Aprendizaje e inferencia en redes bayesianas híbridas con datos incompletos

Ángel T. Sáez-Ruiz[1],
Directores: Ana D. Maldonado[1,2] y Rafael Rumí[1,2]

[1] Departamento de Matemáticas, Universidad de Almería, Almería, España
[2] Centro de Desarrollo y Transferencia de Investigación Matemática a la Empresa, Universidad de Almería, Almería, España
{angelsaez,ana.d.maldonado,rrumi}@ual.es

Resumen Modelar conjuntos de datos híbridos usando redes bayesianas sigue siendo una tarea compleja si se desea prescindir de la discretización. En este ámbito, los modelos de redes bayesianas basados en mixturas de polinomios ofrecen una alternativa flexible al permitir abordar estas situaciones sin imponer restricciones a la estructura de la red ni a la naturaleza de las variables. Cuando se trabaja en el modelado de problemas reales, la presencia de datos incompletos, debido a la existencia de valores faltantes o de variables latentes no observables, es una de las situaciones más habituales.
El objetivo principal de este proyecto de tesis doctoral es desarrollar nuevas metodologías que permitan modelar problemas reales con conjuntos de datos híbridos parcialmente observados mediante el uso de redes bayesianas basadas en mixturas de polinomios. Para promover la transferencia de conocimiento, estas metodologías serán desarrolladas en código abierto, de forma que estén disponibles para la comunidad científica para su uso y consulta. Además, a lo largo de la tesis doctoral se mostrará la aplicabilidad y versatilidad de esta metodología.

Palabras clave: Redes bayesianas híbridas · mixturas de polinomios · datos incompletos.

Datos de la tesis doctoral

- **Título provisional de la tesis doctoral:** Aprendizaje e inferencia en redes bayesianas híbridas con datos incompletos.
- **Fecha de inicio:** Noviembre 2024.
- **Modalidad:** El doctorado se está realizando a tiempo completo.
- **Fecha de defensa estimada:** Octubre de 2028.
- Los nombres del autor y de los directores de la tesis doctoral, así como su información de contacto, se encuentran al inicio de este documento.

1 Antecedentes

El machine learning probabilístico es un campo consolidado que ofrece modelos con una estructura gráfica que los hace fácilmente comprensibles. Estos modelos

se basan principalmente en la utilización de modelos gráficos probabilísticos [8], que permiten razonar mediante probabilidad y están definidos con dos partes diferenciadas. La primera de ellas es un componente cualitativo en forma de grafo que representa las relaciones de independencia entre las variables del dominio que se está modelando; y la segunda, un componente cuantitativo que consta de un conjunto de distribuciones de probabilidad que cuantifican las relaciones especificadas en el grafo.

En este proyecto de tesis doctoral se emplearán las redes bayesianas [6] (BNs). Las BNs son un tipo de modelo gráfico probabilístico en el que la parte cualitativa es un *grafo dirigido acíclico* (DAG, por sus siglas en inglés) y la parte cuantitativa consta de las distribuciones de probabilidad condicionada, una para cada variable dadas sus variables padres. Originalmente, las BNs fueron definidas para modelos en el que todas las variables fuesen discretas. Por otro lado, una BN se dice *híbrida* cuando coexisten variables discretas y continuas en la red. La Figura 1 muestra un ejemplo de red bayesiana híbrida con siete variables. En este caso Y denota a las variables discretas y Z a aquellas que son continuas.

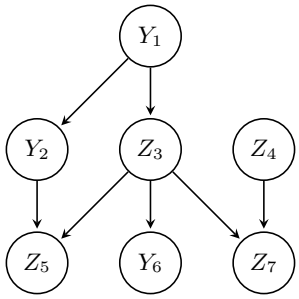


Fig. 1. Ejemplo de red bayesiana híbrida con siete variables. Y y Z denotan a variables aleatorias discretas y continuas, respectivamente.

Una práctica habitual en BNs es discretizar las variables continuas presentes en el problema a modelar. Esta discretización permite emplear la metodología existente para el caso discreto y evitar la complejidad de tratar con conjuntos híbridos. Aunque este proceso conlleva una pérdida de información.

Dentro de las BN hibridas, existen alternativas a la discretización. Un ejemplo de ello es la red bayesiana lineal condicional gaussiana [16], en la cuál se asume que la distribución conjunta de las variables continuas, condicionada a las variables discretas, sigue una distribución normal multivariante. Además, asume que las variables discretas solo pueden tener como padres otras variables discretas para poder aplicar algoritmos de inferencia exacta. Sin embargo, cuando se trabaja con datos reales, asumir que las variables continuas siguen una distribución de este tipo no siempre es posible. Esta serie de hechos ha

supuesto el desarrollo de otros modelos, como las mixturas de polinomios [10, 17] (MoPs, por sus siglas en inglés).

Los modelos MoP se basan en la representación de las densidades de probabilidad mediante expresiones polinómicas, lo que permite una mayor adaptabilidad a la aproximación de gran variabilidad de distribuciones. Una función de densidad MoP condicionada para una variable Z se define mediante la partición del soporte de sus padres, asignando a la variable Z una densidad MoP específica para cada región de dicha partición. De este modo, la influencia de las variables continuas sobre la variable hija se codifica exclusivamente a través de la segmentación de su soporte en hipercubos, resultando en una función de densidad condicionada que es un polinomio definido a trozos, cuya forma depende de la región del espacio soporte en la que se encuentren los padres. Además, las distribuciones MoP son cerradas bajo adición, multiplicación e integración, permitiendo el uso de algoritmos de inferencia exacta. Por último, tampoco imponen restricciones en el grafo de la red, en contraposición al modelo condicional lineal gaussiano.

Cuando se trabaja con conjuntos de datos reales, es habitual que estos conjuntos estén incompletos [7]. Esto puede deberse a dos situaciones, principalmente: (i) valores faltantes, cuando en algunas observaciones hay valores faltantes de alguna variable; (ii) variables latentes, cuando no hay observaciones en alguna de las variables.

Esta tesis doctoral se centra en el desarrollo de métodos de aprendizaje de redes bayesianas usando distribuciones MoP con conjuntos de datos incompletos.

Aprendizaje con datos incompletos

Modelar datos incompletos es una tarea mucho más compleja en comparación con los datos completos. Existen diversas estrategias para tratar los valores faltantes, desde la eliminación de las observaciones con datos faltantes, la imputación de los datos y hasta el aprendizaje de los parámetros empleando los datos incompletos. Algunas de las técnicas tradicionales para la imputación de valores son el uso del valor medio o el empleo de algún modelo de regresión para estimar los valores a partir de la información disponible. Sin embargo, estas técnicas son ampliamente criticadas debido al sesgo que se puede introducir en los datos [18].

Dentro del contexto de las redes bayesianas, han surgido otros métodos iterativos como *Data Augmentation* [19], *Expectation-Maximization* [11] y *Bound and Collapse* [14]. Estos métodos iterativos se basan en la idea de que, al trabajar con datos faltantes, se van a resolver dos problemas simultáneamente: la estimación de los parámetros y la proposición de nuevos valores para los datos faltantes. De esta forma, se emplea un conjunto de datos completo para el aprendizaje de los parámetros y, dados esos parámetros, es posible determinar los valores más probables para aquellos datos no observados.

Además, las redes bayesianas aprovechan de forma particularmente efectiva la integración de estas variables latentes, lo que les confiere importantes ventajas, como (i) la simplificación de su estructura al permitir representar dependencias complejas de forma compacta, ya que al introducir variables ocultas es posible

reducir el número de arcos directos entre las variables observables [8]; (ii) la posibilidad de construir modelos de clasificación no supervisada, ya que las variables ocultas pueden representar factores latentes que ayudan a identificar estructuras complejas subyacentes en los datos [13]; y (iii) la superación de las limitaciones de modelos tradicionales como el Naive Bayes (NB) o Tree Augmented Naive Bayes (TAN) [2].

Si nos centramos en este último aspecto, el modelo naive Bayes (NB) impone una suposición de independencia condicional muy fuerte que es, por lo general, poco realista. La violación de este supuesto también podría interpretarse como un indicador de la presencia de variables latentes. En otras palabras, podemos asumir que los datos observados se generan a partir de un modelo en el que todas las variables, salvo la clase y las predictoras, son no observables.

Dentro de la clase de modelos de clasificación con variables latentes encontramos el llamado naive Bayes jerárquico (HNB, por sus siglas en inglés) [20], donde el uso de variables latentes permite codificar dependencias condicionadas en el modelo, en comparación con el modelo NB. Otra clase de modelos de clasificación con variables ocultas son los latent classification models (LCM) [9], donde se introduce una capa oculta entre la variable clase y las variables predictoras observables, y estas están densamente conectadas. El modelo LCM puede entenderse como una combinación de análisis factorial y naive bayes, donde las variables ocultas funcionan como factores latentes que explican las variables observables. La clase de modelos propuesta en [9] queda limitada a dominios continuos, ya que los nodos ocultos se representan mediante distribuciones condicionadas de tipo CLG, lo que no permite que tengan hijos discretos. Hasta la fecha, no se han explorado modelos de esta clase que permitan tanto variables discretas como continuas en el conjunto de variables predictoras observables.

Análisis de sensibilidad en modelos MoPs

El objetivo del análisis de sensibilidad es estudiar cómo las variaciones en los parámetros afectan a la distribución de probabilidad de una variable de interés. El caso más simple se produce cuando solo se modifica un parámetro del modelo mientras que el resto permanecen fijos.

El análisis de sensibilidad en las redes bayesianas (BNs) discretas ha sido ampliamente estudiado en la literatura [1, 6, 15]. En las BNs discretas, los parámetros están directamente asociados a los valores de las variables. De este modo, la incompletitud de los datos, el conocimiento parcial de los dominios y el conocimiento experto son factores que pueden afectar a la precisión de los parámetros de la BN [4].

Por el contrario, para las variables aleatorias continuas, los parámetros no están directamente vinculados a un único estado de la variable, y cualquier modificación en ellos afecta a la probabilidad de todos los sucesos. En el caso de los CLG, los parámetros representan la media y la varianza de la variable. El análisis de sensibilidad en CLG ha sido estudiado en la literatura [3, 5, 22]. En

estos trabajos, el objetivo es analizar las diferencias en una distribución objetivo mediante distintas perturbaciones de un único parámetro de la BN.

Cuando se emplean las BNs basadas en MoP, las distribuciones de probabilidad condicional (CPDs) se expresan mediante densidades polinómicas. Hasta la fecha, el análisis de sensibilidad de los parámetros de las distribuciones MoP no ha sido estudiado en la literatura. El estudio de la sensibilidad de los parámetros de los modelos MoP puede contribuir a mejorar el rendimiento de los modelos tanto desde el punto de vista predictivo como del coste computacional. Además, se espera que el análisis de sensibilidad permita una mejor integración de las variables latentes en los modelos MoP.

2 Objetivos

La hipótesis de partida es que los modelos MoP pueden extenderse para la representación de patrones no observables. Esto conllevará a la definición de una nueva clase de modelos con variables latentes, en las que las distribuciones condicionadas se modelan empleando MoPs.

Para ello, se desarrollarán métodos de aprendizaje para estimar distribuciones MoP a partir de datos incompletos. Además, estos métodos serán implementados en un paquete de R para facilitar su accesibilidad y fomentar su adopción dentro de la comunidad científica.

Finalmente, se espera la aplicación de estos métodos al análisis de datos reales, como muestra de su potencial para resolver problemas en diferentes campos científicos.

De este modo, los objetivos de este proyecto serán los siguientes:

- **[O1]** Aprendizaje de redes bayesianas mediante modelos con variables latentes basados en distribuciones MoPs. Definición de modelos y desarrollo de métodos de aprendizaje paramétrico.
- **[O2]** Desarrollo de software accesible, intuitivo y eficiente con todos los avances metodológicos del proyecto.
- **[O3]** Aplicación de los desarrollos metodológicos y el software al modelado de datos reales. En particular, se centrará en la clasificación de la duplicabilidad de un gen en el genoma de la especie *Arabidopsis thaliana.*

3 Metodología

En términos metodológicos, para el desarrollo de la tesis doctoral, se empleará el proceso científico estándar, que consta de cinco etapas básicas: (i) identificación de los problemas a abordar; (ii) revisión de las fuentes científicas relevantes; (iii) Propuesta de soluciones para el problema, incluyendo nuevos desarrollos metodológicos, diseño de algoritmos, implementaciones innovadoras o mejoras de las ya existentes; (iv) validación de la solución propuesta mediante la utilización de los conjuntos de datos disponibles; (v) evaluación de las propuestas presentadas. Además, los resultados serán publicados en congresos y revistas de

prestigio, no solo con el fin de su difusión, sino también con la intención de mejorar los resultados a través de la opinión de la comunidad científica.

En cuanto al plan de trabajo, este proyecto de tesis se ha estructurado en tres paquetes de trabajo (PT), que se describen a continuación. Estos PTs se han definido de acuerdo con los objetivos planteados; que a su vez se dividen en tareas que muestran un mayor detalle de las actividades que se llevarán a cabo para alcanzar dichos objetivos.

[PT1] Aprendizaje de estructuras restringidas con datos completos. Este paquete de trabajo está directamente relacionado con el objetivo O1, que se centra en el desarrollo de métodos de aprendizaje estructural de redes bayesianas restringidas, usando las distribuciones MoP en el proceso de aprendizaje a partir de datos completos. Además, se iniciará el objetivo O3 con la implementación de modelos utilizando datos completos para la predicción de la presencia de genes duplicados en el genoma de la especie *Arabidopsis thaliana.*

Las tareas que forman parte del paquete PT1 son las siguientes:

- **Tarea 1.1.** Revisión bibliográfica.
  Se realizará una revisión bibliográfica sobre la metodología existente para el aprendizaje estructural de redes bayesianas. Esta revisión incluirá el aprendizaje de estructuras restringidas.

- **Tarea 1.2.** Aprendizaje de modelos TAN con datos observados.
  En esta tarea se desarrollarán algoritmos para el aprendizaje de estructuras tipo TAN donde las distribuciones MoPs sean relevantes para el aprendizaje estructural.

- **Tarea 1.3.** Aprendizaje de modelos $k-$DB con datos observados.
  Esta tarea generalizará la tarea anterior a otros modelos restringidos que flexibilicen las restricciones estructurales del modelo TAN.

- **Tarea 1.4.** Aplicaciones.
  Esta tarea consistirá en la aplicación de los métodos implementados a lo largo del PT. Como caso de estudio, nos centraremos en el problema de duplicabilidad de genes en el genoma de la especie *Arabidopsis thaliana.* En primer lugar, se realizará la obtención y preprocesamiento de los datos. Se seleccionarán variables cuantitativas y cualitativas, mostrando la versatilidad de los modelos MoPs. Posteriormente, a partir del conjunto de datos seleccionado, se construirán modelos de clasificación que serán validados utilizando validación cruzada. Finalmente, los modelos desarrollados serán comparados con otras metodologías existentes en el estado del arte.

[PT2] Análisis de sensibilidad de redes bayesianas MoP. Este paquete de trabajo está íntegramente relacionado con el objetivo O1. Este paquete tiene la finalidad de entender el comportamiento de los modelos MoP y ser una de

las vías de enlace entre el aprendizaje con datos completos y el aprendizaje con datos incompletos. La principal vía de enlace será la búsqueda de características que deban cumplir las variables para la incorporación de variables latentes a modelos con todas las variables observadas.

Las tareas que forman parte del paquete PT2 son las siguientes:

- **Tarea 2.1.** Revisión bibliográfica.
  Se realizará una revisión bibliográfica sobre el estudio de la sensibilidad de los parámetros en redes bayesianas. Esta revisión estará enfocada tanto a la sensibilidad de las estimaciones de las distribuciones, así como también en estrategias para la mejora del rendimiento de los modelos.

- **Tarea 2.2.** Análisis de sensibilidad con distribuciones univariantes.
  En esta tarea se realizará un estudio sobre la sensibilidad de distribuciones univariantes. Este estudio incluirá la formulación de medidas de sensibilidad y de estrategias para reducir la sensibilidad de las estimaciones durante el proceso de aprendizaje.

- **Tarea 2.3.** Análisis de las perturbaciones en las redes bayesianas.
  Esta tarea extiende la anterior al caso de redes bayesianas con distribuciones MoP. Para ello, se realizarán diferentes análisis sobre el efecto que genera una perturbación en una variable de la red en el proceso de inferencia con el objetivo de mejorar el rendimiento de los modelos.

[PT3] Aprendizaje paramétrico con datos incompletos. Este paquete de trabajo está asociado al objetivo O1 y se centra en el desarrollo de métodos de aprendizaje de las distribuciones MoP a partir de datos incompletos. En él se desarrollan las metodologías para el aprendizaje en situaciones donde los datos presentan ausencia parcial de información, que es el objetivo principal de la tesis doctoral. Este PT también contribuye al objetivo O3 con la implementación de modelos empleando datos incompletos para la predicción de la presencia de genes duplicados en el genoma de la especie *Arabidopsis thaliana.*

Las tareas que forman parte del paquete PT3 son las siguientes:

- **Tarea 3.1.** Revisión bibliográfica.
  Se realizará una revisión bibliográfica sobre la metodología existente para el aprendizaje de redes bayesianas en presencia de datos incompletos. Esta revisión incluirá enfoques basados tanto en la imputación de datos como en estrategias de optimización de parámetros.

- **Tarea 3.2.** Aprendizaje de modelos con datos parcialmente observados.
  En esta tarea se desarrollarán algoritmos para manejar datos incompletos en nodos observables. Esto incluirá la formulación de estrategias específicas para las MoP que permitan estimar parámetros en situaciones donde ciertas observaciones no estén disponibles. Se evaluará la robustez de los métodos propuestos frente a diferentes patrones y niveles de ausencia de datos.

– **Tarea 3.3.** Aprendizaje de modelos con variables latentes.
  Esta tarea abordará el problema de incorporar variables no observables en redes bayesianas basadas en MoPs. Las variables latentes influyen en las relaciones de las variables observables y pueden, potencialmente, simplificar la estructura del modelo. Los desarrollos de la tarea 3.2 pueden ser adaptados al caso anterior siempre que el DAG sea conocido.

– **Tarea 4.3.** Aplicaciones.
  Esta tarea consistirá en la aplicación de los métodos implementados a lo largo del PT. Se continuará con el caso de estudio de duplicabilidad de genes en el genoma de la especie *Arabidopsis thaliana.* Los modelos empleados en la Tarea 1.4 serán comparados con sus versiones que incluyan variables parcialmente observadas y con otros modelos con variables latentes. Además, se harán comparativas con otras metodologías del estado del arte que consideren el aprendizaje con datos incompletos.

[PT4] Desarrollo de software. Este paquete de trabajo está íntegramente relacionado con el objetivo O2, que consiste en la implementación de los métodos desarrollados en los PT anteriores en software de código abierto.

Las tareas que forman parte del paquete PT4 son las siguientes:

– **Tarea 4.1.** Implementación de los métodos.
  La implementación de los métodos desarrollados en PT1 y PT3 se hará en R y se tomará como punto de partida el paquete MoTBFs, disponible en el repositorio oficial de R, CRAN, ya que este implementa algoritmos para el aprendizaje paramétrico e inferencia de MoPs a partir de un conjunto de datos completo.

– **Tarea 4.2.** Distribución y documentación.
  El paquete se desarrollará de forma progresiva a lo largo del proyecto de tesis doctoral. El código estará disponible con licencia de software libre a través de un repositorio de software como GitHub o similar. Las versiones estables del paquete se publicarán a través del CRAN. Esto permitirá poner el software a disposición de la comunidad científica desde las etapas iniciales del proyecto. La documentación o manual de usuario estará disponible online y se actualizará tras cada versión.

Para la consecución de los objetivos fijados, mediante los paquetes de trabajo enumerados anteriormente, se propone una temporalización de 48 meses. La Figura 3, muestra un cronograma orientativo de la temporalización de cada una de las tareas descritas en dichos paquetes de trabajo.

4 Relevancia, resultados e impacto esperados

Las redes bayesianas han mostrado una gran aplicabilidad dentro del ámbito científico [12, 21]. Por este motivo, se espera que este proyecto tenga un impacto

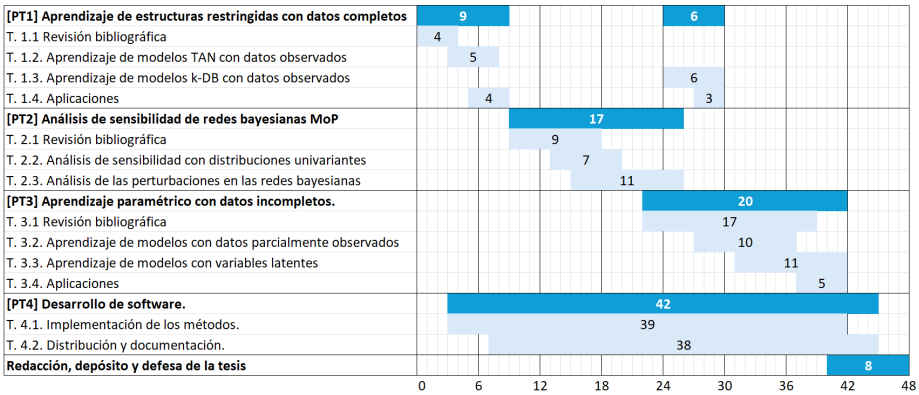


Fig. 2. Temporalización de los paquetes de trabajo (PT) con sus respectivas tareas (T.). Se ha considerado una duración de la tesis doctoral de 48 meses.

significativo en muchos campos del ámbito científico. El desarrollo metodológico será implementado en software de código abierto, lo que permitirá una mayor difusión y aplicabilidad. Los avances metodológicos que se desarrollarán serán aplicables en cualquier ámbito con fuentes de datos similares.

En cuanto a la producción científica, se prevé la publicación de al menos tres artículos en revistas indexadas de relevancia en el campo de estudio. En dichos artículos se abordarán los paquetes de trabajo PT1, PT2 y PT3. Además, los resultados preliminares serán presentados en congresos internacionales.

Agradecimientos: El trabajo de los autores es finaciado por la subvención PID2022-139293NB-C31 fundada por MICIU/AEI/10.13039/501100011033 y por ERDF/EU "A way of Making Europe", por el grupo de investigación FQM-244, financiado por PPIT-UAL, Junta de Andalucía-ERDF 2021-2027, Objetivo RSO1.1. Programa: 54.A.

Declaración de conflicto de intereses: Los autores declaran no tener ningún conflicto de intereses relevante para el contenido de este documento.

Referencias

1. Ballester-Ripoll, R., Leonelli, M.: Global sensitivity analysis of uncertain parameters in Bayesian networks. International Journal of Approximate Reasoning **180**, 109368 (2025). https://doi.org/10.1016/j.ijar.2025.109368
2. Bishop, C.M., Nasrabadi, N.M.: Pattern Recognition and Machine Learning, vol. 4. Springer (2006)
3. Castillo, E., Kjærulff, U.: Sensitivity analysis in Gaussian Bayesian networks using a symbolic-numerical technique. Reliability Engineering & System Safety **79**(2), 139–148 (2003). https://doi.org/10.1016/S0951-8320(02)00225-9

4. Druzdzel, M.J., Van Der Gaag, L.C.: Elicitation of probabilities for belief networks: combining qualitative and quantitative information. In: Proceedings of the Eleventh Conference on Uncertainty in Artificial Intelligence. p. 141–148. Morgan Kaufmann Publishers Inc. (1995)
5. Gómez-Villegas, M.A., Maín, P., Susi, R.: Extreme inaccuracies in Gaussian Bayesian networks. Journal of Multivariate Analysis **99**(9), 1929–1940 (2008). https://doi.org/10.1016/j.jmva.2008.02.027
6. Jensen, F.V., Nielsen, T.D.: Bayesian Networks and Decision Graphs. Springer (2007). https://doi.org/10.1007/978-0-387-68282-2
7. Junninen, H., Niska, H., Tuppurainen, K., Ruuskanen, J., Kolehmainen, M.: Methods for imputation of missing values in air quality data sets. Atmospheric environment **38**(18), 2895–2907 (2004). https://doi.org/10.1016/j.atmosenv.2004.02.026
8. Koller, D., Friedman, N.: Probabilistic graphical models: principles and techniques. MIT press (2009)
9. Langseth, H., Nielsen, T.D.: Classification using hierarchical naive bayes models. Machine learning **63**(2), 135–159 (2006). https://doi.org/10.1007/s10994-006-6136-2
10. Langseth, H., Nielsen, T.D., Rumí, R., Salmerón, A.: Mixtures of truncated basis functions. International Journal of Approximate Reasoning **53**(2), 212–227 (2012). https://doi.org/10.1016/j.ijar.2011.10.004
11. Lauritzen, S.L.: The EM algorithm for graphical association models with missing data. Computational statistics & data analysis **19**(2), 191–201 (1995). https://doi.org/10.1016/0167-9473(93)E0056-A
12. Machado, P.G., Ribeiro, C.O., do Nascimento, C.A.O.: Risk analysis in energy projects using Bayesian networks: A systematic review. Energy Strategy Reviews **47**, 101097 (2023). https://doi.org/10.1016/j.esr.2023.101097
13. McLachlan, G.J., Peel, D.: Finite mixture models. John Wiley & Sons (2000)
14. Ramoni, M., Sebastiani, P.: The use of exogenous knowledge to learn bayesian networks from incomplete databases. In: International Symposium on Intelligent Data Analysis. pp. 537–548. Springer (1997). https://doi.org/10.1007/BFb0052869
15. Rohmer, J.: Uncertainties in conditional probability tables of discrete Bayesian belief networks: A comprehensive review. Engineering Applications of Artificial Intelligence **88**, 103384 (2020). https://doi.org/10.1016/j.engappai.2019.103384
16. Shachter, R.D., Kenley, C.R.: Gaussian influence diagrams. Management Science **35**(5), 527–550 (1989). https://doi.org/10.1287/mnsc.35.5.527
17. Shenoy, P.P., West, J.C.: Inference in hybrid Bayesian networks using mixtures of polynomials. International Journal of Approximate Reasoning **52**(5), 641–657 (2011). https://doi.org/10.1016/j.ijar.2010.09.003
18. Soley-Bori, M., et al.: Dealing with missing data: Key assumptions and methods for applied analysis. Boston University **4**(1), 1–19 (2013)
19. Tanner, M.A., Wong, W.H.: The calculation of posterior distributions by data augmentation. Journal of the American statistical Association **82**(398), 528–540 (1987). https://doi.org/10.1080/01621459.1987.10478458
20. Zhang, N.L., Nielsen, T.D., Jensen, F.V.: Latent variable discovery in classification models. Artificial intelligence in medicine **30**(3), 283–299 (2004). https://doi.org/10.1016/j.artmed.2003.11.004
21. Zheng, J., Zhuo, Y., Jiang, X., Zeng, L., Ge, Z.: Advances in Bayesian networks for industrial process analytics: Bridging data and mechanisms. Expert systems with applications **271** (2025). https://doi.org/10.1016/j.eswa.2025.126670

22. Zhu, M., Liu, S., Jiang, J.: A novel divergence for sensitivity analysis in Gaussian Bayesian networks. International Journal of Approximate Reasoning **90**, 37–55 (2017). https://doi.org/10.1016/j.ijar.2017.07.002