Descubrimiento causal federado mediante cumulantes

Pablo Torrijos[1], Fabio Stella[2], José A. Gámez[1], and José M. Puerta[1]

[1] Departamento de Sistemas Informáticos, Universidad de Castilla-La Mancha, Albacete, 02071, España.
[2] Dipartimento di Informatica, Sistemistica e Comunicazione, Università degli Studi di Milano Bicocca, Milano, 20126, Italia
{Pablo.Torrijos,Jose.Gamez,Jose.Puerta}@uclm.es,Fabio.Stella@unimib.it

Resumen En este artículo estudiamos los modelos lineales acíclicos no gaussianos (LiNGAM) en entornos federados, los cuales permiten ir más allá de la equivalencia de Markov. En muchos dominios los datos escasean, y centralizar los de distintos clientes casi nunca es viable por regulaciones como el Reglamento General de Protección de Datos (RGPD), lo que hace que la federación ofrezca un punto intermedio entre privacidad y precisión. El problema es que DirectLiNGAM, el modelo centralizado de referencia, no se puede federar de forma directa. Para sortear esto, podemos utilizar cumulantes de orden superior: dependen solo de la distribución conjunta de las variables implicadas y son aditivos sobre muestras independientes, de modo que basta una ronda de comunicación independientemente de la partición (horizontal, vertical o híbrida). Sin embargo, FedISHC, el único método federado por esta vía, falla cuando el ruido es casi simétrico. Presentamos la familia FedRCD, con tres variantes que balancean rondas de comunicación y ruido algebraico; dos son las contrapartes federadas exactas de *high-order cumulant* (HC) y HC-LiNGAM. Las variantes que realizan una sola ronda admiten desaprendizaje exacto a cualquier granularidad, desde una observación hasta un cliente entero. Mostramos además que, a tamaños muestrales reales, toda la familia basada en cumulantes no ordena las variables por la asimetría poblacional que los *scores* codifican en el límite, sino por una escalera de varianzas inducida por el grafo acíclico dirigido (DAG): la contraparte en cumulantes de la *varsortability*. La estandarización marginal la colapsa a un orden casi aleatorio, mientras que DirectLiNGAM, invariante a escala, no se ve afectado.

Keywords: Descubrimiento causal federado · federación horizontal, vertical e híbrida · LiNGAM· desaprendizaje federado.

1. Introducción

El aprendizaje estructural causal a partir de datos observacionales tiene aplicaciones en genómica [12], señalización de proteínas [13] o econometría [6]. En algunos dominios, sobre todo en salud, los datos escasean y la única forma

de aumentar el tamaño muestral es combinar cohortes de pacientes de varios hospitales o centros de investigación. Esto mejora la calidad del aprendizaje causal, pero plantea serios problemas de privacidad: regulaciones como el RGPD restringen el envío de datos personales en bruto fuera de la institución que los recogió. El aprendizaje federado [5,14] alivia estas restricciones: los clientes conservan los datos en local e intercambian solo parámetros o estadísticos agregados, normalmente a través de un servidor central. El escenario federado se divide en tres regímenes de partición [14]: *horizontal* (mismas variables, distintas muestras), *vertical* (mismas muestras, distintas variables) e *híbrido*.

Algunos enfoques estándar de descubrimiento causal han sido adaptados al entorno federado. Los métodos basados en restricciones, como las variantes federadas de PC [4], dependen de tests de independencia condicional sobre el conjunto condicionante completo; los de optimización continua [7] dependen de funciones de *score* algebraicas. Ambas familias recuperan solo clases de equivalencia de Markov, y ambas sufren en los regímenes vertical e híbrido: cuando las variables de un test de independencia condicional están repartidas entre clientes, el test no se puede evaluar sin agrupar datos. El marco LiNGAM [9,10] permite ir más allá de la equivalencia de Markov y recuperar un orden causal bajo el supuesto de ruido exógeno no gaussiano. DirectLiNGAM [10], modelo referente en LiNGAM, se apoya en tests de independencia no paramétricos sobre datos centralizados difíciles de federar de forma directa. Métodos centralizados recientes [3] sustituyen esos tests por *scores* por pares en forma cerrada basados en cumulantes de orden superior. Estos son federables de manera natural: los cumulantes dependen solo de la distribución conjunta de las variables implicadas [1], de modo que la ausencia de variables en un cliente no sesga las entradas que ese cliente sí puede calcular (lo que cubre la partición vertical), y son aditivos sobre grupos de muestras independientes [11] (lo que cubre la horizontal). El protocolo es idéntico en los tres regímenes, se ejecuta en una sola ronda y admite desaprendizaje federado exacto mediante la resta de los momentos en bruto correspondientes.

El único método federado por esta vía hasta ahora, FedISHC [2], instancia el protocolo con cumulantes de tercer orden y ejecuta deflación secuencial en el servidor. Su límite con ruido simétrico es esperable: tanto su *score* de identificación como su coeficiente de deflación dividen por la asimetría de la fuente candidata, nula para distribuciones simétricas (Sección 2.4). Arreglarlo exige algo más que pasar de un *score* de tercer orden a uno de cuarto, porque FedISHC solo actualiza *arrays* de tercer orden y un *score* de cuarto orden sobre ellos queda invariante. Dado que la identificación y la deflación deben rediseñarse de forma conjunta, presentamos la familia FedRCD, que combina un criterio de identificación basado en momentos de cuarto orden con un coeficiente de deflación estable, definido a partir de covarianzas y diseñado para evitar valores cercanos a cero.

Cabe además otra pregunta: con tamaños muestrales realistas, ¿qué domina la ordenación? La señal realmente no es la asimetría que los *scores* codifican, que se anula en las fuentes, sino una escalera de varianzas inducida por el DAG, que puede interpretarse como la contraparte en cumulantes de la *varsortability* [8]. La estandarización marginal la elimina y colapsa todos los métodos de cumulantes,

centralizados o federados. Mientras, métodos como DirectLiNGAM que son invariantes a escala no se ven afectados (Sección 4.5).

Las contribuciones de este trabajo son: 1) formalizamos el límite con ruido simétrico de FedISHC y explicamos por qué un *score* de cuarto orden por sí solo no lo repara; 2) presentamos la familia FedRCD, que combina un criterio de fuente de cuarto orden con un coeficiente de deflación estable basado en covarianzas, con tres variantes balanceando rondas de comunicación frente a ruido algebraico; 3) mostramos que toda la familia basada en cumulantes ordena las variables por una escalera de varianzas inducida por el DAG, lo que la sitúa en el régimen dependiente de escala que [8] identificó para los métodos continuos basados en MSE; y 4) aportamos evidencia empírica en DAGs de Erdős-Rényi bajo once familias de ruido distintas.

El resto del artículo se organiza de la siguiente forma: la Sección 2 repasa LiNGAM, la identificación de cuarto orden y el escenario federado. La Sección 3 desarrolla la familia FedRCD. La Sección 4 presenta la evaluación empírica, incluido el diagnóstico de estratificación. La Sección 5 contiene las conclusiones.

2. Antecedentes y formulación del problema

2.1. El modelo LiNGAM

Sea $\mathbf{X} = (x_1, \ldots, x_p)^\top \in \mathbb{R}^p$ un vector de variables observadas. El marco LiNGAM [9,10] modela su estructura causal como $\mathbf{X} = \mathbf{B}\mathbf{X} + \mathbf{E}$, donde $\mathbf{B} \in \mathbb{R}^{p\times p}$ es una matriz de coeficientes causales que puede permutarse de forma estricta a triangular inferior, y $\mathbf{E} = (e_1, \ldots, e_p)^\top$ es un vector de perturbaciones mutuamente independientes y no gaussianas con $\mathbb{E}[e_i] = 0$ y $\mathbb{E}[e_i^2] = \sigma_i^2 > 0$. Despejando $\mathbf{X}$ se obtiene la representación lineal (mezcla) $\mathbf{X} = \mathbf{A}\mathbf{E}$ con $\mathbf{A} = (\mathbf{I} - \mathbf{B})^{-1}$, donde $\mathbf{I}$ es la identidad $p \times p$. LiNGAM ofrece una garantía más fuerte que los métodos basados en restricciones: siempre que como máximo una perturbación sea gaussiana, $\mathbf{B}$ queda identificada de forma única a partir de la distribución conjunta de $\mathbf{X}$; es decir, se recupera el DAG completo y no solo su clase de equivalencia de Markov [9]. Para ello se requieren las siguientes suposiciones: linealidad, aciclicidad, independencia mutua del ruido y ausencia de *confunders* no observados (variables ocultas que influyen causalmente en dos o más variables observadas); se admite ruido heterocedástico. Sus dos estimadores clásicos, ICA-LiNGAM [9] (vía análisis de componentes independientes, ICA, sobre la matriz de mezcla) y DirectLiNGAM [10] (vía regresión y tests de independencia no paramétricos entre fuentes candidatas y residuos), logran buena recuperación estructural pero son complicados de federar exactamente de forma directa: ICA opera sobre los datos en bruto y los tests con kernel necesitan acceso conjunto a las variables que comparan.

2.2. Cumulantes de orden superior e identificación de fuentes

El cumulante marginal de orden m de x_i se denota $\kappa_m(x_i)$ [1], y $\kappa_{m_1,m_2}(x_i, x_j)$ denota el cumulante conjunto con m_1 copias de x_i y m_2 de x_j. Usamos $\kappa_\bullet$ para

cumulantes poblacionales y $\hat{\kappa}_\bullet$ para las estimaciones empíricas; los casos propios κ_3, κ_4 y conjuntos $\kappa_{1,2}, \kappa_{2,1}, \kappa_{1,3}, \kappa_{2,2}, \kappa_{3,1}$ son instancias de esta notación. Los cumulantes son multilineales, aditivos para variables independientes y se anulan en orden ≥ 3 para variables gaussianas. La no gaussianidad rompe la simetría direccional de los cumulantes conjuntos y aporta la señal estadística que distingue causa de efecto. El *score* de asimetría por pares de cuarto orden [3] es

$$\tau_{ij} = \left|\kappa_4(x_i)\,\kappa_{1,3}(x_i, x_j) - \kappa_{2,2}(x_i, x_j)\,\kappa_{3,1}(x_i, x_j)\right|. \tag{1}$$

En el límite poblacional, $\tau_{ij} = 0$ si y solo si x_i es ancestro de x_j (o las dos variables son independientes), y $\tau_{ij} > 0$ en otro caso [3, Teoremas 2–3]. Un *nodo fuente* del DAG es una variable sin aristas entrantes; x_s es fuente si y solo si $\sum_{j \neq s} \tau_{sj} = 0$. La fuente en cada paso de un algoritmo de ordenación se obtiene resolviendo

$$s = \arg\min_{i \in U} \sum_{j \in U,\, j \neq i} \tau_{ij}, \tag{2}$$

donde $U \subseteq \{1, \ldots, p\}$ recoge los índices de las variables aún no colocadas en el orden parcial (el *conjunto activo*). HC [3] identifica cada fuente mediante (2), deflación por mínimos cuadrados ordinarios (OLS) ($x_j \leftarrow x_j - \hat{\beta}_{js}\, x_s$, $\hat{\beta}_{js} = \hat{\Sigma}_{js}/\hat{\Sigma}_{ss}$) y repite, con coste total $O(np^3)$. HC-LiNGAM [3] calcula la matriz τ una vez y ordena globalmente por sumas por filas $T_\tau(x_i) = \sum_{j \neq i} \tau_{ij}$, con coste $O(np^2)$. Ambos son referencias centralizadas para los métodos de la Sección 3.

2.3. Descubrimiento causal federado

Los datos están repartidos entre K clientes que no pueden compartir observaciones en bruto. El cliente k posee el conjunto $\mathcal{D}_k$ sobre el subconjunto de variables $\mathbf{X}_k \subseteq \mathbf{X}$ con n_k muestras, donde $\mathbf{X} = \bigcup_{k=1}^{K} \mathbf{X}_k$ y $N = \sum_{k=1}^{K} n_k$. Siguiendo a [2], exigimos que para cada par (x_i, x_j) al menos un cliente observe ambas variables. La condición es más laxa que pedir un conjunto común completo de variables y es lo que permite cubrir todo el tensor de cumulantes conjuntos; las lagunas de cobertura por pares dejarían las entradas correspondientes no identificables.

Cada cumulante conjunto $\kappa_{m_1,m_2}(x_i, x_j)$ es función de la distribución bivariante de (x_i, x_j), así que los clientes que no observan ambas variables no aportan a esa entrada. Como los cumulantes son aditivos sobre muestras independientes [11], el tensor global se recupera de forma exacta a partir de las sumas de momentos en bruto por cliente, sea la partición horizontal, vertical o híbrida. Cada cliente transmite $O(p_k^2)$ valores en coma flotante con $p_k = |\mathbf{X}_k|$ y el servidor compone estimaciones numéricamente idénticas al cálculo centralizado. El mismo mecanismo descarta la federación exacta de cualquier método que requiera residuos o tests de independencia con kernel sobre datos en bruto.

FedISHC. FedISHC [2] agrega cumulantes de tercer orden en una sola ronda y ejecuta deflación secuencial en el servidor. Identifica las fuentes con el *score* de tercer orden $\tau_{ij}^{(3)} = \left|\kappa_3(x_i)\,\kappa_{1,2}(x_i, x_j) - \kappa_{2,1}(x_i, x_j)\,\kappa_{1,2}(x_j, x_i)\right|$. La

influencia causal de una fuente identificada x_s sobre cada x_j se estima como $\hat{\alpha}_{js} = \hat{\kappa}_{2,1}(x_j, x_s)/\hat{\kappa}_3(x_s)$, y los *arrays* de tercer orden se actualizan mediante $\hat{\kappa}_3(x_j)' = \hat{\kappa}_3(x_j) - \hat{\alpha}_{js}^3\,\hat{\kappa}_3(x_s)$. Tanto el *score* como el coeficiente llevan $\kappa_3(x_s)$ en el denominador, lo que la Sección 2.4 explota para formalizar el límite. En [2] también se introduce FedHC, una variante sin deflación que usa $\tau^{(3)}$ y ordena las variables por sumas por filas en una sola pasada. FedHC hereda la debilidad de $\tau^{(3)}$ con ruido simétrico, pero la ausencia de deflación evita la acumulación de error a lo largo de los $p-1$ pasos.

2.4. Limitaciones de FedISHC con ruido simétrico

Con ruido simétrico el tercer cumulante propio $\kappa_3(x_s)$ se anula en la población y, con muestras finitas, queda dominado por la fluctuación muestral. FedISHC lo coloca en el denominador tanto de su *score* $\tau^{(3)}$ como de su coeficiente de deflación $\hat{\alpha}_{js}$, de modo que ambos se vuelven inestables a medida que $\kappa_3(x_s) \to 0$: el problema no está en el muestreo, sino en dividir por una cantidad que tiende a cero.

El *score* de cuarto orden $\tau^{(4)}$ de (1) evita el problema, ya que $\kappa_4(x_s) \neq 0$ para toda distribución no gaussiana estándar. Pero cambiar solo el *score* no basta. Sea FedISHC($\tau^{(4)}$) el algoritmo que identifica con $\tau^{(4)}$ pero conserva la deflación de tercer orden de FedISHC, que solo actualiza los *arrays* κ_3. Como $\tau^{(4)}$ depende de κ_4 y de $\kappa_{1,3}, \kappa_{2,2}, \kappa_{3,1}$, y ninguno de ellos se modifica en esa actualización, cada iteración evalúa $\tau^{(4)}$ sobre la misma matriz restringida al conjunto activo. Hay que reparar a la vez identificación y deflación.

3. La familia FedRCD

Proponemos FedRCD (*Federated Regression-Directed Cumulants*), una familia LiNGAM federada que aborda los dos problemas indentificados en Sec. 2.4.

3.1. Coeficiente de deflación OLS

La inestabilidad de $\hat{\alpha}_{js}$ procede de su denominador $\hat{\kappa}_3(x_s)$. El método de mínimos cuadrados ordinarios (OLS) ofrece un reemplazo con denominador acotado: cuando x_s es la fuente actual no tiene padres bajo LiNGAM, de modo que $\Sigma_{js} = b_{js}\Sigma_{ss}$ exactamente y el coeficiente OLS $\hat{\beta}_{js} = \hat{\Sigma}_{js}/\hat{\Sigma}_{ss}$ es consistente para el mismo parámetro estructural b_{js} que persigue $\hat{\alpha}_{js}$. Su denominador $\Sigma_{ss} = \mathrm{Var}(x_s)$ es estrictamente positivo para cualquier variable no degenerada, sea cual sea el ruido. La familia FedRCD usa $\tau^{(4)}$ para identificar y $\hat{\beta}_{js}$ donde hace falta deflación. Hay una segunda consecuencia, visible en la Sección 4.5: al no inyectar ruido divergente en los *arrays*, la deflación preserva el orden por profundidad de varianzas y cumulantes que induce el DAG.

3.2. Las variantes FedRCD-(NI/Def/X)

Las tres variantes (Algoritmo 1) comparten el criterio $\tau^{(4)}$, el coeficiente $\hat{\beta}_{js}$ y el protocolo de agregación en cliente. Se diferencian en dónde ocurre la deflación

y cuántas rondas exige. Las tres se aplican a federación horizontal, vertical e híbrida, y recuperan $\hat{\mathbf{B}}$ a partir de la $\hat{\boldsymbol{\Sigma}}$ previa a la deflación mediante Lasso adaptativo con $\lambda = 0.01\bar{\sigma}$ fijo, una vez determinado el orden.

Algoritmo 1 Familia FedRCD

Entrada: Estadísticos globales $\hat{\kappa}_\bullet, \hat{\boldsymbol{\Sigma}}$ agregados de los clientes; $U = \{1, \dots, p\}$; **modo** $\in \{\text{NI}, \text{Def}, \text{X}\}$
Salida: Orden causal π; matriz $\hat{\mathbf{B}}$

$\pi \leftarrow [\,]$; $\hat{\boldsymbol{\Sigma}}_0 \leftarrow \hat{\boldsymbol{\Sigma}}$ ▷ Guardar covarianza original para $\hat{\mathbf{B}}$
si modo = NI **entonces**
  Calcular $\tau_{ij}^{(4)}$ para todo $i \neq j$ con (1)
  $\pi \leftarrow \operatorname{argsort}\big(\sum_{j \neq i} \tau_{ij}^{(4)}\big)_{i=1}^{p}$ ▷ Una pasada, sin deflación
si no
  mientras $|U| > 1$ **hacer**
    Calcular $\tau_{ij}^{(4)}$ para $i, j \in U$ con (1)
    $s \leftarrow \arg\min_{i \in U} \sum_{j \in U,\, j \neq i} \tau_{ij}^{(4)}$ ▷ Identificar fuente
    Añadir s a π
    $\hat{\beta}_{js} \leftarrow \hat{\Sigma}_{js} / \hat{\Sigma}_{ss}$ para cada $j \in U \setminus \{s\}$ ▷ Coeficiente OLS
    si modo = Def **entonces** actualizar $\hat{\kappa}_\bullet, \hat{\boldsymbol{\Sigma}}$ en el servidor (forma cerrada)
    si no si modo = X **entonces**
      El servidor envía $(s, \hat{\boldsymbol{\beta}})$ a los clientes
      Clientes: $x_j \leftarrow x_j - \hat{\beta}_{js}\, x_s$ para $j \in U \setminus \{s\}$; eliminar x_s
      Los clientes envían momentos nuevos; el servidor reagrega $\hat{\kappa}_\bullet, \hat{\boldsymbol{\Sigma}}$
    fin si
    $U \leftarrow U \setminus \{s\}$
  fin mientras
  Añadir el elemento restante de U a π
fin si
$\hat{\mathbf{B}} \leftarrow \text{AdaptiveLasso}(\hat{\boldsymbol{\Sigma}}_0, \pi)$ ▷ Pesos desde la $\hat{\boldsymbol{\Sigma}}$ original

FedRCD-NI calcula $\tau^{(4)}$ una vez y ordena por las sumas por filas $T_\tau(x_i) = \sum_{j \neq i} \tau_{ij}$ en una sola pasada, con coste $O(p^2)$ en el servidor. No hay deflación: por el Teorema 4 de [3], $T_\tau(x_i) < T_\tau(x_j)$ en la población siempre que x_i sea predecesor de x_j, así que la ordenación recupera el orden verdadero (con muestras finitas y grafos densos). FedRCD-NI es la contraparte federada exacta de HC-LiNGAM [3] y admite desaprendizaje federado exacto a nivel de instancia sin necesidad de reentrenar el modelo: como los momentos en bruto son aditivos, también son sustraíbles, así que el servidor resta la contribución del subconjunto a olvidar (una observación, una cohorte o un cliente entero) y reescala por $(N - n_{\text{rm}})^{-1}$, recuperando los estadísticos que se habrían obtenido sin esas muestras. En los regímenes vertical e híbrido se aplica par a par con recuentos N_{ij}.

FedRCD-Def añade deflación algebraica en el servidor tras eliminar cada fuente. La multilinealidad de los cumulantes aplicada al residuo $r_j = x_j - \beta_{js}\, x_s$ da actualizaciones en forma cerrada de todos los *arrays* agregados (extendiendo

el Lema 2 en [2] de tercer a cuarto orden y de $\hat{\alpha}$ a $\hat{\beta}$). El servidor las aplica en una sola ronda con coste total $O(p^3)$. El coeficiente de deflación es estable y las actualizaciones son exactas a nivel poblacional. Con muestras finitas, sin embargo, cada paso algebraico inyecta ruido de estimación que se acumula a lo largo de las $p-1$ actualizaciones. El desaprendizaje exacto se preserva porque el protocolo sigue siendo un intercambio de una sola ronda.

FedRCD-X lleva la deflación a los clientes para evitar la acumulación algebraica. En cada una de las $p-1$ rondas, el servidor identifica la fuente con (2), difunde los coeficientes OLS, y cada cliente aplica deflación a sus datos locales, descarta la fuente identificada y devuelve estadísticos suficientes nuevos. Los cumulantes se recalculan a partir de residuos reales en cada ronda en lugar de aproximarse algebraicamente, y no se acumula error de aproximación. FedRCD-X es la contraparte federada de HC [3]. El precio son $p-1$ rondas con carga total $O(p^3)$. El desaprendizaje ya no es trivial, ya que cada ronda depende de coeficientes derivados del agregado anterior, por lo que eliminar un cliente a posteriori invalidaría todas las rondas siguientes.

4. Evaluación experimental

4.1. Configuración

Generación de datos. Muestreamos DAGs Erdős-Rényi ER-2 (densidad esperada 0.2) con pesos en las aristas en $U([-1.5,-0.5]\cup[0.5,1.5])$. El ruido exógeno procede de once familias continuas (Gaussiana al cubo, Pareto, Logística, Uniforme, Laplace, Poisson, Exponencial, t de Student, Gamma, Exponencial al cubo; más un caso Mixto en el que cada variable toma una familia distinta). Por defecto se utilizan $p=20$, $n=2000$ observaciones por cliente y $K=10$ clientes. El *grid* principal barre $p\in\{10,20,30,50\}$ y $K\in\{1,5,10,20,50\}$ bajo partición horizontal. Los resultados se promedian sobre 10 semillas aleatorias, y se calcula el error estándar de la media.

Métodos de referencia. Centralizados: DirectLiNGAM [10], HC y HC-LiNGAM [3]. Federados: FedISHC y su variante sin deflación FedHC [2]. FedHC es el análogo de tercer orden de FedRCD-NI: misma ordenación de una pasada, mismo coste $O(p^2)$ en servidor, pero identificación con $\tau^{(3)}$ en lugar de $\tau^{(4)}$. Para una comparación justa, todos los métodos recuperan los pesos de las aristas vía Lasso adaptativo sobre la $\hat{\boldsymbol{\Sigma}}$ agregada con $\lambda=0.01\bar{\sigma}$. FedRCD-NI y FedRCD-X producen las mismas ordenaciones que HC-LiNGAM y HC dados los mismos estadísticos agregados, por lo que no duplicamos los resultados.

Métricas. LiNGAM recupera principalmente un orden causal σ. Un F1-score por pares entre órdenes penaliza muchos válidos, porque compara la σ estimada con un único orden topológico arbitrario del DAG de referencia pese a que suelen existir varios válidos. En su lugar evaluamos contra el orden parcial inducido por el DAG verdadero $\mathcal{G}$. La clausura transitiva $\mathcal{G}_+$ de $\mathcal{G}$ es el conjunto de pares (x_i,x_j) para los que existe un camino dirigido de x_i a x_j, es decir, x_i es ancestro

de x_j. Cada relación ancestral $x_i \to x_j \in \mathcal{G}_+$ exige $x_i \prec_\sigma x_j$ en el orden estimado; los pares sin relación ancestral se excluyen. Definimos la *ancestral accuracy* como

$$\textsc{AncAcc}(\sigma, \mathcal{G}) = \frac{\left|\{(x_i, x_j) : x_i \to x_j \in \mathcal{G}_+,\ x_i \prec_\sigma x_j\}\right|}{|\mathcal{G}_+|}. \tag{3}$$

AncAcc premia cualquier orden consistente con $\mathcal{G}$. También reportamos la distancia de Hamming estructural (SHD) para medir error a nivel de aristas.

4.2. Reproducibilidad

Los algoritmos están implementados en Python 3.13 con el framework `Flower` para la orquestación federada. DirectLiNGAM usa la biblioteca `lingam`; HC, HC-LiNGAM, FedISHC y FedHC se reimplementaron desde cero al no haber código público disponible. El código y los scripts están en GitHub[3]. Los experimentos se ejecutan en un AMD Ryzen AI 9 HX 370 con 32 GB de RAM.

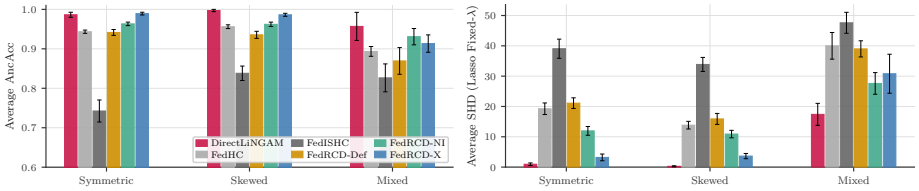
4.3. Robustez ante distintas distribuciones de ruido

La Figura 1 agrega AncAcc y SHD por régimen de ruido: *simétrico* (Logística, Uniforme, t de Student, Laplace), donde $\kappa_3 \approx 0$; *asimétrico* (el resto de familias unimodales) y *mixto*. La jerarquía es consistente en los regímenes simétrico y asimétrico. FedISHC queda claramente por detrás, más aún con ruido simétrico (como anticipa la Sección 2.4) mientras que DirectLiNGAM (centralizado) se acerca a la recuperación perfecta. Entre los métodos federados, FedRCD-X está muy cerca de DirectLiNGAM, seguido de FedRCD-NI, FedHC y FedRCD-Def. La brecha entre regímenes es menor de lo que se esperaría dado $\kappa_3 \to 0$: el rendimiento con muestras finitas se rige por el mecanismo de la Sección 4.5. El régimen mixto rompe el patrón ligeramente: FedISHC recupera algo de exactitud porque la asimetría media no nula aporta señal parcial pero sigue por detrás; los federados se agrupan más, FedRCD-NI supera a FedRCD-X por poco con FedHC cerca, y DirectLiNGAM en este caso no recupera el orden perfecto. Las conclusiones en SHD son idénticas a las obtenidas en AncAcc, lo que demuestra la importancia de obtener el orden correcto de las variables.

4.4. Estudio de ablación

Para aislar la fuente del déficit de FedISHC, implementamos FedISHC($\tau^{(4)}$), que reemplaza solo el criterio de identificación ($\tau^{(3)} \to \tau^{(4)}$) conservando la deflación original de tercer orden (Figura 2a, media sobre todo el ruido). AncAcc sube de 0.802 en FedISHC a 0.959 en FedISHC($\tau^{(4)}$), igualando a FedRCD-NI (0.960), que no hace deflación alguna: el cuello de botella en FedISHC es la identificación, no la deflación, y la deflación de tercer orden de FedISHC($\tau^{(4)}$) deja invariantes los tensores de cuarto orden. FedRCD-Def hace actualizaciones

[3] `github.com/ptorrijos99/FedLiNGAM`



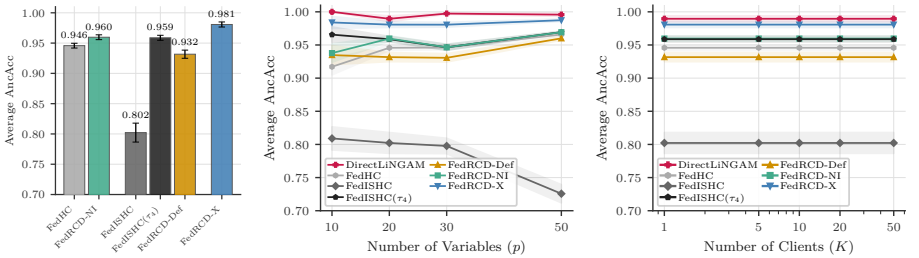
(a) AncAcc ↑ por régimen de ruido (b) SHD ↓ por régimen de ruido

Figura 1: Rendimiento por régimen de ruido (p=20, K=10).

genuinas de cuarto orden, pero la acumulación de ruido algebraico con muestras finitas limita sus ganancias. FedRCD-X, la única variante iterativa que recalcula los cumulantes a partir de residuos reales en cada ronda, es el mejor de media.

La Figura 2b muestra la AncAcc media frente a p con $K = 10$. DirectLiNGAM y FedRCD-X son los más estables. FedISHC($\tau^{(4)}$) es el mejor del resto en $p = 10$ pero converge con FedRCD-NI y FedHC al crecer p: como su deflación de tercer orden deja $\tau^{(4)}$ inalterado, difiere de una ordenación de una pasada solo por la restricción al conjunto activo, que ayuda con p pequeño pero queda dominada por el ruido muestral al crecer p.

La Figura 2c muestra la AncAcc frente a K con $p = 20$: todas las curvas son planas. La agregación de cumulantes no pierde información dado un $\mathcal{D}$ fijo, sea cual sea el reparto de los $\mathcal{D}_k$ (IID o no IID; horizontal, vertical o híbrido), por lo que las curvas se trasladan literalmente a los regímenes vertical e híbrido.



(a) FedISHC($\tau^{(4)}$) (p=20, K=10). (b) Escalabilidad (K=10). (c) Invariancia a la federación (p=20).

Figura 2: Ablación en FedISHC($\tau^{(4)}$), escalabilidad e invariancia a la federación.

4.5. ¿Qué lee en realidad la familia de cumulantes?

Las secciones anteriores defienden FedRCD por sus propios méritos: elimina una divergencia, estabiliza la deflación e iguala a las contrapartes centralizadas

de [3] dados los mismos estadísticos agregados. El diagnóstico que sigue plantea otra pregunta. El *score* $\tau^{(k)}$ es, por construcción, nulo en población en las fuentes verdaderas, y la identificación del orden descansa en sumas por filas de tensores de cumulantes. A los tamaños muestrales que realmente usamos, ¿con qué se alinea esa ordenación? En [8] se define la *varsortability* v como la fracción de caminos dirigidos a lo largo de los cuales la varianza marginal es monótona, y se reporta $v > 0.94$ en simulaciones genéricas de ruido aditivo. Ahí los métodos continuos basados en MSE (NOTEARS, GOLEM-EV, MSE-GDS) alcanzan buenos resultados sobre datos en bruto pero colapsan bajo una estandarización marginal. Mientras, algoritmos como DirectLiNGAM, PC y GES, con criterios invariantes a la escala, no se ven afectados.

El mismo mecanismo actúa sobre las magnitudes de los cumulantes. Bajo LiNGAM, los descendientes acumulan varianza de sus ancestros, $\mathrm{Var}(x_i)$ es monótona en la profundidad del DAG en promedio, y $\kappa_m(x_i)$ hereda esa escala: empíricamente se comporta como σ_i^m a primer orden. Las sumas por filas $\sum_j \tau_{ij}^{(k)}$ que guían la identificación quedan dominadas con N finito por esta escala monótona en profundidad y no por la asimetría poblacional, que se anula en las fuentes. Llamamos a esta propiedad *estratificación de varianza*, la contraparte en cumulantes de la *varsortability*. Para comprobarlo estandarizamos cada variable antes de agregar cumulantes ($\tilde{x}_i = x_i/\sigma_i$), un reescalado diagonal que preserva la identificabilidad de LiNGAM pero elimina la escalera, llevando v de 0.953 a 0.5. La Tabla 1 reporta AncAcc $\pm$ DE en ambos regímenes.

Tabla 1: AncAcc media sobre datos en bruto y estandarizados (Est.) marginalmente ($p = 20$, $K = 10$). La estandarización elimina la escalera de varianzas monótona en profundidad, reduciendo v a 0.5. Los valores por debajo de la base aleatoria de 0.5 indican inversión del orden verdadero.

	Simétrico		Asimétrico	
Método	Bruto (v=0.953)	Est. (v=0.5)	Bruto (v=0.953)	Est. (v=0.5)
DirectLiNGAM (K=1)	**0.996 ± 0.002**	**0.996 ± 0.002**	**0.998 ± 0.001**	**0.998 ± 0.001**
FedHC	0.937 ± 0.003	0.410 ± 0.014	0.933 ± 0.004	0.573 ± 0.016
FedISHC	0.769 ± 0.009	0.470 ± 0.013	0.839 ± 0.008	0.436 ± 0.016
FedISHC($\tau^{(4)}$)	0.940 ± 0.003	0.387 ± 0.013	0.942 ± 0.004	0.521 ± 0.020
FedRCD-Def	0.924 ± 0.005	0.308 ± 0.010	0.917 ± 0.007	0.358 ± 0.017
HC-LiNGAM (K=1) / FedRCD-NI	0.940 ± 0.003	0.303 ± 0.013	0.942 ± 0.004	0.441 ± 0.022
HC (K=1) / FedRCD-X	0.978 ± 0.003	0.368 ± 0.014	0.976 ± 0.004	0.470 ± 0.019

Todos los métodos basados en cumulantes colapsan bajo estandarización, en ambos regímenes, sea cual sea el orden ($\tau^{(3)}$ o $\tau^{(4)}$), la estrategia de deflación y la condición de centralizado o federado. El patrón coincide con los resultados obtenidos en [8] para métodos basados en MSE, extendidos aquí a una nueva clase de estimadores. La AncAcc estandarizada cae a la base aleatoria de 0.5 o por debajo, con varios métodos entre 0.30 y 0.40. Ese rendimiento por debajo del azar es más fuerte que el observado para NOTEARS, y el mecanismo es directo: la estandarización reescala $\kappa_m(x_i)$ por σ_i^{-m}, los nodos más profundos pierden

más magnitud que las fuentes, la escala se invierte y la ordenación por sumas por filas lee con confianza la escalera invertida. DirectLiNGAM no se ve afectado porque sus tests con kernel sobre residuos en bruto acceden directamente a la asimetría poblacional.

La asimetría que codifican $\tau^{(3)}$ y $\tau^{(4)}$ está en los datos, pero no es accesible a la familia de cumulantes con N del orden de miles. Es el precio de la federación exacta bajo agregación de cumulantes, paralelo al que los métodos basados en MSE pagan por la aciclicidad diferenciable [8]. También explica la brecha en bruto entre FedISHC y FedRCD-X con ruido simétrico: la deflación de FedISHC inyecta el ruido divergente de $\hat{\alpha}_{js}$ en su actualización de κ_3 y corrompe la escala a lo largo de las $p-1$ actualizaciones, mientras que FedRCD-X la recalcula desde residuos reales y restaura la estratificación sobre datos cada vez más limpios. FedHC y FedRCD-NI la preservan trivialmente; FedRCD-Def queda en medio.

5. Conclusiones

Partimos del modo de fallo teórico de FedISHC con ruido simétrico (el tercer cumulante propio del denominador tiende a cero) y de que el simple cambio del *score* no basta, ya que la deflación también opera sobre el orden equivocado. La familia FedRCD corrige ambos a la vez trabajando en el dominio de cuarto orden con un coeficiente OLS basado en covarianzas. FedRCD-NI y FedRCD-X son las contrapartes federadas exactas de HC-LiNGAM y HC, igualando el rendimiento centralizado dados los mismos estadísticos agregados. Todos los algoritmos desarrollados se aplican a federación horizontal, vertical e híbrida. Además, las variantes de una sola ronda (FedRCD-NI y FedRCD-Def) admiten desaprendizaje exacto a cualquier granularidad.

El diagnóstico de la Sección 4.5 demuestra que toda la familia federada basada en cumulantes, junto con sus referencias centralizadas, ordena las variables por la escalera de varianzas inducida por el DAG y no por la asimetría poblacional que los *scores* codifican en cero. La estandarización marginal las colapsa a una ordenación casi aleatoria, mientras que DirectLiNGAM, invariante a escala, conserva su rendimiento. Esto sitúa a la familia en el régimen dependiente de escala que se identificó en [8] para métodos basados en MSE. Por tanto, pese a sus múltiples ventajas, este tipo de métodos solo son apropiados en casos reales en los que realmente suceda esa acumulación de varianza. Dentro de este marco, FedRCD-X alcanza el techo de la federación exacta bajo agregación de cumulantes. Como trabajo futuro queda determinar cuándo basta la estratificación, el tamaño muestral necesario para acceder a la asimetría poblacional, y la aplicación de privacidad diferencial a los tensores transmitidos.

Agradecimientos

Este trabajo ha sido financiado por los proyectos PID2022-139293NB-C32 (MICIU/AEI/10.13039/501100011033 y FEDER, UE), FPU21/01074 (MI-

CIU/AEI/10.13039/501100011033 y FSE+) y 2025-GRIN-38476 (Universidad de Castilla-La Mancha y FEDER, una manera de hacer Europa). Fabio Stella ha sido financiado por el MUR mediante la ayuda "Dipartimenti di Eccellenza 2023-2027" del Dipartimento di Informatica, Sistemistica e Comunicazione de la Università degli Studi di Milano Bicocca, Milano, Italia, y del Piano Nazionale per gli investimenti Complementari al PNRR (Proyecto n.º PNC0000003, AdvaNced Technologies for Human-centrEd Medicine, ANTHEM).

Referencias

1. Brillinger, D.R.: Time Series: Data Analysis and Theory. Society for Industrial and Applied Mathematics (2001)
2. Chen, W., Gu, W., Peng, L., et al.: Horizontal and Vertical Federated Causal Structure Learning via Higher-order Cumulants. Proceedings of the AAAI Conference on Artificial Intelligence **40**(24), 20280–20288 (Mar 2026)
3. Chen, W., Peng, L., Huang, Z., et al.: Higher Order Cumulants-Based Method for Direct and Efficient Causal Discovery. IEEE Transactions on Neural Networks and Learning Systems pp. 1–14 (2025)
4. Huang, J., Guo, X., Yu, K., et al.: Towards Privacy-Aware Causal Structure Learning in Federated Setting. IEEE Transactions on Big Data **9**(6), 1525–1535 (Dec 2023)
5. McMahan, B., Moore, E., Ramage, D., et al.: Communication-Efficient Learning of Deep Networks from Decentralized Data. In: Proceedings of the 20th AISTATS. vol. 54, pp. 1273–1282 (2017)
6. Moneta, A., Chlass, N., Entner, D., Hoyer, P.: Causal Search in Structural Vector Autoregressive Models. In: Proceedings of the NeurIPS Mini-Symposium on Causality in Time Series. vol. 12, pp. 95–114. PMLR (2011)
7. Ng, I., Zhang, K.: Towards Federated Bayesian Network Structure Learning with Continuous Optimization. In: Proceedings of the 25th AISTATS. vol. 151, pp. 8095–8111. PMLR (2022)
8. Reisach, A., Seiler, C., Weichwald, S.: Beware of the Simulated DAG! Causal Discovery Benchmarks May Be Easy to Game. In: Advances in Neural Information Processing Systems. vol. 34, pp. 27772–27784 (2021)
9. Shimizu, S., Hoyer, P.O., Hyvärinen, A., Kerminen, A.: A Linear Non-Gaussian Acyclic Model for Causal Discovery. Journal of Machine Learning Research **7**, 2003–2030 (2006)
10. Shimizu, S., Inazumi, T., Sogawa, Y., et al.: DirectLiNGAM: A Direct Method for Learning a Linear Non-Gaussian Structural Equation Model. Journal of Machine Learning Research **12**, 1225–1248 (2011)
11. Speed, T.P.: Cumulants and Partition Lattices. Australian Journal of Statistics **25**(2), 378–388 (1983)
12. Tejada-Lapuerta, A., Bertin, P., Bauer, S., et al.: Causal machine learning for single-cell genomics. Nature Genetics **57**(4), 797–808 (2025)
13. Zhai, J., Ji, J., Liu, J.: Inferring Causal Protein Signaling Networks with Reinforcement Learning via Artificial Bee Colony Neural Architecture Search. In: Proceedings of the Thirty-Fourth IJCAI. pp. 8996–9004 (2025)
14. Zhang, C., Xie, Y., Bai, H., et al.: A survey on federated learning. Knowledge-Based Systems **216**, 106775 (2021)