BIOMETRIC METHODOLOGY

Functional Bayesian networks for discovering causality from multivariate functional data

Fangting Zhou¹,² | Kejun He² | Kunbo Wang³ | Yanxun Xu³ | Yang Ni¹

¹Department of Statistics, Texas A&M University, College Station, Texas, USA
²Center for Applied Statistics, Institute of Statistics and Big Data, Renmin University of China, Beijing, China
³Department of Applied Mathematics and Statistics, Johns Hopkins University, Baltimore, Maryland, USA

Correspondence
Yang Ni, Department of Statistics, Texas A&M University, College Station, TX, USA.
Email: yangni87@gmail.com

Abstract
Multivariate functional data arise in a wide range of applications. One fundamental task is to understand the causal relationships among these functional objects of interest. In this paper, we develop a novel Bayesian network (BN) model for multivariate functional data where conditional independencies and causal structure are encoded by a directed acyclic graph. Specifically, we allow the functional objects to deviate from Gaussian processes, which is the key to unique causal structure identification even when the functions are measured with noises. A fully Bayesian framework is designed to infer the functional BN model with natural uncertainty quantification through posterior summaries. Simulation studies and real data examples demonstrate the practical utility of the proposed model.

KEYWORDS
causal discovery, directed acyclic graphs, multivariate longitudinal/functional data, non-Gaussianity, structure learning

1 | INTRODUCTION

This paper develops a novel functional Bayesian network (BN) for modeling directed conditional independence and causal relationships of multivariate functional data, which arise in a wide range of applications. For example, learning brain effective connectivity networks from electroencephalogram (EEG) records is crucial for understanding brain activities and neuron responses. Another example is longitudinal medical studies where multiple clinical variables are recorded at possibly distinct time points across variables and/or patients. Knowing the causal dependencies of these clinical variables may help physicians decide on the right interventions. Functional data can also go beyond those defined on the time domain, for example, on the spatial domain (environmental data, spatially resolved transcriptomics, etc.).

Joint analysis of multiple functional objects has continually attracted great attention in recent years with focuses mainly on reducing dimensionality and capturing functional dependencies. For instance, Kowal et al. (2017) and Kowal (2019) proposed to model time-ordered functional data through a time-varying parameterization for functional time series. Using basis transformation strategies,
Zhang et al. (2016) built an autoregressive model for spatially correlated functional data, while Lee et al. (2018) modeled functional data in serial correlation semiparametrically. Chiou and Müller (2014) developed a linear manifold model characterizing the functional dependencies between multiple random processes.

1.1 Functional graphical models

In a similar but conceptually different manner, functional graphical models have been recently proposed to model conditional independencies of multivariate functional data. Graphical models give rise to a compact probabilistic representation of high-dimensional data through graph-encoded conditional independence constraints. One key challenge is that the graph is typically unknown and needs to be inferred from data. While graphical models have been extensively studied for vector and matrix-variate data (Leng & Tang, 2012; Ni et al., 2017; Wang & West, 2009; Yuan & Lin, 2007), only recently have there been several developments for the functional data. Zhu et al. (2016) extended Markov and hyper Markov laws of decomposable undirected graphs for random vectors to those for random functions. Qiao et al. (2019) adopted the group lasso penalty on the precision matrix of coefficients extracted from the basis expansion of functions. Zapata et al. (2022) introduced the idea of partial separability to reduce the computational cost of Qiao et al. (2019). Qiao et al. (2020) further extended Qiao et al. (2019) and proposed to characterize the time-varying conditional independencies of random functions through smoothing techniques. To relax the Gaussian process assumption of the aforementioned methods, Li and Solea (2018), Solea and Li (2022), and Lee et al. (2022) proposed models based on additive conditional independence and copula Gaussian models.

Despite these exciting developments of functional undirected graphical models, the work on functional directed graphical models is sparse. Generally, undirected graphs admit a different set of conditional independence constraints from directed graphs. For example, the directed graph in Figure 1A implies $X_2 \perp X_3$ and $X_2 \not\perp X_3 | X_1$, but there exists no undirected counterpart that admits the same set of conditional (in)dependence assertions. More importantly, causal discovery is only possible with directed graphs given additional causal assumptions (Pearl, 2000). To the best of our knowledge, Lee and Li (2021) and Yang and Suzuki (2022) are the only works that infer causal relationships from multivariate functional data. However, as will become evident in Sections 3 and 4, our model differs from theirs in several significant aspects.

1.2 Causal discovery

As hinted earlier, one of the two important problems we intend to address in this work is discovering causality from functional observations. Causal discovery has been used as one of the first steps to investigate the physical mechanism that governs the operation and dynamics of an unknown system. Given the learned causal knowledge, subsequent causal inference (e.g., deriving the interventional and counterfactual distributions) can be conducted under the do-calculus framework (Pearl, 2000). Therefore, inferring causal relationships potentially has more significant scientific impacts than learning noncausal associations since it may help answer fundamental questions about nature.

BNs paired with causal assumptions are among the most popular approaches in identifying the unknown causal structure represented by a directed acyclic graph (DAG). One significant obstacle to using BNs to discover causality from purely observational data is that, in general, only Markov equivalence classes (MEC) can be learned based on conditional independence constraints alone. Causal interpretations of members in the same MEC can be drastically different, and generally, only bounds on causal effects can be calculated (Maathuis et al., 2009). For example, the three DAGs in Figure 1B constitute an MEC with the only conditional independence relationship $X_2 \perp X_3 | X_1$, but the causal directions are completely reversed in the last graph compared to the first one.

Numerous researchers, however, have found that causal discovery is indeed possible with additional distributional assumptions on the data-generating process, at least for
finite-dimensional data. Examples include but are not limited to linear non-Gaussian models (LiNGAM, Shimizu et al. 2006), non-linear additive noise models (Hoyer et al. 2008), and linear Gaussian models with equal error variances (Peters & Bühlmann, 2014). See more related methods in a recent book of Peters et al. (2017). Although remarkable progress has been made in the causal discovery area for traditional finite-dimensional data, what remains largely lacking is methods capable of discovering causality from general, purely observational, multivariate functional data. We remark that given a known causal graph, there are existing approaches that can be used to infer causal effects. For example, Lindquist (2012) developed a causal mediation analysis framework where the treatment and outcome are scalars and the mediator is a univariate random function. Our scope is substantially different from this line of work in that we do not assume the causal graph to be known; in fact, learning the causal graph structure is precisely the focus of this paper.

1.3 Proposed functional Bayesian networks

We propose a novel functional BN model for multivariate functional data for which conditional independence and causal relationships are represented by a DAG. As one would expect, the proposed functional BN factorizes over the DAG and respects all directed Markov properties (i.e., conditional independence constraints) encoded in the DAG via the notion of d-separation. Then for ease of exposition, we reformulate the proposed BN constructed in the functional space to an equivalent BN defined on the space of basis coefficients via basis expansion. Because in practice, functional data are always observed with noises, two essential ingredients are built into the proposed BNs to capture the functional dependencies and to learn the causal structure. First, we capture the within-function dependencies through a set of orthonormal basis functions chosen in a data-driven way. The resulting basis functions are interpretable and computationally efficient. Second, we encode the unknown causal structure by a structural equation model on the basis coefficients. Due to the equivalence of probability measures on the functional space and the space of basis coefficients, conditional independence and causal relationships naturally transform back to the original random functions. To allow for unique DAG identification, we move away from the Gaussian process assumption often adopted by the existing functional graphical models and instead assume our random functions are generated from a discrete scale mixture of Gaussians. We theoretically prove and empirically verify that the unique DAG identification is indeed possible even when the functions are observed with noises.

To conduct inference and uncertainty quantification from a finite amount of data, the proposed model is based on a Bayesian hierarchical formulation with carefully chosen prior distributions. Posterior inference is carried out through Markov chain Monte Carlo (MCMC). We perform simulation studies to demonstrate the capability of the proposed model in recovering causal structures and key parameters of interest. Data analysis with brain EEG records illustrates the applicability of the proposed framework in the real world. We also apply the proposed model to a COVID-19 multivariate longitudinal data set (shown in Section D of the Supporting Information).

2 OVERVIEW OF BNs

2.1 DAGs and BNs

Let \( X = (X_1, \ldots, X_p)^T \in \mathcal{X}_1 \times \cdots \times \mathcal{X}_p \) denote a \( p \)-dimensional random vector. Denote \( \mathcal{M} := \{1, \ldots, m\} \) for any integer \( m \geq 1 \). Let \( X_S = (X_\ell)_{\ell \in S} \) be a subvector of \( X \) with \( S \subseteq [p] \). A DAG \( G = (V, E) \) consists of a set of nodes \( V = [p] \) and a set of directed edges represented by a binary adjacency matrix \( E = (E_{\ell \ell'}) \) where \( E_{\ell \ell'} = 1 \) if and only if \( \ell \rightarrow j \) for \( \ell \neq j \in V \). DAGs do not allow directed cycles \( j_0 \rightarrow j_1 \rightarrow \cdots \rightarrow j_k = j_0 \). Each node \( j \in V \) represents a random variable \( X_j \in \mathcal{X}_j \). Each directed edge \( \ell \rightarrow j \) and the lack there of represent conditional dependence and independence of \( X_\ell \) and \( X_j \), respectively. Note that although \( X_j \) is often a scalar but it need not be. In fact, \( X_j \) is a random function in this paper. Denote \( pa(j) = \{\ell \in V : \ell \rightarrow j\} \) the set of parents of \( j \) in graph \( G \). A BN \( B = (G, P) \) on \( X \) is a probability model where the joint probability distribution \( P \) of \( X \) factorizes with respect to \( G \) in the following manner,

\[
P(X) = \prod_{j=1}^{p} P_j(X_j | X_{pa(j)}), \tag{1}
\]

where \( P_j \) is the conditional distribution of \( X_j \) given \( X_{pa(j)} \) under \( P \). Let \( de(j) = \{\ell \in V : j \rightarrow \cdots \rightarrow \ell\} \) denote the descendants of \( j \) in \( G \) and let \( nd(j) = V \setminus \{de(j) \cup \{j\}\} \) denote the non-descendants of \( j \). The BN factorization (1) implies the local directed Markov property—any variable is conditionally independent of its non-descendants given its parents, \( X_j \perp X_{nd(j) \cup \{pa(j)\}} | X_{pa(j)} \) for any \( j \in [p] \). In fact, the reverse is also true: if a distribution \( P \) respects the local Markov property according to a DAG \( G \), then \( P \) must factorize over \( G \) as in (1). In summary, BN factorization and local Markov property are equivalent.
2.2 Causal DAGs and causal BNs

A causal DAG $G$ is a DAG except that the directed edges are interpreted causally, that is, we say $X_\ell$ is a direct cause (with respect to $V$) of $X_j$ and $X_j$ is a direct effect of $X_\ell$ if $\ell \to j$. For simplicity, we will overload $\text{nd}(j)$ and $\text{pa}(j)$ to denote the noneffects and directed causes of $j$ in a causal DAG. To define a causal BN, we begin by asserting the local causal Markov assumption (Pearl, 2000; Spirtes et al., 2000)—given a causal DAG $G$, a variable is conditionally independent of its noneffects given its direct causes. By noting the correspondence between noneffects and nondescendants, and between direct causes and parents in DAGs and causal DAGs, the local causal Markov assumption states that the distribution $P$ of $X$ respects the local Markov property of the causal DAG $G$, which in turn implies that $P$ must also factorize over $G$ (recall the equivalence between BN factorization and local Markov property). Therefore, a causal BN $B = (G, P)$ is a probability model where $P$ factorizes with respect to a causal DAG $G$ as in (1).

2.3 Structural equation representation of BNs

A BN is often represented by a structural equation model (SEM), $X_j = f_j(X, \epsilon_j)$ for any $j \in [p]$, where the transformation $f_j$ depends on $X$ only through its parents/direct causes $X_{\text{pa}(j)}$, and the exogenous variables $\epsilon = (\epsilon_1, \ldots, \epsilon_p)^T \sim P_\epsilon$ are assumed to be mutually independent. The distribution $P$ of $X$ is then induced from $F = \{f_1, \ldots, f_p\}$ and $P_\epsilon$.

3 FUNCTIONAL BAYESIAN NETWORKS

3.1 General framework

Now we introduce the construction of BNs for multivariate functional data. Denote the space of square integrable functions on domain $D$ with respect to measure $\mu$ as $L^2(D) = \{ h : \int_D h^2(\omega) d\mu(\omega) < \infty \}$. We focus on a compact set $D \subset \mathbb{R}$ (in fact, without loss of generality, $D = [0,1]$) and the Lebesgue measure $\mu$ for simplicity. Let $Y = (Y_1, \ldots, Y_p)^T \in L^2(D_1) \times \cdots \times L^2(D_p)$ be a collection of $p$ random functions. Denote $\mathcal{H} = \times_{j=1}^p \{ (\omega, j) : \omega \in D_j \}$ the joint domain of $Y$ and $(L^2(\mathcal{H}), B(L^2(\mathcal{H})), P)$ its probability space. Similarly, for any subset $A \subset [p]$, denote the joint domain $\mathcal{H}_A = \times_{j \in A} \{ (\omega, j) : \omega \in D_j \}$ and $B(L^2(\mathcal{H}_A))$ the Borel $\sigma$-algebra on $L^2(\mathcal{H}_A)$. Let $A, B, C$ be disjoint subsets of $[p]$. Following Zhu et al. (2016), we say $Y_A$ is conditionally independent of $Y_B$ given $Y_C$ under $P$, if for any measurable set $D_A \subseteq L^2(\mathcal{H}_A)$, $P(Y_A \in D_A | Y_B, Y_C) = B(L^2(\mathcal{H}_C))$ measurable and $P(Y_A \in D_A | Y_B, Y_C) = P(Y_A \in D_A | Y_B, Y_C)$. We introduce a DAG $G = (V, E)$ where each node $j \in V$ represents a random function $Y_j$. First, we give the formal definition of a functional BN.

**Definition 1** (Functional BNs). We say $B = (G, P)$ is a functional BN for a set of random functions $Y$ if $P$ factorizes with respect to the DAG $G$,

$$P(Y_1 \in D_1, \ldots, Y_p \in D_p) = \prod_{j=1}^p P_j(Y_j \in D_j|Y_{\text{pa}(j)} \in D_{\text{pa}(j)})$$

for any measurable sets $D_j \subseteq L^2(D_j)$ for any $j \in [p]$, where $P_j$ is the conditional probability measure of $Y_j$ given $Y_{\text{pa}(j)}$ under $P$.

Just like the ordinary finite-dimensional BN, the functional BN factorization implies the local Markov property and vice versa.

**Definition 2** (Functional Local Directed Markov Property). A probability measure $P$ of $Y$ satisfies the local directed Markov property with respect to $G$ if $Y_j \perp Y_{\text{nd}(j) \setminus \text{pa}(j)} | Y_{\text{pa}(j)}$, that is, $P(Y_j \in D_j | Y_{\text{nd}(j) \setminus \text{pa}(j)}, Y_{\text{pa}(j)}) = B(L^2(H_{\text{pa}(j)}))$ measurable and $P(Y_j \in D_j | Y_{\text{nd}(j) \setminus \text{pa}(j)}, Y_{\text{pa}(j)}) = P(Y_j \in D_j | Y_{\text{pa}(j)})$ for any $D_j \subseteq L^2(D_j)$.

**Proposition 1.** Functional BN factorization is equivalent to functional local directed Markov property.

Proof of Proposition 1 is trivial and thus omitted. For modeling convenience, we use orthonormal basis expansion of random functions to (equivalently) redefine the functional BN in the space of basis coefficients. Let $\{ \phi_{jk}^{\text{vec}} \}_{k=1}^\infty$ be a sequence of orthonormal basis functions of $L^2(D_j)$ and expand $Y_j = \sum_{k=1}^\infty Z_{jk} \phi_{jk}$, where $Z_{jk} = \int_{D_j} Y_j(\omega) \phi_{jk}(\omega) d\omega$. The resulting coefficient sequence $Z_j = (Z_{jk})_{k=1,\ldots,\infty}$ lies in the space of square summable sequences $\ell_2^2 = \{ h : \Sigma_{k=1}^\infty h_k^2 < \infty \}$. The within-function and between-function covariance can then be expressed in terms of the covariance of the coefficient sequences,

$$\text{cov}(Y_j(\omega_j), Y_\ell(\omega_\ell)) = \sum_{k=1}^\infty \sum_{n=1}^\infty \phi_{jk}(\omega_j) \phi_{n\ell}(\omega_\ell) \text{cov}(Z_{jk}, Z_{n\ell}), \quad \forall \omega_j \in D_j, \omega_\ell \in D_\ell, \forall j, \ell \in [p].$$

Because $L^2(D_j)$ and $\ell_2^2$ are isometrically isomorphic for each $j$, for any disjoint subsets $A, B, C \subset [p]$, $Y_A \perp Y_B | Y_C$ if and only if $Z_A \perp Z_B | Z_C$ where $Z = (Z_1, \ldots, Z_p)^T$. Hence,
if \( Y \) follows the proposed BN model \( B = (G, P) \), then the coefficient sequence \( Z \) follows \( B_Z = (G, P_Z) \) for some probability measure \( P_Z \) of \( Z \), and vice versa where each node of the DAG \( G \) either represents a random function \( Y_j \) or, equivalently, its corresponding coefficient sequence \( Z_j \). Moreover, the joint probability \( P \) of \( Y \) factorizes with respect to \( G \) if and only if the joint probability \( P_Z \) of \( Z \) factorizes with respect to \( G \).

**Proposition 2.** Suppose \( Y \sim P \) and let \( Z \) be the corresponding coefficient sequences from the orthonormal basis expansion. Then,

\[
P(Y_1 \in D_1, \ldots, Y_p \in D_p) = \prod_{j=1}^{p} P_j(Y_j \in D_j | Y_{pa}(j) \in D_{pa}(j)),
\]

for any measurable sets \( D_j \subseteq L^2(D_j) \) for any \( j \in [p] \) if and only if

\[
P_Z(Z_1 \in D'_1, \ldots, Z_p \in D'_p) = \prod_{j=1}^{p} P_Z(Z_j \in D'_j | Z_{pa}(j) \in D'_{pa}(j)),
\]

for any measurable sets \( D'_j \subseteq \ell^2 \) for any \( j \in [p] \).

The proof directly follows the preceding paragraph. Just like the ordinary finite-dimensional BN, if one makes the causal Markov assumption, the DAG \( G \) in the proposed functional BN can be interpreted causally. Hereafter, by default, we always make the causal Markov assumption (hence \( G \) is a causal DAG, the edge strength is interpreted as the direct causal effect, etc.) but all the results are simply reduced to those of a directed conditional independence model when the causal Markov assumption is dropped.

### 3.2 Functional linear non-Gaussian BNs

Section 3.1 introduces a general framework for modeling directed conditional independence and causal relationships for multivariate functional data. In this subsection, we discuss in detail one specific case of the proposed general framework, namely, the Functional Linear Non-Gaussian (FLiNG) BNs. Specifically, the FLiNG-BN assumes \( Z \) follows a linear SEM,

\[
Z_j = \sum_{\ell=1}^{p} B_{j\ell} Z_{\ell} + \epsilon_j, \quad \forall j \in [p],
\]

where \( \epsilon_j \) is an infinite-dimensional exogenous vector, \( B_{j\ell} = (B_{j\ell}(k_j, k_\ell))_{k_\ell=1}^{\infty} \) is an infinite-dimensional direct causal effect matrix from \( Z_\ell \) to \( Z_j \), and \( \ell \rightarrow j \) is present in \( G \) (i.e., \( Z_\ell \) is a direct cause of \( Z_j \)) if there exist \( k_\ell \) and \( k_j \) such that \( B_{j\ell}(k_j, k_\ell) \neq 0 \). We assume \( \epsilon_1, \ldots, \epsilon_p \) to be independent, which means there are no latent confounders. Neither causal effects nor the causal graph is assumed to be known; they will be inferred from observational data. Because \( L^2(D_j) \) and \( \ell^2 \) are isometrically isomorphic for all \( j \in [p] \), the causal relationships of \( Z \) encoded in DAG \( G \) directly transfer to the causal relationships of \( Y \), that is, \( Z_\ell \) is a direct cause of \( Z_j \) if and if only if \( Y_\ell \) is a direct cause of \( Y_j \). Under very mild conditions, the infinite sequence \( Z_j \) is guaranteed to be square summable in our model; see Section E of the Supporting Information.

In practice, the random functions \( Y \) can only be measured on finite grids with random noises. In other words, we do not observe realizations of \( Y \) but instead we observe realizations of \( W = (W_1, \ldots, W_p)^T \) where \( W_j = (W_j(1), \ldots, W_j(m_j)) \), which is the set of measurements of \( Y_j \) on a finite grid \( D_j = \{\omega_j(1), \ldots, \omega_j(m_j)\} \subseteq D_j \) with independent white noises \( \epsilon_j(m) \sim N(0, \sigma_j) \) for any \( m \in [m_j] \).

\[
W_j(m) = Y_j(\omega_j(m)) + e_j(m).
\]

Note that \( D_j \) can be different across \( j \) (also across realizations).

One seemingly inconsequential element of the FLiNG-BN but turning out to be crucial for discovering causality is the specification of the probability distribution of the exogenous variables \( \epsilon_j = (\epsilon_{j\ell})_{k_\ell=1}^{\infty} \) in (2). A tempting choice may be Gaussian but it is the non-Gaussianity of \( \epsilon_j \) that allows causal identification as we will show in Section 3.3. Specifically, we assume \( \epsilon_{jk} \) to follow a finite scale mixture of Gaussian distributions, \( \epsilon_{jk} \sim \sum_{m=1}^{M_{jk}} r_{jk}^m N(0, \sigma_{jk}^m) \), where \( M_{jk} \) is the number of mixture components. The non-Gaussian exogenous variables lead to non-Gaussian coefficient sequences \( Z \), which in turn lead to non-Gaussian-process distributed random functions \( Y \). In addition to enabling causal identification, non-Gaussian processes are robust against outlying curves (Zhu et al., 2011). For finite-sample inference, we reduce the dimension of functional data by truncating the orthonormal basis functions at level \( K \) such that \( \phi_j = (\phi_j(1), \ldots, \phi_j(K))^T \), as commonly done in existing functional data analysis literature; see Section F of the Supporting Information for more discussion regarding the truncation in our context. Consequently, (3) is turned into

\[
W_j(m) = \sum_{k=1}^{K} Z_{jk} \phi_j(\omega_j(m)) + e_j(m).
\]

### 3.3 Causal identifiability

The proposed functional BNs are useful representations of directed conditional independence and causal
relationships for multivariate functional data. The big remaining question is the learning of the underlying (causal) DAGs from observational data. Constraint-based methods, which are often model-free, have been popular for DAG learning. For the proposed functional BNs, we, in principle, can also use constraint-based methods, which test for conditional independence of pairs of functions. However, conditional independence tests are notoriously difficult and inefficient even for scalar random variables. Furthermore, even if we have access to oracle conditional independence tests for random functions, we can only hope for identifying the best MEC by definition (recall that an MEC contains DAGs with exactly the same set of conditional independence relationships). This may be acceptable if one is only interested in learning conditional independence relationships. But as mentioned in Section 1, for causal discovery, this is clearly unsatisfactory because the directionality of a potentially large number of Markov equivalent DAGs may be left undetermined and hence the causal interpretations of these edges are unclear. Because the proposed FLiNG-BN is a proper probability model, we can exploit certain features of the model, namely the non-Gaussianity, to uniquely identify the underlying causal DAG.

**Definition 3** (Causal Identifiability). Suppose \( Y \) follows the FLiNG-BN \( B = (G, P) \), and suppose \( W \) is a noisy version of \( Y \) with noise variances \( \sigma = (\sigma_1, \ldots, \sigma_p) \) as defined in (3). Let \( P_W \) denote the distribution of \( W \) induced from FLiNG-BN and the noises. We say that the causal DAG of FLiNG-BN is identifiable from \( W \) if there does not exist another BN \( B’ = (G’, P’) \) where \( G’ \neq G \) and noise variances \( \sigma’ = (\sigma’_1, \ldots, \sigma’_p) \) such that the induced distributions on \( W, P_W \), is equivalent to \( P_W \), that is, \( P_W(W) \equiv P_W'(W) \).

**Theorem 1** (Causal Identifiability). The causal DAG of FLiNG-BN is identifiable if the number of Gaussian mixture components \( M_{jk} > 1 \) for any \( j \) and \( k \).

Theorem 1 signifies that by examining the probability distribution \( P_W \), to which we have access through the observational data alone, one can gauge the likelihood that a given causal DAG is the data-generating DAG. With a finite data set, we shall focus on weighing different candidate causal DAGs by their posterior probabilities. Here, we provide the outline of the proof; the complete proof is given in Section A of the Supporting Information. Given a chosen set of basis functions, we show the result in the space of basis coefficients. The problem then transforms to prove that, given \( Z = BZ + \epsilon \) and observe \( W = Z + \epsilon \), there does not exist another equivalent parameterization \( Z’ = B’Z’ + \epsilon’ \) and \( W = Z’ + \epsilon’ \). Since we assume each component of \( \epsilon \) follows a Gaussian scale mixture, the induced distribution on \( W \) is a multivariate Gaussian mixture (with different precision matrices). We prove that the causal effect matrix \( B \) is uniquely identifiable from such mixture model by combining the identification of Gaussian mixture components, uniqueness of LDL decomposition (a variant of the Cholesky decomposition where a positive-definite matrix is decomposed as \( A = L D L^T \) with a lower unit triangular matrix \( L \) and a diagonal matrix \( D \)), and proof of causal ordering identification. We demonstrate the identifiability result with a toy example.

**Example 1.** Consider a true functional causal graph \( 1 \rightarrow 2 \) and the corresponding data generating model \( Z_1 = e_1 \) with \( e_1 \sim 0.5N(0, 0.5) + 0.5N(0, 1) \) and \( Z_2 = Z_1 + e_2 \) with \( e_2 \sim 0.5N(0, 0.5) + 0.5N(0, 1) \); note for simplicity, we assume in this example that the number of basis functions is \( K = 1 \). This leads to four Gaussian mixture components in the joint distribution of \( (Z_1, Z_2) \), namely,

\[
P(Z_1, Z_2) = 0.25 \times N \begin{bmatrix} 0 & 0.5 & 0.5 \\ 0 & 0.5 & 1.0 \\ 0 & 0.5 & 1.5 \end{bmatrix} + 0.25 \times N \begin{bmatrix} 0 & 1.0 & 1.0 \\ 0 & 1.0 & 1.5 \\ 0 & 1.0 & 2.0 \end{bmatrix} + 0.25 \times N \begin{bmatrix} 0 & 1.0 & 1.0 \\ 0 & 1.0 & 1.5 \\ 0 & 1.0 & 2.0 \end{bmatrix}
\]

If we regress \( Z_2 \) on \( Z_1 \) separately for each mixture component, the regression coefficients will be the same on the population level, that is, \( \beta_1 = \beta_2 = \beta_3 = \beta_4 = \beta = 1 \); however, in the anticausal direction, if we regress \( Z_1 \) on \( Z_2 \), the regression coefficients are different: \( \beta’_1 = 1/2, \beta’_2 = 1/3, \beta’_3 = 2/3, \) and \( \beta’_4 = 1/2 \). That is to say, the global regression in the anticausal direction cannot be optimal for each local component, showing that the causal model would be preferable to the anticausal model, which leads to causal identification.

Now assume we observe with noises \( W_1 = Z_1 + e_1 \) and \( W_2 = Z_2 + e_2 \) with \( e_1, e_2 \sim N(0, 0.1) \). We sample \( n = 1000 \) observations from this model and index them by the subscript \( i = 1, \ldots, n \). For the purpose of illustration, suppose we know the mixture component assignment of each observation and define four groups of observations based on the combination of variances of \( e_1 \) and \( e_2 \),

\[
C_1 = \{i : \text{Var}(e_i) = 0.5 \text{ and Var}(e_i) = 0.5\}, \ C_2 = \{i : \text{Var}(e_i) = 0.5 \text{ and Var}(e_i) = 1\},
\]

\[
C_3 = \{i : \text{Var}(e_i) = 1 \text{ and Var}(e_i) = 0.5\}, \ C_4 = \{i : \text{Var}(e_i) = 1 \text{ and Var}(e_i) = 1\}.
\]

We fit linear regression separately to all observations and to observations in each of the four groups with the true causal direction \( 1 \rightarrow 2 \) (regressing \( W_2 \) on \( W_1 \)) and the anticausal direction \( 2 \rightarrow 1 \) (regressing \( W_1 \) on \( W_2 \)), which are shown in Figure 2. We observe that the fitted lines are almost identical in the causal direction for all groups whereas they can
be quite different across groups in the anti-causal direction. Therefore, only the true causal graph gives a unique regression coefficient among all groups. Notice that if there is only one mixture component (i.e., degeneration to the Gaussian case), no comparison can be made between the causal and anticausal directions since there will be only one regression line.

The next counterexample illustrates the necessity of the non-Gaussian assumption for causal identification.

**Example 2.** Consider a similar bivariate case to Example 1 but now the exogenous variables are Gaussian instead of a mixture of Gaussians. Suppose the true functional causal graph $1 \rightarrow 2$ and the corresponding data-generating model $Z_1 = \epsilon_1 \sim N(0, \tau_1)$ and $Z_2 = bZ_1 + \epsilon_2 \sim N(0, \tau_2)$; note again for simplicity, we assume in this example that the number of basis functions is $K = 1$. Assume we observe with noises $W_1 = Z_1 + \epsilon_1$ and $W_2 = Z_2 + \epsilon_2$ with $\epsilon_1 \sim N(0, \sigma_1)$ and $\epsilon_2 \sim N(0, \sigma_2)$. The induced joint distribution on $W = (W_1, W_2)$ is then bivariate Gaussian with mean 0 and covariance matrix

$$
\begin{pmatrix}
\tau_1 + \sigma_1 & b\tau_1 \\
\frac{b\tau_1}{b\tau_1 + \tau_2 + \sigma_2}
\end{pmatrix}
$$

Further consider the anticausal model $2 \rightarrow 1$ where $Z_2' = \epsilon_2' \sim N(0, \tau_2')$ and $Z_1' = b'Z_2' + \epsilon_1' \sim N(0, \tau_1')$. Suppose $W_1 = Z_1' + \epsilon_1'$ and $W_2 = Z_2' + \epsilon_2'$ with $\epsilon_1' \sim N(0, \sigma_1')$ and $\epsilon_2' \sim N(0, \sigma_2')$. The induced joint distribution on $W = (W_1, W_2)$ is still bivariate Gaussian with mean 0 and covariance matrix

$$
\begin{pmatrix}
\tau_1^2 + \sigma_1^2 & b\tau_1' \\
\frac{b\tau_1'}{b\tau_1' + \tau_2' + \sigma_2'}
\end{pmatrix}
$$

For any chosen $\tau_1', \sigma_1' > 0$ such that $\tau_1' + \sigma_1' < \tau_1 + \sigma_1 - b^2(\tau_1'/\tau_1 + \sigma_2')$, if we set

$$
b' = \frac{(\tau_1 + \sigma_1 - \tau_1' - \sigma_1')}{b\tau_1},
\tau_2' = \frac{b\tau_1'}{(\tau_1 + \sigma_1 - \tau_1' - \sigma_1')},
\sigma_2' = \frac{b^2\tau_1 + \tau_2 + \sigma_2 - b^2\tau_1'/(\tau_1 + \sigma_1 - \tau_1' - \sigma_1')},
$$
then the induced distribution coincides with that under the true causal model (i.e., Gaussian with mean 0 and the same covariance). Therefore, causal identification fails in this case.

## 4 Bayesian Inference

The inference of the proposed FLiNG-BN framework can be carried out in either a frequentist (e.g., maximizing penalized likelihood) or a Bayesian (e.g., sampling from
posterior distribution) fashion. Existing frequentist functional graphical models (Lee et al., 2022; Lee & Li, 2022; Qiao et al., 2019, 2020; Solea & Li, 2022; Zapata et al., 2022) often estimate graphs in two separate steps—estimate the basis coefficient sequence of each function marginally via functional principle component analysis, and learn a graph based on the estimated coefficient sequences. However, the eigenfunctions that marginally explain the most variation of each individual function do not necessarily explain well the conditional/causal relationships among a set of functions. Moreover, the estimation uncertainty is not propagated from the first step to the second, which may result in overly confident inference. To mitigate these potential drawbacks of the two-step approaches, we propose a Bayesian inference procedure that jointly infers basis coefficient sequences and the DAG structure. This joint inference approach constructs orthonormal basis functions adaptive to their conditional/causal relationships and allows for finite-sample inference and uncertainty quantification.

4.1 Adaptive orthonormal basis functions

We assume the basis functions to be shared across all random functions (Kowal et al., 2017; Zapata et al., 2022), \( \phi_{jk} (\omega) := \phi_k (\omega) \) for all \( j \in \{p\} \), which are more parsimonious than models based on function-specific basis functions. As mentioned above, we do not prespecify a fixed set of orthonormal basis functions but instead, they are learned adaptively from data by further expanding them with spline basis functions (Kowal et al., 2017), 
\[
\phi_k (\omega) = \sum_{l=1}^{p} A_{kl} b_l (\omega),
\]

where \( b = (b_1, \ldots, b_p)^T \) is a set of cubic B-spline basis functions with equally spaced knots and \( A_k = (A_{k1}, \ldots, A_{kL})^T \) with \( \forall k \in [K] \) are spline coefficients. Because \( A_k \)s are not fixed a priori, so are \( \phi_k \)s.

We summarize the main steps of prior specification and refer the details to Kowal et al. (2017). First, to regularize the roughness of \( \phi_k \) in a frequentist framework, one would consider a penalized likelihood with the roughness penalty,
\[
\lambda_k \int (\phi''(\omega))^2 d\omega = \lambda_k A_k^T \Omega A_k,
\]

where \( \lambda_k > 0 \) is the regularization parameter and \( \Omega = \int (b''(\omega) b'(\omega))^T d\omega \). As a Bayesian counterpart, the regularization term is equivalent to a prior on the B-spline coefficients \( A_k \sim N(0, \lambda_k^{-1} \Omega^{-1}) \), where \( \Omega^{-1} \) is a pseudo-inverse (since \( \Omega \) is rank-deficient by 2). Let \( \Omega = U D U^T \) be the singular value decomposition of \( \Omega \). To facilitate efficient computation, we follow Wand and Ormerod (2008) and reparameterize \( \phi_k = \sum_{l=1}^{p} A_{kl} b_l = \sum_{l=1}^{L} \tilde{A}_{kl} \tilde{b}_l \) with \( \tilde{b}(\omega) = (1, \omega, b_1(\omega) U_p D_p^{-1/2})^T \) where \( D_p \) is the \((L-2) \times (L-2)\) submatrix of \( D \) corresponding to nonzero singular values and \( U_p \) is the corresponding \((L-2) \times (L-2)\) submatrix of \( U \). The reparameterization induces a prior on \( \tilde{A}_k = (\tilde{A}_{k1}, \ldots, \tilde{A}_{kL})^T \sim N(0, S_k) \) where \( S_k = \text{diag}(\infty, \infty, \lambda_k^{-1}, \ldots, \lambda_k^{-1}) \) with the first two dimensions corresponding to the unpenalized constant and linear terms. In practice, we replace \( \infty \) by \( 10^8 \).

Second, we constrain the regularization parameters \( \lambda_1 > \cdots > \lambda_K > 0 \) to identify the ordering of basis functions, which sorts the basis functions by decreasing smoothness. Unlike the functional principal component analysis (PCA) where the principal components are ordered by the proportion of variance explained, the adopted Bayesian approach is less prone to rough functions. Given the ordering constraint, a uniform prior is imposed such that \( \lambda_k \sim U(L_k, U_k) \), where \( U_l = 10^8, L_k = \lambda_{k+1} \) for \( k = 1, \ldots, K-1 \), \( U_k = \lambda_{K-1} \) for \( k = 2, \ldots, K \), and \( L_K = 10^{-8} \).

Finally, consider the orthonormal constraint
\[
\int \phi_k(\omega) \phi_h(\omega) d\omega = \int \tilde{A}_k^T \tilde{b}(\omega) \tilde{b}^T(\omega) \tilde{A}_h d\omega = \tilde{A}_k^T J = I(k = h),
\]

(5)

with \( J = \int \tilde{b}(\omega) \tilde{b}^T(\omega) d\omega \). This constraint can be enforced by projection and normalization during the course of MCMC; see Section B of the Supporting Information for details.

4.2 Prior model

4.2.1 Prior on B-spline coefficients \( A \)

The prior on \( A_k \) serves three purposes. First, it forces \( \phi_k \)s to be orthonormal, that is, \( \int \phi_k(\omega) \phi_h(\omega) d\omega = I(k = h) \), \( \forall k, h \in [K] \). Second, it regularizes the roughness of \( \phi_k \)s to prevent overfitting and sorts the orthonormal basis functions by increasing roughness. Third, it enables posterior inference on the orthonormal basis functions simultaneously with the graph estimation without having to fix them a priori.

The key problem we aim to address in this paper is causal structure learning, that is, inferring the adjacency matrix,
\( \mathbf{E} = (\mathbf{E}_{\ell j}) \) (recall \( \mathbf{E}_{\ell j} = 1 \) if and only if \( \ell \to j \)). We propose to use a beta-Bernoulli-like prior \( \mathbf{E}_{\ell j} \sim \text{Bernoulli}(r) \) with \( r \sim \text{Beta}(a_r, b_r) \), subject to the acyclicity constraint,

\[
P(\mathbf{E} | r) \propto \prod_{j \neq \ell} r^{E_{\ell j}} (1 - r)^{1 - E_{\ell j}} \mathbb{I}(G \text{ is a DAG}).
\]

We set \( a_r = b_r = 1 \). Scott and Berger (2010) showed that the beta-Bernoulli prior allows automatic multiplicity adjustment in sparse regression problems. In our context, the marginal distribution of \( \mathbf{E} \) with \( r \) integrated out equals

\[
P(\mathbf{E}) \propto \text{Beta}
\left(
\sum_{j \neq \ell} E_{\ell j} + 1,
\sum_{j \neq \ell} (1 - E_{\ell j}) + 1
\right)
\mathbb{I}(G \text{ is a DAG}).
\]

(6)

The marginal distribution strongly prevents false discoveries by increasing the penalty against additional edges as the dimension \( p \) grows. For example, the marginal (6) favors an empty graph over a graph with one edge by a factor of \( p^2 - p \), which increases with \( p \).

Conditional on \( \mathbf{E} \), we assume independent matrix-variate spike-and-slab priors on the direct causal effects,

\[
\mathbf{B}_{\ell j} | E_{\ell j} \sim (1 - E_{\ell j}) \mathbf{\delta}_0(\mathbf{B}_{\ell j}) + E_{\ell j} \mathbf{N}(\mathbf{B}_{\ell j} | \mathbf{\gamma} \mathbf{I}, \mathbf{I}),
\]

where \( \mathbf{\delta}_0(\cdot) \) is a point mass at a \( K \times K \) zero matrix \( \mathbf{O} \) and \( \mathbf{N}(\cdot | \mathbf{O}, \mathbf{\gamma} \mathbf{I}, \mathbf{I}) \) is a centered matrix-variate normal distribution with row and column covariance matrices \( \mathbf{\gamma} \mathbf{I} \) and \( \mathbf{I} \) where \( \mathbf{I} \) is a \( K \times K \) identity matrix. The hyperparameter \( \mathbf{\gamma} \) indicates the overall causal effect size and is assumed to follow a conjugate inverse-gamma prior, \( \mathbf{\gamma} \sim \text{IG}(a_\gamma, b_\gamma) \) with \( a_\gamma = b_\gamma = 1 \).

### 4.2.3 Prior on the Gaussian scale mixture

We choose conjugate priors,

\[
\pi_{jk} = (\pi_{jk}^1, \ldots, \pi_{jk}^M) \sim \text{Dirichlet}(\alpha, \ldots, \alpha), \quad \tau_{jk}^m \sim \text{IG}(a_\tau, b_\tau),
\]

\[
\forall j \in [p], k \in [K], m \in [M],
\]

which allow for straightforward Gibbs sampling. As default, we set \( \alpha = 1 \) and \( a_\tau = b_\tau = 1 \).

### 4.2.4 Prior on observation noises

We complete the prior specification with a conjugate inverse-gamma prior on the variance of noises, \( \sigma_j \sim \text{IG}(a_\sigma, b_\sigma) \) for \( \forall j \in [p] \) with \( a_\sigma = b_\sigma = 0.01 \).

We summarize the proposed Bayesian hierarchical model in Figure 3. We simulate posterior samples through MCMC. Details are given in Section B of the Supporting Information. The per-iteration computational complexity for our MCMC algorithm, which is dominated by sampling the basis coefficient sequences \( \mathbf{Z} \) and the direct causal effects \( \mathbf{B} \), is \( O(\max(n(PK)^3, (PK)^4)) \).

### 5 Simulation Studies

We conducted simulation studies to evaluate the proposed FLiNG-BN model. We considered two scenarios. In the first scenario, the functions were observed on an evenly spaced grid; this is the scenario commonly studied in the existing functional undirected graphical models (Qiao et al., 2019) and is also similar to our later EEG application. In the second scenario (the details and results are provided in Section C of the Supporting Information), the functions were observed on an unevenly spaced grid, similar to the COVID-19 longitudinal application. We compared the proposed FLiNG-BN with a functional undirected graphical model (FGLASSO; Qiao et al. 2019). We did not make comparisons with Lee and Li (2022) and Yang and Suzuki (2022) due to the lack of publicly available codes at the time of submission. In addition, we compared FLiNG-BN with approaches based on two-step estimation procedures. In the first step, we extracted basis coefficients obtained from functional PCA using the package fdaSpace (Carroll et al., 2021). In the second step, given the estimated basis coefficients, we constructed causal graphs using either the LiNGAM (Shimizu et al., 2006) algorithm (termed FPCA-LiNGAM) or the PC (Spirtes & Glymour, 1991) algorithm (termed FPCA-PC). LiNGAM estimates a causal DAG based on the linear non-Gaussian assumption whereas PC generally returns only an equivalence class of DAGs based on conditional independence tests. Their implementations are available from the R package pcalg (Kalisch et al., 2020).

To mimic the EEG data application, we simulated data from FLiNG-BN with all the combinations of sample size \( n \in \{50, 100, 200\} \), number of functions \( p \in \{30, 60, 90\} \), and grid size \( d \in \{125, 250\} \). The grid spanned the unit interval [0, 1]. We set the true number of basis functions to be \( K = 5 \). To generate basis functions, we first simulated nonorthonormal functions \( \phi_k^U \) \( \forall k \in [K] \) from a set of \( L = 6 \) cubic B-spline basis functions with evenly spaced knots, \( \phi_k^U = \sum_{\ell=1}^L A_{k\ell} b_\tau \), where \( A_{k\ell} \)'s were generated from a standard normal distribution. We then empirically orthogonalized \( \phi_k^U \) to get the orthonormal basis functions \( \phi_k \). The simulation true graph \( G \) was generated from the Erdős–Rényi model with connection
probability \(2/p\), subject to the acyclicity constraint. Given the true graph \(G\), each block of nonzero direct causal effects \(B_{j\ell}\) was generated independently from a standard matrix-variate normal distribution. Then, the basis coefficient sequences \(Z\) were generated from (2) where the exogenous variables \(\varepsilon_s\) were generated from a centered Laplace distribution with scale \(b = 0.5\). Note that when we fit FLiNG-BN to the simulated data, we still assumed the exogenous variables to be discrete scale Gaussian mixture although the simulation true exogenous variables were Laplace. Finally, noisy observations were simulated following (4) with the signal-to-noise ratio, that is, the mean value of \(|y_j^{(i)}(\omega_j^{(i)}(m))/\sigma_j\) across all samples \(i \in [n]\) and grid points \(m \in [m_j^{(i)}]\), set to 5.

For implementing the proposed FLiNG-BN, we set the number of mixture components to \(M = 5\) and the number of B-spline basis functions to \(L = 20\) (note that the simulation truth was \(L = 6\)), and ran MCMC for 5000 iterations (discarding the first half as burn-in and retaining every fifth iteration). The causal graph \(G\) was estimated by thresholding the posterior probability of inclusion at 0.5 (i.e., the median probability model; Barbieri & Berger, 2004). Parameters of competing methods were set to their default values. To assess the graph recovery performance, we calculated true positive rate (TPR), false discovery rate (FDR), and Matthews correlation coefficient (MCC),\(\quad\text{TPR} = TP(TP + FN)^{-1},\quad\text{FDR} = FP(TP + FP)^{-1},\quad\text{MCC} = (TP \times TN - FP \times FN)((TP + FP) \times (TP + FN) \times (TN + FP) \times (TN + FN))^{-1/2},\)

where \(TP, TN, FP,\) and \(FN\) stand for the numbers of true positives, true negatives, false positives, and false negatives, respectively. MCC ranges from \(-1\) to 1 with 0 indicating a random guess and 1 a perfect recovery. Since FGLASSO learns an undirected graph, we compared it with a moralization of the true graph. Graph moralization converts a DAG to an undirected graph by first marrying all the unmarried parents and then removing all the directions. A probability distribution that respects the Markov property of a DAG must respect the Markov property of its moral graph. Similarly, since the PC algorithm returns the MEC representation, we compared it with the MEC of the true causal graph. The MEC representation is shown in an essential graph, where any edge presented between two nodes is directed if it has the same direction in all members of this MEC and is undirected otherwise.

The results based on 50 repeat simulations are summarized in Table 1, from which we conclude that the proposed FLiNG-BN significantly outperformed all the competitors FGLASSO, FPCA-LiNGAM, and FPCA-PC across all
Since LiNGAM is not applicable to cases where \( q > n \), with \( q = pK \) being the total number of extracted basis coefficients across all functions, the results from those cases are not available and indicated by -.

| \( p \) | \( d \) | \( n \) | FLiNG-BN | FGlasso | FPCA-LiNGAM | FPCA-PC |
|-------|-------|-------|----------|----------|-------------|---------|
|       |       |       | TPR      | FDR      | MCC         | TPR      | FDR      | MCC         | TPR      | FDR      | MCC         |
| 30    | 125   | 50    | 0.62 (0.07) | 0.14 (0.07) | 0.72 (0.07) | 0.58 (0.02) | 0.88 (0.02) | 0.16 (0.02) | - | - | - |
| 30    | 125   | 100   | 0.71 (0.08) | 0.19 (0.05) | 0.75 (0.06) | 0.63 (0.03) | 0.85 (0.03) | 0.20 (0.03) | 0.84 (0.02) | 0.85 (0.01) | 0.31 (0.02) | 0.30 (0.01) | 0.89 (0.01) | 0.13 (0.01) |
| 30    | 125   | 200   | 0.73 (0.05) | 0.13 (0.08) | 0.79 (0.06) | 0.69 (0.03) | 0.84 (0.05) | 0.19 (0.03) | 0.92 (0.04) | 0.87 (0.01) | 0.30 (0.01) | 0.13 (0.02) | 0.96 (0.01) | 0.02 (0.01) |
| 30    | 250   | 50    | 0.68 (0.05) | 0.25 (0.08) | 0.73 (0.06) | 0.57 (0.02) | 0.88 (0.04) | 0.16 (0.04) | - | - | - |
| 30    | 250   | 100   | 0.75 (0.04) | 0.26 (0.03) | 0.74 (0.03) | 0.64 (0.03) | 0.85 (0.04) | 0.18 (0.03) | 0.88 (0.04) | 0.86 (0.02) | 0.34 (0.02) | 0.18 (0.02) | 0.92 (0.02) | 0.08 (0.01) |
| 30    | 250   | 200   | 0.85 (0.01) | 0.30 (0.06) | 0.79 (0.04) | 0.69 (0.02) | 0.83 (0.04) | 0.21 (0.02) | 0.97 (0.05) | 0.85 (0.03) | 0.35 (0.03) | 0.22 (0.02) | 0.94 (0.01) | 0.08 (0.02) |
| 60    | 125   | 50    | 0.68 (0.03) | 0.05 (0.03) | 0.80 (0.02) | 0.57 (0.04) | 0.89 (0.06) | 0.11 (0.05) | - | - | - |
| 60    | 125   | 100   | 0.68 (0.04) | 0.12 (0.04) | 0.75 (0.04) | 0.60 (0.03) | 0.85 (0.05) | 0.15 (0.04) | - | - | - |
| 60    | 125   | 200   | 0.74 (0.02) | 0.11 (0.02) | 0.82 (0.02) | 0.61 (0.03) | 0.82 (0.04) | 0.17 (0.03) | 0.86 (0.03) | 0.89 (0.02) | 0.25 (0.02) | 0.22 (0.01) | 0.95 (0.01) | 0.11 (0.01) |
| 60    | 250   | 50    | 0.70 (0.02) | 0.15 (0.02) | 0.77 (0.01) | 0.59 (0.04) | 0.82 (0.04) | 0.16 (0.03) | - | - | - |
| 60    | 250   | 100   | 0.70 (0.01) | 0.13 (0.10) | 0.79 (0.05) | 0.62 (0.04) | 0.80 (0.04) | 0.17 (0.03) | - | - | - |
| 60    | 250   | 200   | 0.76 (0.02) | 0.11 (0.01) | 0.85 (0.01) | 0.69 (0.05) | 0.80 (0.03) | 0.19 (0.04) | 0.91 (0.02) | 0.85 (0.02) | 0.33 (0.01) | 0.17 (0.01) | 0.84 (0.05) | 0.15 (0.03) |
| 90    | 125   | 50    | 0.63 (0.04) | 0.10 (0.04) | 0.75 (0.03) | 0.52 (0.02) | 0.89 (0.03) | 0.10 (0.03) | - | - | - |
| 90    | 125   | 100   | 0.66 (0.03) | 0.12 (0.03) | 0.74 (0.02) | 0.55 (0.04) | 0.87 (0.03) | 0.15 (0.02) | - | - | - |
| 90    | 125   | 200   | 0.67 (0.02) | 0.13 (0.02) | 0.76 (0.01) | 0.57 (0.03) | 0.85 (0.04) | 0.17 (0.03) | - | - | - |
| 90    | 250   | 50    | 0.58 (0.03) | 0.09 (0.02) | 0.68 (0.03) | 0.54 (0.05) | 0.87 (0.04) | 0.11 (0.04) | - | - | - |
| 90    | 250   | 100   | 0.65 (0.05) | 0.13 (0.04) | 0.73 (0.04) | 0.58 (0.04) | 0.82 (0.03) | 0.15 (0.03) | - | - | - |
| 90    | 250   | 200   | 0.70 (0.02) | 0.12 (0.02) | 0.78 (0.01) | 0.61 (0.06) | 0.80 (0.04) | 0.18 (0.05) | - | - | - |

TABLE 1 Functions observed on the evenly spaced grid. Average operating characteristics based on 50 repetitions are reported; standard deviations are given within the parentheses.
combinations of $n$, $p$, and $d$. This is not surprising because (i) FGLASSO is not designed for learning directed graphs; they were compared with the proposed FLiNG-BN because of the lack of alternative functional BN implementation. (ii) Although FPCA-LiNGAM and FPCA-PC are capable of learning directed graphs, they still performed poorly because they are implemented in a two-step procedure where there is little reason to believe that the basis coefficients extracted by the functional PCA in the first step are useful to capture the functional dependencies in the second step. (iii) Unlike the proposed approach, none of the competing methods controls for false discovery and some impose the stringent Gaussian assumption, resulting in high FDR and/or low TPR.

The proposed FLiNG-BN has a few hyperparameters $L$, $M$, $\alpha$, $(a_-, b_1)$, $(a_+, b_2)$, $(a_-, b_3)$, and $(a_+, b_4)$. We performed sensitivity analyses of these parameters at four different values with $(n, p, d) = (100, 30, 250)$ in Section C of the Supporting Information. Our model appeared to be relatively robust within the tested ranges of hyperparameters.

6 | APPLICATIONS

We applied the proposed FLiNG-BN to the brain EEG data set obtained from Begleiter (1999). The data set consists of 122 subjects with 77 in the alcoholic group and 45 in the control group, and was previously used to demonstrate functional undirected graphical models by Zhu et al. (2016) and Qiao et al. (2019). The 64 electrodes placed on subjects’ scalps (standard positions) measuring voltage values were sampled at 256 Hz for 1 s. Each subject completed 120 trials under one stimulus or two stimuli. See Zhang et al. (1995) for details of the data collection procedure. We averaged all trials for each subject under one stimulus condition. We separately analyzed these two groups to find their commonalities and differences in brain activity. Hence, we had $n = 77$ or $n = 45$ subjects and $p = 64$ functions representing the brain EEG signals at different scalp positions recorded at $d = 256$ time points. We focused on EEG signals filtered at $\alpha$ frequency bands between 8 and 12.5 Hz using the $\text{eegfilt}$ function in the EEGLAB toolbox from MATLAB (Delorme et al., 2004).

To check the Gaussianity of the observed functions, we performed Shapiro–Wilk normality test (Shapiro & Wilk, 1965) to each of $p = 64$ scalp positions at each of $d = 256$ time points. The null hypothesis (i.e., the observations are marginally Gaussian) was rejected for many combinations of scalp position and time point, and therefore, the non-Gaussianity of the proposed model is deemed appropriate.

Five orthonormal basis functions were selected for both the alcoholic and control groups according to the procedure described in Section B of the Supporting Information.

We ran MCMC for 10,000 iterations, discarded the first half as burn-in, and retained every 10th iteration after burn-in. The estimated basis functions are shown in Figure 4. As evident from the plot, they are very similar across the two groups. The causal networks estimated by thresholding the posterior probability of inclusion at 0.9 (by controlling the Bayesian false discovery rate (Müller et al., 2006) at 0.05) are shown in Figure 5. The sparsity level is approximately 3.0% for the alcoholic group and 2.5% for the control group.

Our results reveal several interesting patterns. First, the connection is relatively dense in the frontal region for both groups. Second, the alcoholic group has more directed connections detected in the left temporal and occipital regions. Third, most brain locations tend to connect to adjacent positions, while distant locations are much less connected. Figure 6 shows the common and differential networks for the two groups, where a substantial connectivity difference is observed between the two groups.

In addition, we demonstrated the proposed FLiNG-BN model with an additional application to COVID-19 multivariate longitudinal data, which have unevenly spaced measurements, in Section D of the Supporting Information.

7 | DISCUSSION

In this paper, we have proposed a functional BN model for causal discovery from multivariate functional data. We have discussed in detail a specific case of the functional BN, namely the functional linear non-Gaussian model, and proved the underlying causal structure is identifiable even if the functions are purely observational and observed with noises. A fully Bayesian inference procedure has been proposed to implement our framework. Through simulation studies and real data applications, we have demonstrated the ability of our model in causal discovery.

Lee and Li (2022) and Yang and Suzuki (2022) also addressed the causal discovery problem for functional data but they are significantly different from the proposed FLiNG-BN. First, they assumed their functions to be noiseless whereas we consider the scenario where functions are observed with noises. The causal identifiability theory is more complicated when functions are noisy. Second, Lee and Li (2022) assumed their functions to be Gaussian whereas our functions are non-Gaussian; this difference leads to different learning algorithms and identifiability theory. Yang and Suzuki (2022) only showed causal identification theoretically in the bivariate case. Third, their inference is a two-step procedure based on causal ordering identification and sparse function-on-function regression, while the proposed Bayesian hierarchical model admits a one-step inference procedure, which learns the graph...
structure by directly searching in the graph space without having to learn the causal ordering first.

We briefly discuss several possible directions to extend our current work. First, we may replace the underlying DAG with cyclic graphs, chain graphs, or ancestral graphs for more general causal and conditional independence structures. We have chosen a linear non-Gaussian SEM on the basis coefficients but this model can be replaced with a nonlinear SEM. Second, instead of fixing the number of basis functions, one could resort to increasing shrinkage...
Common (left panel) and differential (right panel) connections for the two groups. Black arrows indicate common connections, red arrows indicate connections detected by the alcoholic group only, and green arrows indicate connections detected by the control group only. This figure appears in color in the electronic version of this paper, and any mention of color refers to that version.

Prior (Bhattacharya & Dunson, 2011) to adaptively truncate redundant functions. Finally, since we have two groups of observations in the EEG application, it would be interesting to jointly estimate the brain networks or estimate the differential network.

Acknowledgments
Ni’s research was supported by NIH 1R01GM148974-01 and NSF DMS-2112943. Xu’s research was supported by NIH R01MH128085, NSF DMS-1918854, and NSF OAC-1940107. He’s research was supported by NSFC No. 11801560, and by the Public Computing Cloud, Remin University of China.

Data Availability Statement
The data that support the findings in this paper are openly available in the University of California Irvine Machine Learning Repository at http://doi.org/10.24432, reference number C5TS3D.

ORCID
Kejun He https://orcid.org/0000-0002-3996-6246
Yanzun Xu https://orcid.org/0000-0001-5554-8637
Yang Ni https://orcid.org/0000-0003-0636-2363

References
Barbieri, M.M. & Berger, J.O. (2004) Optimal predictive model selection. The Annals of Statistics, 32, 870–897.
Begleiter, H. (1999) EEG database. UCI Machine Learning Repository. DOI: https://doi.org/10.24432/C5TS3D.
Bhattacharya, A. & Dunson, D.B. (2011) Sparse Bayesian infinite factor models. Biometrika, 98, 291–306.
Carroll, C. et al. (2021) flapace: Functional data analysis and empirical dynamics.
Chiu, J.M. & Müller, H.G. (2014) Linear manifold modelling of multivariate functional data. Journal of the Royal Statistical Society: Series B, 76, 605–626.
Delorme, A. & Makeig, S. (2004) EEGLAB: an open source toolbox for analysis of single-trial EEG dynamics including independent component analysis. Journal of Neuroscience Methods, 134, 9–21.
Hoyer, P., Janzing, D., Mooij, J. M., Peters, J. & Schölkopf, B. (2008) Nonlinear causal discovery with additive noise models. Advances in Neural Information Processing Systems, 21, 689–696.
Kalisch, M., Hauser, A., Maathuis, M.H. & Mächler, M. (2020) An overview of the pcalg package for R.
Kowal, D.R. (2019) Integer-valued functional data analysis for measles forecasting. Biometrics, 75, 1321–1333.
Kowal, D.R., Matteson, D.S. & Ruppert, D. (2017) A Bayesian multivariate functional dynamic linear model. Journal of the American Statistical Association, 112, 723–744.
Lee, K.Y., Li, L., Li, B. & Zhao, H. (2022) Nonparametric functional graphical modeling through functional additive regression operator. Journal of the American Statistical Association, 1–15.
Lee, K.Y. & Li, L. (2022) Functional structural equation model. Journal of the Royal Statistical Society: Series B, 84, 600–629.
Lee, W., Miranda, M.F., Rausch, P., Baladandayuthapani, V., Fazio, M., Downs, J.C. & Morris, J.S. (2018) Bayesian semiparametric functional mixed models for serially correlated functional data, with application to glaucoma data. Journal of the American Statistical Association, 114, 495–513.
Leng, C. & Tang, C.Y. (2012) Sparse matrix graphical models. Journal of the American Statistical Association, 107, 1187–1200.
Li, B. & Solea, E. (2018) A nonparametric graphical model for functional data with application to brain networks based on fMRI. Journal of the American Statistical Association, 113, 1637–1655.
Lindquist, M.A. (2012) Functional causal mediation analysis with an application to brain connectivity. *Journal of the American Statistical Association*, 107, 1297–1309.

Maathuis, M.H., Kalisch, M. & Bühlmann, P. (2009) Estimating high-dimensional intervention effects from observational data. *The Annals of Statistics*, 37, 3133–3164.

Müller, P., Parmigiani, G. & Rice, K. (2006) FDR and Bayesian multiple comparisons rules. In: *Proceedings of the 8th Valencia World Meeting on Bayesian Statistics*.

Ni, Y., Stingo, F.C. & Baladandayuthapani, V. (2017) Sparse multidimensional graphical models: a unified Bayesian framework. *Journal of the American Statistical Association*, 112, 779–793.

Pearl, J. (2000) *Causality: models, reasoning, and inference*. Cambridge, MA: Cambridge University Press.

Peters, J. & Bühlmann, P. (2014) Identifiability of Gaussian structural equation models with equal error variances. *Biometrika*, 101, 219–228.

Peters, J., Janzing, D. & Schölkopf, B. (2017) *Elements of causal inference: foundations and learning algorithms*. Cambridge, MA: The MIT Press.

Qiao, X., Guo, S. & James, G.M. (2019) Functional graphical models. *Journal of the American Statistical Association*, 114, 211–222.

Qiao, X., Qian, C., James, G.M. & Guo, S. (2020) Doubly functional graphical models in high dimensions. *Biometrika*, 107, 415–431.

Scott, J.G. & Berger, J.O. (2010) Bayes and empirical-Bayes multiplicity adjustment in the variable-selection problem. *The Annals of Statistics*, 38, 2587–2619.

Shapiro, S.S. & Wilk, M.B. (1965) An analysis of variance test for normality (complete samples). *Biometrika*, 52, 591–611.

Shimizu, S., Hoyer, P.O., Hyvärinen, A. & Kerminen, A. (2006) A linear non-Gaussian acyclic model for causal discovery. *Journal of Machine Learning Research*, 7, 2003–2030.

Solea, E. & Li, B. (2022) Copula Gaussian graphical models for functional data. *Journal of the American Statistical Association*, 117, 781–793.

Spirtes, P., Glymour, C. & Scheines, R. (2000) *Causation, prediction, and search*. Cambridge, MA: MIT Press.

Spirtes, P. & Glymour, C. (1991) An algorithm for fast recovery of sparse causal graphs. *Social Science Computer Review*, 9, 62–72.

Wand, M.P. & Ormerod, J.T. (2008) On semiparametric regression with O’Sullivan penalized splines. *Australian & New Zealand Journal of Statistics*, 50, 179–198.

Wang, H. & West, M. (2009) Bayesian analysis of matrix normal graphical models. *Biometrika*, 96, 821–834.

Yang, T. & Suzuki, J. (2022) The functional LiNGAM. In: *International Conference on Probabilistic Graphical Models*, volume 186, pages 25–36.

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

Zapata, J., Oh, S.Y. & Petersen, A. (2022) Partial separability and functional graphical models for multivariate Gaussian processes. *Biometrika*, 109, 665–681.

Zhang, L., Baladandayuthapani, V., Zhu, H., Baggerly, K.A., Majewski, T., Czerniak, B.A. & Morris, J.S. (2016) Functional CAR models for large spatially correlated functional datasets. *Journal of the American Statistical Association*, 111, 772–786.

Zhang, X.L., Begleiter, H., Porjesz, B., Wang, W. & Litke, A. (1995) Event related potentials during object recognition tasks. *Brain Research Bulletin*, 38, 531–538.

Zhu, H., Brown, P.J. & Morris, J.S. (2011) Robust, adaptive functional regression in functional mixed model framework. *Journal of the American Statistical Association*, 106, 1167–1179.

Zhu, H., Straw, N. & Dunson, D.B. (2016) Bayesian graphical models for multivariate functional data. *Journal of Machine Learning Research*, 17, 1–27.

**SUPPORTING INFORMATION**

Proofs, model and MCMC details, additional simulation experiments, and real data examples referenced in Sections 3-6, and codes for MCMC simulations are available with this paper at the Biometrics website on Wiley Online Library.

**How to cite this article**: Zhou, F., He, K., Wang, K., Xu, Y. & Ni, Y. (2023) Functional Bayesian networks for discovering causality from multivariate functional data. *Biometrics*, 79, 3279–3293. [https://doi.org/10.1111/biom.13922](https://doi.org/10.1111/biom.13922)