27 Congreso Nacional de Estadística e Investigación Operativa
Lleida, 8–11 de abril de 2003

ANÁLISIS DEL ALUMNADO DE LA UNIVERSIDAD DE ALMERÍA MEDIANTE REDES BAYESIANAS

María Morales, Antonio Salmerón

Departamento de Estadística y Matemática Aplicada
Universidad de Almería, 04120 Almería, España
E-mail: {Maria.Morales,Antonio.Salmeron}@ual.es

RESUMEN

Realizamos un análisis del alumnado de la Universidad de Almería correspondiente al curso académico 2000/2001 utilizando modelos gráficos probabilísticos. Estos modelos permiten el manejo de bases de datos con numerosas variables de carácter heterogéneo. Los modelos propuestos pueden ser utilizados para tareas tales como la clasificación o la extracción de perfiles.

Palabras y frases clave: Redes bayesianas, análisis multivariante, aplicaciones.
Clasificación AMS: 62H99.

1. Introducción

Las redes bayesianas son una representación compacta de una distribución de probabilidad multivariante. Formalmente, una *red bayesiana* es un grafo dirigido acíclico donde cada nodo representa una variable aleatoria y las dependencias entre las variables quedan codificadas en la propia estructura del grafo según el criterio de d-separación (Pearl 1988). Asociada a cada nodo de la red hay una distribución de probabilidad condicionada a los padres de ese nodo, de manera que la distribución conjunta factoriza como el producto de las distribuciones condicionadas asociadas a los nodos de la red.
Diferentes tipos de inferencias pueden llevarse a cabo sobre estos modelos. La tarea más frecuente es la llamada *propagación de probabilidad*, para la cual existen diversos algoritmos que sacan partido de las independencias codificadas por la red para realizar los cálculos de manera eficiente (Cano, Moral, and Salmerón 2000; Lauritzen and Spiegelhalter 1988; Madsen and Jensen 1999). La propagación de probabilidad consiste en la obtención de las probabilidades a posteriori de ciertas variables de la red dado que se conoce el valor que toman algunas otras variables observadas.

Otra tarea muy habitual con redes bayesianas consiste en la extración de la explicación o explicaciones más probables a una determinada observación (Gámez 1998; Nilsson 1998).
Desde el punto de vista del análisis de datos, las redes bayesianas son una potente herramienta por varios motivos:

1. No suponen un determinado modelo subyacente.
2. Son fácilmente interpretables.
3. Son adaptables y permiten la incorporación de conocimiento a priori de forma cualitativa.

Nuestro objetivo es construir una red bayesiana que modelice a los alumnos matriculados en la Universidad de Almería en el curso 2000/2001 y poder realizar inferencias sobre dicha red.
En la sección 2 describimos los datos utilizados en el estudio. En la sección 3 presentamos el programa con el cual se ha construido el modelo y realizado el análisis. El modelo obtenido (la red bayesiana) se describe en la sección 4, mientras que las secciónes 5 y 6 se dedican, respectivamente, al análisis de los datos mediante propagación de probabilidad e inferencia abductiva.

2. Los datos

Disponemos de una base de datos con 13.747 alumnos de los cuales hemos considerado las siguientes variables de interés:

- Facultad o Escuela,
- Estudio en el que se matricula el alumno,
- Tipo de estudio (ciclo corto, ciclo largo o sólo segundo ciclo),
- Edad del alumno; a partir de los treinta años de edad se han agrupado en los valores 31-35, 36-40, 41-50 y MAS50,
- Nacionalidad del alumno, tenemos alumnos de once nacionalidades: Francia, Alemania, Argentina, Italia, Marruecos, Bélgica, Portugal, Rusia, Reino Unido, Venezuela y España. En la base de datos no están incluidos los alumnos Erasmus,
- Provincia y comunidad autónoma del domicilio familiar,
- Año en el que comenzó los estudios universitarios,
- Año en el que comenzó a estudiar en la Universidad de Almería,

- Año en el que comenzó el estudio del que se encuentra matriculado,
- Año en que accedió al centro,
- Estudios del padre y de la madre, agrupados en seis categorias:
  1. Sin estudios,
  2. Primarios completos,
  3. Bachillerato elemental,
  4. Bachillerato superior,
  5. Diplomado o estudios profesionales de grado medio,
  6. Licenciado, arquitecto, e. superior militar, etc.
- Trabajo del padre y de la madre agrupados en:
  1. Directores o Gerentes de empresas de la administración pública
  2. Técnicos o profesionales asociados a titulaciones universitarias o no universitarias de carácter postsecundario
  3. Empleados administrativos y trabajadores de los servicios
  4. Trabajadores cualificados en agricultura y pesca
  5. Trabajadores cualificados y operadores de máquinas en la industria
  6. Trabajadores no cualificados
  7. Profesionales de las fuerzas armadas
  8. Personal que no han tenido un trabajo remunerado
- Trabajo del alumno.
- Modo de ingreso en la Universidad, en el estudio y en el centro,
- Si el alumno entró por distrito compartido o no,
- Modo de acceso al segundo ciclo,
- Estudios del alumno
- Curso más alto en el que se matricula el alumno,
- Número de créditos o asignaturas matriculados,
- Si se matricula en asignaturas o créditos de complementos de formación,
- Nota de acceso,
- Si la familia del alumno es numerosa.

Los datos los hemos adaptado al formato exigido por el programa Elvira, que es el utilizado para conseguir la red deseada.

3. Programa Elvira

Este programa, cuyo nombre oficial es *Entorno de Desarrollo para Modelos Gráficos Probabilísticos*, fue desarrollado entre los años 1997 y 2000, como objetivo de un Proyecto Coordinado de I+D, financiado por la CICYT. En este proyecto participaron 25 profesores de 8 universidades españolas agrupadas en cuatro subproyectos: Granada, Almería, País Vasco y UNED. Una descripción detallada del programa puede encontrarse en Elvira consortium (2002).
El programa presenta tres modos básicos:

- Edición
- Inferencia
- Aprendizaje

El modo *Aprendizaje* toma la base de datos y construye las tablas de probabilidad y la estructura de la red bayesiana, para ello, el algoritmo elegido ha sido el PC de Spirtes, Glymour, and Scheines (1993).
Para poder ver mejor la estructura gráfica de la red, utilizamos el modo *Edición*, modificando la red de forma que puedan verse todos los nodos y las relaciones entre ellos. Una vez creada la red, utilizamos el modo *Inferencia* para propagar probabilidades y extraer perfiles.

4. La red bayesiana

La red obtenida fue la que se muestra en la figura 1.
Debido a la complejidad de la red, hemos eliminado variables para poder observar mejor las que definen al alumno, por ello consideraremos sólo las variables: edad, sexo, titulación en la que se matricula, estudios y trabajo del alumno, provincia y comunidad autónoma del domicilio familiar, estudios y trabajo de ambos padres, modo de ingreso en la universidad y en el estudio, año de ingreso en la UAL, año en el que comenzó el estudio en el que se encuentra matriculado y si el alumno pertenece a una familia numerosa. Con estas variables, la red que se obtiene es la mostrada en la figura 2.
Gracias a la red podemos descubrir las relaciones de independencia entre las variables, por ejemplo, los estudios del padre y el trabajo del padre influyen en los estudios de la madre del alumno; para que el estudio de la madre sea independiente del trabajo del padre se han de conocer el trabajo que realiza la madre y los estudios del padre.
Cuando un alumno elige una titulación, esta influido directamente por su sexo, los estudios que posea, el modo de ingreso y el año en el que se matriculó por primera vez. Además, la situación del domicilio familiar influye en el estudio elegido salvo

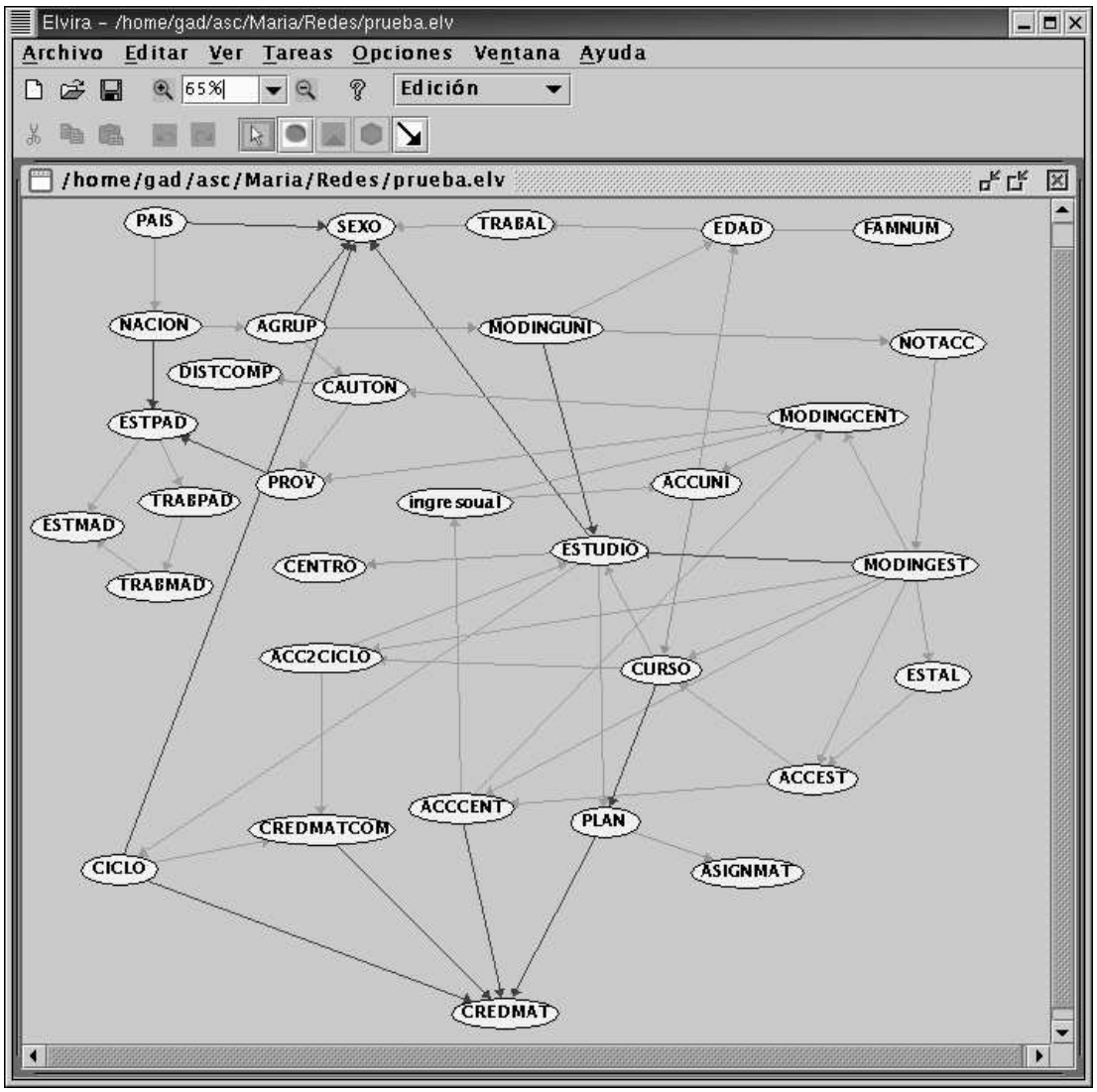

Figura 1: Modelización del alumnado de la UAL en el curso 2000/01.

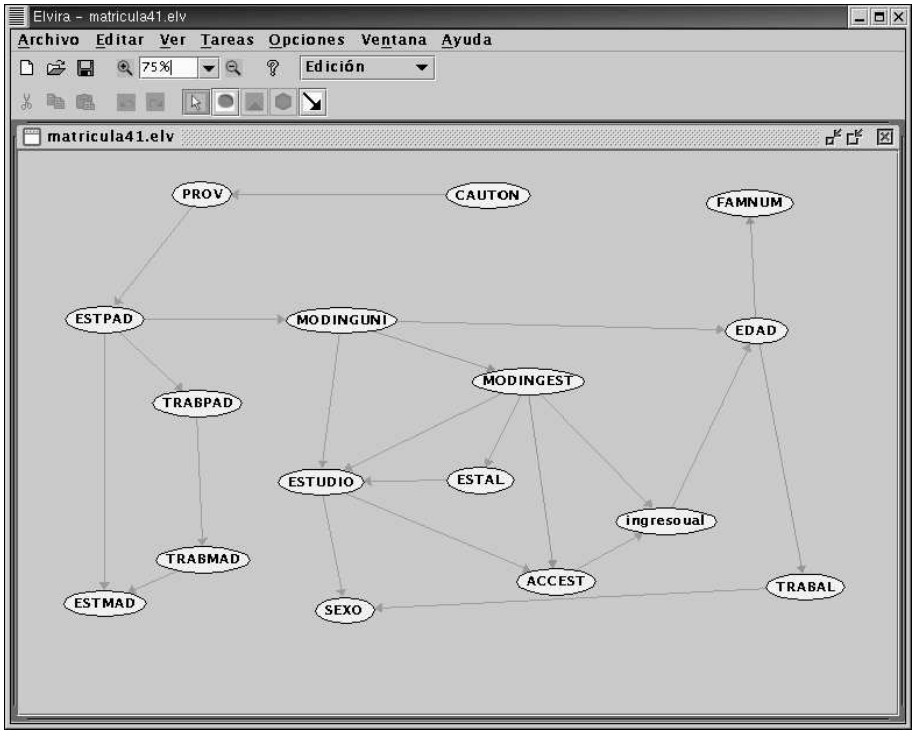

Figura 2: Modelización del alumnado reduciendo el número de variables de interés.

que se tenga alguna evidencia sobre los estudios del padre o el modo de acceso a la Universidad, en cuyo caso ambas variables serán independientes.
Si se conoce la forma con la que el alumno accede a la Universidad, el estudio elegido es independiente de los estudios del padre. La edad también está relacionada con el trabajo del alumno y el número de miembros de su familia en conexión divergente. Ambas variables estarán relacionadas siempre que no se conozca la edad del alumno; en caso de conocerla serán independientes.
Podemos ver cómo el sexo del alumno, el número de horas que trabaja y el estudio elegido están relacionados mediante una conexión convergente, por lo que, si conocemos el sexo del alumno, las otras dos variables son dependientes.
Por último, debido a los múltiples caminos existentes entre el modo de acceso al estudio y la edad del alumno, éstos serán independientes cuando las variables modo de acceso a la Universidad, año de comienzo en la UAL y estudio en el que se matricula, estén observadas. Otra posible combinación que provoca la independencia entre edad y acceso al estudio es cuando se conoce el trabajo del alumno en vez del estudio.

5. Análisis de datos mediante propagación de probabilidades

Como ya hemos comentado anteriormente, la propagación de probabilidades consiste en calcular las probabilidades a posteriori de las variables de interés una vez observado el valor de otras variables. En esta sección realizamos un estudio del alumnado mediante esta técnica.

5.1. Estudios de los padres

Al observar la relación existente entre los estudios de los padres del alumno, hemos propagado fijando el estudio del padre. Las probabilidades obtenidas para los estudios de la madre se muestran en la tabla 1.

PADRE SIN ESTUDIOS

Sin estudios	**0.71704316**
Primarios	0.21975710
Bachill. elem.	0.02095862
Bachill. sup.	0.01723570
Diplomada/sim	0.01877380
Licenciada/sim	0.00623163

PADRE CON PRIMARIOS

Sin estudios	0.10021954
Primarios	**0.7784169**
Bachill. elem.	0.05482064
Bachill. sup.	0.02964386
Diplomada/sim	0.03297228
Licenciada/sim	0.00392677

PADRE CON BACHILL. ELEM.

Sin estudios	0.05755850
Primarios	**0.41764974**
Bachill. elem.	0.36678904
Bachill. sup.	0.07920664
Diplomada/sim	0.06744488
Licenciada/sim	0.01135120

PADRE CON BACHILL. SUP.

Sin estudios	0.05435043
Primarios	0.29644014
Bachill. elem.	0.1891694
Bachill. Sup.	**0.31574579**
Diplomada/sim	0.12198044
Licenciada/sim	0.02231380

PADRE DIPLOMADO/SIM

Sin estudios	0.04092554
Primarios	0.21423414
Bachill. elem.	0.17000457
Bachill. sup.	0.18204787
Diplomada/sim	**0.35177430**
Licenciada/sim	0.04101357

PADRE LICENCIADO/SIM

Sin estudios	0.02824530
Primarios	0.15471582
Bachill. elem.	0.15825322
Bachill. Sup.	0.19590209
Diplomada/sim	**0.24209182**
Licenciada/sim	0.22079174

Cuadro 1: Probabilidades de los estudios de la madre fijado el estudio del padre

Como puede verse en las tablas, el estudio de la madre está muy relacionado con el del padre. En general, el estudio de la madre es del mismo nivel que el del padre,

seguido de cerca de los estudios primarios.
A continuación fijamos el valor de los estudios de la madre y calculamos las probabilidades del estudio del padre. Como puede verse en la tabla 2 el comportamiento es distinto al del caso del padre; la madre tiene, en general, una pareja del mismo nivel de estudios que ella o superior, dejando muy atrás los estudios inferiores a los suyos. Estas diferencias quedan patentes sobre todo en madres con estudios de Bachiller superior o de mayor nivel.

MADRE SIN ESTUDIOS

Sin estudios	**0.60480871**
Primarios	0.26944058
Bachill. elem.	0.04733067
Bachill. sup.	0.03919559
Diplomado/sim	0.02583400
Licenciado/sim	0.01339044

MADRE CON PRIMARIOS

Sin estudios	0.06089480
Primarios	**0.68752335**
Bachill. elem.	0.11282619
Bachill. sup.	0.07023213
Diplomado/sim	0.04442736
Licenciado/sim	0.02409617

MADRE CON BACHILL. ELEM.

Sin estudios	0.02250733
Primarios	0.18764777
Bachill. elem.	**0.38400612**
Bachill. sup.	0.17368960
Diplomado/sim	0.13663015
Licenciado/sim	0.095519022

MADRE CON BACHILL. SUP.

Sin estudios	0.02443914
Primarios	0.13397677
Bachill. elem.	0.10949113
Bachill. Sup.	**0.38278598**
Diplomado/sim	0.19318219
Licenciado/sim	0.15612480

MADRE DIPLOMADA/SIM

Sin estudios	0.02708106
Primarios	0.15160043
Bachill. elem.	0.09484686
Bachill. sup.	0.15044068
Diplomado/sim	**0.37975389**
Licenciado/sim	0.19627708

MADRE LICENCIADA/SIM

Sin estudios	0.03059484
Primarios	0.06144969
Bachill. elem.	0.05433109
Bachill. Sup.	0.09366589
Diplomado/sim	0.15069493
Licenciado/sim	**0.60926356**

Cuadro 2: Probabilidades de los estudios del padre según los estudios de la madre

5.2. Titulación escogida por el alumno

La siguiente pregunta que nos hemos realizado es si algunas características del alumno, como sexo, situación del domicilio familiar, estudios de los padres o situación socioeconómica de éstos, influyen en el estudio que el alumno elige. Para responder a esta pregunta hemos fijado los valores de dichas variables y hemos obtenido las probabilidades de las cuatro titulaciones más probables obteniendo los siguientes resultados:

El sexo del alumno influye claramente en la titulación elegida, mientras que en los hombres las titulaciones con mayor probabilidad son Diplomado en Empresariales (con un 12.91 % del alumnado masculino), Derecho (9.79 %), I.T.A. Hortofruticultura y Jardinería (8.82 %) y Magisterio especialidad en Educación Física (7.70 %), las mujeres prefieren estudiar Diplomado en Empresariales (11.47 %), Derecho (11.33 %), Psicología (10.7 %) y Magisterio especialidad en Educación Infantil (8.63 %).
En la tabla 3 pueden verse las probabilidades de cada titulación dependiendo del sexo del alumno.

TITULACIÓN	PROBABILIDAD	
	Hombre	Mujer
ITA	0.00135400	0.00049414
Explotaciones	0.03102497	0.01068436
Hortofruticultura	0.08838798	0.03784593
Industrias	0.02138901	0.01339349
Mecanización	0.02889537	0.00491129
Infor. Gestión	0.04248047	0.01220966
Infor. Sistemas	0.07074159	0.00652672
Agronomo	0.01848407	0.00479293
Ing. Informática	0.01125051	0.00183311
Enfermería	0.01599980	0.04551432
D. Empresariales	0.13481752	0.11907127
Turismo	0.01737058	0.03847637
LADE	0.06906182	0.06252503
Ing. Químico	0.00991444	0.00928712
Ambientales	0.03422945	0.03461162
Matemáticas	0.02119834	0.01811059
Química	0.02641434	0.02296958
Gestión y A.P.	0.01276531	0.02015814
Derecho	0.09920338	0.11409701
Ed.Física	0.07731960	0.02656397
Ed. Infantil	0.00890975	0.08632264
Ed. Musical	0.01598487	0.01860073
Ed. Primaria	0.02833632	0.06125039
Lengua Extr.	0.01260771	0.03118545
F.Hispánica	0.01230431	0.02726732
F. Inglesa	0.01213027	0.02573779
Humanidades	0.01938299	0.01824956
Psicología	0.03952249	0.10693198
Psicopedagogía	0.00826330	0.01710151

Cuadro 3: Probabilidad de cada titulación según el sexo del alumno

Para estudiar la influencia del domicilio familiar en los estudios del alumno hemos separado a los alumnos almerienses de los alumnos procedentes del resto de provincias de nuestra comunidad autónoma y del resto de España. A los alumnos de Almería los hemos clasificado en distritos postales (los de Almería capital), y los alumnos procedentes de pueblos de la provincia los hemos agrupado en tres zonas: Poniente, Levante e Interior. En la tabla 4 podemos ver cómo apenas varían los porcentajes de las cuatro carreras con más alumnos al cambiar la zona en donde viven.

Distrito	Dip. Empresariales	Derecho	Psicología	LADE
D1	0.1194	0.1094	0.0802	0.0672
D2	0.123	0.1078	0.0786	0.066
D3	0.1205	0.1078	0.0786	0.066
D4	0.1198	0.1090	0.0798	0.069
D5	0.1204	0.1084	0.0792	0.0664
D6	0.1206	0.1075	0.0784	0.0657
D7	0.1206	0.1074	0.0783	0.0656
D8	0.1213	0.1059	0.0769	0.0645
D9	0.1211	0.1062	0.0771	0.0647
Poniente	0.1215	0.1054	0.0763	0.0640
Levante	0.1217	0.1051	0.0760	0.0638
Interior	0.1216	0.1052	0.0762	0.0639

Cuadro 4: Porcentajes de alumnos de Almería en las cuatro titulaciones más frecuentes

Vemos cómo las titulaciones más probables son las mismas en todos los casos y que las probabilidades apenas varían al cambiar de zona. Sin embargo, si añadimos la variable sexo sí se produce una pequeña variación en las dos titulaciones con probabilidad más alta dentro de la tabla de las alumnas (tabla 5).
Nótese cómo en el caso de las mujeres, en el centro de Almería predomina, aunque con poca diferencia, la diplomatura de Empresariales sobre Derecho, mientras que en el resto de los distritos y en la provincia la probabilidad de Derecho aumenta ocupando la primera posición.
En los hombres (tabla 6) no ocurre esto, aunque puede verse cómo la diferencia entre la titulación principal, Empresariales, y las otras tres es mayor que en el caso femenino, lo cual indica una mayor homogeneidad entre los hombres.
Para el caso de los alumnos que proceden de otras provincias andaluzas, los resultados de la propagación, expuestos en la tabla 7, muestran cómo las probabilidades de las titulaciones en cada provincia son muy semejantes; en general las titulaciones elegidas por estos alumnos son Psicología (que es cursada por un 14 % aproximadamente), e ITA. Hortofruticultura y Jardinería (sobre un 11 % del alumnado procedente las otras provincias de Andalucía), seguidas por Magisterio en la especialidad

Distrito	Dip. Empresariales	Derecho	Psicología	Ed. Infantil
D1	0.1165	0.1133	0.1107	0.0843
D2	0.1147	0.1141	0.1085	0.0855
D3	0.1146	0.1142	0.1086	0.0854
D4	0.1161	0.1136	0.1102	0.0845
D5	0.1153	0.1139	0.1094	0.0849
D6	0.1143	0.1144	0.1082	0.0856
D7	0.1142	0.1144	0.1081	0.0857
D8	0.1125	0.1150	0.1061	0.0868
D9	0.1129	0.1148	0.1066	0.0865
Poniente	0.1119	0.1151	0.1053	0.0873
Levante	0.1117	0.1153	0.1040	0.0875
Interior	0.1118	0.1152	0.1052	0.0876

Cuadro 5: Titulaciones más probables escogidas por las alumnas según zona

Distrito	Dip. Empresariales	Derecho	Hortofruticultura	Ed. Física
D1	0.1274	0.1004	0.0876	0.0762
D2	0.1283	0.0989	0.0879	0.0767
D3	0.1286	0.0989	0.0880	0.0768
D4	0.1278	0.1000	0.0877	0.0764
D5	0.1282	0.0995	0.0879	0.0766
D6	0.1287	0.0987	0.0881	0.0769
D7	0.1288	0.0986	0.0881	0.0769
D8	0.1295	0.0973	0.0884	0.0772
D9	0.1293	0.0975	0.0883	0.0771
Poniente	0.1297	0.0968	0.0884	0.0773
Levante	0.1300	0.0966	0.0885	0.0774
Interior	0.1300	0.0967	0.0885	0.0773

Cuadro 6: Titulaciones más probables escogidas por los alumnos varones según zona

de Educación Física, Enfermería y Explotaciones Agropecuarias.

Titulación	Cádiz	Córdoba	Granada	Huelva	Jaén	Málaga	Sevilla
Agronomo	0.04305	0.04303	0.04350	0.04048	0.04330	0.04328	0.04287
Ambientales	0.04873	0.04874	0.04853	0.04991	0.04862	0.04862	0.04881
Derecho	0.02235	0.02244	0.02077	0.03149	0.02147	0.02153	0.02299
Dip. Empres.	0.04423	0.04427	0.04349	0.04854	0.04382	0.04384	0.04453
Dip. Infor.	0.00189	0.00189	0.00189	0.00192	0.00189	0.00189	0.00189
Enfermería	0.07648	0.07644	0.07737	0.07137	0.07698	0.07695	0.07613
Explotac.	0.08880	0.08874	0.08993	0.08231	0.08943	0.08939	0.08835
Gestión y A.P.	0.00508	0.00509	0.00495	0.00582	0.00501	0.00501	0.00513
Hortofrut.	0.11437	0.11431	0.11544	0.10820	0.11496	0.11493	0.11394
Industrias	0.05373	0.05371	0.05416	0.05126	0.05397	0.05396	0.05356
Inf. Gestión	0.02364	0.02364	0.02365	0.02360	0.02364	0.02364	0.02364
Inf. Sistemas	0.01269	0.01270	0.01246	0.01401	0.01256	0.01257	0.01278
Ing. Informática	0.00204	0.00205	0.00199	0.00232	0.00202	0.00202	0.00206
Ing. Quimico	0.00294	0.00295	0.00289	0.00327	0.00291	0.00291	0.00297
LADE	0.01400	0.01403	0.01355	0.01663	0.01375	0.01376	0.01418
Química	0.00653	0.00654	0.00633	0.00768	0.00641	0.00642	0.00661
F. Hispánica	0.00295	0.00296	0.00274	0.00416	0.00283	0.00284	0.00303
F.Inglesa	0.00096	0.00097	0.00070	0.00245	0.00082	0.00083	0.00106
Humanidades	0.01565	0.01565	0.01565	0.01563	0.01565	0.01565	0.01565
Psicología	0.13938	0.13934	0.14018	0.13474	0.13983	0.13980	0.13906
Psicoped.	0.00678	0.00679	0.00670	0.00728	0.00674	0.00674	0.00682
Matemát.	0.00507	0.00507	0.00492	0.00592	0.00498	0.00499	0.00513
Mecanizacion	0.01064	0.01065	0.01060	0.01087	0.01062	0.01062	0.01066
Ed Física	0.09539	0.09537	0.09587	0.09262	0.09566	0.09564	0.09520
Ed Infantil	0.04992	0.04993	0.04981	0.05057	0.04986	0.04986	0.04997
Ed. Musical	0.01358	0.01359	0.01332	0.01507	0.01343	0.01344	0.01368
Ed. Primaria	0.05947	0.05948	0.05947	0.05953	0.05947	0.05947	0.05948
Lengua Extr.	0.01144	0.01145	0.01111	0.01331	0.01126	0.01127	0.01157
Turismo	0.02362	0.02363	0.02350	0.02436	0.02355	0.02356	0.02368

Cuadro 7: Probabilidades de las titulaciones según la provincia del domicilio del alumno

Por último, la tabla 8 recoge los resultados obtenidos para los alumnos procedentes de otras comunidades autónomas. Podemos observar cómo este grupo de alumnos es más heterogéneo: la probabilidad se reparte entre todas las titulaciones destacándose únicamente Derecho en donde se ha matriculado un 14.52 % de los alumnos con domicilio familiar fuera de Andalucía.
Los siguientes experimentos realizados consisten en calcular la probabilidad a posteriori de los estudios que cursan los alumnos dado los estudios y situación socioeconómica de los padres. En la tabla 9 se resumen las probabilidades de las titulaciones más probables según el estudio del padre. Como puede verse, las titulaciones son las mismas que para el caso en el que se desconocen los estudios del padre, y las probabilidades apenas varían al cambiar de un valor a otro de la variable. Se puede apreciar, cómo los alumnos con padres con estudios superiores (diplomados, licenciados o similares) se decantan un poco más hacia Derecho, Psicología y LADE, disminuyendo el porcentaje de estos alumnos que se matriculan en la diplomatura de Empresariales. Resultados análogos se han obtenido al calcular las probabilidades introduciendo como evidencia los estudios de la madre o el trabajo de ambos padres. Por lo tanto, estas variables no parecen influir en el alumnado de nuestra

Titulación	Probabilidad
Explotaciones.	0.01193
Hortofruticultura.	0.02780
Industrias	0.03719
Mecanización	0.01046
Infor. Gestión	0.02058
Infor. Sistemas	0.01896
Agronomo	0.01844
Ing. Informática	0.00419
Enfermería	0.00307
Dip. Empresariales	0.05223
Turismo	0.03497
LADE	0.01713
Ing. Químico	0.00305
Ambientales	0.09505
Matemáticas	0.00887
Química	0.01329
Gestión y A. P.	0.00893
Derecho	0.14521
Ed. Física	0.08310
Ed. Infantil	0.06277
Ed. Musical	0.04563
Ed. Primaria	0.07588
Lengua Extr.	0.04367
F. Hispánica	0.01139
F. Inglesa	0.01508
Humanidades	0.01152
Psicología	0.09932
Psicopedagogía	0.01129

Cuadro 8: Probabilidad de las titulaciones elegidas por los alumnos procedentes de fuera de Andalucia

Universidad a la hora de matricularse en un estudio.

Est.Padre	Dip. Empresariales	Derecho	Psicología	LADE
Sin Estudios	0.1231	0.1023	0.0723	0.0606
Primarios	0.1217	0.1043	0.0757	0.0636
Bachiller elem.	0.1213	0.1080	0.0807	0.0678
Bachiller sup.	0.1209	0.1096	0.0807	0.0678
Diplomado/sim.	0.1178	0.1119	0.0824	0.0689
Licenciado/sim	0.1167	0.1130	0.0832	0.0696

Cuadro 9: Titulaciones con probabilidad más alta según estudios del padre

5.3. Una curiosidad

Para finalizar incluimos, a modo de curiosidad, los resultados obtenidos al propagar dado el valor del año de inicio de los estudios. Con esto obtenemos las titulaciones donde los alumnos permanecen más años. Los resultados se recogen en la tabla 10.

Año Inicio	Titulaciones con mayor Probabilidad			
2000	D.Empresariales 0.1298	Psicología 0.0815	LADE 0.0798	Derecho 0.0655
1999	D.Empresariales 0.1000	Psicología 0.0763	Derecho 0.0740	Ed. Infantil 0.0664
1998	D.Empresariales 0.0915	Derecho 0.0812	Ed. Infantil 0.0774	LADE 0.0753
1997	D.Empresariales 0.1251	Derecho 0.0847	Psicología 0.0740	LADE 0.0692
1996	D.Empresariales 0.1358	Derecho 0.1221	Psicología 0.0950	LADE 0.0807
1995	D.Empresariales 0.2375	Derecho 0.1292	LADE 0.0678	Psicología 0.0613
1994	Derecho 0.2244	Hortofruticultura 0.1772	Matemáticas 0.0738	Psicología 0.0727
1993	Derecho 0.2949	Matemáticas 0.1634	Psicología 0.1009	Química 0.0936

Cuadro 10: Titulaciones con los alumnos más antíguos

6. Extracción de perfiles

Una tarea relacionada con la propagación de probabilidades que se puede llevar a cabo sobre redes bayesianas, y que las hace especialmente interesantes para el análisis de datos, es la llamada *inferencia abductiva* o la búsqueda de las explicaciones más probables (ver Gámez (1998, Nilsson (1998)), que más adelante definiremos formalmente. En términos de análisis de datos, podemos enterder la inferencia abductiva como la búsqueda del perfil más probable de los individuos de una población, bajo determinadas condiciones impuestas por las variables observadas. Este perfil puede venir dado por algunas (abducción parcial) o todas (abducción total) las variables del sistema.

6.1. Abducción total

Definiremos en esta sección los conceptos de explicación y explicación más probable, y veremos cómo se pueden obtener usando el algoritmo HUGIN de Jensen, Lauritzen, and Olesen (1990).

Definición 1. *Sea $G = (\Upsilon, \varepsilon)$ una red bayesiana y x_O una observación del conjunto de variables $X_O \subset \Upsilon$. Se dice que $x \in U$ es una* explicación *de x_O si $x^{\downarrow X_O} = x_O$.*

Ante una observación x_O, está claro que pueden darse muchas explicaciones, pero quizás sólo algunas de ellas sean de interés. Una manera de elegir entre ellas puede ser en función de su probabilidad 'a posteriori'.

Definición 2. *Sea $G = (\Upsilon, \varepsilon)$ una red bayesiana y x_O una observación del conjunto de variables $X_O \subset \Upsilon$. Se dice que $x \in U$ es la* explicación más probable (EMP) *de x_O si*

$$x = \arg \max_V P(\Upsilon | x_O) \ . \tag{1}$$

Obsérvese que la búsqueda de la explicación más probable no es lo mismo que obtener la configuración formada por los valores de cada variable independientemente que tienen mayor probabilidad a posteriori, pues esto sólo sería válido, en general, en el caso de que todas las variables fueran independientes dos a dos.
Una forma de extraer la explicación más probable (Gámez 1998) consiste en aplicar el algoritmo HUGIN pero sustituyendo el operador de marginalización (suma) por el del máximo; es decir, cuando se aplica la marginalización sobre un separador, en lugar de tomar la suma de los valores de probabilidad para los distintos valores de la variables que se está borrando, se toma el máximo valor de probabilidad. A continuación, una vez que se ha hecho un paso de mensajes sobre el árbol de cliques, la configuración más probable viene determinada por la unión de las configuraciones de máxima probabilidad en cada clique.

6.2. Abducción parcial

En muchos casos, especialmente cuando el número del variables del problema es muy alto, estaremos interesados en perfiles en cuanto a un número reducido de características o variables. En ese caso, hablaremos de abducción parcial, que se puede definir formalmente como sigue.

Definición 3. *Sea $G = (\Upsilon, \varepsilon)$ una red bayesiana y x_O una observación del conjunto de variables $X_O \subset \Upsilon$. Sea $X_E \subset \Upsilon$ el conjunto de variables de interés o conjunto explicación. Sea $X_R = \Upsilon \backslash X_E$. Decimos que $x_E \in U_{X_E}$ es la* explicación más probable (EMP)*, en términos de X_E, de x_O, si*

$$x = \arg \max_{X_E} \sum_{X_R} P(X_E, X_R | x_O) \ . \tag{2}$$

6.3. Perfiles de los alumnos matriculados

En nuestro experimento hemos obtenido los cuatro perfiles más probables de un alumno de la Universidad de Almería matriculado en el curso 2000/01. El perfil que hemos definido está compuesto por siete variables: sexo, edad, domicilio familiar, estudios y trabajo de ambos padres o tutores.
Con estas siete variables, la probabilidad a posteriori de primer perfil es muy pequeña, de 0.00146583, y muy próxima a los tres siguientes, con probabilidades 0.00143001, 0.00142169 y 0.0013870 lo que nos indica que el alumnado de nuestra Universidad forma un grupo muy heterogéneo.
El perfil obtenido en los cuatro casos, ha sido el de una mujer, con domicilio en Almería capital y los estudios de ambos padres son primarios; la edad y la situación socioeconómica de los padres varía del siguiente modo :

1. Los dos perfiles más probables corresponden a hijos de trabajadores no cualificados y de edad 20 años (con probabilidad 0.00146583) y 21 (con probabilidad 0.00143001)

2. Los perfiles siguientes corresponden a alumnos cuyo padre es un trabajador cualificado u operador de maquinaria en una industria y la madre no posee trabajo remunerado, sus edades son 20 y 21 años (con probabilidades 0.00142169 y 0.0013870 respectivamente)

A continuación nos hemos interesado por el perfil de los estudiantes varones y el de las mujeres, por lo que hemos introducido una variable observada: el sexo del alumno. El resultado obtenido es el de cuatro perfiles con probabilidades muy pequeñas y próximas entre sí, por lo que ambos grupos son también muy heterogéneos.
En el caso de los alumnos varones, el perfil es semejante al obtenido antes de diferenciar el sexo del alumno: domicilio en Almería capital, sus padres, trabajadores no

cualificados,con estudios primarios y la edad del alumno entre 20 y 22 años (probabilidades 0.00235197, 0.00233564, 0.00231015), el cuarto perfil obtenido, con probabilidad 0.00228116 varía en el trabajo de los padres: el padre trabajador cualificado y la madre sin trabajo remunerado.
En cuanto a las mujeres que estudian en nuestra universidad, las probabilidades de los cuatro perfiles siguen siendo bajas aunque algo más diferenciadas que en el caso de los hombres: 0.00260023, 0.00253668, 0.00252194 y 0.00246031. Los perfiles han sido idénticos a los obtenidos en el caso general.
El siguiente perfil con el que hemos experimentado es el de los alumnos de cada titulación, para ello hemos observado la variable ESTUDIO y fijado como variables de interés el sexo y la edad del alumno, la situación del domicilio familiar y los estudios de ambos padres.
Las probabilidades obtenidas son también bajas, aunque mayores que en el caso general, oscilando entre 0.00718966 de Magisterio especialidad en Lengua Extranjera a 0.02179936 de la especialidad en Educación Musical. Observando estas probabilidades, la facultad cuyos alumnos tienen características más homogéneas es la Facultad de Humanidades y CC. de la Educación, seguida de la Escuela Politécnica Superior, aunque también son las que más dispersión tienen de un estudio a otro; la facultad con alumnos más heterogéneos es la de CC. Económicas y Empresariales y también es la que menor dispersión presenta entre sus titulaciones.
En general, los alumnos de todas las titulaciones proceden de Almería capital y ambos padres poseen estudios primarios, los datos que más varían de una titulación a otra es el sexo y la edad. En las tablas 11 a 14 podemos ver los perfiles de los alumnos de las distintas titulaciones de nuestra Universidad, en orden creciente del valor la probabilidad del perfil .
En los perfiles señalados con *, el domicilio familiar del alumno no se encuentra en Almería capital, sino en un pueblo del interior de la provincia, en el caso del cuarto perfil de Magisterio esp. Ed. Musical, y en un pueblo del poniente almeriense en el caso de las titulaciones Ingeniero Químico, Turismo y Magisterio esp. Lengua Extranjera.
Por último comentar que, si eliminamos la edad del conjunto de las variables de interés, las probabilidades de los perfiles aumenta notablemente, así como la diversidad entre los distintos perfiles dentro de una misma titulación, por ejemplo, los cuatro perfiles más probables para las titulaciones Informática de Sistemas, Informática de Gestión y Mecanización, sin incluir la edad serían los mostrados en la tabla 15.
Nótese cómo aumenta la representatividad del perfil, por ejemplo en el caso de Informática de Sistemas, el primer perfil ha pasado de representar al 1.66 % del alumnado al de esta titulación 11.07 %. Además, la diferencia de probabilidad entre los dos primeros perfiles es mucho mayor que la que se tenía cuando se incluía la edad. Todo esto nos hace suponer que la edad es uno de los factores causantes de la gran heterogeneidad entre los alumnos.

TITULACIÓN	PROBABILIDAD	SEXO	EDAD
Ed. Musical	0.02179936	Mujer	19
	0.01383841	Hombre	19
	0.01247986	Mujer	20
	0.01079565*	Mujer*	19*
Ed. Infantil	0.02007065	Mujer	20
	0.01706847	Mujer	21
	0.01655484	Mujer	19
	0.011343372	Mujer	22
Informática de Sistemas	0.01663724	Hombre	20
	0.01539843	Hombre	19
	0.01464627	Hombre	21
	0.01298762	Hombre	22
Filología Inglesa	0.01633261	Mujer	20
	0.01420155	Mujer	19
	0.01152656	Mujer	21
	0.01039937	Mujer	22
Enfermería	0.01619355	Mujer	20
	0.01353293	Mujer	21
	0.01351515	Mujer	19
	0.01094367	Mujer	22
Ed. Física	0.01453391	Hombre	20
	0.01251103	Hombre	19
	0.01234848	Hombre	21
	0.00982415	Hombre	22
I. T. A Mecanización y construcciones rurales	0.01481497	Hombre	22
	0.01360481	Hombre	23
	0.01204123	Hombre	21
	0.01115793	Hombre	24
Filología Hispánica	0.01432215	Mujer	22
	0.01193358	Mujer	20
	0.01100377	Mujer	23
	0.01081186	Mujer	21

Cuadro 11: Perfiles del alumnado por titulación

TITULACIÓN	PROBABILIDAD	SEXO	EDAD
Psicología	0.01353684	Mujer	20
	0.01278722	Mujer	19
	0.01269323	Mujer	22
	0.01243839	Mujer	21
Ingeniería Informática	0.01243555	Hombre	23
	0.01203537	Hombre	24
	0.01133211	Hombre	22
	0.01080294	Hombre	25
Informática de Gestión	0.01217422	Hombre	19
	0.01214321	Hombre	20
	0.01141193	Hombre	22
	0.01129544	Hombre	21
Gestión y Admón. Pública	0.01176801	Mujer	22
	0.01171350	Mujer	21
	0.01022963	Mujer	20
	0.00954103	Mujer	23
ITA Explotaciones Agropecuarias	0.01101348	Hombre	22
	0.01027528	Hombre	24
	0.01023980	Hombre	21
	0.00980237	Hombre	23
Psicopedagogía	0.01057929	Mujer	24
	0.01006166	Mujer	23
	0.00963072	Mujer	25
	0.00895383	Mujer	22
Ingeniero Agrónomo	0.01053784	Hombre	24
	0.01018148	Hombre	25
	0.00898143	Hombre	23
	0.00812651	Hombre	26
Ambientales	0.01041729	Mujer	20
	0.01014067	Mujer	21
	0.00941829	Mujer	19
	0.00916326	Mujer	22

Cuadro 12: Perfiles del alumnado por titulación

TITULACIÓN	PROBABILIDAD	SEXO	EDAD
Licenciado en Química	0.01016503	Mujer	22
	0.00884710	Mujer	23
	0.00821033	Hombre	22
	0.00743694	Mujer	21
ITA Hortofruticultura y Jardinería	0.01018598	Hombre	21
	0.00989709	Hombre	20
	0.00982630	Hombre	22
	0.00864730	Hombre	24
LADE	0.00973474	Mujer	22
	0.00901181	Mujer	21
	0.00862535	Mujer	20
	0.00817608	Hombre	22
Ingeniero Químico	0.00953066	Mujer	19
	0.00808927*	Mujer(PT)	19*
	0.00722588	Hombre	19
	0.00697319	Mujer	17 o 18
Matemáticas	0.00919575	Mujer	25
	0.00842074	Hombre	25
	0.00735760	Mujer	24
	0.00687814	Mujer	26
ITA Industrias Agrarias y Alimentarias	0.00877693	Hombre	21
	0.00841933	Hombre	22
	0.00803190	Hombre	20
	0.00743640	Hombre	24
Derecho	0.00851583	Mujer	22
	0.00787564	Mujer	20
	0.00768034	Mujer	21
	0.00741057	Mujer	23
Turismo	0.00844524	Mujer	21
	0.00716800*	Mujer (PT)	21*
	0.00654911	Mujer	20
	0.00555940	Mujer	22

Cuadro 13: Perfiles del alumnado por titulación

TITULACIÓN	PROBABILIDAD	SEXO	EDAD
Diplomado Empresariales	0.00791412	Mujer	21
	0.00757226	Mujer	19
	0.00738868	Mujer	22
	0.00738451	Mujer	20
Humanidades	0.00758871	Mujer	22
	0.00733674	Mujer	21
	0.00684438	Mujer	23
	0.00657251	Mujer	20
Lengua Extranjera	0.00718966	Mujer	20
	0.00690315	Mujer	22
	0.00671837	Mujer	21
	0.00610231*	Mujer (PT)	20*

Cuadro 14: Perfiles del alumnado por titulación

ESTUDIO	PROBABILIDAD	SEXO	DOMICILIO	ESTPADRE	ESTMADRE
Inform. Sistemas	0.11072215	H	Capital	Primarios	Primarios
	0.05483267	H	Interior	Primarios	Primarios
	0.04629487	H	Poniente	Primarios	Primarios
	0.02650214	H	Capital	Bachiller El.	Primarios
Inform. Gestión	0.09134038	H	Capital	Primarios	Primarios
	0.04523428	H	Interior	Primarios	Primarios
	0.03819101	H	Poniente	Primarios	Primarios
	0.03415719	H	Capital	Primarios	Primarios
ITA Mecaniz. y Construc. Rurales	0.10138294	H	Capital	Primarios	Primarios
	0.05020764	H	Interior	Primarios	Primarios
	0.04238998	H	Poniente	Primarios	Primarios
	0.02460575	H	Capital	Bachiller El.	Primarios

Cuadro 15: Perfil de los alumnos sin incluir la variable edad

6.4. Perfiles de los alumnos con domicilio familiar fuera de Almería

Para terminar nos hemos preguntado el perfil de los alumnos que, viviendo fuera de Almería, eligen nuestra Universidad para cursar sus estudios. Así hemos diferenciado a los alumnos procedentes de la comunidad Andaluza (exceptuando la provincia de Almería) de los del resto de España, obteniendo en ambos casos el perfil de una **mujer y padres con estudios primarios**, las edades y las probabilidades de cada perfil pueden verse en la tabla 16.

	PROBABILIDAD	EDAD
Resto de Andalucía	0.02771550	20
	0.02630984	21
	0.02444633	22
	0.02431895	19
Resto de España	0.02979146	20
	0.02842081	21
	0.02702040	19
	0.02661795	22

Cuadro 16: Perfiles del alumnado procedente de Andalucía y del resto de España

En ambos casos, el perfil del alumno es el de una mujer, de 20 años de edad y ambos padres con estudios primarios.
Dentro de nuestra comunidad autónoma, en la tabla 17, hemos obtenido el perfil de los alumnos procedentes de cada provincia.
Hay que resaltar que los perfiles de las siete provincias son idénticos entre sí, y las diferencias entre sus probabilidades insignificantes.

7. Agradecimientos

Este trabajo ha sido parcialmente subvencionado por el Ministerio de Ciencia y Tecnología a través del proyecto TIC2001-2973-C05-02 y por la Junta de Andalucía, grupo FQM244 del Plan Andaluz de Investigación.

PROVINCIA	PROBABILIDAD	EDAD
Cádiz	0.027730283	20
	0.026332163	21
	0.024474704	22
	0.024344152	19
Córdoba	0.027730949	20
	0.026333170	21
	0.024475982	22
	0.024345287	19
Granada	0.027718006	20
	0.026313617	21
	0.024451138	22
	0.024323223	19
Huelva	0.027801243	20
	0.026439365	21
	0.024610916	22
	0.024465119	19
Málaga	0.027723866	20
	0.026322470	21
	0.024462387	22
	0.024333214	19
Sevilla	0.027735210	20
	0.026339607	21
	0.024484161	22
	0.024352551	19

Cuadro 17: Perfiles más frecuentes de los alumnos procedentes de otras provincias andaluzas

Referencias

Cano, A., S. Moral, and A. Salmerón (2000). Penniless propagation in join trees. *International Journal of Intelligent Systems 15, 1027–1059.*

Elvira consortium (2002). Elvira: An environment for probabilistic graphical models. In J. Gámez and A. Salmerón (Eds.), Proceedings of the First European Workshop on Probabilistic Graphical Models (PGM'02)*, pp. 222–230.*

Gámez, J. (1998). Abducción en modelos gráficos. In J. Gámez and J. Puerta (Eds.), Sistemas expertos probabilísticos*, pp. 113–140.*

Jensen, F., S. Lauritzen, and K. Olesen (1990). Bayesian updating in causal probabilistic networks by local computation. Computational Statistics Quarterly 4, 269–282.

Lauritzen, S. and D. Spiegelhalter (1988). Local computations with probabilities on graphical structures and their application to expert systems. *Journal of the Royal Statistical Society, Series B 50, 157–224.*

Madsen, A. and F. Jensen (1999). Lazy propagation: a junction tree inference algorithm based on lazy evaluation. Artificial Intelligence 113, 203–245.

Nilsson, D. (1998). An efficient algorithm for finding the m most probable configurations in probabilistic expert systems. *Statistics and Computing 8, 159–173.*

Pearl, J. (1988). Probabilistic reasoning in intelligent systems*. Morgan-Kaufmann (San Mateo).*

Spirtes, P., C. Glymour, and R. Scheines (1993). Causation, prediction and search*, Volume 34 of* Lecture Notes in Statistics*. Springer Verlag.*