Supporting Information for Functional Bayesian Networks for Discovering Causality from Multivariate Functional Data by

Fangting Zhou, Kejun He, Kunbo Wang, Yanxun Xu, and Yang Ni

A. Proof of Theorem 1

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.*

Proof. Recall from the main text that the noisy functional observations $W_j(\cdot)$ are assumed to be drawn from a mean function $Y_j(\cdot)$ plus a white noise process $e_j(\cdot)$ with variance σ_j, and the mean function $Y_j(\cdot)$ is expanded by a set of orthonormal basis functions $\phi_{j1}(\cdot), \ldots, \phi_{jK}(\cdot)$, that is,

$$W_j(\cdot) = Y_j(\cdot) + e_j(\cdot) = \sum_{k=1}^{K} Z_{jk}\phi_{jk}(\cdot) + e_j(\cdot),\ \forall j \in [p].$$

We project each $W_j(\cdot)$ onto the space spanned by the basis functions $\phi_{j1}(\cdot), \ldots, \phi_{jK}(\cdot)$,

$$W_{jk} = \int W_j(\omega)\phi_{jk}(\omega)d\omega = Z_{jk} + e_{jk},\ \forall k \in [K], \tag{S.1}$$

where $e_{jk} = \int e_j(\omega)\phi_{jk}(\omega)d\omega \sim N(0, \sigma_j)$ and each $\boldsymbol{Z}_j = (Z_{j1}, \ldots, Z_{jK})^T$ follows a structural equation model,

$$\boldsymbol{Z}_j = \sum_{\ell=1}^{p} \boldsymbol{B}_{j\ell}\boldsymbol{Z}_\ell + \boldsymbol{\epsilon}_j,\ \forall j \in [p], \tag{S.2}$$

where $\boldsymbol{B}_{j\ell} = [B_{j\ell}(k_j, k_\ell)] \in \mathbb{R}^{K\times K}$ and $\boldsymbol{\epsilon}_j = (\epsilon_{j1}, \ldots, \epsilon_{jK})^T$. Now let $\boldsymbol{Z} = (\boldsymbol{Z}_1^T, \ldots, \boldsymbol{Z}_p^T)^T$ and $\boldsymbol{\epsilon} = (\boldsymbol{\epsilon}_1^T, \ldots, \boldsymbol{\epsilon}_p^T)^T$ be the $p \times K$ dimensional vectors, and stack $\boldsymbol{W}_j = (W_{j1}, \ldots, W_{jK})^T$ and $\boldsymbol{e}_j = (e_{j1}, \ldots, e_{jK})^T$ similarly into $\boldsymbol{W} = (\boldsymbol{W}_1^T, \ldots, \boldsymbol{W}_p^T)^T$ and $\boldsymbol{E} = (\boldsymbol{e}_1^T, \ldots, \boldsymbol{e}_p^T)^T$. Define the block matrix $\boldsymbol{B}$ with $\boldsymbol{B}_{j\ell}$ its (j, ℓ)-th block. Let $\boldsymbol{\Omega} = (\boldsymbol{I} - \boldsymbol{B})^{-1}$ and partition $\boldsymbol{\Omega}$ the same way as $\boldsymbol{B}$. Note that $\boldsymbol{\Omega}$ has unit diagonal elements and is a full rank matrix because of the acyclicity constraint, and $\boldsymbol{\Omega}_{j\ell} \neq \boldsymbol{O}$ if and only if $\ell \in an(j)$. See, for example, Shojaie and Michailidis (2010)), where $an(j) = \{\ell : \ell \to \cdots \to j\}$ denotes the set of ancestors of j.

Now we can rewrite Equations (S.1) and (S.2) in a matrix form,

$$\begin{aligned} \boldsymbol{W} &= \boldsymbol{Z} + \boldsymbol{E} \\ &= \boldsymbol{\Omega\epsilon} + \boldsymbol{E}, \end{aligned} \tag{S.3}$$

where $\boldsymbol{E} \sim N(\boldsymbol{0}, \boldsymbol{\Sigma})$ with $\boldsymbol{\Sigma} = \text{diag}(\tau_1, \ldots, \tau_1, \ldots, \tau_p, \ldots, \tau_p)$ a diagonal matrix. Note the mixture of Gaussian distributions $\epsilon_{jk} \sim \sum_{m=1}^{M_{jk}} \pi_{jk}^m N(\mu_{jk}^m, \tau_{jk}^m)$ can be equivalently written as

$$\epsilon_{jk}|\xi_{jk} = m \sim N(\mu_{jk}^m, \tau_{jk}^m),\ P(\xi_{jk} = m) = \pi_{jk}^m.$$

Stack $\boldsymbol{\xi}_j = (\xi_{j1}, \ldots, \xi_{jK})^T$ the same way as $\boldsymbol{\epsilon}_j$ into $\boldsymbol{\xi}$. Then conditional on $\boldsymbol{\xi}$, the distribution of $\boldsymbol{\Omega\epsilon}$ is given by

$$\boldsymbol{\Omega\epsilon}|\boldsymbol{\xi} \sim N(\boldsymbol{\Omega\mu_\xi}, \boldsymbol{\Omega T_\xi \Omega}^T),$$

where $\boldsymbol{\mu_\xi} = [\mu_{jk}^{\xi_{jk}}]$ is a $p \times K$ dimensional vector and $\boldsymbol{T_\xi} = \text{diag}(\boldsymbol{\tau_\xi})$ is a diagonal matrix with entries $\boldsymbol{\tau_\xi} = [\tau_{jk}^{\xi_{jk}}]$ defined similarly as $\boldsymbol{\mu_\xi}$. Hence, from (S.3), we have

$$\boldsymbol{W}|\boldsymbol{\xi} \sim N(\boldsymbol{\Omega\mu_\xi}, \boldsymbol{\Omega T_\xi \Omega}^T + \boldsymbol{\Sigma}).$$

Now consider two BNs $\mathcal{B} = (G, P)$ and $\mathcal{B}' = (G', P')$ with $G \neq G'$. Without loss of generality, assume the nodes of G and G' are labeled according to the topological/causal ordering of G, i.e., $\ell \not\to j$ in G if $\ell > j$. Consequently, $\boldsymbol{\Omega}$ is a lower triangular matrix because $\boldsymbol{B}$ is a lower triangular matrix and $\boldsymbol{\Omega} = (\boldsymbol{I} - \boldsymbol{B})^{-1}$. We prove by contradiction and mathematical induction that $\mathcal{B}$ and $\mathcal{B}'$ are not equivalent if $G \neq G'$.

Suppose $\mathcal{B}$ and $\mathcal{B}'$ are equivalent, and therefore $P(\boldsymbol{W}) \equiv P'(\boldsymbol{W})$. Then because of the identifiability of finite Gaussian mixture models (Yakowitz and Spragins, 1968) up to label permutation, we must have, for any cluster assignment $\boldsymbol{\xi}$,

$$\boldsymbol{\Omega T_\xi \Omega}^T + \boldsymbol{\Sigma} = \boldsymbol{\Omega}' \boldsymbol{T}'_{\boldsymbol{\xi}} \boldsymbol{\Omega}'^T + \boldsymbol{\Sigma}'.$$

Therefore, for any two clusters $\boldsymbol{\xi}$ and $\widetilde{\boldsymbol{\xi}}$,

$$\boldsymbol{\Omega}(\boldsymbol{T}_{\boldsymbol{\xi}} - \boldsymbol{T}_{\widetilde{\boldsymbol{\xi}}})\boldsymbol{\Omega}^T = \boldsymbol{\Omega}'(\boldsymbol{T}'_{\boldsymbol{\xi}} - \boldsymbol{T}'_{\widetilde{\boldsymbol{\xi}}})\boldsymbol{\Omega}'^T. \tag{S.4}$$

For each j and k, we pick $\xi_{jk} \neq \widetilde{\xi}_{jk}$ such that $\tau_{jk}^{\xi_{jk}} \neq \tau_{jk}^{\widetilde{\xi}_{jk}}$ (the Lebesgue measure of $\{(\tau_{jk}^1, \ldots, \tau_{jk}^{M_{jk}}) | \tau_{jk}^m = \tau_{jk}^{m'}, \forall m \neq m'\}$ is zero). The resulting $\boldsymbol{\xi}$ and $\widetilde{\boldsymbol{\xi}}$ then give rise to a full rank diagonal matrix $\boldsymbol{T}_{\boldsymbol{\xi}} - \boldsymbol{T}_{\widetilde{\boldsymbol{\xi}}}$. Due to (S.4), $\boldsymbol{T}'_{\boldsymbol{\xi}} - \boldsymbol{T}'_{\widetilde{\boldsymbol{\xi}}}$ must also be a full rank matrix. Therefore, it suffices to show $\boldsymbol{\Omega}'$ must be lower triangular and then the uniqueness of LDL decomposition would lead to the contradiction that $G \neq G'$. To prove $\boldsymbol{\Omega}'$ is lower triangular, we use mathematical induction with respect to the its columns.

First, consider the base case – the last column $\ell = p \times K$ of $\boldsymbol{\Omega}'$. We pick $\boldsymbol{\xi} = \widetilde{\boldsymbol{\xi}}$ except for $\xi_{pK} \neq \widetilde{\xi}_{pK}$ such that $\tau_{pK}^{\xi_{pK}} \neq \tau_{pK}^{\widetilde{\xi}_{pK}}$. Then, $\boldsymbol{T}_{\boldsymbol{\xi}} - \boldsymbol{T}_{\widetilde{\boldsymbol{\xi}}}$ is a diagonal matrix with only the (ℓ, ℓ)-th entry no-zero. In addition, $\boldsymbol{T}_{\boldsymbol{\xi}} - \boldsymbol{T}_{\widetilde{\boldsymbol{\xi}}}$ is rank one, and due to (S.4), $\boldsymbol{T}'_{\boldsymbol{\xi}} - \boldsymbol{T}'_{\widetilde{\boldsymbol{\xi}}}$ must also be a diagonal matrix with rank one and $(\boldsymbol{T}'_{\boldsymbol{\xi}} - \boldsymbol{T}'_{\widetilde{\boldsymbol{\xi}}})_{tt} \neq 0$ for exactly one $t \in [\ell]$. Recall that $\boldsymbol{\Omega}$ is a lower unit triangular matrix: $\Omega_{ss} = 1$ and $\Omega_{st} = 0$ for any $t > s$. Hence,

$$\begin{aligned}
0 = \Omega_{r\ell}(\boldsymbol{T}_{\boldsymbol{\xi}} - \boldsymbol{T}_{\widetilde{\boldsymbol{\xi}}})_{\ell\ell}\Omega_{q\ell} &= \sum_{s=1}^{\ell} \Omega_{rs}(\boldsymbol{T}_{\boldsymbol{\xi}} - \boldsymbol{T}_{\widetilde{\boldsymbol{\xi}}})_{ss}\Omega_{qs} = \boldsymbol{\Omega}_{r\star}(\boldsymbol{T}_{\boldsymbol{\xi}} - \boldsymbol{T}_{\widetilde{\boldsymbol{\xi}}})\boldsymbol{\Omega}_{q\star}^T = \boldsymbol{\Omega}'_{r\star}(\boldsymbol{T}'_{\boldsymbol{\xi}} - \boldsymbol{T}'_{\widetilde{\boldsymbol{\xi}}})\boldsymbol{\Omega}_{q\star}'^T \\
&= \sum_{s=1}^{\ell} \Omega'_{rs}(\boldsymbol{T}'_{\boldsymbol{\xi}} - \boldsymbol{T}'_{\widetilde{\boldsymbol{\xi}}})_{ss}\Omega'_{qs} = \Omega'_{rt}(\boldsymbol{T}'_{\boldsymbol{\xi}} - \boldsymbol{T}'_{\widetilde{\boldsymbol{\xi}}})_{tt}\Omega'_{qt}, \ \forall r < \ell \text{ or } q < \ell,
\end{aligned}$$

$$\begin{aligned}
0 \neq \Omega_{\ell\ell}(\boldsymbol{T}_{\boldsymbol{\xi}} - \boldsymbol{T}_{\widetilde{\boldsymbol{\xi}}})_{\ell\ell}\Omega_{\ell\ell} &= \sum_{s=1}^{\ell} \Omega_{\ell s}(\boldsymbol{T}_{\boldsymbol{\xi}} - \boldsymbol{T}_{\widetilde{\boldsymbol{\xi}}})_{ss}\Omega_{\ell s} = \boldsymbol{\Omega}_{\ell\star}(\boldsymbol{T}_{\boldsymbol{\xi}} - \boldsymbol{T}_{\widetilde{\boldsymbol{\xi}}})\boldsymbol{\Omega}_{\ell\star}^T = \boldsymbol{\Omega}'_{\ell\star}(\boldsymbol{T}'_{\boldsymbol{\xi}} - \boldsymbol{T}'_{\widetilde{\boldsymbol{\xi}}})\boldsymbol{\Omega}_{\ell\star}'^T \\
&= \sum_{s=1}^{\ell} \Omega'_{\ell s}(\boldsymbol{T}'_{\boldsymbol{\xi}} - \boldsymbol{T}'_{\widetilde{\boldsymbol{\xi}}})_{ss}\Omega'_{\ell s} = \Omega'_{\ell t}(\boldsymbol{T}'_{\boldsymbol{\xi}} - \boldsymbol{T}'_{\widetilde{\boldsymbol{\xi}}})_{tt}\Omega'_{\ell t},
\end{aligned}$$

where the subscript $r\star$ denotes the rth row of a matrix as a row vector. This further implies

$$\Omega'_{rt} = 0, \ \forall r < \ell,$$

$$\Omega'_{\ell t} \neq 0.$$

Recall that due to the acyclic constraint $\Omega'_{\ell\ell} = 1$. Therefore, $t = \ell$ and $\Omega'_{r\ell} = 0$ for any $r < \ell$, which completes the proof of the base case.

Now, suppose we have proved that for any $\ell > i$, it holds that $\Omega'_{r\ell} = 0$ for any $r < \ell$. We need to show this is true for $\ell = i$. As before, pick $\boldsymbol{\xi} = \widetilde{\boldsymbol{\xi}}$ except for $\xi_{jk} \neq \widetilde{\xi}_{jk}$ such that $\tau_{jk}^{\xi_{jk}} \neq \tau_{jk}^{\widetilde{\xi}_{jk}}$, where $\ell = (j-1) \times K + k$. Then, $\boldsymbol{T}_{\boldsymbol{\xi}} - \boldsymbol{T}_{\widetilde{\boldsymbol{\xi}}}$ is a diagonal matrix with only the (ℓ, ℓ)-th entry no-zero. In addition, $\boldsymbol{T}_{\boldsymbol{\xi}} - \boldsymbol{T}_{\widetilde{\boldsymbol{\xi}}}$ is rank one, and due to (S.4), $\boldsymbol{T}'_{\boldsymbol{\xi}} - \boldsymbol{T}'_{\widetilde{\boldsymbol{\xi}}}$ must also be a diagonal matrix with rank one and $(\boldsymbol{T}'_{\boldsymbol{\xi}} - \boldsymbol{T}'_{\widetilde{\boldsymbol{\xi}}})_{tt} \neq 0$ for exactly one $t \in [\ell]$. Note that by the induction assumption, t must be less than or equal to ℓ, because otherwise the first ℓ by ℓ block of $\boldsymbol{\Omega}'(\boldsymbol{T}'_{\boldsymbol{\xi}} - \boldsymbol{T}'_{\widetilde{\boldsymbol{\xi}}})\boldsymbol{\Omega}'^T$ would be zero, whereas the corresponding block of $\boldsymbol{\Omega}(\boldsymbol{T}_{\boldsymbol{\xi}} - \boldsymbol{T}_{\widetilde{\boldsymbol{\xi}}})\boldsymbol{\Omega}^T$ cannot be zero, because $(\boldsymbol{\Omega}(\boldsymbol{T}_{\boldsymbol{\xi}} - \boldsymbol{T}_{\widetilde{\boldsymbol{\xi}}})\boldsymbol{\Omega}^T)_{\ell\ell} = \boldsymbol{\Omega}_{\ell\star}(\boldsymbol{T}_{\boldsymbol{\xi}} - \boldsymbol{T}_{\widetilde{\boldsymbol{\xi}}})\boldsymbol{\Omega}^T_{\ell\star} = (\boldsymbol{T}_{\boldsymbol{\xi}} - \boldsymbol{T}_{\widetilde{\boldsymbol{\xi}}})_{\ell\ell} \neq 0$. Hence,

$$0 = \boldsymbol{\Omega}_{r\star}(\boldsymbol{T}_{\boldsymbol{\xi}} - \boldsymbol{T}_{\widetilde{\boldsymbol{\xi}}})\boldsymbol{\Omega}^T_{q\star} = \boldsymbol{\Omega}'_{r\star}(\boldsymbol{T}'_{\boldsymbol{\xi}} - \boldsymbol{T}'_{\widetilde{\boldsymbol{\xi}}})\boldsymbol{\Omega}'^T_{q\star} = \Omega'_{rt}(\boldsymbol{T}'_{\boldsymbol{\xi}} - \boldsymbol{T}'_{\widetilde{\boldsymbol{\xi}}})_{tt}\Omega_{qt}, \ \forall r < \ell \text{ or } q < \ell,$$

$$0 \neq (\boldsymbol{T}_{\boldsymbol{\xi}} - \boldsymbol{T}_{\widetilde{\boldsymbol{\xi}}})_{\ell\ell} = \boldsymbol{\Omega}_{\ell\star}(\boldsymbol{T}_{\boldsymbol{\xi}} - \boldsymbol{T}_{\widetilde{\boldsymbol{\xi}}})\boldsymbol{\Omega}^T_{\ell\star} = \boldsymbol{\Omega}'_{\ell\star}(\boldsymbol{T}'_{\boldsymbol{\xi}} - \boldsymbol{T}'_{\widetilde{\boldsymbol{\xi}}})\boldsymbol{\Omega}'^T_{\ell\star} = \Omega'_{\ell t}(\boldsymbol{T}'_{\boldsymbol{\xi}} - \boldsymbol{T}'_{\widetilde{\boldsymbol{\xi}}})_{tt}\Omega'_{\ell t},$$

which respectively imply

$$\Omega'_{rt} = 0, \ \forall r < \ell,$$

$$\Omega'_{\ell t} \neq 0.$$

Due to the acyclic constraint $\Omega'_{\ell\ell} = 1$. Therefore, $t = \ell$ and $\Omega'_{r\ell} = 0$ for any $r < \ell$, which completes the proof.

B. Posterior Inference

The closed form of the posterior distribution is not available; we use MCMC to draw posterior samples. We assume n realizations of multivariate random functions are available and use a superscript (i) to index the realization. For updating the Gaussian scale mixture parameters, we introduce for each realization a latent vector $\boldsymbol{c}_i = (\boldsymbol{c}_1^{(i)}, \ldots, \boldsymbol{c}_p^{(i)})$ with $\boldsymbol{c}_j^{(i)} = (c_{j1}^{(i)}, \ldots, c_{jK}^{(i)})$ where $c_{jk}^{(i)} = m$ if the exogenous variable $\epsilon_{jk}^{(i)}$ belongs to mixture component

m. Let $\boldsymbol{T}_i = \text{diag}(\boldsymbol{\tau}_1^{(i)}, \ldots, \boldsymbol{\tau}_p^{(i)})$ be the diagonal covariance matrix for exogenous variables $\boldsymbol{\epsilon}^{(i)} = (\boldsymbol{\epsilon}_1^{(i)}, \ldots, \boldsymbol{\epsilon}_p^{(i)})$, where $\boldsymbol{\tau}_j^{(i)} = (\tau_{j1}^{(i)}, \ldots, \tau_{jK}^{(i)})$ and $\tau_{jk}^{(i)} = \sum_{m=1}^{M} \tau_{jk}^{m} I(c_{jk}^{(i)} = m)$.

Choosing the Number K of Basis Functions. While we could put a prior on K to learn the number of basis functions jointly with other parameters through reversible jump MCMC, or adaptively truncate and eliminate redundant functions through shrinkage priors like in Bhattacharya and Dunson (2011); Legramanti et al. (2020), they can lead to considerable computational burden and potential Markov chain mixing problems. Therefore, in this article, we take a simple heuristic approach (Kowal et al., 2017). First, we impute the functional observations and arrange the data into a $(n \times p) \times d$ matrix where $d = |\cup_{i=1}^{n} \cup_{j=1}^{p} D_j^{(i)}|$ is the size of the union of the measurement grid over all realized random functions. Then, we perform singular value decomposition and select the minimum K such that the proportion of variance explained exceeds 90%. The value is fixed throughout MCMC. Note that although K is fixed, the basis functions are adaptively inferred.

Updating the Transformed B-Spline Coefficients $\widetilde{\boldsymbol{A}}_k$. Recursively for each $k \in [K]$, denote $\widetilde{w}_{j,-k}^{(i)}(m) = w_j^{(i)}(m) - \sum_{h \neq k} z_{jh}^{(i)} \phi_h(\omega_j^{(i)}(m))$ for any $i \in [n]$, $j \in [p]$, and $m \in [m_j^{(i)}]$. Conditional on all other parameters, first calculate the unconstrained mean and variance from its full conditional distribution

$$\boldsymbol{V}_k = \sum_{i=1}^{n} \sum_{j=1}^{p} \left(\sum_{m=1}^{m_j^{(i)}} \boldsymbol{b}(\omega_j^{(i)}(m)) \boldsymbol{b}^T(\omega_j^{(i)}(m)) \right) \times \sigma_j^{-1} \times (z_{jk}^{(i)})^2 + \text{diag}(10^{-8}, 10^{-8}, \lambda_k, \ldots, \lambda_k),$$

$$\boldsymbol{\mu}_k = \boldsymbol{V}_k^{-1} \left[\sum_{i=1}^{n} \sum_{j=1}^{p} \left(\sum_{m=1}^{m_j^{(i)}} \widetilde{w}_{j,-k}^{(i)}(m) \boldsymbol{b}(\omega_j^{(i)}(m)) \right) \times \sigma_j^{-1} \times z_{jk}^{(i)} \right].$$

Then sample $\widetilde{\boldsymbol{A}}_k^U$ from $N(\boldsymbol{\mu}_k, \boldsymbol{V}_k^{-1})$. Denote $\boldsymbol{P}_k = \boldsymbol{J} \widetilde{\boldsymbol{A}}_{-k}$, where $\boldsymbol{J}$ is the penalty matrix from Equation (5) in the main text. Finally, transform and normalize the unconstrained sample to $\widetilde{\boldsymbol{A}}_k^N = \widetilde{\boldsymbol{A}}_k^U - \boldsymbol{V}_k^{-1} \boldsymbol{P}_k (\boldsymbol{P}_k^T \boldsymbol{V}_k^{-1} \boldsymbol{P}_k)^{-1} \boldsymbol{P}_k \widetilde{\boldsymbol{A}}_k^U$ and $\widetilde{\boldsymbol{A}}_k = \widetilde{\boldsymbol{A}}_k^N \times ([\widetilde{\boldsymbol{A}}_k^N]^T \boldsymbol{J} \widetilde{\boldsymbol{A}}_k^N)^{-1/2}$.

Updating the Regularization Parameter λ_k. Independently for each $k \in [K]$, conditional on all other parameters, denote $\beta_k = 2^{-1} \times \sum_{\ell=1}^{L} A_{k\ell}^2$ and $\alpha = 2^{-1} \times (L-2)$. We sample each λ_k from a gamma distribution $G(\alpha, \beta_k)$ truncated at (L_k, U_k).

Updating the Variance of Observation Noises. Independently for each $j = 1, \ldots, p$, denote the residual term as $\widetilde{w}_j^{(i)}(m) = w_j^{(i)}(m) - \sum_{k=1}^{K} z_{jk}^{(i)} \phi_k(\omega_j^{(i)}(m))$. Conditional on all other parameters, we sample each σ_j from an inverse-gamma distribution $IG(\alpha_j, \beta_j)$, where $\alpha_j = 10^{-2} + 2^{-1} \times \sum_{i=1}^{n} m_j^{(i)}$ and $\beta_j = 10^{-2} + 2^{-1} \times \sum_{i=1}^{n} \sum_{m=1}^{m_j^{(i)}} (\widetilde{w}_j^{(i)}(m))^2$.

Updating Basis Coefficient Sequences. Denote $\boldsymbol{z}_i = (\boldsymbol{z}_1^{(i)}, \ldots, \boldsymbol{z}_p^{(i)})^T$ where $\boldsymbol{z}_j^{(i)} = (z_{j1}^{(i)}, \ldots, z_{jK}^{(i)})$, for each $i = 1, \ldots, n$. Let

$$\begin{aligned}
\boldsymbol{V}_i &= \text{block-diag}(\boldsymbol{V}_1^{(i)}, \ldots, \boldsymbol{V}_p^{(i)}) + (\boldsymbol{I} - \boldsymbol{B})\boldsymbol{T}_i^{-1}(\boldsymbol{I} - \boldsymbol{B})^T, \\
&\qquad \text{where } \boldsymbol{V}_j^{(i)} = \sigma_j^{-1} \times \left(\sum_{m=1}^{m_j^{(i)}} \boldsymbol{\phi}(\omega_j^{(i)}(m)) \boldsymbol{\phi}^T(\omega_j^{(i)}(m)) \right), \\
\boldsymbol{\mu}_i &= \boldsymbol{V}_i^{-1}(\boldsymbol{\mu}_1^{(i)}, \ldots, \boldsymbol{\mu}_p^{(i)}), \text{ where } \boldsymbol{\mu}_j^{(i)} = \sigma_j^{-1} \times \left(\sum_{m=1}^{m_j^{(i)}} w_j^{(i)}(m) \boldsymbol{\phi}(\omega_j^{(i)}(m)) \right),
\end{aligned}$$

where block-diag$(\boldsymbol{V}_1, \ldots, \boldsymbol{V}_p)$ is a block-diagonal matrix with diagonal blocks given by $\boldsymbol{V}_1, \ldots, \boldsymbol{V}_p$. One then independently samples $\boldsymbol{z}_i \sim N(\boldsymbol{\mu}_i, \boldsymbol{V}_i^{-1})$.

Updating the Adjacency Matrix $\boldsymbol{E}$. Recursively for each $E_{j\ell}$, we perform a birth/death move such that $\boldsymbol{E}' = \boldsymbol{E}$ except $E'_{j\ell} = 1 - E_{j\ell}$. If the resulting $\boldsymbol{E}'$ is not acyclic, proceed directly to the next step. Otherwise, calculate

$$\begin{aligned}
&V_j^{(i)} = \gamma \times \sum_{\ell=1}^{p} I(E_{j\ell} = 1) \sum_{k=1}^{K} (z_{\ell k}^{(i)})^2, \quad V_j^{(i)'} = \gamma \times \sum_{\ell=1}^{p} I(E'_{j\ell} = 1) \sum_{k=1}^{K} (z_{\ell k}^{(i)})^2, \\
&\alpha = \sum_{i=1}^{n} \sum_{k=1}^{K} \left[\text{LN}(z_{jk}^{(i)}; 0, \tau_{jk}^{(i)} + V_j^{(i)}) - \text{LN}(z_{jk}^{(i)}; 0, \tau_{jk}^{(i)} + V_j^{(i)'}) \right] + (1 - 2E_{j\ell}) \log(r^{-1}(1-r)),
\end{aligned}$$

where $\text{LN}(x; \mu, s)$ denotes the log-normal density at x with mean μ and variance s. Then sample $u \sim U(0,1)$ and take the new proposal if $\alpha < \log(u^{-1} - 1)$.

Updating the Direct Causal Effects $\boldsymbol{B}$. For each $\ell \neq j \in [p]$, if $E_{j\ell} = 0$, then take $\boldsymbol{B}_{j\ell} = \boldsymbol{O}$. Otherwise, denote N_j the set of indices ℓ such that $E_{j\ell} = 1$. Let $\boldsymbol{B}_{N_j} = (\boldsymbol{B}_{j\ell})_{\ell \in N_j}$ and $\boldsymbol{b}_{jk}$ be

the k-th row of $\boldsymbol{B}_{N_j}$. Similarly, denote $\boldsymbol{z}_{N_j}^{(i)} = (\boldsymbol{z}_\ell^{(i)})_{\ell \in N_j}$ the subvector of $\boldsymbol{z}^{(i)}$ corresponds to non-zero connections. Then for each $k = 1, \ldots, K$, denote

$$\boldsymbol{V}_{jk} = \gamma^{-1}\boldsymbol{I} + \sum_{i=1}^{n}\left[(\tau_{jk}^{(i)})^{-1} \times \boldsymbol{z}_{N_j}^{(i)}(\boldsymbol{z}_{N_j}^{(i)})^T\right], \quad \boldsymbol{\mu}_{jk} = \boldsymbol{V}_{jk}^{-1}\sum_{i=1}^{n}\left[(\tau_{jk}^{(i)})^{-1} \times z_{jk}^{(i)} \times \boldsymbol{z}_{N_j}^{(i)}\right].$$

We sample each $\boldsymbol{b}_{jk}$ independently from $N(\boldsymbol{\mu}_{jk}, \boldsymbol{V}_{jk}^{-1})$.

Updating the Causal Effect Size γ. Denote the shape $\alpha = 1 + K^2 \times 2^{-1} \times \sum_{j=1}^{p}\sum_{\ell=1}^{p} E_{j\ell}$ and scale $\beta = 1 + 2^{-1} \times \sum_{j=1}^{p}\sum_{\ell=1}^{p}\sum_{s=1}^{K}\sum_{t=1}^{K} B_{j\ell}^2(s,t)$. Given other parameters, sample γ from $IG(\alpha, \beta)$.

Updating the Edge Probability r. Let $a = 1+\sum_{j=1}^{p}\sum_{\ell=1}^{p} E_{j\ell}$ and $b = 1-p+\sum_{j=1}^{p}\sum_{\ell=1}^{p}(1-E_{j\ell})$. Given other parameters, sample the edge probability r from Beta(a, b).

Updating the Gaussian Scale Mixture. Denote $\boldsymbol{\varepsilon}^{(i)} = \boldsymbol{z}_i - \boldsymbol{B}\boldsymbol{z}_i$ the current exogeneous variable for each sample $i = 1, \ldots, n$. For each $j \in [p]$ and $k \in [K]$, denote $\alpha_m = 1+\sum_{i=1}^{n} I(c_{jk}^{(i)} = m)$ and sample the cluster probability vector $\boldsymbol{\pi}_{jk}$ from Dirichlet$(\alpha_1, \ldots, \alpha_M)$. For each $i = 1, \ldots, n$, calculate $\log \pi_m^{(i)} = \log(\pi_{jk}^m) + \text{LN}(\varepsilon_{jk}^{(i)}; 0, \tau_{jk}^m)$. Sample $c_{jk}^{(i)}$ from $\{1, \ldots, M\}$ with class probability $(\pi_1^{(i)}, \ldots, \pi_M^{(i)})$. Finally, we update the variance parameter for each cluster from an inverse-gamma distribution $IG(\alpha_m, \beta_m)$, where $\alpha_m = 1 + 2^{-1} \times \sum_{i=1}^{n} I(c_{jk}^{(i)} = m)$ and $\beta_m = 1 + 2^{-1} \times \sum_{i=1}^{n} I(c_{jk}^{(i)} = m) \times (\varepsilon_{jk}^{(i)})^2$.

Initialization. We initialize the DAG G to be an empty graph (i.e., set the adjacency matrix E to a zero matrix O), the direct causal effect matrix $B = O$, the edge inclusion probability $r = 0.5$, the causal effect size $\gamma = 1$, and the observation noise variances $\sigma_j = 1$ for all j. For the Gaussian mixture, we initialize the component weight vectors $\pi_{jk} = (1/M, \ldots, 1/M)$ for each j and k, and the component-specific variance parameters $\tau_{jk}^m = 1$ for all m. The basis coefficients z_i are estimated from the singular value decomposition, for each i, and the regularization parameter $\lambda_k = 1e^{-8}$ for each k.

C. Simulations: Unevenly Spaced Grid and Sensitivity Analysis

C.1 *Functions Observed on Unevenly Spaced Grid*

We simulated data with $(n, p) \in \{500, 800\} \times \{20, 50\}$. Different from the first scenario in Section 5 of the main text, we first randomly sampled $m = 200$ points from Unif$(0, 1)$; let D denote the set of these 200 points. Then each realization i of the function j was assumed to be measured at a random subset $D_j^{(i)} \subset D$ of points with size $m_j^{(i)} = 10$. The causal graph, direct causal effect matrix, orthonormal basis functions, basis coefficient sequences, and observations were generated the same way as in Section 5 of the main text. We repeated each scenario for 50 repetitions and compared with FGLASSO, FPCA-LiNGAM, and FPCA-PC. Implementation details and accuracy evaluation were the same as in Section 5 of the main text. The results (shown in Table S.1) are similar to those in Section 5 of the main text (Table 1), which demonstrate the fact that the proposed FLiNG-BN is capable of and superior in learning DAGs for general multivariate functional data.

[Table 1 about here.]

C.2 *Sensitivity Analysis*

The proposed FLiNG-BN has a few hyperparameters L, M, α, (a_r, b_r), (a_γ, b_γ), (a_τ, b_τ), and (a_σ, b_σ). We performed sensitivity analyses of these parameters at four different values under scenario 1 with $(n, p, d) = (100, 30, 250)$. Results are summarized in Table S.2. Our model appeared to be relatively robust within the tested ranges of hyperparameters.

[Table 2 about here.]

D. Application: Demonstration with COVID-19 Clinical Data

We applied the proposed FLiNG-BN model to the clinical data of COVID-19 patients who were admitted to five hospitals of the Johns Hopkins Medicine healthcare system after January 1, 2021 from the Johns Hopkins University CROWN database. The dataset contains

demographic information, clinical variable measurements (laboratory results and vital signs), treatment information, and clinical outcomes for each patient (Ignatius et al., 2021; Garibaldi et al., 2021). In this experiment, we randomly selected $n = 100$ patients with length of stays in hospital between 3 to 14 days. We chose $p = 9$ functions including 5 vital signs (respiratory rate, pulse oximetry (SpO2), fraction of inspired oxygen (FiO2), systolic blood pressure (SBP), and diastolic blood pressure (DBP)) and 4 laboratory results (creatinine, lymphocyte count, white blood cell count (WBC), and glomerular filtration rate (GFR)). Measurement times were rounded to the nearest hour from admission, resulting in 302 unique measurement time points ($d = 302$). We took at most 1 measurement per variable and hour for each patient and used mean values if multiple measurements exist within an hour. The medians and interquartile ranges (IQR) of the number of measurements per patient for different variables are presented in Table S.3.

[Table 3 about here.]

Before applying the proposed method, we scaled each functional variable into zero mean and unit variance, and fitted linear regressions with function values being responses and demographic variables (age, race, and sex) being predictors. The fitted residuals were then used to demonstrate the performance of the FLiNG-BN model. We set the number of mixture components $M = 5$, the number of B-spline basis functions $L = 20$, hyperparameters of beta-Bernoulli prior $a_r = 20$, and $b_r = 10$. The other hyperparameters were set to default values as described in Section 4 of the main text. We ran MCMC for 10,000 iterations, discarded the first 7,500 iterations as burn-in, and retained every 10th iteration after burn-in. The causal network was estimated by thresholding the posterior probability of inclusion at 0.9. Our results show several patterns of the selected clinical variables. First, the connection is sparse with the sparsity level of 1.4%. Second, a strong directed connection from creatinine to GFR is detected. This result is consistent with clinical observations that GFR differs by

age, race, and sex, and has strong non-linear relationships with creatinine. And in practice, when GFR is not measured directly, creatinine combined with demographics can be used to estimate GFR.

E. Condition for Square Summable Sequences

To provide the condition under which the infinite sequence $\boldsymbol{Z}_j$ is square summable, we first introduce some notations. Denote $(u, v) \in E$ the directed edge from node u to v. A directed path $\pi = \langle v_1, \cdots, v_T \rangle$ is a set of distinct nodes such that $(v_{t-1}, v_t) \in E$ for all $t = 2, \ldots, T$, and we denote $E(\pi)$ the set of directed edges along the path π. Let $pa(j) = \{k : (k, j) \in E\}$ be the set of parents and $an(j) = \{k : \exists \pi = \langle v_1 = k, v_2, \cdots, v_{T-1}, v_T = j \rangle\}$ be the set of ancestors of node j, respectively. We denote $ph(k \to j)$ the set of all directed paths from node k to j.

To find a condition under which the infinite sequence is square summable, notice that

$$\sum_{k=1}^{\infty} (\pm 1) Z_{jk} < \infty \text{ for any choices of the value } \pm 1 \Rightarrow \sum_{k=1}^{\infty} Z_{jk}^2 < \infty.$$

See, for example, Resnick (2019). We rewrite

$$\boldsymbol{Z}_j = \sum_{\ell=1}^{p} \boldsymbol{B}_{j\ell} \boldsymbol{Z}_\ell + \boldsymbol{\epsilon}_j = \sum_{\ell=1}^{p} \boldsymbol{A}_{j\ell} \boldsymbol{\epsilon}_\ell + \boldsymbol{\epsilon}_j, \ \forall j = 1, \ldots, p,$$

where $\boldsymbol{B}_{j\ell} \neq \boldsymbol{O}$ if and only if $\ell \in pa(j)$ and $\boldsymbol{A}_{j\ell} \neq \boldsymbol{O}$ if and only if $\ell \in an(j)$, with

$$\boldsymbol{A}_{j\ell} = \sum_{\pi \in ph(\ell \to j)} \prod_{(u,v) \in E(\pi)} \boldsymbol{B}_{vu}.$$

It can be seen that

$$\begin{aligned} \sum_{k=1}^{\infty} S_{jk} Z_{jk} &= \sum_{k=1}^{\infty} S_{jk} \sum_{\ell=1}^{p} \sum_{h=1}^{\infty} A_{j\ell}(k, h) \epsilon_{\ell h} + \sum_{k=1}^{\infty} S_{jk} \epsilon_{jk} \\ &= \sum_{\ell=1}^{p} \sum_{h=1}^{\infty} \left(\sum_{k=1}^{\infty} S_{jk} A_{j\ell}(k, h) \right) \epsilon_{\ell h} + \sum_{k=1}^{\infty} S_{jk} \epsilon_{jk}, \end{aligned}$$

for any choice of $S_{jk} = \pm 1$, where $\epsilon_{\ell h}$ and ϵ_{jk} are independent random variables. By the

Kolmogorov convergence criterion (Resnick, 2019), if

$$\sum_{\ell=1}^{p}\sum_{h=1}^{\infty}\left(\sum_{k=1}^{\infty}S_{jk}A_{j\ell}(k,h)\right)^2\mathbb{E}(\epsilon_{\ell h}^2)+\sum_{k=1}^{\infty}\mathbb{E}(\epsilon_{jk}^2)<\infty,$$

for any choice of $S_{jk}=\pm 1$, then $\sum_{k=1}^{\infty}\pm Z_{jk}$ converges almost surely, which gives a condition under which the infinite sequence is square summable. In practice, we take a finite truncated model, where the summation is on finite terms (which amounts to assuming $Z_{jk}=0$ for $k>K$), and hence convergence is always guaranteed if the coefficients are finite and second moments of exogenous variables exist. Since we used proper priors (normal and inverse-gamma) for these parameters, the condition is satisfied almost surely.

F. The Effect of Finite Truncation

Consider the infinite-dimensional functional object $Y_j=\sum_{k=1}^{\infty}Z_{jk}\phi_{jk}$. Denote $\boldsymbol{Z}_j=(Z_{j1},\ldots,Z_{jk},\ldots)$ whose causal relationships are modeled by the linear non-Gaussian Bayesian network

$$\boldsymbol{Z}_j=\sum_{\ell=1}^{p}\boldsymbol{B}_{j\ell}\boldsymbol{Z}_\ell+\boldsymbol{\epsilon}_j,\ \forall j=1,\ldots,p,$$

where $\boldsymbol{B}_{j\ell}=[B_{j\ell}(k,h)]\neq\boldsymbol{O}$ if and only if $\ell\in pa(j)$. We assume that for each j, only a finite number K_j of basis coefficients contribute to the causal relationships and, without loss of generality, they are the first K_j basis coefficients, i.e., $B_{j\ell}(k,h)=0$ for any $k>K_j$ or $h>K_\ell$. Under this assumption, we can write the joint distribution of $\boldsymbol{Z}=(\boldsymbol{Z}_1,\ldots,\boldsymbol{Z}_p)$ as

$$P(\boldsymbol{Z})=P(\boldsymbol{Z}_1,\ldots,\boldsymbol{Z}_p)=P(\boldsymbol{Z}_1^+,\ldots,\boldsymbol{Z}_p^+)\prod_{j=1}^{p}P(\boldsymbol{Z}_j^-),$$

where $\boldsymbol{Z}_j^+=(Z_{j1},\ldots,Z_{jK_j})$ and $\boldsymbol{Z}_j^-=(Z_{j,K_j+1},Z_{j,K_j+2},\ldots)$. The full causal relationships are captured by $\boldsymbol{Z}^+=(\boldsymbol{Z}_1^+,\ldots,\boldsymbol{Z}_p^+)$ with the linear non-Gaussian Bayesian network,

$$\boldsymbol{Z}_j^+=\sum_{\ell=1}^{p}\boldsymbol{B}_{j\ell}^+\boldsymbol{Z}_\ell^++\boldsymbol{\epsilon}_j^+,\ \forall j=1,\ldots,p,$$

where $\boldsymbol{B}_{j\ell}^+$ is a $K_j\times K_\ell$ dimensional matrix.

In practice, we may not know K_j. Suppose we choose $\widetilde{K}_j$ basis functions where $\widetilde{K}_j>K_j$

for all j. Denote $\widetilde{\boldsymbol{Z}} = (\widetilde{\boldsymbol{Z}}_1, \ldots, \widetilde{\boldsymbol{Z}}_p)$ with $\widetilde{\boldsymbol{Z}}_j = (Z_{j1}, \ldots, Z_{j\widetilde{K}_j})$. The joint distribution of $\widetilde{\boldsymbol{Z}}$ writes

$$P(\widetilde{\boldsymbol{Z}}) = P(\widetilde{\boldsymbol{Z}}_1, \ldots, \widetilde{\boldsymbol{Z}}_p) = P(\boldsymbol{Z}_1^+, \ldots, \boldsymbol{Z}_p^+) \prod_{j=1}^{p} P(\widetilde{\boldsymbol{Z}}_j^-),$$

where $\widetilde{\boldsymbol{Z}}_j^- = (Z_{j,K_j+1}, \ldots, Z_{j\widetilde{K}_j})$, which induces the same marginal distribution on $\boldsymbol{Z}^+$. Therefore, from our identifiability theorem, the causal structure is identifiable from the induced marginal distribution $P(\boldsymbol{Z}_1^+, \ldots, \boldsymbol{Z}_p^+)$.

However, if we choose a smaller $\widetilde{K}_j$ number of basis functions where such that $\widetilde{K}_j < K_j$, our causal identifiability theorem does not apply anymore. To see this, let $\widetilde{\boldsymbol{Z}}_j^+ = (Z_{j,\widetilde{K}_j+1}, \ldots, Z_{jK_j})$, we have

$$P(\widetilde{\boldsymbol{Z}}) = P(\widetilde{\boldsymbol{Z}}_1, \ldots, \widetilde{\boldsymbol{Z}}_p) = \int P(\boldsymbol{Z}_1^+, \ldots, \boldsymbol{Z}_p^+)\ d(\widetilde{\boldsymbol{Z}}_1^+, \ldots, \widetilde{\boldsymbol{Z}}_p^+).$$

Since Bayesian networks are not closed under marginalization (Richardson and Spirtes, 2002), the causal identification problem becomes substantially more challenging even for Gaussian non-functional models. Hence, we instead perform simulations to test the robustness of our model with respect to K_j. We found that if we under-specify K_j, both false positives and false negatives can occur whereas an over-specification of K_j has much smaller effects. Therefore, according to the above discussions, we suggest choosing a relatively large K_j in practice.

[Table 4 about here.]

References

Bhattacharya, A. and Dunson, D. B. (2011). Sparse Bayesian infinite factor models. *Biometrika* **98,** 291–306.

Garibaldi, B. T., Wang, K., Robinson, M. L., Zeger, S. L., Bandeen-Roche, K., Wang, M.-C., et al. (2021). Comparison of time to clinical improvement with vs without remdesivir

treatment in hospitalized patients with COVID-19. *JAMA Network Open* **4,** e213071–e213071.

Ignatius, E. H., Wang, K., Karaba, A., Robinson, M., Avery, R. K., Blair, P., et al. (2021). Tocilizumab for the treatment of COVID-19 among hospitalized patients: A matched retrospective cohort analysis. In *Open Forum Infectious Diseases*, volume 8, page ofaa598.

Kowal, D. R., Matteson, D. S., and Ruppert, D. (2017). A Bayesian multivariate functional dynamic linear model. *Journal of the American Statistical Association* **112,** 733–744.

Legramanti, S., Durante, D., and Dunson, D. B. (2020). Bayesian cumulative shrinkage for infinite factorizations. *Biometrika* **107,** 745–752.

Resnick, S. (2019). *A Probability Path*. Springer.

Richardson, T. and Spirtes, P. (2002). Ancestral graph Markov models. *The Annals of Statistics* **30,** 962–1030.

Shojaie, A. and Michailidis, G. (2010). Penalized likelihood methods for estimation of sparse high-dimensional directed acyclic graphs. *Biometrika* **97,** 519–538.

Yakowitz, S. J. and Spragins, J. D. (1968). On the identifiability of finite mixtures. *The Annals of Mathematical Statistics* **39,** 209–214.

Table S.1: Functions observed on an unevenly spaced grid. Average operating characteristics based on 50 repetitions are reported; standard deviations are given within the parentheses.

p	n	FLiNG-BN			FGLASSO			FPCA-LiNGAM			FPCA-PC		
		TPR	FDR	MCC	TPR	FDR	MCC	TPR	FDR	MCC	TPR	FDR	MCC
20	500	0.94 (0.04)	0.29 (0.06)	0.80 (0.04)	0.59 (0.06)	0.73 (0.03)	0.18 (0.03)	0.71 (0.04)	0.79 (0.02)	0.33 (0.03)	0.29 (0.04)	0.81 (0.02)	0.22 (0.02)
20	800	0.92 (0.05)	0.14 (0.05)	0.89 (0.03)	0.61 (0.05)	0.68 (0.04)	0.21 (0.03)	0.79 (0.03)	0.73 (0.02)	0.36 (0.03)	0.36 (0.05)	0.80 (0.04)	0.27 (0.04)
50	500	0.75 (0.02)	0.44 (0.04)	0.65 (0.03)	0.68 (0.06)	0.87 (0.07)	0.13 (0.04)	0.84 (0.02)	0.86 (0.04)	0.27 (0.03)	0.32 (0.03)	0.85 (0.04)	0.22 (0.03)
50	800	0.87 (0.03)	0.33 (0.02)	0.79 (0.05)	0.69 (0.06)	0.86 (0.06)	0.15 (0.04)	0.92 (0.03)	0.78 (0.04)	0.33 (0.03)	0.39 (0.05)	0.73 (0.02)	0.28 (0.03)

Table S.2: Sensitivity analysis with respect to the choices of hyperparameters. Average operating characteristics based on 50 repetitions are reported; standard deviations are given within the parentheses.

Parameter	$L = 5$	$M = 3$	$\alpha = 0.1$	$a_r, b_r = 0.5, 0.5$	$a_\gamma, b_\gamma = 0.1, 0.1$	$a_\tau, b_\tau = 0.1, 0.1$	$a_\sigma, b_\sigma = 0.1, 0.1$
TPR	0.643 (0.028)	0.731 (0.041)	0.760 (0.030)	0.720 (0.038)	0.720 (0.024)	0.735 (0.021)	0.709 (0.029)
FDR	0.130 (0.037)	0.234 (0.025)	0.222 (0.056)	0.237 (0.050)	0.204 (0.070)	0.213 (0.035)	0.251 (0.092)
MCC	0.735 (0.028)	0.744 (0.013)	0.759 (0.020)	0.737 (0.038)	0.752 (0.029)	0.757 (0.025)	0.723 (0.048)
Parameter	$L = 10$	$M = 7$	$\alpha = 0.5$	$a_r, b_r = 0.2, 0.8$	$a_\gamma, b_\gamma = 1, 10$	$a_\tau, b_\tau = 0.1, 1$	$a_\sigma, b_\sigma = 0.01, 0.01$
TPR	0.720 (0.017)	0.716 (0.049)	0.763 (0.024)	0.716 (0.016)	0.705 (0.041)	0.735 (0.016)	0.705 (0.014)
FDR	0.140 (0.032)	0.226 (0.047)	0.161 (0.036)	0.217 (0.038)	0.198 (0.064)	0.183 (0.048)	0.133 (0.102)
MCC	0.784 (0.019)	0.736 (0.019)	0.763 (0.024)	0.745 (0.021)	0.748 (0.042)	0.771 (0.019)	0.791 (0.050)
Parameter	$L = 15$	$M = 10$	$\alpha = 2.0$	$a_r, b_r = 0.9, 0.1$	$a_\gamma, b_\gamma = 1, 0.1$	$a_\tau, b_\tau = 10, 1$	$a_\sigma, b_\sigma = 0.001, 0.001$
TPR	0.714 (0.025)	0.702 (0.024)	0.702 (0.020)	0.727 (0.018)	0.718 (0.021)	0.731 (0.008)	0.782 (0.009)
FDR	0.173 (0.033)	0.251 (0.068)	0.205 (0.039)	0.250 (0.069)	0.199 (0.075)	0.279 (0.048)	0.139 (0.021)
MCC	0.765 (0.023)	0.721 (0.043)	0.743 (0.030)	0.734 (0.031)	0.754 (0.030)	0.721 (0.023)	0.824 (0.013)
Parameter	$L = 30$	$M = 20$	$\alpha = 5.0$	$a_r, b_r = 0.1, 0.1$	$a_\gamma, b_\gamma = 5, 25$	$a_\tau, b_\tau = 25, 5$	$a_\sigma, b_\sigma = 1, 1$
TPR	0.668 (0.036)	0.691 (0.049)	0.716 (0.009)	0.706 (0.015)	0.716 (0.016)	0.731 (0.015)	0.773 (0.015)
FDR	0.309 (0.075)	0.237 (0.031)	0.232 (0.043)	0.223 (0.054)	0.168 (0.034)	0.349 (0.037)	0.124 (0.034)
MCC	0.684 (0.043)	0.722 (0.038)	0.738 (0.022)	0.736 (0.021)	0.769 (0.022)	0.682 (0.021)	0.811 (0.021)

Table S.3: Medians and IQRs of the number of measurements per patient.

Name	Median (IQR)	Name	Median (IQR)	Name	Median (IQR)
Respiratory Rate	9 (6 to 11)	SpO2	9 (6 to 11)	GFR	9 (6 to 12)
FiO2	8 (4 to 12)	SBP	9 (6 to 11)	Lymphocyte	9 (7 to 13)
DBP	9 (6 to 11)	Creatinine	10 (7 to 13)	WBC	7 (4 to 11)

Table S.4: Sensitivity analysis with respect to the choice of K. Average operating characteristics based on 50 repetitions are reported; standard deviations are given in parentheses.

K	Adaptive	2	4	5	6	8	12
TPR	0.75 (0.04)	0.57 (0.03)	0.63 (0.05)	0.72 (0.02)	0.49 (0.06)	0.35 (0.04)	0.27 (0.05)
FDR	0.26 (0.03)	0.23 (0.03)	0.29 (0.03)	0.31 (0.04)	0.34 (0.12)	0.52 (0.03)	0.74 (0.08)
MCC	0.74 (0.03)	0.62 (0.02)	0.66 (0.03)	0.69 (0.03)	0.46 (0.09)	0.28 (0.03)	0.19 (0.04)