j})p(D_{<j}),
\]

where

\[
D_{<j} \sim \text{I-Ga} \left( \frac{1}{2} a_{j^D}^D, \frac{1}{2} U_{j|<j}(\mathbf{1}) \right),
\]

\[
L_{<j|} | D_{<j} \sim \mathcal{N}_{|pa_{\eta}(j)|} \left( -U_{<j|}^{-1} U_{<j|}^{-1} D_{<j} U_{<j|}^{-1}, D_{<j} U_{<j|}^{-1} \right),
\]

\[
\eta_j | L_{<j|}, D_{<j} \sim \mathcal{N} \left( m_j + L_{<j|}^{T} m_{pa_{\eta}(j)} D_{<j}/a_{\mu}, D_{<j}/a_{\mu} \right)
\]

\[
\text{and } a_{j^D}^D = a_{\Omega} + |pa_{\eta}(j)| - q + 1.
\]

### 3.2 Prior on DAG structures

Let \( S_q \) be the space of all DAGs on \( q \) nodes. For a given DAG \( D = (V, E) \in S_q \), let \( S^B \) be the 0-1 adjacency matrix of its skeleton, that is the underlying undirected graph obtained after removing the orientation of all its edges. Accordingly,
for each \((u, v)\)-element in \(S^D\), \(S^D_{u, v} = 1\) if and only if \((u, v) \in E\) or \((v, u) \in E\), and zero otherwise. Conditionally on a prior probability of inclusion \(\pi \in (0, 1)\) we assume \(S^D_{u,v} \mid \pi \overset{\text{iid}}{\sim} \text{Ber}(\pi)\) for each \(u > v\). Therefore,

\[
p(S^D \mid \pi) = \pi^{|S^D|}(1 - \pi)^{|S^D| - |S^D|},
\]

where \(|S^D|\) is the number of edges in \(D\) (equivalently in its skeleton) and \(q(q - 1)/2\) corresponds to the maximum number of edges in a DAG on \(q\) nodes.

We then proceed hierarchically by assigning \(\pi \sim \text{Beta}(a, b)\). Integrating out \(\pi\), the resulting prior on \(S^D\) is

\[
p(S^D) \propto \Gamma \left( |S^D| + a \right) \Gamma \left( \frac{q(q-1)}{2} - |S^D| + b \right) \frac{\Gamma(a+b) \Gamma(a) \Gamma(b)}{\Gamma \left( \frac{q(q-1)}{2} + a + b \right)}.
\]

A similar prior was introduced by Scott and Berger\(^{37}\) for variable selection in linear models, where it was also shown to account for multiplicity correction. Finally, we set

\[
p(D) \propto p(S^D), \quad D \in S_q.
\]

Hyperparameters \(a\) and \(b\) can be chosen to reflect a prior knowledge of sparsity in the graph, if available; in the next section we fix for instance \(a = 1, b = (2q - 2)/3\), which is consistent with an expected prior probability of edge inclusion smaller than 0.5; see also Peters and Bühlmann.\(^{38}\) The default choice \(a = b = 1\), which corresponds to \(\pi \sim \text{Unif}(0, 1)\), can be instead adopted in the absence of substantive prior information.

Finally, the prior on the precision parameter is taken to be \(a_0 \sim \text{Gamma}(c, d)\). Hyperparameters \(c, d > 0\) control the prior number of clusters.\(^{39}\) A sensitivity analysis on a grid of values for \(c\) and \(d\) led to the choice (hereafter employed) \(c = 3, d = 1\), which results in a moderate expected number of groups, and a 90\% approximate prior credible interval \(1 < a_0 < 6\); see also Murugiah and Sweing\(^{40}\) for empirical approaches driving the choice of \(c\) and \(d\).

## 4 | POSTERIOR INFERENCE: CLUSTERING, STRUCTURAL LEARNING, AND CAUSAL EFFECTS

We implement an MCMC algorithm to sample from the posterior distribution of the DP mixture model (11). Our proposal relies on a slice sampler\(^{10}\) which is based on the number of explicitly represented mixture components and maintains the structure of a blocked Gibbs sampler. Full details are provided in Appendix S1.

The output of our MCMC scheme is a collection of \(S\) draws approximately sampled from the (augmented) posterior of \((\xi, \Theta^*)\) where \(\Theta^*\) is the triple \((\mu^*, \Omega^*, D^*)\). Specifically, for each MCMC iteration \(t = 1, \ldots, S\) our algorithm returns the \(n\)-dimensional vector of individual allocations \(\xi(t) = \left( \xi_1(t), \ldots, \xi_n(t) \right)\), with \(\xi_i(t) \in \{1, \ldots, K(t)\}\) where \(K(t)\) is the number of distinct clusters, together with the collection of \(K(t)\) distinct cluster-specific parameters \(\{\theta_1(t), \ldots, \theta_{K(t)}(t)\}\). From the MCMC output we can construct an \((n, n)\) posterior similarity matrix \(S\) whose \((i, i')\)-element represents the posterior probability that subjects \(i\) and \(i'\) belong to the same cluster, namely

\[
\hat{p}(\xi_i = \xi_{i'} \mid X) = \frac{1}{S} \sum_{t=1}^S 1 \left\{ \xi_i(t) = \xi_{i'}(t) \right\},
\]

The latter can be used to obtain an estimate \(\hat{c}\) of the partition induced by the DP, for example by including subject \(i\) and \(i'\) in the same cluster whenever \(\hat{p}(\xi_i = \xi_{i'} \mid X)\) exceeds a given threshold, say 0.5, as we do in the simulation results of Section 5. Alternative strategies for point estimation of clustering structures are available; see for instance Dahl et al.\(^{41}\) and references therein. In particular, we also considered the method of Wade and Ghahramani\(^{42}\) which takes as input the estimated posterior similarity matrix and estimates the clustering minimizing the posterior expectation of the Binder Loss (BL) and the Variation of Information (VI). In addition, the SALSO method was introduced
by Dahl et al.\textsuperscript{43} to improve on the computational scalability and accuracy w.r.t. previous methods. With regard to our simulation study we did not observe substantial differences among the clustering structures estimated by the various methods.

MCMC samples may also be used to provide subject-specific estimates of DAGs and parameters that are needed to estimate causal effects for each individual as set out in the motivations described in Section 1.1. To this end, we start by defining for each subject $i$ and edge $(u, v)$, $u \neq v$, the posterior probability of edge inclusion

\[ \hat{p}_i(u \to v \mid X) = \frac{1}{S} \sum_{s=1}^{S} \mathbb{I}\left\{(u, v) \in D_s^{(i)} \right\}. \]  

where, with a slight abuse of notation, $\mathbb{I}\{(u, v) \in D\} = 1$ if $D$ contains the edge $u \to v$, and zero otherwise. If we include only those edges $u \to v$ for which $\hat{p}_i(u \to v \mid X) > w$, the resulting graph can be adopted as a DAG estimate $\hat{D}_i$ provided that the graph is acyclic. In the following we fix $w = 0.5$.

Recall now the definition of causal effect $\gamma_s$ as a function of the precision (inverse-covariance) matrix in Equation (19). A BMA estimate of $\gamma_s$ for individual $i$ can be recovered from the MCMC output as

\[ \hat{\gamma}_{s,i} = \frac{1}{S} \sum_{s=1}^{S} \hat{\gamma}_{s,i}^{(s)}. \]

where $\hat{\gamma}_{s,i}^{(s)}$ is computed as in (10) by setting $\Sigma = \left[ \Omega_{s,i}^{(s)} \right]^{-1}$.

5 | SIMULATIONS

In this section we evaluate the performance of our method through simulation studies. Additional simulation results are included in Section 3 of Appendix S1.

5.1 | Settings

We consider settings with $q = 20$ nodes and $K = 2$ clusters. For cluster $k \in \{1, 2\}$ the sample size $n_k$ takes values in \{50, 100, 200, 500\}, and for each instance we set $n_1 = n_2$. In addition, cluster-specific parameters $\theta_k = (D_k, L_k, \eta_k, D_k)$ are generated under two scenarios. In Equal DAGs (scenario) we randomly generate a sparse DAG $D_1$ by fixing a probability of edge inclusion equal to 0.1 and set $D_2 = D_1$, which implies that the two DAG models are structurally equal; in Different DAGs (scenario) we instead generate $D_1$ and $D_2$ independently, so that $D_1$ and $D_2$ are different in general. DAG parameters $(D_k, L_k, \eta_k)$ are generated independently across $k = 1, 2$ by setting $D_k = I_q$, while uniformly sampling the nonzero elements of $L_k$ in $[-1, -0.1] \cup [0.1, 1]$; in addition, we sample the elements of each $\eta_k$ in the interval $[-b, b]$, with $b \in \{1, 2, 5\}$. Intuitively, higher values of $b$ lead to stronger separation between the means of the two groups, and this should improve cluster identification. We then set $U = I_q$, $a_\mu = 1$, $m = 0$, $a_{\omega} = q$ in the Normal-DAG-Wishart prior so that the prior is weakly informative because its weight corresponds to a sample of size one; see also Rodriguez et al.\textsuperscript{24} for a comparison. Furthermore, to favor sparsity, we fix $a = 1$, $b = (2q - 2)/3$ in the Beta prior on $\pi$ leading to the prior on DAGs (16). From further simulation experiments not reported for brevity it also appeared that results are quite insensitive to these hyperparameter choices, especially for large sample sizes. Under each scenario we then perform $N = 20$ simulations. Our MCMC scheme is implemented for a number of MCMC iteration $S = 25,000$, after having assessed its convergence through some pilot runs.

5.2 | Clustering

We first evaluate the performance of our method with regard to cluster allocation. To this end, we compare the true partition $c$ with the estimated partition $\hat{c}$ by means of the BL\textsuperscript{43} and the VI.\textsuperscript{44} The two metrics, normalized in $[0, 1]$ are
respectively defined as
\[
\text{BL}(c, \hat{c}) = \frac{2}{n(n-1)} \sum_{i<j} \left\{ 1(c_i = c_j, \hat{c}_i \neq \hat{c}_j) + 1(c_i \neq c_j, \hat{c}_i = \hat{c}_j) \right\},
\]
\[
\text{VI}(c, \hat{c}) = \frac{1}{\log(n)} \left\{ H(c) + H(\hat{c}) - 2I(c, \hat{c}) \right\},
\]
with \( H(c) = -\sum_{k=1}^K p(k) \log p(k) \) and \( I(c, \hat{c}) = \sum_{k=1}^K \sum_{h=1}^H p(k, h) \log \left( \frac{p(k, h)p(h)}{p(k)h} \right) \) representing the entropy associated to clustering \( c \), and the mutual information between the two clusterings \( c, \hat{c} \), where \( p(k) = \sum_i 1(c_i = k)/n \) and \( p(k, h) = \sum_i 1(c_i = k, \hat{c}_i = h)/n \); see also Meilă.\textsuperscript{44} Intuitively, lower values of the two indexes correspond to better performances in the clustering allocation. We compute BL and VI under each simulated dataset and scenario. Results are summarized in the plots of Figure 1. Each sequence of points joined by a dotted line represents the average values (w.r.t. the \( N = 20 \) simulations) of an index computed for increasing sample sizes \( n_k \) and one value of \( b \) (with increasing values of \( b \) from dark to light grey). It appears that the performance of the method improves as \( n_k \) grows under each scenario both in terms of BL and VI. Moreover, higher values of \( b \) make cluster identification easier even for moderate sample sizes, for example, \( n_k = 50 \). In addition, the clustering performance is better under Different DAGs scenario which corresponds to settings with DAGs generated independently and therefore also “structurally” different.

### 5.3 Structural learning

We now evaluate the performance of our method in learning the graph structures. To this end, under each simulation, we measure the Structural Hamming Distance (SHD) between each (individual) estimated DAG \( \hat{D}_i \), \( i = 1, \ldots, n \) and the corresponding true DAG. SHD corresponds to the number of edge insertions, deletions or flips needed to transform the estimated DAG into the true one; accordingly, lower values of SHD correspond to better performances.

For comparison purposes, we also include two alternative, yet opposed, learning strategies which are not based on DP mixture models. The first one corresponds to an oracle setting wherein the true two-group clustering is known beforehand. We call this benchmark Two-group oracle. The second instead wrongly assumes that all observations are conditionally iid from the same one-component model, and we name it One-group naive. Both benchmarks try to evaluate differential performance in structural learning: the former assesses the gain afforded by removing imperfect knowledge on clustering; the latter instead captures decay due to naively neglecting heterogeneity. In both benchmark strategies, while not running a DP mixture model, we use the same specifications for the prior on DAG- and parameter space.

Results, for each of the two scenarios Equal DAGs and Different DAGs, are summarized in the plots of Figure 2. Each box-plot represents the distribution of SHD, averaged with respect to individuals belonging to the same true cluster, with increasing group sample size \( n_k \in \{50, 100, 200, 500\} \), and increasing values of \( b \in \{1, 2, 5\} \) (from left to right panels). It appears that One-group naive performs worse w.r.t. the other two methods under both scenarios. In particular its performance worsens as \( n_k \) increases and for larger values of \( b \), because under both circumstances the two clusters become better separated. Reassuringly, Two-group oracle performs only slightly better than our DP mixture method even in the setting \( b = 1 \), where cluster identification is more difficult, with results nearly indistinguishable for \( n_k \in \{200, 500\} \). In addition, both methods improve their performance as \( n_k \) grows.

### 5.4 Causal effect estimation

We finally consider causal effect estimation. Under each scenario, we compare the collection of subject-specific BMA causal effect estimates \( \hat{\gamma}_{s,i} \), \( i = 1, \ldots, n \), \( s = 1, \ldots, q \), in Equation (19) with the corresponding true causal effects \( \gamma_{s,i} \) by means of the absolute-value distance
\[
d_{s,i} = |\hat{\gamma}_{s,i} - \gamma_{s,i}|.
\]
The distribution of \( d_{s,i} \) across subjects and simulated datasets is summarized through the average distance whose percentage values, computed under each scenario, are reported in Tables 1 and 2; the two tables refer to scenarios Equal DAGs and Different DAGs, respectively. For comparisons, we also compute the same collection of causal effect estimates
**FIGURE 1** Simulations. Average (w.r.t. 20 simulations) Variation of Information (VI) and Binder Loss (BL) index under Equal DAGs and Different DAGs scenarios, for increasing sample sizes $n_k \in \{50, 100, 200, 500\}$. Dark, middle, and light grey dots correspond to values of $b \in \{1, 2, 5\}$ respectively.
**Figure 2** Simulations. Structural Hamming Distance (SHD) between estimated and true Directed Acyclic Graphs (DAGs) under *Equal DAGs* and *Different DAGs* scenarios, for increasing sample sizes \( n_k \in \{50, 100, 200, 500\} \) and increasing values of \( b \in \{1, 2, 5\} \) (from left to right panels). Dark, medium, and light grey box-plots correspond to *DP mixture*, *Two-group oracle* and *One-group naive* strategies, respectively.

**Table 1** Simulations

| Method            | \( n_k = 50 \) | \( n_k = 100 \) | \( n_k = 200 \) | \( n_k = 500 \) |
|-------------------|----------------|----------------|----------------|----------------|
| \( b = 5 \)      |                |                |                |                |
| *DP mixture*      | 3.18           | 2.75           | 2.15           | 1.60           |
| *Two-group oracle*| 2.72           | 2.56           | 2.00           | 1.47           |
| \( b = 2 \)      |                |                |                |                |
| *DP mixture*      | 3.41           | 2.90           | 2.25           | 1.72           |
| *Two-group oracle*| 2.80           | 2.70           | 2.05           | 1.63           |
| \( b = 1 \)      |                |                |                |                |
| *DP mixture*      | 3.65           | 2.91           | 2.25           | 1.63           |
| *Two-group oracle*| 2.90           | 2.56           | 2.04           | 1.41           |

*Note:* *Equal DAGs* scenario. Average absolute-value distance (computed across simulations and subjects) between estimated and true subject-specific causal effect. Results reported for values of \( n_k \in \{50, 100, 200, 500\} \) and \( b = \{1, 2, 5\} \).

Abbreviations: DAG, Directed Acyclic Graph; DP, Dirichlet Process.
under Two-group oracle. It appears that, while both methods improve their performances as \( b \) and \( n_k \) grows, \textit{DP mixtures} performs only slightly worse than \textit{Two-group oracle}. In addition, such differences are more evident under scenarios with moderate sample size \( n_k \) and smaller values of \( b \), for example, \( b = 1 \), where indeed cluster allocation was also more difficult.

### 6 ANALYSIS OF AML DATA

In this section we apply our methodology to the protein dataset of patients affected by AML described in Section 1. This dataset contains the level of \( q = 18 \) proteins and phosphoproteins involved in apoptosis and cell cycle regulation according to the KEGG database\(^45\) for 256 newly diagnosed AML patients and is provided as a supplement to Kornblau et al.\(^4\) Classification of AML patients is commonly based on the French-American-British (FAB) system and relies on morphologic features, along with flow cytometric analysis of surface marker expression, cytogenetics, and assessment of recurrent molecular abnormalities. In particular, 11 of the FAB subtypes are present in the dataset, besides one group of patients with unknown subtype; see also Table 3.

The same dataset was analyzed by Peterson et al\(^{20}\) and Castelletti et al\(^{21}\) from a multiple graphical model-perspective. These authors included in their analysis four AML subtypes for which a reasonable sample size is available: M0 (17 subjects), M1 (34 subjects), M2 (68 subjects), and M4 (59 subjects). Specifically, assuming the four groups were given, both papers developed a Bayesian analysis which allows potentially common features in graphical structures—undirected in the first paper and directed in the second one—to be shared among groups. In the end, both methods revealed strong similarities between groups in terms of the estimated protein-network structures, but at the same time they were able to identify a few protein interactions specific to a given subtype. In this paper we take a different approach, and apply our mixture model to the \textit{full} dataset (including all the \( n = 256 \) subjects) without grouping the patients a priori, but rather letting the model cluster the observations as it learns the graphical structure of protein-protein interactions. Finally, we also evaluate the effect of interventions on proteins in the network at a subject-specific level.

Among the proteins included in the study, AKT belongs to the phosphoinositide 3-kinase (PI3K)-Akt-mammalian target of rapamycin (mTOR) pathway (PI3K-Akt-mTOR pathway) which is one of the intracellular pathways aberrantly up-regulated in AML.\(^{46}\) Activation of this pathway (eg, induced by AKT regulation) has been established to play an important role in leukemogenesis. In addition, targeting the PI3K-Akt-mTOR pathway with specific inhibitors may produce different effects on AML patients, reflecting biological heterogeneity in the intracellular signaling status.\(^{47}\) Because of the

### Table 3 Acute myeloid leukemia (AML) data

| Subtype | M0 | M1 | M2 | M4 | M4EOS | M5 | M5A | M5b | M6 | M7 | RAEBT | Unknown |
|---------|----|----|----|----|-------|----|-----|-----|----|----|-------|---------|
| Size    | 17 | 34 | 68 | 59 | 9     | 6  | 13  | 9   | 7  | 5  | 5     | 24      |

\textit{Note:} AML subtypes defined according to the French-American-British (FAB) system and corresponding number of subjects (size) in the dataset.
role played in AML progression and response to therapy, we therefore consider the AKT protein and phosphoproteins (AKT, AKT.p308, AKT.p473 in the following) as responses of interest for our causal-effect analysis.

We implement the proposed model by running our MCMC scheme for $S = 120,000$ iterations, which includes a burn-in period of 20,000 runs. We fix hyper-parameters $c = 4$, $d = 1$ in the Gamma prior on the precision parameter $a_0$, which is consistent with a moderate expected number of clusters $K$. In addition, we set $U = I_q$, $a_\mu = 1$, $m = 0$, $a_\Omega = q$ in the Normal-DAG-Wishart prior as in the simulation settings of Section 5.1. With regard to the Beta prior on the probability of edge inclusion $\pi$ we instead fix $a = 1$, $b = 10$, which reflects a moderate degree of sparsity in the network. To assess the convergence of our algorithm we also run independent MCMC chains with results suggesting a highly satisfactory agreement in terms of clustering, graph structure learning and causal effect estimation; see also the Supplementary material for further details.

Starting from the MCMC output we first produce the $(n, n)$ similarity matrix based on the posterior probabilities (17) computed for each pair of subjects $(i, i')$, $i \neq i'$. In particular, fixing a threshold for clustering inclusion equal to 0.5, we obtain a partition with two groups of size $n_1 = 105$, $n_2 = 151$. For ease of interpretation we numbered subjects in cluster 1 from 1 to 105, followed by those in cluster 2 from 106 to 256. Results are summarized in the heat map of Figure 3 where the axes report the ordered subjects. We note that in both estimated clusters there exist a few subjects appearing to be borderline in the sense that they barely qualify for membership to the assigned group (their inclusion probabilities in groups 1 and 2 are approximately equal).

We now focus on graph structure learning. Specifically, we construct for each subject $i = 1, \ldots, n$ a $(q, q)$ matrix collecting the posterior inclusion probabilities of each (directed) edge $(u, v)$, $u \neq v$. Results for two randomly chosen subjects, whose membership is estimated to be cluster 1 and cluster 2 respectively, are reported in Figure 4. The two heat maps reveal an appreciable degree of sparsity in each of the two underlying DAG structures, together with some noticeable differences in the network links.

Using Equation (19), we can provide a subject-specific BMA estimate of the causal effect on each of the responses AKT, AKT.p308, AKT.p473 following an intervention on any other protein $s$ in the network leading to the collection $\gamma_{i,s}$, $s = 1, \ldots, 18$ for each subject $i = 1, \ldots, 256$. Results are summarized in the heat maps of Figure 5 where each plot refers to one of the three response variables. The pattern already observed in Figure 3 is also apparent. Subjects assigned to the same cluster reveal broadly similar causal effect estimates, with some notable exceptions in both groups for a few individuals.

More interestingly, the effect of an intervention varies across the two groups, showing that the cluster structure produced by our analysis has a causal counterpart. Take for instance the effect of protein PTEN onto AKT: this is more strongly positive in group 1 than in group 2. Conversely, if the response is AKT.p308 the effect in both groups is negative, and more pronounced in group 2. This finding is of potential interest as protein PTEN has been identified as a tumor suppressor because it is capable of breaking the PI3K-Akt-mTOR pathway, and therefore represents a common target.
for inactivation in cancers. \cite{48,49} From a personalized therapy perspective, one can then argue that AKT regulation can be induced through interventions which however should be selected at subject-specific level, because they can result in heterogeneous causal effects—and corresponding levels of efficacy—across different subjects.

In addition, to appreciate the role played by population heterogeneity in causal effect estimation, we compare our results with those based on the alternative One-group naive strategy (Section 5.3). Results are included in the right-side maps of Figure 5.

Clearly, in this setting, causal effects following any intervention are equal across subjects. This output reveals substantial differences relative to our previous analysis suggesting that methods which neglect population heterogeneity can produce misleading estimates of causal effects. In particular, the one-group assumption has in some cases a “dilution” effect on coefficients’ estimates. This means that each causal effect obtained from One-group naive is akin to an average of cluster-specific causal estimates which are substantially different among groups. This happens for instance with regard to response AKT for causal effects associated with protein PTEN. Here, the causal effect obtained from One-group naive corresponds to a value in between the collection of causal effects resulting from DP mixture. As a consequence the ensuing causal effect coefficient provides an inadequate quantification of the underlying effect because it under- and over-estimates causal effects for individuals in clusters 1 and 2, respectively.

Common practice would suggest to first cluster patients into groups according to selected covariates, whenever available, and then infer DAGs and causal effects for each group separately. In this connection, the leukemia subtypes (Table 3) may be used for this purpose. A comparison with this alternative approach is provided in Appendix S1, where it is shown that patients belonging to some subtypes, in particular M0, M1, M2, are not primarily assigned to either estimated cluster. As a consequence, for subjects belonging to these subtypes there are substantial differences in the resulting causal effect estimates. This finding suggests that methods which only rely on a predetermined group classification may produce inaccurate and possibly misleading estimates of causal effects.

7 | DISCUSSION

In this paper we present a Bayesian framework to evaluate heterogeneous causal effects based on DAGs. We model heterogeneity through an infinite mixture model which accounts both for structure and parameter uncertainty. Because of the discreteness of the process governing the generation of the individual parameters, a posterior distribution on the clustering structure of the units is also available.

Our method does not assume any variable ordering because this would impose a strong restriction on the edge orientations and consequently on the set of potential DAGs. Nevertheless, even with cross-sectional observational data, the directions of some edges can be established based only on conditional independence relationships embodied in the
FIGURE 5  Acute myeloid leukemia data. Heat maps of causal effects on responses AKT, AKT.p308, AKT.p473, following an intervention on one target protein among the 18 in the network (AKT, . . . , XIAP); left-side heat maps refer to DP-mixture; right-side heat maps to One-group naive.

sample. As a consequence, the orientation of the remaining edges is undetermined, because of Markov equivalence between DAGs encoding the same conditional independence statements. Two Markov equivalent DAGs however need not be equivalent from a causal perspective, in which case two different causal effects are obtained from the same intervention, and this is the reason why our method works at DAG level. We also stress that our setup does not select a single DAG but rather a whole collection of DAGs, so that each subject-specific causal effect is a Bayesian model average of causal effects across DAGs.

When substantive knowledge suggests the existence of directions for specific edges, our method could still be implemented by suitably limiting the DAG space compatibly with the assumed ordering. This operationally translates to a restricted set of moves between DAGs in the accompanying MCMC algorithm.
Our analysis is based on a DP mixture of Gaussian DAG models. While more general Bayesian nonparametric models might be adopted, for example, Müller and Mitra,\textsuperscript{8} and Barrios et al.\textsuperscript{50} we believe that the main content of our contribution, namely causal inference under heterogeneity based on DAGs, is best captured by the current DP mixture model because of its popularity, interpretability, and simplicity of implementation.

Protein and gene expression levels are known to be affected by genetic, environmental, demographic, and other factors. This means that, in addition to the measured variables, there will typically be sources of heterogeneity which are hidden or latent, and failing to incorporate these sources may have detrimental effects on the study.\textsuperscript{51} Currently we do not consider latent variables in our model. In principle they could be made part of our setup, along the lines of Frot et al\textsuperscript{52} for structure learning, and of Shpitser and Tchetgen Tchetgen\textsuperscript{53} for the identification of causal effects, although this would add a significant layer of complexity to the whole procedure.

The type of interventions we have considered may be called perfect, meaning that they eliminate dependencies between targeted variables and their direct causes; the identifiability of causal DAGs under perfect interventions was characterized by Hauser and Bühlmann.\textsuperscript{54} More recently the broader notion of general intervention has been introduced, which may modify the dependencies between targeted variables and their causes without eliminating them (see Yang et al\textsuperscript{55}).

ACKNOWLEDGMENTS

This research was partially supported by UCSC (D1 and 2019-D.3.2 research grants). We thank two reviewers, an Associate Editor and the Editor for constructive comments that helped improve the paper. We also thank Raffaele Argiento for useful discussions on the slice sampler and Leonardo Bottolo for suggestions about the choice of the application.

CONFLICT OF INTEREST

The authors declare no potential conflict of interests.

DATA AVAILABILITY STATEMENT

The AML data of Kornblau et al\textsuperscript{4} are provided as supplement to the original paper and publicly available at http://bioinformatics.mdanderson.org/supplements.html (under “RPPA Data in AML”). Codes implementing our method are publicly available at https://github.com/FedeCastelletti/bnp_mixture_causal_dags.

ORCID

Federico Castelletti https://orcid.org/0000-0001-7911-2942

REFERENCES

1. Wilkerson MD, Hayes DN. Consensus cluster plus: a class discovery tool with confidence assessments and item tracking. Bioinformatics. 2010;26(12):1572-1573.
2. Shen R, Olshen AB, Ladanyi M. Integrative clustering of multiple genomic data types using a joint latent variable model with application to breast and lung cancer subtype analysis. Bioinformatics. 2009;25(22):2906-2912.
3. Zhang W, Le TD, Liu L, Zhou ZH, Li J. Mining heterogeneous causal effects for personalized cancer treatment. Bioinformatics. 2017;33(15):2372-2378.
4. Kornblau S, Tibes R, Qiu Y, et al. Functional proteomic profiling of AML predicts response and survival. Blood. 2009;1(113):154-164.
5. Lauritzen SL. Graphical Models. Oxford, UK: Oxford University Press; 1996.
6. Pearl J. Causality: Models, Reasoning, and Inference. Cambridge, UK: Cambridge University Press; 2000.
7. Maathuis M, Nandy P. A review of some recent advances in causal inference. In: Bühlmann P, Drineas P, Kane M, van de Laan M, eds. Handbook of Big Data. New York: Chapman and Hall/CRC Press; 2016:387-408.
8. Müller P, Mitra R. Bayesian nonparametric inference – Why and how. Bayesian Anal. 2013;8(2):269-302.
9. Cao X, Khare K, Ghosh M. Posterior graph selection and estimation consistency for high-dimensional Bayesian DAG models. Ann Stat. 2019;47(1):319-348.
10. Walker SG. Sampling the Dirichlet mixture model with slices. Commun Stat Simul Comput. 2007;36(1):45-54.
11. Godsill SJ. On the relationship between Markov chain Monte Carlo methods for model uncertainty. J Comput Graph Stat. 2012;10(2):230-248.
12. Rubin DB. Estimating causal effects of treatments in randomized and nonrandomized studies. J Educ Psychol. 1974;66(5):688-701.
13. Athey S, Imbens G. Recursive partitioning for heterogeneous causal effects. Proc Natl Acad Sci. 2016;113(27):7353-7360.
14. Wager S, Athey S. Estimation and inference of heterogeneous treatment effects using random forests. J Am Stat Assoc. 2018;113(523):1228-1242.
15. Lee K, Bargagli-Stoffi FJ, Dominici F. Causal rule ensemble: interpretable inference of heterogeneous treatment effects. arXiv preprint, 2020.
16. Hahn PR, Murray JS, Carvalho CM. Bayesian regression tree models for causal inference: regularization, confounding, and heterogeneous effects (with discussion). *Bayesian Anal.* 2020;15(3):965-1056.
17. Shpitser I, Sherman E. Identification of personalized effects associated with causal pathways. Proceedings of the 34th Annual Conference on Uncertainty in Artificial Intelligence UAI-18; 2018; IEEE.
18. Castelletti F, Consonni G. Bayesian inference of causal effects from observational data in Gaussian graphical models. *Biometrics.* 2021;77(1):136-149.
19. Castelletti F, Consonni G. Bayesian causal inference in probit graphical models. *Bayesian Anal.* 2021;16(4):1113-1137.
20. Petersen C, Stingo FC, Vannucci M. Bayesian inference of multiple Gaussian graphical models. *J Am Stat Assoc.* 2015;110(509):159-174.
21. Castelletti F, La Rocca L, Peluso S, Stingo FC, Consonni G. Bayesian learning of multiple directed networks from observational data. *Stat Med.* 2020;39(30):4745-4766.
22. Ni Y, Stingo FC, Baladandayuthapani V. Sparse multi-dimensional graphical models: a unified Bayesian framework. *J Am Stat Assoc.* 2017;112(518):779-793.
23. Ni Y, Stingo FC, Baladandayuthapani V. Bayesian graphical regression. *J Am Stat Assoc.* 2019;114(525):184-197.
24. Rodriguez A, Lenkoski A, Dobra A. Sparse covariance estimation in heterogeneous samples. *Electr J Stat.* 2011;5:981-1014.
25. Talluri R, Baladandayuthapani V, Mallick BK. Bayesian sparse graphical models and their mixtures. *Stat.* 2014;3(1):109-125.
26. Ickstadt K, Bornkamp B, Grzegorczyk M, et al. Nonparametric Bayesian networks. In: Bernardo JM, Bayarri MJ, Berger JO, et al., eds. *Bayesian Statistics.* Vol 9. Oxford: Oxford University Press; 2011:283-316.
27. Ni Y, Müller P, Zhu Y, Ji Y. Heterogeneous reciprocal graphical models. *Biometrics.* 2018;74(2):606-615.
28. Finegold M, Drton M. Robust graphical modeling of gene networks using classical and alternative t-distributions. *Ann Appl Stat.* 2011;5(2A):1057-1080.
29. Finegold M, Drton M. Robust Bayesian graphical modeling using Dirichlet t-distributions. *Bayesian Anal.* 2014;9(3):521-550.
30. Cremaschi A, Argiento R, Shoemaker K, Peterson C, Vannucci M. Hierarchical normalized completely random measures for robust graphical modeling. *Bayesian Anal.* 2019;14(4):1271-1301.
31. Rubin DB. Causal inference using potential outcomes. *J Am Stat Assoc.* 2005;100(469):322-331.
32. Roy J, Lum KJ, Daniels MJ. A Bayesian nonparametric approach to marginal structural models for point treatments and a continuous or survival outcome. *Biostatistics.* 2016;18(1):32-47.
33. Oganisian A, Mitra N, Roy JA. A Bayesian nonparametric model for zero-inflated outcomes: prediction, clustering, and causal estimation. *Biometrics.* 2021;77(1):125-135.
34. Buhlmann P. Invariance, causality and robustness. *Stat Sci.* 2020;35(3):404-426.
35. Witte J, Hencckel L, Maathuis MH, Didelez V. On efficient adjustment in causal graphs. *J Mach Learn Res.* 2020;21(246):1-45.
36. Geiger D, Heckerman D. Parameter priors for directed acyclic graphical models and the characterization of several probability distributions. *Ann Stat.* 2002;30(5):1412-1440.
37. Scott JG, Berger JO. Bayes and empirical-Bayes multiplicity adjustment in the variable-selection problem. *Ann Stat.* 2010;38(5):2587-2619.
38. Peters J, Buhlmann P. Identifiability of Gaussian structural equation models with equal error variances. *Biometrika.* 2014;101(1):219-228.
39. Escobar MD, West M. Bayesian density estimation and inference using mixtures. *J Am Stat Assoc.* 1995;90(430):577-588.
40. Murugiah S, Sweeting T. Selecting the precision parameter prior in Dirichlet process mixture models. *J Am Stat Assoc.* 2015;110(509):159-174.
55. Yang K, Katcoff A, Uhler C. Characterizing and learning equivalence classes of causal DAGs under interventions. In: Dy J, Krause A, eds. *Proceedings of the 35th International Conference on Machine Learning. Volume 80 of Proceedings of Machine Learning Research*. Stockholm: PMLR; 2018:5541-5550.

**SUPPORTING INFORMATION**

Additional supporting information can be found online in the Supporting Information section at the end of this article.

**How to cite this article:** Castelletti F, Consonni G. Bayesian graphical modeling for heterogeneous causal effects. *Statistics in Medicine*. 2023;42(1):15-32. doi: 10.1002/sim.9599