Butlletí de la Societat Catalana de Matemàtiques
Vol. 39, núm. 1-2, 2024. Pàg. 131–169. DOI: 10.2436/20.2002.01.117

Tensors en estadística algebraica

Luis Sierra, Marta Casanellas i Piotr Zwiernik

Resum: En estadística i anàlisi de dades trobem tensors contínuament. Els objectes centrals que vinculen la ciència de dades amb la teoria de tensors i l'àlgebra són els models amb variables latents. En aquest article proporcionem una visió general de la teoria de tensors, posant èmfasi especial en les seves aplicacions en estadística algebraica. Reforcem aquest enfocament amb nombrosos exemples amb l'objectiu d'il·lustrar els conceptes clau. A més, s'inclou una revisió extensa de la literatura per guiar els lectors cap a estudis més detallats sobre la matèria.

Paraules clau: tensors, estadística algebraica, models gràfics, descomposició tensorial.

Classificació MSC2020: 14M99, 15A69, 62H22, 62R01, 62R07.

1 Introducció

Els tensors, com a concepte que generalitza les matrius a dimensions superiors, van ser introduïts a finals del segle xix per autors com William Rowan Hamilton, Woldemar Voigt o Gregorio Ricci-Curbastro, i es van fer populars amb el treball [69]. Tot i que els tensors van ser desenvolupats per al càlcul diferencial i la geometria, avui en dia tenen un paper crucial en estadística i ciència de dades, ja que ofereixen un marc matemàtic potent per entendre fenòmens complexos, capturar relacions intricades i extreure intuïcions valuoses de conjunts de dades d'alta dimensió (vegeu el resum [46], per exemple). En aquest article volem mostrar la importància dels tensors en estadística i ciència de dades, centrant-nos en els seus fonaments matemàtics i les connexions amb l'estadística algebraica.

Aplicacions bàsiques. Els tensors tenen aplicacions extenses en diversos àmbits, incloent-hi l'aprenentatge automàtic, la neurociència i la genòmica. En l'aprenentatge automàtic, els tensors faciliten l'anàlisi de dades multidimensionals i permeten tasques com el reconeixement d'imatges, el resum automàtic de vídeos i el processament del llenguatge natural. Per exemple, en

el reconeixement d'imatges s'utilitzen tensors per representar imatges d'alta resolució, capturar relacions espacials i permetre el desenvolupament de models sofisticats d'aprenentatge profund (vegeu [42, 9]).

En neurociència, els tensors tenen un paper crucial per desxifrar els patrons de connectivitat cerebral i comprendre l'organització funcional del cervell (vegeu, per exemple, [28]). Les xarxes cerebrals, representades com a tensors d'ordre superior, capturen les interaccions entre diverses regions cerebrals i proporcionen informació sobre processos cognitius, trastorns neurològics i plasticitat cerebral. Els tensors permeten l'extracció de característiques significatives de les dades d'imatges per ressonància magnètica funcional, cosa que ajuda els investigadors a identificar les regions cerebrals associades a tasques o condicions específiques; vegeu [16].

La genòmica és un altre àmbit en què els tensors ofereixen un marc potent per analitzar dades biològiques d'alta dimensió. En el camp de les dades òmiques, s'empren tensors per integrar diverses característiques moleculars com ara l'expressió gènica, la metilació de l'ADN i les modificacions d'histones (vegeu, per exemple, [94, 60, 74]). Mitjançant la modelització d'aquestes relacions complexes amb tensors, els investigadors són capaços d'identificar signatures moleculars associades a malalties, predir resultats de pacients i orientar estratègies de tractament personalitzades.

Modelització tensorial. La complexitat geomètrica dels tensors planteja reptes en la seva anàlisi i interpretació. Per superar-los s'han desenvolupat mètodes que aprofiten estructures de tensors més simples. Molts dels models aplicats es basen en altres de més bàsics: tensors diagonals i tensors de rang u o de baix rang. Aquesta simplificació millora la interpretabilitat i facilita la computació eficient i la inferència en altes dimensions.

Com veurem, tant els tensors diagonals com els de rang u apareixen de manera natural en el context d'independència de variables aleatòries. En general, molts dels models estadístics que fan servir independència de variables (per exemple, l'anàlisi de components independents, o la independència condicionada) es basen a trobar una descomposició tensorial en termes de tensors de rang inferior.

El camp de l'*estadística algebraica* s'erigeix com el marc adequat per estudiar els tensors dins dels models estadístics (la bibliografia sobre estadística algebraica inclou els llibres [66, 61, 34, 85, 99]). Aprofitant conceptes algebraics, aquesta nova disciplina proporciona nous punts de vista sobre les propietats i el comportament dels tensors en contexts estadístics diversos. Les eines algebraiques s'han aplicat amb èxit per abordar reptes relacionats amb la identificabilitat de models de variables latents, l'estimació de paràmetres, l'ajust de models, el disseny d'experiments i altres problemes estadístics fonamentals.

Variables latents i models gràfics. En ciència de dades, la inclusió de variables *latents* (o *ocultes*, no observades) és una metodologia habitual per a millorar l'expressivitat d'un model sense incrementar dràsticament la seva complexitat

computacional. Malgrat el seu immens potencial, la presència de variables latents introdueix reptes importants en l'estimació i interpretació del model. Per abordar aquests reptes sovint es requereixen algorismes sofisticats i tècniques estadístiques complexes. Això ha constituït una motivació important per al desenvolupament de la teoria d'aprenentatge singular (vegeu [92]) i una gran part dels avenços recents en estadística algebraica se centra en problemes relacionats.

Aprofitant l'estructura algebraica dels tensors, s'ha aconseguit una comprensió més profunda dels models estadístics amb variables latents i s'han desenvolupat algorismes d'estimació eficients. La modelització de la dependència entre les variables aleatòries (tant les observades com les latents) se sol fer mitjançant un *model gràfic*. En un model gràfic, els vèrtexs d'un graf representen variables aleatòries i les seves arestes codifiquen les relacions de dependència condicionada. Quan aquesta representació utilitza arestes dirigides en un graf acíclic, el model també és conegut com a *xarxa bayesiana* o *graf acíclic dirigit*.

Els arbres latents formen la família més manejable de xarxes bayesianes amb variables latents i poden utilitzar-se per modelitzar estructures de dependència quan s'esperen variables de confusió (*confounders*) no observades (vegeu, per exemple, [62, secció 2]. No obstant això, hi ha diverses altres raons per les quals els models d'arbres latents han esdevingut populars en ciències. En primer lloc, els models d'arbres latents engloben un rang més ampli de distribucions de probabilitat que els models d'arbres totalment observats, però mantenen avantatges computacionals, especialment en escenaris d'alta dimensió. A més, l'algorisme *max-product* permet una inferència eficient d'estats no observats. En segon lloc, els arbres poden representar processos evolutius i s'utilitzen com a tals en anàlisi filogenètica (vegeu, per exemple, [80] i la secció 6.1), en lingüística, modelitzant l'evolució del llenguatge (vegeu [70, 76]), i en tomografia de xarxes per inferir estructures d'Internet ([19, 36]). Els models d'arbres latents gaussians (per exemple, els models d'arbres de moviment brownià) capturen correlacions de forma natural per a la tomografia de xarxes gràcies a la disminució de correlacions amb la distància en l'arbre; vegeu la secció 4.3. En tercer lloc, els models d'arbres latents capturen l'estructura jeràrquica en conjunts de dades complexes i estan estretament relacionats amb mètodes de clusterització jeràrquics ([10, 50, 96, 59]). Aquest marc també troba aplicació en visió per computador; vegeu [93] i [24].

Els models d'arbres latents generalitzen els models de Markov ocults (HMM), coincidint amb aquests en arbres eruga ([68]). Els models d'arbres de Markov ocults (HMTM; vegeu [29]) relaxen les restriccions dels HMM i permeten qualsevol estructura d'arbre latent. Les aplicacions inclouen el processament del senyal ([23, 72]), la biomedicina ([52, 65]) i la lingüística ([95]). Aquests models inclouen models filogenètics, models de Bayes primaris, models d'arbres de moviment brownià i un model d'anàlisi factorial, fet que condueix a un marc unificat per a diverses classes de models ([90]).

Un altre exemple important de model de variables latents és l'*anàlisi factorial*, que es fa servir popularment en l'estudi de dades multivariades per

identificar factors latents que influeixen en les variables observades. Aquest camp té aplicacions en finances, psicologia i ciències socials, entre d'altres. Una classe de models relacionats és l'*anàlisi de components principals* (PCA), una tècnica àmpliament utilitzada per a la reducció de la dimensionalitat i la compressió de dades. La versió probabilística, coneguda com a *PCA probabilístic*, introdueix variables latents per tenir en compte el soroll de les dades, i el fa més flexible en aplicacions reals (vegeu la secció 6.3).

Una altra classe important de models gràfics és la de les *màquines de Boltzmann restringides* (RBM; vegi's la secció 6.2), que són components essencials en l'aprenentatge profund, particularment en l'entrenament de xarxes de creença profundes (*deep belief networks*) i altres models generatius. Les RBM involucren variables latents que capturen patrons ocults en les dades i es poden estendre a models gràfics probabilístics amb múltiples capes de variables latents. Han revolucionat camps com la visió per computador i el reconeixement de veu i, des de la perspectiva algebraica, aquesta classe de models ha estat àmpliament estudiada per Montúfar i col·laboradors en [58, 57].

En aquest article expositiu expliquem el paper que tenen els tensors en l'estadística algebraica i en mostrem diversos exemples. A la secció 2 presentem tensors que apareixen de manera natural en estadística i proporcionem els primers resultats que donen resposta a preguntes estadístiques en termes de propietats algebraiques dels tensors. A la secció 3 aprofundim en els models gràfics i presentem exemples de models gràfics dirigits i no dirigits, amb variables discretes o contínues, fent èmfasi especial en els models d'arbres latents. La secció 4 la dediquem al cas de variables gaussianes, que requereix un tractament específic. Basant-nos en els exemples anteriors, a la secció 5 il·lustrem els conceptes de rang tensorial i descomposició tensorial tot explorant el mètode dels moments en aquest context. Finalment, la secció 6 està dedicada a aplicacions addicionals dels tensors en estadística algebraica, en què ens centrem en l'anàlisi filogenètica, les màquines de Boltzmann restringides, i l'anàlisi de components independents.

2 Primers exemples i definicions

En aquesta secció introduïm breument els principals objectes d'estudi: tensors, tensors simètrics, tensors diagonals, tensors de rang u, i els seus noms corresponents en estadística. Ens centrem sobretot en el cas de tensors discrets per a aquesta secció, però alguns resultats també valen per al cas general. Les referències bàsiques per a aquests conceptes inclouen [48, 99, 13, 56].

Una matriu real $r_1 \times r_2$ és una taula dos dimensional, $A = (a_{i_1,i_2})$, on $i_1 \in \{0,\dots,r_1-1\}$, $i_2 \in \{0,\dots,r_2-1\}$, i $a_{i_1,i_2} \in \mathbb{R}$. Donats $m \in \mathbb{N}$ i enters $r_1,\dots,r_m \geq 1$, una *matriu* $r_1 \times \cdots \times r_m$*-dimensional* és una col·lecció de nombres reals $a_{i_1,\dots,i_m}$ on $i_j \in \{0,\dots,r_j-1\}$ per a $j = 1,\dots,m$. Com explicarem a continuació, aquestes matrius multidimensionals s'identifiquen de manera natural amb els *tensors*.

Començant pel cas de les matrius usuals 2-dimensionals, sigui $\{e_i^j\}_{i=0,\dots,r_j-1}$ la base canònica a $\mathbb{R}^{r_j}$, $j = 1,\dots,m$. Una matriu A es pot identificar amb una

aplicació bilineal de $\mathbb{R}^{r_1} \times \mathbb{R}^{r_2}$ en $\mathbb{R}$, $(x, y) \mapsto x^\top A y$ (on l'element (e_i^1, e_j^2) s'envia a $a_{i,j}$). Si $E^{i,j}$ denota la matriu formada per zeros excepte un 1 en l'entrada en la posició (i, j), la seva aplicació bilineal corresponent es denota amb el tensor $e_i^1 \otimes e_j^2$. D'aquesta manera, una matriu A com la d'abans es correspon amb el tensor $\sum a_{i,j} e_i^1 \otimes e_j^2$ en l'espai vectorial de tensors $\mathbb{R}^{r_1} \otimes \mathbb{R}^{r_2}$ (pensat com a l'espai de formes bilineals en $\mathbb{R}^{r_1} \times \mathbb{R}^{r_2}$).

Donats dos vectors $v^1 = (v_0^1, \ldots, v_{r_1-1}^1) \in \mathbb{R}^{r_1}$ i $v^2 = (v_0^2, \ldots, v_{r_2-1}^2) \in \mathbb{R}^{r_2}$, el seu *producte tensorial* es defineix com el tensor $v^1 \otimes v^2 = \sum_{i_1,i_2} v_{i_1}^1 v_{i_2}^2 e_{i_1}^1 \otimes e_{i_2}^2 \in \mathbb{R}^{r_1} \otimes \mathbb{R}^{r_2}$. Per exemple, si $v^1 = (2, 0) \in \mathbb{R}^2$ i $v^2 = (1, 1) \in \mathbb{R}^2$, aleshores $v^1 \otimes v^2 = 2e_0^1 \otimes e_0^2 + 2e_0^1 \otimes e_1^2$ és un tensor diferent de $v^2 \otimes v^1 = 2e_0^1 \otimes e_0^2 + 2e_1^1 \otimes e_0^2$.

Més generalment, un *tensor* $r_1 \times \cdots \times r_m$ sovint es defineix com l'aplicació multilineal $\mathbb{R}^{r_1} \times \cdots \times \mathbb{R}^{r_m} \to \mathbb{R}$: el tensor $e_{i_1}^1 \otimes \cdots \otimes e_{i_m}^m$ es correspon amb l'aplicació multilineal

$$\begin{aligned} \mathbb{R}^{r_1} \times \cdots \times \mathbb{R}^{r_m} &\longrightarrow \mathbb{R} \\ (e_{j_1}^1, \ldots, e_{j_m}^m) &\longmapsto \begin{cases} 1 & \text{si } (j_1, \ldots, j_m) = (i_1, \ldots, i_m), \\ 0 & \text{altrament} \end{cases} \end{aligned}$$

i aquests tensors formen una base natural de l'espai vectorial $\mathbb{R}^{r_1} \otimes \cdots \otimes \mathbb{R}^{r_m}$ dels tensors $r_1 \times \cdots \times r_m$. En aquest sentit, el tensor $t = \sum_{i_1,\ldots,i_m} a_{i_1,\ldots,i_m} e_{i_1}^1 \otimes \cdots \otimes e_{i_m}^m$ en aquest espai es correspon amb la matriu multidimensional $A = [a_{i_1,\ldots,i_m}]_{i_j=0,\ldots,r_j-1}$ de dimensions $r_1 \times \cdots \times r_m$. Quan parlem de coordenades d'un tensor, ens referim a les seves coordenades en la base natural.

Com abans, donats m vectors

$$v^1 = (v_0^1, \ldots, v_{r_1-1}^1) \in \mathbb{R}^{r_1}, \ldots, v^m = (v_0^m, \ldots, v_{r_m-1}^2) \in \mathbb{R}^{r_m},$$

el seu *producte tensorial* és el tensor $v^1 \otimes \cdots \otimes v^m = \sum_{i_1,\ldots,i_m} v_{i_1}^1 v_{i_2}^2 \cdots v_{i_m}^m e_{i_1}^1 \otimes \cdots \otimes e_{i_m}^m$.

Diem que un tensor T de $\otimes^r \mathbb{R}^n = \mathbb{R}^n \otimes \overset{r}{\cdots} \otimes \mathbb{R}^n$ és *simètric* si $T_{i_1 \cdots i_r} = T_{j_1 \cdots j_r}$ per a tota permutació $(j_1, \ldots, j_r)$ de $(i_1, \ldots, i_r)$. L'espai de tensors simètrics de $\otimes^r \mathbb{R}^n$ es denota per $S^r(\mathbb{R}^n)$. La dimensió de $S^r(\mathbb{R}^n)$ és $\binom{n+r-1}{r}$, donat que un tensor $T \in S^r(\mathbb{R}^n)$ es pot codificar amb les entrades $T_{i_1 \cdots i_r}$, on $1 \le i_1 \le \cdots \le i_r \le n$.

2.1 Mesures discretes com a tensors

Considerem una col·lecció de m variables aleatòries discretes, cadascuna amb r_i possibles estats, $X_i \in \mathcal{X}_i = \{0, \ldots, r_i - 1\}$ amb $i = 1, \ldots, m$. El vector aleatori $X = (X_1, \ldots, X_m)$ pren valors a $\mathcal{X} = \mathcal{X}_1 \times \cdots \times \mathcal{X}_m$ i una distribució de probabilitat $p = (p(x))_{x \in \mathcal{X}}$ de X es pot identificar amb un punt a

$$\mathbb{R}^{\mathcal{X}} := \mathbb{R}^{\mathcal{X}_1} \otimes \cdots \otimes \mathbb{R}^{\mathcal{X}_m},$$

on $\mathbb{R}^{\mathcal{X}_i}$ és una còpia de $\mathbb{R}^{r_i}$ amb coordenades indexades per $\mathcal{X}_i$. Les coordenades de l'espai tensorial $\mathbb{R}^{\mathcal{X}}$ es denoten amb p_x, $x \in \mathcal{X}$. Per exemple, si $m = 2$,

$r_1 = r_2 = 2$, aleshores cada X_i és binària i les coordenades de $p \in \mathbb{R}^{\mathcal{X}}$ són p_{00}, p_{01}, p_{10}, p_{11}, $p = p_{00}e_0 \otimes e_0 + p_{01}e_0 \otimes e_1 + p_{10}e_1 \otimes e_0 + p_{11}e_1 \otimes e_1$ (ometem els superíndexs perquè en aquest cas els dos espais són el mateix).

Per definició, totes les distribucions de probabilitat discretes es poden representar al símplex $\Delta^{\mathcal{X}} \subseteq \mathbb{R}^{\mathcal{X}}$, on

$$\Delta^{\mathcal{X}} := \left\{ p \in \mathbb{R}^{\mathcal{X}} : p_x \geq 0, \sum_{x \in \mathcal{X}} p_x = 1 \right\}.$$

Qualsevol model estadístic per a X és, per definició, una família de distribucions de probabilitat i, per tant, una família de punts de $\Delta^{\mathcal{X}}$. Això dona una identificació bàsica dels models estadístics discrets amb subconjunts de l'espai tensorial $\mathbb{R}^{\mathcal{X}}$.

Quan es fan consideracions geomètriques referents als tensors, sol ser més convenient treballar sobre l'espai projectiu. Denotem amb $\mathrm{Proj}(V)$ la projectivització d'un espai vectorial V i sigui $\mathbb{P}^{\mathcal{X}}_{\mathbb{R}} := \mathrm{Proj}(\mathbb{R}^{\mathcal{X}})$. Per definició, aquest és el conjunt de punts $\mathbb{R}^{\mathcal{X}} \setminus \{0\}$, identificant aquells que es troben sobre la mateixa recta que passa per l'origen. Observem que el símplex de probabilitat $\Delta^{\mathcal{X}}$ està en correspondència amb la part no negativa de $\mathbb{P}^{\mathcal{X}}_{\mathbb{R}}$. En efecte, la bijecció envia un punt $q \in \mathbb{P}^{\mathcal{X}}_{\mathbb{R}}$ amb coordenades q_x no negatives a $p = (p_x)_{x \in \mathcal{X}} \in \Delta^{\mathcal{X}}$, on $p_x = q_x / \sum_{y \in \mathcal{X}} q_y$.

La immersió en el projectiu és sovint convenient perquè molts models de dependències complexes sovint estan definits mòdul una constant de normalització (vegeu, per exemple, els models definits a (2)).

2.2 Model d'independència i tensors de rang u

Donada una col·lecció de variables discretes $X = (X_1, \ldots, X_m) \in \mathcal{X}$, diem que les components de X són independents (o que X pertany al *model d'independència total*) si el tensor $p \in \Delta^{\mathcal{X}}$ corresponent a la distribució de X es pot escriure com el producte tensorial de les distribucions individuals $p^i \in \mathbb{R}^{\mathcal{X}_i}$, $p = p^1 \otimes \cdots \otimes p^m$. En altres paraules,

$$p_{i_1 i_2 \cdots i_m} = p^1_{i_1} p^2_{i_2} \cdots p^m_{i_m}, \quad \text{per tot } x = (i_1, \ldots, i_m) \in \mathcal{X}. \tag{1}$$

Per tenir un equivalent geomètric d'aquesta construcció, diem que $p \in \mathbb{R}^{\mathcal{X}}$ és un *tensor de rang u* (o un *tensor pur o descomponible*) si és el producte tensorial d'una m-tupla de vectors $p^i \in \mathbb{R}^{\mathcal{X}_i}$ per a $i = 1, \ldots, m$. El model d'independència per al vector X està contingut en el conjunt de tensors de rang u. Quan $m = 2$, els tensors de rang u es corresponen amb les matrius de rang u. En efecte, $p = p^1 \otimes p^2$ és el tensor $\sum_{i,j} p^1_i p^2_j e^1_i \otimes e^2_j$, que, amb la correspondència introduïda a dalt, es pot identificar amb la matriu de rang u

$$\begin{pmatrix} p^1_0 \\ \vdots \\ p^1_{r_1-1} \end{pmatrix} \begin{pmatrix} p^2_0 & \cdots & p^2_{r_2-1} \end{pmatrix}.$$

En el context de la geometria algebraica projectiva, el conjunt de tensors de rang u es correspon amb la varietat de Segre real

$$\mathrm{Seg}(\mathbb{P}_{\mathbb{R}}^{\mathcal{X}_1} \times \cdots \times \mathbb{P}_{\mathbb{R}}^{\mathcal{X}_m}) \subset \mathbb{P}_{\mathbb{R}}^{\mathcal{X}} = \mathrm{Proj}(\mathbb{R}^{\mathcal{X}_1} \otimes \cdots \otimes \mathbb{R}^{\mathcal{X}_m});$$

vegeu [13, 10.5.12]. Per exemple, $\mathrm{Seg}(\mathbb{P}_{\mathbb{R}}^1 \times \mathbb{P}_{\mathbb{R}}^1)$ és la immersió de $\mathbb{P}_{\mathbb{R}}^1 \times \mathbb{P}_{\mathbb{R}}^1$ a $\mathbb{P}_{\mathbb{R}}^3 = \mathrm{Proj}(\mathbb{R}^2 \otimes \mathbb{R}^2)$ donada per l'aplicació

$$\begin{aligned} \mathbb{P}_{\mathbb{R}}^1 \times \mathbb{P}_{\mathbb{R}}^1 &\longrightarrow \mathbb{P}_{\mathbb{R}}^3 \\ ([p_0^1 : p_1^1], [p_0^2 : p_1^2]) &\longmapsto [p_0^1 p_0^2 : p_0^1 p_1^2 : p_1^1 p_0^2 : p_1^1 p_1^2]. \end{aligned}$$

Aquesta varietat està formada pels punts $[p_{00} : p_{01} : p_{10} : p_{11}]$ a $\mathbb{P}_{\mathbb{R}}^3$ que satisfan l'equació $p_{00}p_{11} - p_{01}p_{10} = 0$ i es correspon amb els tensors de rang u, $p^1 \otimes p^2$, o matrius

$$\begin{pmatrix} p_{00} & p_{01} \\ p_{10} & p_{11} \end{pmatrix}$$

de rang u. Anàlogament, si pensem els elements de $\mathbb{R}^{\mathcal{X}_1} \otimes \mathbb{R}^{\mathcal{X}_2}$ com a matrius, $\mathrm{Seg}(\mathbb{P}_{\mathbb{R}}^{\mathcal{X}_1} \otimes \mathbb{P}_{\mathbb{R}}^{\mathcal{X}_2})$ és el conjunt de matrius $r_1 \times r_2$ de rang u (definides pels menors 2×2 nuls). De fet, es pot provar fàcilment que la part no negativa de $\mathrm{Seg}(\mathbb{P}_{\mathbb{R}}^{\mathcal{X}_1} \times \cdots \times \mathbb{P}_{\mathbb{R}}^{\mathcal{X}_m})$ és isomorfa al model d'independència total.

2.3 Tensors de moments - cas discret

Considerem les mesures discretes de la secció 2.1. Donats $u = (u_1, \ldots, u_m) \in \mathbb{N}^m$ i un vector $x = (x_1, \ldots, x_m) \in \mathcal{X}$, definim el monomi

$$x^u := x_1^{u_1} \cdots x_m^{u_m},$$

on fem servir la convenció $0^0 = 1$. Si X té distribució p, aleshores el *moment* corresponent és

$$\mu_u = \mathbb{E}[X^u] = \sum_{x \in \mathcal{X}} p_x x^u$$

i es diu que és un *moment d'ordre k* si $k = u_1 + \cdots + u_m$. En particular, $\mu_{0\cdots0} = 1$.

EXEMPLE 1. Considerant el cas binari $(X_1, X_2) \in \{0,1\}^2$, es té $\mu_{00} = 1$, $\mu_{10} = p_{10} + p_{11}$, $\mu_{01} = p_{01} + p_{11}$, $\mu_{11} = p_{11}$.

PROPOSICIÓ 1. *L'aplicació* $\mathbb{R}^{\mathcal{X}} \to \mathbb{R}^{\mathcal{X}}$ *definida com a*

$$\mu_u = \mathbb{E}[X^u], \quad \textit{per a tot } u \in \mathcal{X},$$

és una bijecció lineal que envia el conjunt $\{\sum_{x \in \mathcal{X}} p_x = 1\}$ *al conjunt* $\{\mu_{0\cdots0} = 1\}$.

Hi ha altres maneres de codificar una distribució de probabilitat discreta que no sigui només amb els seus moments. Un exemple important és el dels *cumulants*. Des del punt de vista geomètric, tractar amb cumulants correspon a fer un canvi de variables (no lineal) que pot ser més adient per estudiar la geometria de segons quins models; vegeu [83, 98, 25]. Si considerem l'exemple 1, podem emprar la transformació en cumulants: $\kappa_{00} = 0$, $\kappa_{10} = \mu_{10}$, $\kappa_{01} = \mu_{01}$, i $\kappa_{11} = \mu_{11} - \mu_{10}\mu_{01}$. En aquest cas la imatge del model d'independència per a aquesta aplicació és el subespai lineal donat per $\kappa_{11} = 0$.

3 Models gràfics

Molts models en estadística i aprenentatge automàtic involucren modelitzar la dependència entre diverses components d'un vector aleatori. Entre els múltiples exemples trobem l'anàlisi factorial, les màquines de Boltzmann restringides, o el model de Bayes primari. Els models gràfics constitueixen el llenguatge natural per a aquestes tasques de modelització multivariant, ja que representen les relacions de dependència mitjançant grafs. En aquest context, els vèrtexs d'un graf representen variables aleatòries i les arestes codifiquen les relacions de dependència condicionada entre elles. Segons el tipus d'arestes, podem distingir dos tipus principals de models gràfics: els que tenen arestes dirigides i els que no, cadascun adient per a diferents tasques.

Per a ambdós tipus de models gràfics és convenient començar per la noció de *distribució factoritzada*. Donat un vector aleatori $X = (X_1, \dots, X_m)$ que pren valors a $\mathcal{X}$ com abans i donat $\mathcal{F}$ un conjunt de subconjunts de $[m]$, podem considerar el conjunt de distribucions sobre $\mathcal{X}$ que satisfan

$$p(x) = \frac{1}{Z} \prod_{F \in \mathcal{F}} \phi_F(x_F), \quad \text{per tot } x \in \mathcal{X}, \tag{2}$$

on ϕ_F són funcions no negatives sobre $\mathbb{R}^{\mathcal{X}_F}$ (anomenades *potencials*) i Z és la constant normalitzadora.

Un exemple trivial és el de (1), on la distribució factoritza en components individuals. En aquest cas, els factors es corresponen amb distribucions marginals, però aquesta propietat no és certa en general.

Exemple (Model d'Ising). El model d'Ising considera distribucions a $\mathcal{X} = \{0,1\}^m$ definides per

$$p(x) = \frac{1}{Z} \prod_{i=1}^{m} e^{h_i x_i} \prod_{i<j} e^{J_{ij} x_i x_j} = \frac{1}{Z} e^{h^\top x + \frac{1}{2} x^\top J x},$$

on $h = [h_i] \in \mathbb{R}^m$, $J = [J_{ij}] \in \mathbb{R}^{m \times m}$, és simètrica amb zeros a la diagonal, i Z és la constant normalitzadora

$$Z = \sum_{x \in \{0,1\}^n} e^{h^\top x + \frac{1}{2} x^\top J x}.$$

Es tracta, per tant, d'una distribució factoritzada. En moltes aplicacions d'aquest model, J té suport en un graf G (i en aquest cas, $J_{ij} = 0$, sempre que no hi hagi una aresta connectant i amb j a G). Des del punt de vista algebraic, aquest model es coneix com a *model gràfic binari* i ha estat estudiat a [31].

Si $\mathcal{X}$ és finit com a la secció 2.1, està clar que (2) imposa restriccions algebraiques sobre el tensor de probabilitat $p \in \mathbb{R}^{\mathcal{X}}$. Més endavant estudiarem algunes de les formes que poden prendre aquestes restriccions. En aquest punt només val la pena assenyalar que aquestes propietats de factorització són equivalents als models de xarxa de tensors, àmpliament considerats en aplicacions industrials diverses. L'article [71] explora aquesta connexió amb més detall.

3.1 Grafs acíclics dirigits

Sigui $\mathcal{G} = (V, E)$ un graf dirigit. Si $(a, b) \in E$ és una aresta del graf, també designada per $\{a \to b\}$, diem que a és un *predecessor* de b, i b un *successor* de a. Un graf dirigit $\mathcal{G}$ és acíclic (breument DAG) si no conté cicles dirigits. Si $v \in V$ és un vèrtex d'un DAG $\mathcal{G}$, es denota amb $\mathbf{pa}(v)$ el conjunt dels *predecessors* de v i amb $\mathbf{de}(v)$ el conjunt de *descendents* de v format pels vèrtexs w tals que existeix un camí dirigit $\{v \to w\} \subseteq \mathcal{G}$. Contràriament, també podem definir $\mathbf{nd}(v)$ com el conjunt de *no descendents* $V \setminus (\{v\} \cup \mathbf{de}(v))$.

Donat un DAG $\mathcal{G} = (V, E)$ amb $V = \{1, \ldots, m\}$, assignem una variable aleatòria X_v a cada node $v \in V$. El model probabilístic associat al vector aleatori $X = (X_1, \ldots, X_m)$ i al DAG $\mathcal{G}$, també anomenat *xarxa bayesiana*, pren valors a $\mathcal{X} = \prod_{i=1}^{m} \mathcal{X}_i$ i conté totes les distribucions sobre $\mathcal{X}$ que satisfan la factorització recursiva següent:

$$p(x) = \prod_{i=1}^{m} p(x_i | \mathbf{pa}(x_i)), \tag{3}$$

on, per a $X_C = \mathbf{pa}(X_i)$, tenim que $p(x_i|x_C)$ denota la funció de densitat de la probabilitat condicionada de X_i donat $X_C = x_C$, especificada per

$$p(x_i|\mathbf{pa}(x_i)) = \frac{p(x_i, \mathbf{pa}(x_i))}{p(\mathbf{pa}(x_i))},$$

si $p(\mathbf{pa}(x_i)) > 0$.

NOTA 1. Per simplificar la notació, utilitzem la mateixa lletra per denotar totes les densitats condicionades. Per evitar confusions, un estat denotat en minúscula amb el mateix nom que una variable aleatòria (en majúscula) es refereix a tal observació de la variable aleatòria.

En aquest article ens centrem en la situació en què el vector aleatori X és exclusivament o bé discret, o bé continu, però també són d'interès les situacions mixtes; vegi's [49].

Donades tres variables aleatòries X, Y, Z, diem que X és independent de Y donada Z (denotat amb $X \perp\!\!\!\perp Y|Z$), si la distribució condicionada satisfà $p(x,y|z) = p(x|z)p(y|z)$ per a tot x, y, z (o, equivalentment, $p(x|y,z) = p(x|z)$). La factorització (3) permet una interpretació important en termes d'independència condicionada coneguda com a *propietat de Markov dirigida local*:

$$X_v \perp\!\!\!\perp X_{\mathbf{nd}(v)\setminus\mathbf{pa}(v)} \mid X_{\mathbf{pa}(v)}, \quad v \in V. \tag{4}$$

EXEMPLE 2. Considerem l'estructura d'una *cadena de Markov*, en què les variables aleatòries $X_1,\ldots,X_m$ satisfan la propietat $p(x_k|x_{k-1},\ldots,x_1) = p(x_k|x_{k-1})$ per a tot $k = 2,\ldots,m$ (col·loquialment l'interpretem com que el futur només es veu afectat pel present immediat, però no pel passat). Degut a aquesta propietat d'independència condicionada, una cadena de Markov es pot representar com un model probabilístic sobre un graf línia en què cada node representa un estat (modelitzat com una variable aleatòria) X_k i X_{k-1} és el seu únic predecessor directe, donat que X_k és independent del «passat» quan coneixem X_{k-1}. Si $m = 3$, el graf associat és el que il·lustrem a la figura 1(a).

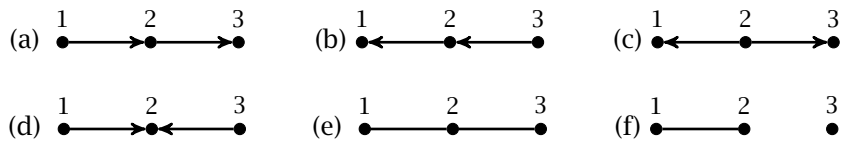

FIGURA 1: Diversos grafs sobre tres nodes.

L'anàlisi de models DAG pot portar a observacions inesperades. Per exemple, el graf (a) de la figura 1 defineix exactament el mateix model que els grafs (b) i (c). En particular, sense hipòtesis addicionals, no existeix una bijecció entre models en DAG i els seus grafs corresponents [89, 21].

EXEMPLE (COL·LISIONADORS). Considerem un vector aleatori format per tres variables aleatòries discretes X_1, X_2, X_3, en què el graf de dependència està donat per la figura 1(d). En aquest model, la funció de densitat es pot factoritzar com a $p(x_1,x_2,x_3) = p(x_2|x_1,x_3)p(x_1)p(x_3)$. Observeu que la distribució marginal de (X_1,X_3) es factoritza com a $p(x_1,x_3) = p(x_1)p(x_3)$ i, per tant, aquestes dues variables són independents. No obstant això, en condicionar X_1, X_3 respecte de $X_2 = x_2$, podem observar que es trenca la propietat d'independència, ja que

$$p(x_1,x_3|x_2) = \frac{p(x_2|x_1,x_3)p(x_1)p(x_3)}{p(x_2)}.$$

Així doncs, tot i que $X_1 \perp\!\!\!\perp X_3$, tenim que $X_1 \not\perp\!\!\!\perp X_3|X_2 = x_2$. Intuïtivament, el fenomen es correspon amb la idea que, malgrat que dos progenitors siguin (genèticament) independents, si observem un fill en comú, tot el que no vingui explicat per un dels progenitors haurà de ser explicat per l'altre, i, per tant,

deixen de ser independents. Aquesta configuració s'anomena *col·lisionador* (*collider* en anglès) i és una coneguda font de confusió en inferència causal, també coneguda com a *paradoxa de Berkson*.

EXEMPLE (DAG ESTRELLA). Considerem un DAG senzill sobre quatre nodes amb arestes dirigides $(0,1)$, $(0,2)$ i $(0,3)$ com el de la figura 2. El model probabilístic associat correspon a la factorització

$$p(x) = p(x_0)p(x_1|x_0)p(x_2|x_0)p(x_3|x_0), \quad \text{per a tot } x \in \mathcal{X}.$$

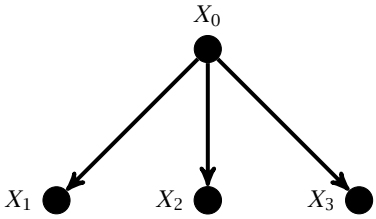

FIGURA 2: Model gràfic sobre quatre nodes amb un predecessor i tres descendents.

Suposem que $\mathcal{X} = \{0,1\}^4$. Cada distribució condicionada $p(x_i|x_0)$, $i = 1,2,3$, es pot representar per una *matriu estocàstica* 2×2 (una matriu no negativa en què totes les files sumen 1)

$$A^i = \begin{pmatrix} p(x_i = 0|x_0 = 0) & p(x_i = 1|x_0 = 0) \\ p(x_i = 0|x_0 = 1) & p(x_i = 1|x_0 = 1) \end{pmatrix}.$$

Denotant amb $\pi_x = p(X_0 = x)$ i $a^i_{x,y}$ l'entrada (x,y)-èsima de la matriu A_i, veiem que el model està parametritzat per

$$p_{x_0x_1x_2x_3} = \pi_{x_0} a^1_{x_0,x_1} a^2_{x_0,x_2} a^3_{x_0,x_3}. \tag{5}$$

Observem que la dimensió de l'espai dels catorze paràmetres π_x, $a^i_{x,y}$ és set perquè les sumes de les files de A^i és u i $\pi_0 + \pi_1 = 1$. De l'equació (5) és senzill veure que, fixant l'estat a X_0 igual a i, les matrius

$$\begin{pmatrix} p_{i000} & p_{i001} \\ p_{i010} & p_{i011} \\ p_{i100} & p_{i101} \\ p_{i110} & p_{i111} \end{pmatrix}, \quad \begin{pmatrix} p_{i000} & p_{i010} \\ p_{i001} & p_{i011} \\ p_{i100} & p_{i110} \\ p_{i101} & p_{i111} \end{pmatrix}, \quad \begin{pmatrix} p_{i000} & p_{i100} \\ p_{i001} & p_{i101} \\ p_{i010} & p_{i110} \\ p_{i011} & p_{i111} \end{pmatrix}$$

són de rang u. D'aquí, podem verificar fàcilment les condicions

$$p_{ijkl}p_{ij'k'l'} = p_{ijkl'}p_{ij'k'l} = p_{ijk'l}p_{ij'kl'} = p_{ij'kl}p_{ijk'l'}, \quad \text{per a } i,j,j',k,k' \in \{0,1\}.$$

D'aquesta manera, les condicions d'independència es tradueixen en relacions polinomials entre les entrades del vector de probabilitats p. En termes de geometria algebraica, podem pensar en la factorització (5) com l'aplicació polinomial

$$\varphi\colon \mathbb{R}^7 \longrightarrow \otimes^4 \mathbb{R}^2,$$

que envia cada conjunt de set paràmetres lliures a la distribució p. Tant des del punt de vista estadístic com de geometria algebraica, és d'interès conèixer un conjunt generador minimal d'equacions polinomials que defineixen el model (és a dir, que s'anul·lin precisament en la imatge de φ). En aquest cas, és conegut que l'ideal de polinomis que s'anul·len a $\mathrm{Im}(\varphi)$ és generat minimalment per divuit equacions quadràtiques de les que hem llistat a dalt; vegeu [40].

Aquests són el tipus de preguntes que es troben en el centre de l'estadística algebraica: per models d'estadística algebraica, utilitzar tècniques de geometria algebraica i àlgebra commutativa per trobar equacions en polinomis i restriccions semialgebraiques que defineixin el model. Conèixer-les pot ser útil per esbrinar si els paràmetres són identificables o no, per estimar els paràmetres, per inferència de versemblança i per ajustar el model.

3.2 Models gràfics no dirigits

Sigui $\mathcal{G}$ un graf no dirigit i sigui C el conjunt de cliques (subgrafs complets) maximals de $\mathcal{G}$. Un *model gràfic no dirigit* és un conjunt de distribucions de probabilitat de tipus

$$f(x) = \frac{1}{Z} \prod_{C \in C} \psi_C(x_C), \quad \text{on } Z = \int_{\mathcal{X}} \prod_{C \in C} \psi_C(x_C)\, \mathrm{d}x, \tag{6}$$

i $\psi_C\colon \mathcal{X}_C \to [0, \infty)$ es coneixen com a *funcions potencials* degut a la seva interpretació en el camp de la física estadística. Aquest és un cas particular de l'expressió a (2) prenent com a $\mathcal{F}$ el conjunt C de cliques de $\mathcal{G}$.

Aquesta definició pot semblar una mica allunyada de les nocions d'independència condicionada que hem descrit per als grafs dirigits, però el 1971, en un article no publicat, John Hammersley i Peter Clifford varen demostrar que, sota la condició $p(x) > 0$ per a tot $x \in \mathcal{X}$, la factorització (6) en un graf no dirigit $\mathcal{G}$ és equivalent a les *propietats globals de Markov*: per a qualssevol subconjunts $A, B, C \subset V$ es té $X_A \perp\!\!\!\perp X_B | X_C$ sempre que C *separi* A, B en $\mathcal{G}$ (és a dir, que tots els camins entre A i B passin per C).

Reprenem el cas de variables discretes de la secció 2.1. Sigui $i_C := (i_j)_{j \in C} \in \mathcal{X}_C$ un estat del vector X_C per una clica $C \subseteq V$. Llavors, la funció potencial $\psi_C\colon \mathcal{X}_C \to [0, \infty)$ es pot veure com un vector $\theta^{(C)}$ amb entrades $\theta^{(C)}_{i_C} = \psi_C(i_C) \in [0, \infty)$. Tal com vam veure a l'exemple DAG estrella, això ens dona la parametrització

$$p_{i_1 i_2 \cdots i_n} = \frac{1}{Z(\theta)} \prod_{C \in C} \theta^{(C)}_{i_C}. \tag{7}$$

D'aquesta manera, aconseguim de nou un model paramètric amb una estructura semialgebraica. Podríem representar una distribució que satisfés (7) com un punt a $\mathbb{P}_{\mathbb{R}}^{\mathcal{X}}$ (ignorant la constant normalitzadora). Veiem que la família de tots aquests punts admet una parametrització monomial i com a tal forma una varietat tòrica en terminologia de geometria algebraica; vegi's [85] per a més detalls.

3.3 Models gràfics latents

Els models gràfics latents estan dissenyats per gestionar situacions en què determinades variables Z no es poden observar de manera directa però tenen un paper crucial a l'hora d'explicar patrons i relacions en les dades observades X. De fet, l'ús de variables latents permet, fent servir un menor nombre de variables, assolir una major expressibilitat del model, és a dir, capturar una major complexitat i variabilitat de les dades. És per això que aquestes variables *latents*, també conegudes com a variables *ocultes*, estan presents en moltes àrees i són essencials per modelitzar fenòmens complexos. Malgrat això, com ja hem esmentat anteriorment, aquests models presenten dificultats a l'hora de fer inferència.

EXEMPLE (MODEL BAYES PRIMARI). Considerem un problema de classificació sobre unes dades de variables discretes $X = (X_1, X_2, \ldots, X_m)$ amb una altra variable discreta Z codificant a quina de les k classes $C = \{c_1, c_2, \ldots, c_k\}$ pertany una observació. Si la classe és desconeguda per a noves observacions, la podem modelitzar com una variable latent. La hipòtesi que es fa servir en el model Bayes primari (també anomenat *classificador bayesià naïf*) és que els atributs són condicionalment independents si es coneix la classe, és a dir, $X_i \perp\!\!\!\perp X_j \mid Z$. Aquesta propietat es pot representar com un model gràfic acíclic dirigit (DAG) en el graf següent:

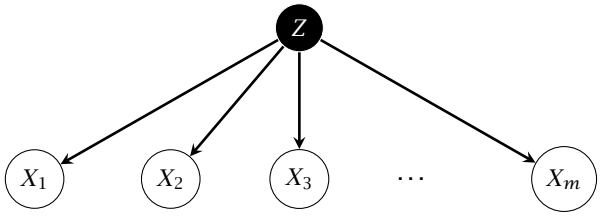

FIGURA 3: Model gràfic amb una variable latent Z i variables observables $X_1, \ldots, X_m$ condicionalment independents donat $Z = z$.

Per tal de predir la classe $z \in C$ per una observació concreta de les dades $\mathbf{x} = (x_1, \ldots, x_m)$, podem fer servir la regla de Bayes per calcular la probabilitat *a posteriori* de cada classe:

$$p(z|x) \propto p(z) \prod_{i=1}^{m} p(x_i|z),$$

i, segons aquesta probabilitat, la classe $z \in C$ s'escull segons la regla de decisió bayesiana, la qual maximitza aquesta probabilitat *a posteriori*:

$$z = \arg\max_{z \in C} \left\{ p(z) \prod_{i=1}^{m} p(x_i|z) \right\}.$$

Aquesta estratègia s'utilitza àmpliament en el processament de llenguatge natural i per a tasques de classificació de texts. Tot i ser un model tan simple, el model Bayes primari obté resultats sorprenentment bons a la pràctica, especialment per a tasques de detecció de correu brossa, anàlisi d'opinions i classificació temàtica.

EXEMPLE (MODEL EVOLUTIU EN UN TRÍPODE). Com s'ha mencionat anteriorment, els models d'arbres latents formen un exemple important de models de variables latents i han estat aplicats amb èxit a la biologia. A la secció 5, considerem processos de Markov en arbres filogenètics, però aquí considerem un cas senzill mitjançant el model Bayes primari amb $m = 3$, en què totes les variables aleatòries X_1, X_2, X_3 i Z són discretes i prenen valors a $S = \{0, 1, 2, 3\}$. Aquest conjunt d'estats representa els quatre nucleòtids (adenina, citosina, timina i guanina) que es troben en les seqüències d'ADN. En aquest cas, les variables aleatòries X_1, X_2, X_3 corresponen a observacions de nucleòtids en seqüències d'ADN associades a espècies biològiques actuals (per exemple, humà, ximpanzé i goril·la), mentre que la variable latent Z codifica nucleòtids en una seqüència d'ADN d'un ancestre comú a aquestes espècies. La propietat de Markov local en aquest DAG dona

$$p(x_1, x_2, x_3, z) = p(z)p(x_1|z)p(x_2|z)p(x_3|z)$$

(cf. exemple DAG estrella), i, marginalitzant sobre la variable latent, obtenim que

$$p(x_1, x_2, x_3) = \sum_{z \in S} p(z)p(x_1|z)p(x_2|z)p(x_3|z),$$

per a qualssevol $(x_1, x_2, x_3) \in S^3$. Aquestes probabilitats $p(x_1, x_2, x_3)$ representen la probabilitat d'observar els nucleòtids x_1, x_2, x_3 a les fulles de l'arbre i es poden estimar a partir de les observacions de les dades a les espècies actuals, mentre que $p(z)$ i $p(x_i|z)$ són paràmetres (desconeguts) del model. En filogenètica, l'estimació d'aquests paràmetres és rellevant perquè mesuren la quantitat de substitucions de nucleòtids transcorregudes des de l'ancestre comú fins a cada espècie actual, que dona una noció de *distància evolutiva* entre una espècie ancestral i els seus descendents. A la secció 5 explorem mètodes basats en moments per estimar aquests paràmetres, i a la secció 6 comentem aquests models amb més detall.

EXEMPLE (BOSSA DE PARAULES). Considerem ara una versió restringida del model Bayes primari en què les X_i condicionades respecte de la variable latent Z són idènticament distribuïdes. Com a exemple pràctic, considerem el model

de la bossa de paraules (*bag-of-words* en anglès), en què un document format per m paraules ha de ser classificat en un dels k *temes* ocults. Suposem que el vocabulari té una mida d i que cada tema $c_h \in C$ té el seu propi vector de probabilitat $\mu_h \in \Delta^d$ per emetre paraules del vocabulari: primer s'escull un tema $c_h \in C$ amb probabilitat $p(z = c_h)$, i després es mostregen m paraules $(x_1, \ldots, x_m)$ de manera independent amb vector de probabilitat μ_h.

En aquest exemple, l'ordre de les paraules en cada document és irrellevant, només es té en compte si hi són presents o no, i, per tant, els mots es consideren com a variables aleatòries *intercanviables*. Això significa que la seva distribució conjunta $p(x_1, \ldots, x_m)$ és invariant per permutacions dels índexs i, en conseqüència, el tensor p corresponent és un tensor simètric.

L'estimació de paràmetres en un model latent no és una tasca fàcil. A l'exemple *mètode dels moments per a la bossa de paraules* explicarem un mètode específic per al model de bossa de paraules basat en el mètode dels moments. Una altra aproximació és optimitzar la versemblança.

NOTA 2. Com hem apuntat en seccions anteriors, una distribució $p(x, z|\,\theta)$ amb variables latents on z denota un estat latent, la inferència de paràmetres pot ser complicada donat que z no és conegut. L'aproximació més clàssica és la de mirar de maximitzar la versemblança marginal sobre les variables observades

$$p(x|\theta) = \sum_z p(x, z|\,\theta),$$

on podem substituir la suma per una integral si Z també inclou variables contínues. Cal remarcar que tota la informació coneguda sobre les variables latents Z es troba a través de la seva distribució *a posteriori* $p(\cdot\,|\,x, \theta)$. En no disposar de la versemblança per al conjunt de dades complet, com a alternativa es procedeix a estimar-la prenent el valor esperat de la distribució *a posteriori* (pas E), i seguidament aquesta versemblança estimada es maximitza respecte als paràmetres θ (pas M). Aquest procediment rellevant per tractar les dades incompletes que provenen de variables latents es coneix com a *algorisme EM* (*expectation maximization* en anglès); vegeu [30].

4 El cas gaussià

En aquesta secció posem el focus en variables aleatòries contínues i, en particular, en variables normals multivariants. Generalitzem primer els tensors de moments introduïts abans en el cas discret.

4.1 Tensors de moments - cas general

Els tensors simètrics apareixen de manera natural com a derivades d'ordre superior de funcions diferenciables $f\colon \mathbb{R}^n \to \mathbb{R}$: les derivades parcials d'ordre r de f es poden organitzar com un tensor $T \in S^r(\mathbb{R}^n)$ amb coordenades

$$T_{i_1 \cdots i_r} = \frac{\partial^r}{\partial x_{i_1} \cdots \partial x_{i_r}} f(x), \quad \text{per } 1 \le i_1, \ldots, i_r \le n.$$

En relació amb l'estadística i l'aprenentatge automàtic, hi ha dos tipus concrets de funcions i les seves derivades parcials que són particularment rellevants. Considerem un vector aleatori $X = (X_1, \dots, X_m)$ en què ara permetem que els espais d'estats $\mathcal{X}_i$ siguin subconjunts arbitraris de $\mathbb{R}$. Siguin $M_X(\boldsymbol{s}) = \mathbb{E}[e^{\boldsymbol{s}^\top X}]$ i $K_X(\boldsymbol{s}) = \log \mathbb{E}[e^{\boldsymbol{s}^\top X}]$ les funcions generatrius de moments i de cumulants, respectivament (vegeu [54]). El *tensor de moment* d'ordre r es denota amb $\mu_r(X)$ i és la multitaula de dimensions $m \times \overset{r}{\cdots} \times m$, l'entrada $(i_1, \dots, i_r)$ de la qual és

$$\mu_{i_1 \cdots i_r}(X) = \mathbb{E}[X_{i_1} \cdots X_{i_r}] = \frac{\partial^r}{\partial s_{i_1} \cdots \partial s_{i_r}} M_X(\boldsymbol{s}) \Big|_{\boldsymbol{s}=0}.$$

De manera similar, el *tensor cumulant* $\kappa_r(X)$ es defineix en coordenades com a

$$\kappa_{i_1 \cdots i_r}(X) = \operatorname{cum}(X_{i_1}, \dots, X_{i_r}) = \frac{\partial^r}{\partial s_{i_1} \cdots \partial s_{i_r}} K_X(\boldsymbol{s}) \Big|_{\boldsymbol{s}=0}.$$

La relació entre $\mu_r(X)$ i $\kappa_r(X)$ per a una r qualsevol és rebuscada però està molt ben entesa; vegi's [79, 54]. Directament, per construcció, $\mu_r(X)$ i $\kappa_r(X)$ són tensors simètrics de l'espai $S^r(\mathbb{R}^m)$.

NOTA 3. Aquí estem fent servir una notació estàndard per a tensor de moments i cumulants. En aquesta convenció tenim, per exemple, que $\mu_1(X) = \mathbb{E}[X_1]$, $\mu_{13}(X) = \mathbb{E}[X_1 X_3]$, $\kappa_{12}(X) = \mathbb{E}[X_1 X_2] - \mathbb{E}[X_1]\mathbb{E}[X_2]$. Això contrasta amb la notació que emprem en la secció 2.3, que també es fa servir regularment en l'estudi algebraic de models discrets.

Com a exemple, noteu que, si $X = (X_1, X_2)$ són dues variables aleatòries independents, aleshores $\mu_{12} = \mu_1 \mu_2$ i $\kappa_{12} = 0$.

No obstant això, els moments i cumulants d'ordre baix no són suficients per a l'estudi d'un model estadístic, en general. Una excepció important és el cas de distribucions normals multivariants (vegi's la subsecció següent), que es parametritzen pel vector de mitjanes μ i per la matriu de variàncies i covariàncies Σ, que són moments/cumulants d'ordre 1 i 2, respectivament.

4.2 Distribució normal multivariant

La distribució normal o gaussiana multivariada és la distribució més important en l'estadística multivariant i en ciència de dades en general. També té un paper important en moltes tècniques de reducció de dimensionalitat per a tensors que esmentarem en aquest document.

Denotant amb $S^2_+(\mathbb{R}^n)$ el conjunt de matrius simètriques $n \times n$ definides positives, la distribució gaussiana n-variada amb mitjana $\mu \in \mathbb{R}^n$ i matriu de covariància $\Sigma \in S^2_+(\mathbb{R}^n)$ és una distribució sobre $\mathbb{R}^n$ amb funció de densitat

$$f(x) = \det(2\pi\Sigma)^{-1/2} \exp\left(-\frac{1}{2}(x-\mu)^\top \Sigma^{-1}(x-\mu)\right). \tag{8}$$

Les funcions generadores de moments i cumulants d'un vector aleatori normal X són

$$M_X(\boldsymbol{s}) = e^{\mu^\top \boldsymbol{s} + \frac{1}{2}\boldsymbol{s}^\top \Sigma \boldsymbol{s}}, \quad K_X(\boldsymbol{s}) = \mu^\top \boldsymbol{s} + \tfrac{1}{2}\boldsymbol{s}^\top \Sigma \boldsymbol{s}.$$

Observem que la funció generatriu de cumulants és un polinomi quadràtic i, per tant, tots els cumulants d'ordre tres o superior, com ara l'asimetria o la curtosi, que són els cumulants d'ordre tres i quatre, respectivament, són nuls. Per un resultat clàssic de Józef Marcinkiewicz, la distribució normal és l'*única* distribució tal que la seva funció generatriu de cumulants és polinomial [53, 51]. El fet que aquesta funció sigui un polinomi té conseqüències importants en probabilitat i estadística (vegeu, per exemple, [44]) i el fet que cap altra distribució no tingui aquesta propietat també és rellevant per justificar el resultat d'identificabilitat de [26] que apareixerà en la secció 6.3.

Una de les propietats fonamentals de la distribució normal és que és tancada per la consideració de marginals i condicionades. De manera més precisa, per a un vector aleatori X amb distribució normal n-variant de mitjana μ i matriu de covariància Σ, considerant una divisió arbitrària en dos blocs (X_A, X_B), la distribució del subvector X_A és també normal amb mitjana $\mu_A = (\mu_i)_{i\in A}$ i matriu de covariància $\Sigma_{A,A} = [\Sigma_{ij}]_{i,j\in A}$. Així mateix, la distribució condicionada de X_A donat $X_B = x_B$ és normal amb mitjana $\mu_A + \Sigma_{A,B}\Sigma_{B,B}^{-1}(x_B - \mu_B)$ i covariància $\Sigma_{A,A} + \Sigma_{A,B}\Sigma_{B,B}^{-1}\Sigma_{B,A}$. Per tant, treballar amb distribucions marginals i condicionades d'una normal multivariant implica una transformació purament algebraica en termes dels paràmetres.

4.3 Models gràfics gaussians

Les restriccions dels models gràfics tant per al cas dirigit com no dirigit permeten una descripció algebraica senzilla quan el vector X segueix una distribució normal multivariant.

Suposem que $\mathcal{G} = (V, E)$ és un graf no dirigit amb vèrtexs $V = \{1, \ldots, m\}$ i que $X = (X_1, \ldots, X_m)$ segueix una distribució normal multivariant. Aleshores, desenvolupant (8), és fàcil veure que la distribució factoritza com un model (6) en $\mathcal{G}$ sempre que la matriu de covariància Σ satisfaci

$$(\Sigma^{-1})_{ij} = 0, \quad \text{per tot } (i,j) \notin E,\ i \neq j.$$

Per tant, parametritzar el model en termes de la inversa de la matriu de covariància $K = \Sigma^{-1}$ ens permet definir el model a partir de senzilles restriccions lineals. Aquesta observació ha motivat una direcció de recerca en estadística algebraica per entendre els subespais lineals de matrius simètriques, per exemple, [87, 81, 82, 100].

Les distribucions gaussianes definides per DAG $\mathcal{G}$ també tenen una descripció simple, la qual és equivalent al model d'equacions lineals estructurals (vegeu la secció 6.3) a base de considerar el sistema d'equacions lineals estocàstiques

$$X_i = \sum_{j\in \mathbf{pa}(X_i)} \lambda_{ij} X_j + \varepsilon_i, \quad \text{per } i \in V, \tag{9}$$

on els ε_i per a $i \in V$ se suposen independents entre si i independents dels $\mathbf{pa}(X_i)$ corresponents. Suposem que els ε_i són variables normals de mitjana

zero, i els coefficients λ_{ij} prenen valors reals arbitraris. En forma matricial podem escriure-ho com a $X = \Lambda X + \varepsilon$, on Λ és triangular superior (reordenant X, si cal). Com que $\mathcal{G}$ és acíclic, $I_m - \Lambda$ és invertible i tenim que $X = (I_m - \Lambda)^{-1}\varepsilon$. D'aquí segueix que la inversa de la matriu de covariància $K = \Sigma^{-1}$ satisfà

$$K = (I_m - \Lambda)^\top \Omega^{-1} (I_m - \Lambda),$$

on Ω és la matriu diagonal amb les variàncies dels ε_i com a entrades. Per tant, aquest model es pot descriure de manera equivalent amb condicions d'anul·lació sobre la factorització de Cholesky de K. Aquests models han estat estudiats des d'una perspectiva algebraica a [84].

EXEMPLE (ANÀLISI FACTORIAL). Per a un conjunt de m variables observables $X = (X_1, \ldots, X_m)$ amb distribució conjunta normal multivariant $X \sim \mathcal{N}_m(\mu, \Sigma)$, ens pot interessar reduir aquestes variables en un espai de $d < m$ dimensions mitjançant una projecció lineal. Equivalentment, busquem modelitzar m variables observades gaussianes que depenen linealment de d variables gaussianes latents independents $Z = (Z_1, \ldots, Z_d)$. Podem escriure, doncs, que $X = \mu + WZ + \varepsilon$ on $\mu \in \mathbb{R}^d$, W és una matriu $m \times d$, i el terme de soroll ε podem suposar que té covariància $\Omega := \mathrm{Cov}(\varepsilon)$ diagonal. La matriu de covariància Σ en aquest cas ve donada per

$$\mathbb{E}[(WZ + \varepsilon)(WZ + \varepsilon)^\top] = WW^\top + \mathrm{Cov}(\varepsilon).$$

Així doncs, Σ viu al conjunt

$$F_{m,d} = \{\Omega + WW^\top \in \mathbb{R}^{m \times m} : \Omega \succ 0 \text{ diagonal}, W \in \mathbb{R}^{m \times d}\},$$

on fem servir $\succ 0$ per dir que la matriu és definida positiva. Observi's que $F_{m,d}$ és un conjunt semialgebraic, ja que es pot veure com la imatge de $\mathbb{R}^m_{>0} \times \mathbb{R}^{m \times d}$ sota una aplicació racional. Per tant, el model d'anàlisi factorial format per les distribucions normals multivariants X amb matriu de covariància $\Sigma \in F_{m,d}$ és un exemple de model estadístic semialgebraic. Entendre quines són les equacions algebraiques i semialgebraiques que defineixen el conjunt $F_{m,d}$ és clau per saber quines distribucions poden encaixar en aquest model.

5 Descomposició tensorial en estadística algebraica

De manera anàloga a la descomposició de matrius, la descomposició de tensors té com a objectiu representar un tensor d'altes dimensions com una sèrie d'operacions elementals en tensors més simples. Les descomposicions de tensors s'han utilitzat en diferents camps, com ara en la visió per computador, la neurociència computacional, la filogenètica o la psicometria. S'utilitzen per obtenir separació cega de fonts (*blind source separation*), l'anàlisi de components independents i per proporcionar estimadors en models de variables latents. Remetem el lector a [46] i [45], on trobarà excel·lents compilacions sobre la descomposició de tensors i les seves aplicacions en aprenentatge automàtic.

5.1 Rang d'un tensor

Un exercici senzill d'àlgebra lineal (usant la descomposició en valors singulars, per exemple) consisteix a veure que una matriu M de dimensions $n_1 \times n_2$ té rang k si i només si es pot escriure com a

$$M = u^1 \otimes v^1 + \cdots + u^k \otimes v^k$$

per a determinats vectors $u^i \in \mathbb{R}^{n_1}$, $v^i \in \mathbb{R}^{n_2}$, i no és possible escriure aquesta descomposició amb menys sumands (aquí $u_i \otimes v_i$ s'ha de pensar com a la matriu de rang u corresponent).

Una manera de generalitzar aquest concepte del rang per tensors és la següent.

DEFINICIÓ 2. El *rang* d'un tensor $T \in \mathbb{R}^{r_1} \otimes \cdots \otimes \mathbb{R}^{r_m}$ és el menor enter k tal que T es pot escriure com a

$$\sum_{j=1}^{k} u^{1,j} \otimes u^{2,j} \otimes \cdots \otimes u^{m,j} \tag{10}$$

per a determinats $u^{i,j} \in \mathbb{R}^{r_i}$, $i \in [m]$, $j \in [k]$. Dit d'una altra manera, un tensor té rang k si es pot escriure com una combinació lineal de k tensors de rang u (purs), però no com a suma d'un nombre menor de k tensors purs.

El *rang simètric* d'un tensor simètric $T \in S^r(\mathbb{R}^m)$ es defineix com l'enter k més petit tal que $T = \sum_{j=1}^{k} u^j \otimes u^j \otimes \cdots \otimes u^j$ per a determinats $u^j \in \mathbb{R}^m$, $j \in [k]$.

Ja hem vist que els tensors que apareixen en el model d'independència total són tensors de rang u. A continuació donem exemples de tensors de rang superior. En termes de geometria algebraica, donat que la varietat $\mathrm{Seg}(\mathbb{P}_{\mathbb{R}}^{X_1} \times \cdots \times \mathbb{P}_{\mathbb{R}}^{X_m})$ està formada pels tensors de rang u, el conjunt de tensors de rang k es troba a la seva k-èsima varietat secant (de fet, la varietat secant és la clausura en la topologia de Zariski d'aquest conjunt de tensors).

EXEMPLE (RANG TENSORIAL). En els tensors que codifiquen la distribució de X en el model de l'exemple *model Bayes primari* amb una variable latent Z que pren valors en un conjunt C de cardinal k, el rang és com a màxim k. En efecte, la propietat de Markov dona la factorització de la distribució conjunta, i, marginalitzant sobre la variable latent Z, obtenim

$$p(x_1, \ldots, x_m) = \sum_{z \in C} p(z) \prod_{i=1}^{m} p(x_i|z)$$

per a qualsevol conjunt d'observacions $x = (x_1, \ldots, x_m) \in \mathcal{X}$. Llavors, considerant els vectors

$$u^{i,z} = (p(x_i = 0|z), \ldots, p(x_i = r_i - 1|z)) \in \mathbb{R}^{X_i}, \quad \text{on } z \in C,\ i = 1, \ldots, m,$$

podem escriure la distribució $p = (p(x))_{x \in \mathcal{X}} \in \mathbb{R}^{\mathcal{X}}$ de $X = (X_1, \ldots, X_m)$ com a

$$p = \sum_{z \in C} p(z) u^{1,z} \otimes \cdots \otimes u^{m,z}. \tag{11}$$

Per tant, p té rang tensorial menor o igual que k. En particular, la distribució conjunta p a les fulles del model evolutiu sobre el trípode de l'exemple *model evolutiu en un trípode* és un tensor de rang menor o igual a quatre.

EXEMPLE 3. Considerem el model introduït a la bossa de paraules. En aquest cas, els vectors $u_{i,z}$ coincideixen amb μ_h si z és el tema c_h (i no depèn de i). Llavors, l'expressió (11) esdevé

$$p = \sum_{h=1}^{k} p(z = c_h)\mu_h \otimes \cdots \otimes \mu_h$$

i obtenim que p és un tensor simètric de rang com a màxim k.

5.2 Descomposició tensorial

Una descomposició com la de (10) ha estat introduïda en diversos àmbits. Es coneix com a *descomposició de rang tensorial* o com a descomposició *poliàdica canònica* (CP en anglès) o també *factor paral·lel* (PARAFAC); vegeu [43, 41].

Aquesta descomposició no és única en general: es poden reescalar els vectors $u_{i,j}$ o permutar els sumands per obtenir una altra descomposició. Aquestes són ambigüitats inherents que solen poder resoldre's per a cada aplicació particular en què treballem. Restringint-nos a descomposicions especials, podem obtenir la unicitat mòdul aquestes ambigüitats. Com en el cas del teorema espectral per a matrius simètriques, l'ortogonalitat és un requisit comú per descompondre tensors simètrics, però no tots els tensors simètrics tenen una descomposició ortogonal. La unicitat en la descomposició CP (mòdul permutacions i reescalatge) es pot obtenir per a tensors de baix rang, com veurem a continuació.

EXEMPLE 4. Considerem el model Bayes primari de l'exemple *model Bayes primari* amb $m = 3$ i deixem que k sigui el cardinal del conjunt d'estats de la variable latent. A [47], l'autor va donar condicions explícites per a la identificabilitat dels sumands (mòdul permutacions i escalars) en la descomposició de rang del tensor distribució conjunta. Aquesta condició explícita es compleix per a tensors *genèrics* del model si k és menor que el nombre r_i d'estats de cada variable observada, i contrasta fortament amb la no unicitat de descomposicions de matrius de rang k com a sumes de matrius de rang u. A [1], els autors mostren com aquest resultat bàsic es pot usar de manera potent per establir la identificabilitat d'una gran classe de models de variables latents.

En el context dels arbres filogenètics com a l'exemple *model evolutiu en un trípode*, aquest resultat es va redescobrir a [20]. En aquest cas, el resultat diu que els paràmetres del model (la distribució de nucleòtids en la variable latent i les probabilitats condicionades) es poden recuperar de manera única (mòdul permutacions) a partir de la distribució conjunta de les variables observades i es dona una construcció explícita per a la seva recuperació utilitzant descomposicions espectrals de determinades matrius extretes de la distribució conjunta.

5.3 Rang tensorial no negatiu

Els tensors que provenen de distribucions de probabilitat, ja sigui discreta o contínua, formen un subconjunt especial dins de la classe de tots els tensors. Per aquesta raó, i especialment per al cas discret, sovint és útil parlar no del seu rang com en (10), sinó d'una variant coneguda com a *rang no negatiu*: diem que un tensor T té *rang no negatiu com a màxim* k si es pot escriure com la suma de k tensors de rang u com abans

$$T = \sum_{j=1}^{k} u^{1,j} \otimes \cdots \otimes u^{m,j}, \quad u^{i,j} \in \mathbb{R}^{r_i}_{\geq 0}, \quad i \in [m],\, j \in [k], \tag{12}$$

on ara requerim que tots els vectors $u^{i,j}$ tinguin entrades no negatives. El *rang no negatiu* denotat amb rank_+ és el menor enter k tal que existeix una expressió com la de (12). El rang no negatiu no ha de coincidir necessàriament amb el rang usual d'un tensor, i per definició es té $\operatorname{rank} T \leq \operatorname{rank}_+ T$. Per a més generalitzacions i variants del rang de tensors, vegeu [48].

NOTA 4. És un exercici senzill comprovar que la matriu següent és de rang tres, però el seu rang no negatiu és quatre.

$$\begin{pmatrix} 1 & 1 & 0 & 0 \\ 0 & 1 & 1 & 0 \\ 0 & 0 & 1 & 1 \\ 1 & 0 & 0 & 1 \end{pmatrix}.$$

Aquest és un exemple senzill sobre com el rang clàssic i el rang no negatiu poden diferir.

Determinar si els rangs clàssic i no negatiu coincideixen és un problema NP-difícil. De fet, la factorització no negativa de matrius és per si mateixa un problema NP-difícil, tot i que mitjançant heurístiques de cerca local [88] es pot aconseguir un rendiment polinomial. Afortunadament, hi ha condicions sota les quals aquestes descomposicions de rang no negatiu són úniques, i aquestes estan estretament lligades la identificabilitat del model subjacent [67].

La descomposició no negativa de tensors ha estat estudiada en el context PARAFAC incloent-hi restriccions de no negativitat en quimiometria [14, 15], i es pot interpretar també com una descomposició de Bayes primari de distribucions de probabilitat respecte a condicionals com es fa a [40].

Tenint en compte l'expressió a (1), el model d'independència total és un exemple d'un tensor de rang no negatiu ≤ 1. Per exemple, el subconjunt de matrius $r_1 \times r_2$ amb $\operatorname{rank}_+ \leq 1$ és la part no negativa de la varietat de Segre $\mathbb{P}^{\mathcal{X}_1} \times \mathbb{P}^{\mathcal{X}_2}$. Les distribucions del model d'independència de dues variables discretes corresponen a matrius en la intersecció de $\Delta^{\mathcal{X}}$ amb el conjunt de matrius de $\operatorname{rank}_+ \leq 1$.

Els models de mixtura o de barreja (*mixture model* en anglès) fan ús implícit del rang tensorial no negatiu. Per a un model estadístic $\mathcal{P} \subset \Delta^X$, el k-èsim *model de mixtura* $\mathrm{Mixt}^k(\mathcal{P})$ ve donat per

$$\mathrm{Mixt}^k(\mathcal{P}) = \left\{ \sum_{i=1}^{k} \pi_i p^i : \pi \in \Delta^{k-1},\, p^i \in \mathcal{P} \right\}.$$

Està format per totes les distribucions de probabilitat de rang no negatiu menor o igual que k i és també un model gràfic amb variables ocultes. Un tractament complet sobre aquesta relació es pot trobar a [34].

Un cas important és el dels tensors de rang no negatiu ≤ 2, que són clau en l'estudi de models de Markov en arbres filogenètics amb estats binaris. El model Bayes primari en 4 amb variables observades binàries i dues classes latents es pot representar per tensors de dimensions $2 \times 2 \times \cdots \times 2$ de rang no negatiu com a màxim 2 i corresponen a mixtures de dues distribucions. Vegeu [3] per a una caracterització d'aquests tensors amb $\mathrm{rank}_+ \leq 2$ i la seva geometria.

EXEMPLE (MODELS TEMÀTICS). El model descrit a bossa de paraules és un classificador basat en el contingut del document, però no és un model generatiu i el seu potencial representatiu és limitat. Els models temàtics (*topic models* en anglès), [11], s'utilitzen quan els documents poden exhibir més d'una temàtica i la classificació del model *bossa de paraules* seria una aproximació massa restrictiva. Un exemple d'aquest tipus de model és el de l'assignació latent de Dirichlet (LDA) [12], que és un model generatiu que funciona com una xarxa bayesiana (secció 3.1), on es té una col·lecció de n documents i k temes, amb la propietat que els elements poden pertànyer a diverses classes (és a dir, un document pot tractar sobre més d'un tema).

Cada *topic* o tema es representen com a distribucions de probabilitat sobre les paraules, i els documents es representen, al seu torn, com a distribucions sobre temes latents. Per a cada document, s'assigna un escalar $\theta_{ij} \in [0,1]$ a cada tema $j \in \{1,2,\ldots,k\}$ segons la seva adequació al document, és a dir, la probabilitat que el tema j aparegui en el document i, com un model de mixtura en què assumim que el nombre total de temes k és conegut. Els pesos associats als temes del document segueixen una distribució de Dirichlet $\mathrm{Dir}(\alpha)$, amb una funció de densitat definida com a

$$p(\theta_i;\alpha) = f(\theta_{i1},\ldots,\theta_{ik};\alpha_1,\ldots,\alpha_k) = \frac{1}{\mathbf{B}(\alpha)} \prod_{j=1}^{k} \theta_{ij}^{\alpha_j - 1},$$

on $\mathbf{B}$ és la funció Beta que normalitza la distribució. Els pesos $\theta_i = (\theta_{i1},\ldots,\theta_{ik}) \in \Delta^{k-1}$ sumen u i, per tant, formen una mesura discreta com teníem a la secció 2.1. D'altra banda, una paraula i té una probabilitat φ_{ij} de pertànyer al tema j, on φ_{ij} segueix una distribució de Dirichlet $\mathrm{Dir}(\beta)$. Els tres conjunts de variables aleatòries que representen paraules, temes i documents es poden interpretar com un model gràfic. Concatenant aquests tres conjunts de variables, obtenim de manera natural una xarxa bayesiana amb capes de variables en l'ordre esmentat de predecessors a descendents.

5.4 Mètode dels moments

Fer aprenentatge en els tipus de models discutits a la secció 3.3 genera molts problemes. Els mètodes de versemblança, que són una opció popular per a molts models simples, ràpidament es tornen intractables en presència de variables latents. Amb poques alternatives bones, el mètode dels moments [64] ara s'utilitza rutinàriament en mixtures gaussianes [5] i en models d'arbres latents [7], [6], [73]. Com que aquest mètode és intrínsecament algebraic, el descrivim amb més detall en aquest manuscrit.

La idea del mètode és senzilla. Suposem que la parametrització del model indueix una parametrització explícita d'alguns moments $\mu = f(\theta)$ de la distribució subjacent. Un estimador pel mètode dels moments procedeix en dues etapes: primer, es troba un bon estimador $\hat{\mu}$ dels moments a partir de la mostra i després es defineix $\hat{\theta}$ com una solució de $\hat{\mu} = f(\theta)$.

Un problema evident d'aquest mètode és que, per a un $\hat{\mu}$ donat, no tenim garanties que l'equació $\hat{\mu} = f(\theta)$ tingui solució. En certa mesura, això es pot solucionar escollint adequadament uns moments determinats per a l'anàlisi. Però, fins i tot si existeix una solució, pot ser que no compleixi algunes de les restriccions necessàries. Per aquest motiu, el mètode dels moments normalment s'aplica cas per cas i requereix una bona comprensió de la classe de models que estem tenint en compte.

A tall d'exemple, en el treball clàssic [63], es va utilitzar una mixtura de dues variables gaussianes per modelitzar les mides de les parts del cos d'una població de crancs. Per fer-ho, Pearson va proposar un sistema d'equacions que involucrava els primers cinc moments de la distribució, i seguidament va aconseguir eliminar les variables i va obtenir un polinomi de grau nou en una única variable. Resolent el polinomi de grau nou resultant, va poder obtenir els paràmetres de la distribució de mixtura. Per a més detalls d'aquest primer exemple bàsic del mètode dels moments, consulteu [4].

EXEMPLE (MÈTODE DELS MOMENTS PER A LA BOSSA DE PARAULES). Considerem la bossa de paraules de l'exemple *bossa de paraules* i suposem que el nombre m de paraules en el document és almenys 3. Anandkumar *et al.* ([7]) van mostrar que, fent servir moments d'ordre com a màxim tres, hom pot recuperar els paràmetres del model (és a dir, els vectors de les distribucions condicionades μ_h i les probabilitats $w_h = p(z = c_h)$ per a $h = 1, \ldots, k$). Aquí expliquem breument aquest mètode dels moments seguint [6, apèndix D].

Podem representar les paraules del vocabulari amb els vectors de la base canònica $e_1, \ldots, e_d \in \mathbb{R}^d$, i considerem X tal que l'espai d'estats de X_i sigui $\mathcal{D} = \{e_1, \ldots, e_d\}$. Considerem les variables X_i com a indicadors de la manera següent:

$$X_i = e_j \iff \text{la } i\text{-èsima paraula del document és } j.$$

Aleshores els moments creuats d'aquests vectors aleatoris es corresponen amb les probabilitats conjuntes:

$$\mathbb{E}[X_1 \otimes X_2] = \sum_{i,j} p(x_1 = e_i, x_2 = e_j) e_i \otimes e_j$$

i obtenim

$$\mathbb{E}[X_1 \otimes X_2] = \sum_{h=1}^{k} \sum_{i,j} p(z = c_h) p(x_1 = e_i | z = c_h) p(x_2 = e_j | z = c_h) e_i \otimes e_j.$$

Amb la notació introduïda es pot escriure com a

$$\mathbb{E}[X_1 \otimes X_2] = \sum_{h=1}^{k} w_h \mu_h \otimes \mu_h. \tag{13}$$

De manera anàloga obtenim

$$\mathbb{E}[X_1 \otimes X_2 \otimes X_3] = \sum_{h=1}^{k} w_h \mu_h \otimes \mu_h \otimes \mu_h.$$

Si prenem $V = (\mu_1 | \ldots | \mu_k)$ i $D = \operatorname{diag}(w_1, \ldots, w_k)$, aleshores, considerant $M_2 = \mathbb{E}[X_1 \otimes X_2]$ com una matriu, podem expressar (13) com a

$$M_2 = VDV^\top. \tag{14}$$

No es tracta d'una diagonalització de $\mathbb{E}[X_1 \otimes X_2]$ (llevat de quan V és ortogonal), però podem emprar tècniques espectrals incorporant els moments de tercer ordre. De fet, considerem $M_3 = \mathbb{E}[X_1 \otimes X_2 \otimes X_3]$ i per a qualsevol $\eta \in \mathbb{R}^d$ definim el 2-tensor $M_3 \bullet \eta$ com a

$$(M_3 \bullet \eta)_{i,j} = \sum_l (M_3)_{i,j,l} \eta_l.$$

Llavors, com a matrius tenim $M_3 \bullet \eta = VD\, d(\eta) V^\top$, on $d(\eta) = \operatorname{diag}(\mu_1^\top \eta, \ldots, \mu_k^\top \eta)$, i si $M = (M_3 \bullet \eta) M_2^{-1}$, obtenim $M = V d(\eta) V^{-1}$. Prenent η amb entrades no repetides, les columnes de V estan determinades pels vectors propis de M (llevat de permutacions i multiplicació per escalars). Un cop es recupera V, $w_1 \ldots, w_k$ són fàcilment recuperables mitjançant l'expressió (14).

Existeixen enfocaments més robusts per a l'estimació de paràmetres utilitzant moments, però van més enllà de l'abast d'aquest estudi. Val la pena assenyalar que els paràmetres estimats pel mètode dels moments també es poden utilitzar com a pas inicial d'un algorisme EM (vegeu la nota 2); en aquest cas, l'EM sovint convergeix a un màxim local després d'un sol pas; vegeu [97].

6 Aplicacions de l'estadística algebraica

6.1 Anàlisi filogenètica

Un *arbre filogenètic* T sobre un conjunt L d'espècies (o altres entitats biològiques) és un arbre amb les fulles etiquetades pels elements de L. La figura 4 mostra tal arbre amb fulles $L = \{$humà, ximpanzé, goril·la, orangutan$\}$.

L'objectiu principal de la filogenètica és reconstruir la història evolutiva del conjunt L d'espècies actuals a partir de la informació proporcionada per una col·lecció de molècules d'ADN associades a elles. Degut a l'estructura de doble hèlix de l'ADN, aquestes molècules es poden interpretar com a paraules o seqüències en l'alfabet que representen els quatre nucleòtids, $\{\mathtt{A},\mathtt{C},\mathtt{G},\mathtt{T}\}$. Sota certes hipòtesis biològiques es pot suposar que cada posició d'aquestes seqüències evoluciona independentment de les altres posicions i que les posicions estan idènticament distribuïdes.

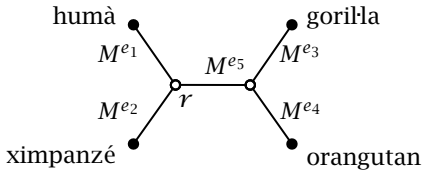

FIGURA 4: Un arbre filogenètic amb conjunt de fulles donat per $L = \{$humà, ximpanzé, goril·la, orangutan$\}$, node arrel r, i matrius estocàstiques per files associades al procés de Markov.

Si T és un arbre filogenètic, modelem la substitució de nucleòtids en T seleccionant primer un node interior r que faci el paper d'arrel de l'arbre (que podria representar l'ancestre comú a totes les espècies de L) i dirigint totes les arestes des de r per veure $T = (V, E)$ com un DAG. A cada $v \in V$ li associem una variable aleatòria X_v prenent valors a $S = \{0, 1, 2, 3\}$ (en representació dels quatre nucleòtids) i considerem la propietat de Markov local sobre el DAG. Aquesta propietat és equivalent al fet que cada variable aleatòria sigui independent de les seves variables no descendents donades les observacions al seu node pare immediat, com s'expressà a (4). És conegut que, per obtenir la identificabilitat dels paràmetres, cal que l'arbre no tingui nodes de grau 2. Per a un arbre de tres fulles, aquest model és el que es dona a l'exemple *model evolutiu en un trípode*.

Per escriure la factorització de la distribució conjunta (1) sota aquest model de Markov, codifiquem les distribucions condicionades en matrius estocàstiques per files: a cada aresta $e = \{u \to v\} \in E$ associem M^e, una matriu 4×4 amb entrades $M^e_{x,y} = p(X_v = y | X_u = x)$, la probabilitat condicionada que un estat x en el node ancestral u es vegi substituït per l'estat y en el descendent v.

Aleshores, la probabilitat conjunta d'observar estats (nucleòtids) x_v a les variables X_v, per a $v \in V$, és

$$p(\{x_v\}_{v\in V}) = p(x_r) \prod_{e=\{u\to u'\}\in E} M^e_{x_{u'},x_u}. \tag{15}$$

En filogenètica només tenim observacions per a les variables aleatòries trobades a les fulles de l'arbre, de manera que les variables aleatòries als nodes interiors són latents (ja que representen espècies extintes). Així, si el conjunt de fulles L és $[n]$, la probabilitat $p_{i_1\cdots i_n}$ d'observar el nucleòtid i_j a la fulla j per $j \in [n]$ es pot obtenir marginalitzant (15) sobre els nodes interiors:

$$p_{i_1\cdots i_n} = \sum_{\{x_v\}_{v\in V} | x_j = i_j,\, j\in[n]} p(\{x_v\}_v).$$

Per cada arbre filogenètic T sobre $L = [n]$ definim com a $p^\top \in \mathbb{R}^{\mathcal{X}} = \mathbb{R}^4 \otimes \overset{n)}{\cdots} \otimes \mathbb{R}^4$ el tensor de la distribució conjunta $(p_{0\cdots 0}, p_{0\cdots 1}, \ldots, p_{3\cdots 3})$ obtingut d'aquesta manera. Aquest model s'anomena *procés de Markov general sobre l'arbre filogenètic*.

Si l'arbre filogenètic que explica la història evolutiva del conjunt L és conegut i tenim unes seqüències d'ADN observades sobre el conjunt L, els paràmetres del procés de Markov corresponent se solen estimar mitjançant un mètode de màxima versemblança. No obstant això, un dels problemes principals en filogenètica és obtenir l'arbre T que s'ajusti millor a les dades: com que el nombre d'arbres filogenètics creix superexponencialment en el nombre de fulles, no és possible buscar exhaustivament per tot l'espai d'arbres. Aquí l'estadística algebraica té un paper important, tal com expliquem a continuació (consulteu també [17]).

A finals dels anys vuitanta, els biòlegs Cavender, Felsenstein i Lake es van adonar que algunes equacions polinomials satisfetes per les coordenades de $p^\top$ permeten distingir entre els diferents arbres que donen lloc a la distribució. És a dir, hi ha equacions polinomials sobre les coordenades d'un tensor $p \in \mathbb{R}^4 \otimes \overset{n}{\cdots} \otimes \mathbb{R}^4$ que es compleixen si $p = p^\top$ per a alguns arbres T però no per a altres. Un exemple d'aquestes equacions prové de l'*aplanament* (*flattening* en anglès) del tensor, com detallem a continuació.

Considerem una bipartició $A|B$ de $L = [n]$: un subconjunt A i el seu complement $B = L \setminus A$. Com teníem a la secció 2.2, aquesta divisió indueix un isomorfisme de l'espai de tensors a l'espai de matrius

$$\begin{aligned} \mathbb{R}^{\mathcal{X}} \cong \mathbb{R}^{\mathcal{X}_A} \otimes \mathbb{R}^{\mathcal{X}_B} &\longrightarrow \mathcal{M}_{|\mathcal{X}_A|\times|\mathcal{X}_B|}(\mathbb{R}) \\ p = (p_{x_1\cdots x_n}) &\longmapsto \mathrm{flatt}_{A|B}(p), \end{aligned} \tag{16}$$

on l'entrada (x_A, x_B) de $\mathrm{flatt}_{A|B}(p)$ és la coordenada de p que es correspon amb la probabilitat d'observar estats $x_A = \{x_l\}_{l\in A}$ a les fulles de A i $x_B = \{x_l\}_{l\in B}$ a les fulles de B. Per exemple, a l'arbre filogenètic de la figura 4, si les fulles

humà, ximpanzé, goril·la i orangutan es denoten amb 1, 2, 3, 4, respectivament, tenim que

$$\operatorname{flatt}_{12|34}(p) = \begin{pmatrix} p_{0000} & p_{0001} & p_{0002} & \cdots & p_{0033} \\ p_{0100} & p_{0101} & p_{0102} & \cdots & p_{0133} \\ p_{0200} & p_{0201} & p_{0202} & \cdots & p_{0233} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ p_{3300} & p_{3301} & p_{3302} & \cdots & p_{3333} \end{pmatrix}.$$

Suposem que $p = p^\top$ per a algun arbre filogenètic T, i considerem una bipartició $A|B$ del conjunt de fulles induïda per la supressió d'una aresta e de T, que anomenarem *bipartició d'aresta*. Llavors $\operatorname{flatt}_{A|B}(p^\top)$ es pot considerar com una taula de probabilitat conjunta de les variables aleatòries $X_A = (X_l)_{l\in A}$ i $X_B = (X_l)_{l\in B}$. Per la propietat de Markov en T sabem que $X_A \perp\!\!\!\perp X_B | X_u$, on X_u és la variable oculta corresponent a un dels vèrtexs de e. Aquest és un model de Bayes primari amb dues variables aleatòries observades X_A, X_B ($m = 2$) i una de latent. Per tant, $p^\top$ és un tensor de rang com a màxim quatre segons l'exemple rang tensorial. En particular, $\operatorname{flatt}_{A|B}(p^\top)$ és una matriu de rang com a màxim quatre perquè (16) envia una suma de tensors de rang u a una suma de matrius de rang u. Així doncs, hem obtingut la primera afirmació del següent teorema:

TEOREMA 3 ([2]). *Sigui T un arbre filogenètic i $A|B$ una divisió del seu conjunt de fulles L. Sigui $p^\top$ un tensor de distribució obtingut per un procés de Markov en T. Aleshores, si $A|B$ és una bipartició d'aresta en T,* $\operatorname{flatt}_{A|B}(p^\top)$ *té rang menor o igual a quatre. A més, si $A|B$ no és una bipartició d'aresta en T, el rang de* $\operatorname{flatt}_{A|B}(p^\top)$ *és més gran que quatre (sempre que $p^\top$ compleixi determinades condicions genèriques).*

Així doncs, l'anul·lació dels menors 5×5 de $\operatorname{flatt}_{A|B}(p)$ defineix equacions polinomials que es compleixen per a distribucions que provenen d'arbres que tenen $A|B$ com a bipartició d'aresta, però que no es compleixen per a distribucions en altres arbres. En lloc d'utilitzar directament aquestes equacions algebraiques, és més natural calcular la distància entre matrius de rang quatre directament utilitzant valors singulars. Aquesta estratègia s'ha aprofitat per proporcionar diversos mètodes per a la reconstrucció filogenètica; consulteu, per exemple, [22, 37, 18].

6.2 Màquines de Boltzmann restringides

Considerem un conjunt de variables aleatòries binàries $V = \{X_1, \ldots, X_n, Y_1, \ldots, Y_m\}$ en els vèrtexs d'un model gràfic no dirigit, on $X = (X_1, \ldots, X_n)$ són les variables observades i $Y = (Y_1, \ldots, Y_m)$ són les variables ocultes. Aquests models gràfics, coneguts com a *màquines de Boltzmann restringides* (*restricted Boltzmann machines* o RBM en anglès) venen dotats d'una estructura de graf bipartit de manera que únicament hi ha arestes entre variables observades i

variables latents; vegeu la figura 5. L'*energia* d'una configuració donada de variables aleatòries amb els seus paràmetres en aquest graf es defineix com la funció

$$E(x, y; \theta) = -\sum_{i=1}^{n} b_i x_i - \sum_{j=1}^{m} c_j y_j - \sum_{\{i,j\}} W_{ij} x_i y_j,$$

on $\theta = (W, b, c)$ són els paràmetres del model. La notació que usem per als paràmetres està influenciada per la notació en la literatura de perceptrons i xarxes neuronals. A partir de l'energia, podem definir la probabilitat d'una realització del vector aleatori com a

$$p(x, y; \theta) = \frac{1}{Z(\theta)} \exp(-E(x, y; \theta)), \quad (x, y) \in \{0, 1\}^{n+m}, \tag{17}$$

on $Z(\theta)$ és la constant normalitzadora. Com podem veure per la connectivitat del graf, les variables observades són independents entre elles condicionades sobre les latents, i viceversa. La distribució marginal de X es pot obtenir de l'equació (17) sumant sobre els estats de les variables latents $y \in \{0, 1\}^m$, i després, mitjançant les manipulacions adients, obtindrem el producte següent (vegeu [57] per a més detalls sobre la derivació i les màquines de Boltzmann en general):

$$\begin{aligned} p(x; \theta) &= \frac{1}{Z(\theta)} \exp\left(b^\top x\right) \prod_{j=1}^{m} \left(1 + \exp\left(w_j x + c_j\right)\right) \\ &= \frac{1}{Z(\theta)} \prod_{j=1}^{m} \left(\lambda_j \prod_{i=1}^{n} p'_{ji}(x_i) + (1 - \lambda_j) \prod_{i=1}^{n} p''_{ji}(x_i) \right) \end{aligned} \tag{18}$$

per a λ_j entre zero i u, on w_j és la fila j-èsima de W i p'_{ji}, p''_{ji} són distribucions que depenen exclusivament de x_i. D'aquesta manera trobem una expressió que ens recorda una distribució producte $q(x_1, \ldots, x_n) = \prod q_i(x_i)$ que lliga amb un producte tensorial. Concretament, l'expressió de l'equació (18) mostra una estructura de producte tensorial de rang no negatiu dos, del tipus que hem esmentat a la secció 5.3. Ho veiem en l'equació següent, que fa explícita aquesta connexió fent servir notació tensorial:

$$p = \frac{1}{Z(\theta)} \prod_{j=1}^{m} (q'_{j1} \otimes \cdots \otimes q'_{jn} + q''_{j1} \otimes \cdots \otimes q''_{jn}),$$

per a determinats q'_{ji}, q''_{ji}. Així doncs, les RBM són un exemple d'un producte de m factors tensorials de rang no negatiu com a màxim dos.

Aquests models han estat explorats i utilitzats en processament de la informació a [78] i a [39] com a precursor de les xarxes neuronals de dues capes, de manera que les seves propietats com a aproximadors de funcions són especialment rellevants en aplicacions d'aprenentatge automàtic. Els tensors de rang no negatiu dos han estat descrits en termes de restriccions de submodularitat a [3].

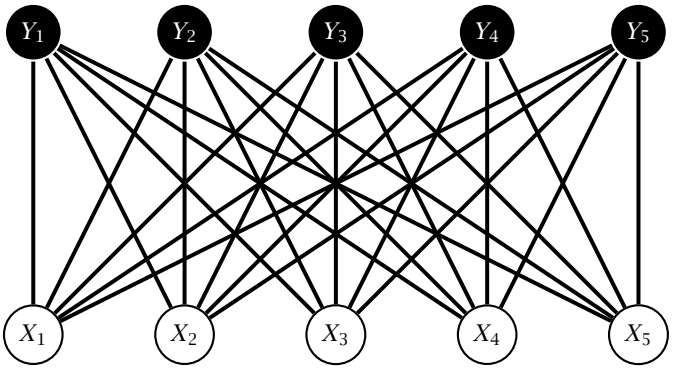

FIGURA 5: Diagrama que mostra una màquina de Boltzmann restrictiva com un model gràfic amb cinc variables latents i cinc variables observades.

NOTA 5. L'anàlisi de les distribucions que sorgeixen d'aquests tensors de rang dos és remarcadament difícil. No obstant això, per al cas en què el model és la suma de dos tensors $2 \times 2 \times 2$, Seigal i Montúfar ([75]) han demostrat que el model de distribucions no degenerades (és a dir, les que cauen a l'interior del símplex) són equivalents a un model gràfic en un arbre amb una única variable oculta de tres estats.

6.3 Components independents i models d'equacions estructurals

Considerem el model següent per a un vector aleatori X,

$$X = A\varepsilon, \tag{19}$$

on $A \in \mathbb{R}^{m\times m}$ i ε és un vector aleatori (latent) amb una estructura de dependència senzilla. Els models d'aquesta forma s'utilitzen extensament en aplicacions. Depenent del context, podem imposar restriccions addicionals a la matriu A o al vector aleatori (latent) ε. Per exemple, en separació cega de fonts és habitual suposar que A és invertible i que ε té components independents, [27]. La proposició ens ajudarà en aquesta discussió:

PROPOSICIÓ 4. *Les components de X són independents si i només si el tensor cumulant $\kappa_r(X)$ és diagonal per a tot $r \geq 2$.*

El resultat anterior té conseqüències importants en ciència de dades. En l'*anàlisi de components independents* (ICA), es postula que el vector aleatori observat $X = (X_1, \ldots, X_m)$ es pot escriure com una transformació lineal $A\varepsilon$, on $\varepsilon = (\varepsilon_1, \ldots, \varepsilon_m)$ té components independents i $A \in \mathbb{R}^{m\times m}$ és invertible. Sota condicions febles sobre la distribució del vector no observat ε, podem recuperar la matriu A a partir de les observacions X (llevat de permutacions de files i canvis en els seus signes), [26]. A la pràctica aquesta recuperació es realitza amb els cumulants d'ordre baix $\kappa_3(Y)$ i $\kappa_4(Y)$, aprofitant el fet que aquests són tensors diagonals.

Donat que el vector ε no s'observa directament, la pregunta principal és si la matriu A es pot recuperar únicament a partir de les observacions de X. Sense pèrdua de generalitat, suposem que ε té mitjana zero i matriu de covariància identitat. Sota aquesta hipòtesi, la covariància de X satisfà $\mathrm{Cov}(X) = AA^\top$, cosa que ens permet identificar A mòdul l'acció del grup ortogonal sobre l'espai de les columnes de A ($AA^\top = (AQ)(AQ)^\top$ per a qualsevol matriu ortogonal Q). Aquest resultat d'identificabilitat es pot millorar si ε és no gaussià. Concretament, en un treball clau, Comon ([26]) va demostrar que A es pot recuperar llevat de permutacions i reescalaments de les columnes per a ± 1, sempre que com a màxim una component de ε sigui normal. Sota condicions genèriques febles, A es pot recuperar a partir dels moments de tercer o quart ordre de X. A la pràctica, molts mètodes utilitzen aquests moments d'ordre inferior.

Per connectar el model (19) a la nostra discussió sobre tensors, considerem $\mu_r(X), \mu_r(\varepsilon) \in S^r(\mathbb{R}^m)$ com els moments tensorials de X i ε, respectivament, tal com hem introduït a la secció 4.1. Suposem que ε té components independents i, per tant, per la proposició 4, $\mu_r(\varepsilon)$ és un tensor diagonal. Així mateix, per la propietat multilineal dels moments, tenim que

$$\mu_r(X) = \mu_r(A\varepsilon) = A \bullet \mu_r(\varepsilon),$$

on la notació $A \bullet \mu_r(\varepsilon) \in S^r(\mathbb{R}^m)$ és l'acció multilineal de $A \in \mathbb{R}^{m\times m}$ sobre $S^r(\mathbb{R}^m)$:

$$(A \bullet \mu_r(\varepsilon))_{i_1,\dots,i_r} = \sum_{j_1,\dots,j_r} A_{i_1 j_1} \cdots A_{i_r j_r} (\mu_r(\varepsilon))_{j_1,\dots,j_r}.$$

En altres paraules, l'equació (19) estableix que, per a tot $r \geq 2$, el tensor moment r-èsim de X és de la forma $A \bullet D$, per a $A \in \mathbb{R}^{m\times m}$ i un tensor diagonal $D \in S^r(\mathbb{R}^m)$. La pregunta es converteix, doncs, en: en quins casos permet aquesta condició, per a r fixada, identificar la matriu A? Com que la matriu de covariància ja ens permetia identificar A mòdul l'acció del grup ortogonal, podem reformular-ho de la manera següent. Suposem $A \bullet D = (AQ) \bullet D'$ per a A invertible i Q ortogonal, on D, D' són tensors diagonals; equivalentment, $Q \bullet D'$ és diagonal. Llavors, podem concloure que Q ha de ser una matriu de permutació llevat de signes? Per a més detalls sobre aquest enfocament, vegeu [55].

Exemple 5. Considerem el cas $r = 3$, $m = 2$. La condició que $Q \bullet D'$ sigui diagonal per a un tensor diagonal D' es pot expressar pel sistema d'equacions cúbic en les entrades de Q:

$$(Q \bullet D')_{112} = Q_{11}^2 Q_{21} D'_{111} + Q_{12}^2 Q_{22} D'_{222} = 0,$$
$$(Q \bullet D')_{122} = Q_{11} Q_{21}^2 D'_{111} + Q_{12} Q_{22}^2 D'_{222} = 0.$$

En forma matricial això és

$$Q \cdot \begin{bmatrix} Q_{11} & 0 \\ 0 & Q_{22} \end{bmatrix} \cdot \begin{bmatrix} Q_{21} & 0 \\ 0 & Q_{12} \end{bmatrix} \cdot \begin{bmatrix} D'_{111} \\ D'_{222} \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}.$$

Com que Q és ortogonal, cadascuna de les matrius diagonals anteriors han de ser o bé la matriu de zeros, o bé invertibles. Si alguna de les dues és idènticament zero, aleshores Q ha de ser una matriu de permutació llevat de signes, i l'equació se satisfà trivialment. Si ambdues matrius són invertibles, l'equació només es pot satisfer si $D'_{111} = D'_{222} = 0$, i en aquest cas D' es correspon amb el tensor idènticament zero. Això ens mostra que, mentre D' sigui un tensor diagonal diferent de zero, $Q \bullet D'$ és diagonal només si Q és una matriu de permutació signada.

En la majoria d'aplicacions suposem que ε té components independents, però també és possible relaxar aquestes condicions (per exemple, en el model de variància comuna). Per a més motivació i resultats bàsics, vegeu [55]. Addicionalment, el cas sobredeterminat on A és una matriu rectangular és altament rellevant a la pràctica; vegeu [91] per als detalls.

Els models de la forma (19) també s'han estudiat extensivament en el context d'anàlisi de la causalitat, i, en aquest àmbit, la matriu A pren una estructura especial. Pearl ([62]) considera el model següent d'*equacions estructurals* definit sobre un graf dirigit:

$$X_i = f_i(\mathbf{pa}(i), \varepsilon_i), \quad i = 1, \ldots, m,$$

és a dir que cada variable del sistema es pot expressar com una funció dels seus nodes predecessors i un terme d'error ε_i. Aquest sistema és no lineal i no paramètric, però generalitza la versió lineal àmpliament utilitzada en els camps de l'economia i les ciències socials que hem presentat a (9). Per al cas del model lineal, es pot escriure de forma compacta com a

$$X = \Lambda X + \varepsilon,$$

on $\Lambda_{ij} = 0$ tret que $j \to i$ sigui una aresta de $\mathcal{G}$. Prenent $A = (I_m - \Lambda)^{-1}$, recuperem el model que teníem a (19).

Donat que, en aquest context, la matriu A té una estructura particular, podríem aspirar a identificar-la directament amb la matriu de covariàncies de X. Per a resultats en aquesta direcció, vegeu [33, 35, 38]. L'exemple següent ha estat adaptat de [32].

EXEMPLE 6. Fumar durant l'embaràs afecta el pes en néixer del nadó? Per respondre aquesta pregunta, suposem que un estudi registra el nivell de tabaquisme matern durant l'embaràs (X_1) i el pes en néixer d'un nadó (X_2). Suposant que existeix un efecte causal del tabaquisme sobre el pes en néixer, pretenem quantificar aquest efecte. A la pràctica, se centren les dades i es fa servir un model lineal:

$$X_1 = \varepsilon_1, \quad X_2 = \lambda_{12} X_1 + \varepsilon_2;$$

l'objectiu és inferir l'efecte λ_{12}. Si ε_1, ε_2 són variables no correlacionades de mitjana 0, podem fer servir que $\operatorname{Cov}(X_1, X_2) = \lambda_{12} \operatorname{Var}(X_1)$ per recuperar λ_{12} a partir de la distribució observada.

Una altra generalització important és la d'incloure variables latents a (9) per modelitzar potencials variables de confusió. Aquestes poden distorsionar les relacions de causalitat observades entre les components de X; vegeu [8, 86]. L'exemple a continuació ha estat adaptat també de [32].

EXEMPLE 7. L'efecte λ_{12} de l'exemple anterior pot no ser directe i pot veure's distorsionat per altres variables de confusió latents. Per exemple, factors de predisposició genètica o socioeconòmics poden afectar tant l'hàbit de fumar com el pes en néixer dels nadons. En aquesta situació, la inferència sobre el quocient $\mathrm{Cov}(X_1, X_2)/\mathrm{Var}(X_1)$ no seria vàlida. Una possible solució és la d'introduir una «variable instrumental» que influenciï directament el nivell de tabaquisme però no el pes en néixer, i que no estigui influenciada per la variable latent. Seguint la bibliografia, es pot emprar una variable instrumental X_3 que mesuri els impostos sobre el tabac. Les equacions lineals que defineixen el model amb la variable latent H vindrien donades per:

$$\begin{aligned} X_1 &= \lambda_{01} + \lambda_{31}X_3 + \lambda_{H1}H + \varepsilon_1, \\ X_2 &= \lambda_{02} + \lambda_{12}X_1 + \lambda_{H2}H + \varepsilon_2, \\ X_3 &= \lambda_{03} + \varepsilon_3, \\ H &= \lambda_{0H} + \varepsilon_H, \end{aligned}$$

on els errors $\{\varepsilon_1, \varepsilon_2, \varepsilon_3, \varepsilon_H\}$ són independents dos a dos. D'aquí se segueix que l'equació $\mathrm{Cov}(X_2, X_3) = \lambda_{12}\,\mathrm{Cov}(X_1, X_3)$ se satisfà per a aquest model i, sempre que $\mathrm{Cov}(X_1, X_3) \neq 0$, podrem deduir propietats de λ_{12} a partir del quocient $\mathrm{Cov}(X_2, X_3)/\mathrm{Cov}(X_1, X_3)$.

Un altre problema conegut en l'anàlisi de causalitat és si l'estructura del graf es pot recuperar a partir de les observacions de X. Un altre cop, això requereix moments d'ordre superior de les dades, la no gaussianitat, i els resultats d'identificabilitat com els que hem presentat per ICA; vegeu [77].

Referències

[1] ALLMAN, E. S.; MATIAS, C.; RHODES, J. A. «Identifiability of parameters in latent structure models with many observed variables». *Ann. Statist.*, 37 (6A) (2009), 3099–3132.

[2] ALLMAN, E. S.; RHODES, J. A. «Phylogenetic ideals and varieties for the general Markov model». *Adv. in Appl. Math.*, 40 (2) (2008), 127–148.

[3] ALLMAN, E. S.; RHODES, J. A.; STURMFELS, B.; ZWIERNIK, P. «Tensors of nonnegative rank two». *Linear Algebra Appl.*, 473 (2015), 37–53.

[4] AMÉNDOLA, C.; CASANELLAS, M.; GARCÍA PUENTE, L. D. «Tapas of algebraic statistics». *Notices Amer. Math. Soc.*, 65 (8) (2018), 936–938.

[5] AMÉNDOLA, C.; FAUGÈRE, J.-C.; STURMFELS, B. «Moment varieties of Gaussian mixtures». *J. Algebr. Stat.*, 7 (1) (2016), 14–28.

[6] Anandkumar, A.; Ge, R.; Hsu, D.; Kakade, S. M.; Telgarsky, M. «Tensor decompositions for learning latent variable models». *J. Mach. Learn. Res.*, 15 (2014), 2773–2832.

[7] Anandkumar, A.; Hsu, D.; Kakade, S. M. «A method of moments for mixture models and hidden Markov models». A: *Proceedings of the 25th Annual Conference on Learning Theory*. Vol. 23 (2012), 33.1–33.34.

[8] Barber, R. F.; Drton, M.; Sturma, N.; Weihs, L. «Half-trek criterion for identifiability of latent variable models». *Ann. Statist.*, 50 (6) (2022), 3174–3196.

[9] Bertolini, M.; Turrini, C. «Problems and related results in algebraic vision and multiview geometry». *Rend. Circ. Mat. Palermo (2)*, 73 (6) (2024), 2205–2231.

[10] Bishop, C. M.; Tipping, M. E. «A hierarchical latent variable model for data visualization». *IEEE Trans. Pattern Anal. Mach. Intell.*, 20 (3) (1998), 281–293.

[11] Blei, D. M. «Probabilistic topic models». *Comm. ACM*, 55 (4) (2012), 77–84.

[12] Blei, D. M.; Ng, A. Y.; Jordan, M. I. «Latent Dirichlet allocation». *J. Mach. Learn. Res.*, 3 (2003), 993–1022.

[13] Bocci, C.; Chiantini, L. *An Introduction to Algebraic Statistics with Tensors*. La Matematica per il 3+2. Cham: Springer, 2019. (Unitext; 118)

[14] Bro, R.; de Jong, S. «A fast non-negativity-constrained least squares algorithm». *Journal of Chemometrics*, 11 (5) (1997), 393–3401.

[15] Bro, R.; Sidiropoulos, N. D. «Least squares algorithms under unimodality and non-negativity constraints». *Journal of Chemometrics*, 12 (4) (1998), 223–247.

[16] Calamante, F.; Mørup, M.; Hansen, L. K. «Defining a local arterial input function for perfusion MRI using independent component analysis». *Magnetic Resonance in Medicine*, 52 (4) (2004), 789–797.

[17] Casanellas, M. «Phylogenetic reconstruction based on Algebra». A: *Cutting-Edge Mathematics*. Cham: Springer, 2024, 26–44. (RSME Springer Ser.; 13)

[18] Casanellas, M.; Fernández-Sánchez, J.; Garrote-López, M. «SAQ: semialgebraic quartet reconstruction». *IIEEE/ACM Trans. Comput. Biol. Bioinform.*, 18 (6) (2021), 2855–2861.

[19] Castro, R.; Coates, M.; Liang, G.; Nowak, R.; Yu, B. «Network tomography: recent developments». *Statist. Sci.*, 19 (3) (2004), 499–517.

[20] Chang, J. T. «Full reconstruction of Markov models on evolutionary trees: identifiability and consistency». *Math. Biosci.*, 137 (1) (1996), 51–73.

[21] Chickering, D. M. «A transformational characterization of equivalent Bayesian network structures». A: *Uncertainty in Artificial Intelligence* (Montreal, PQ, 1995). San Francisco, CA: Morgan Kaufmann, 1995, 87–98.

[22] Chifman, J.; Kubatko, L. «Quartet inference from SNP data under the coalescent model». *Bioinformatics*, 30 (23) (2014), 3317–3324.

[23] Choi, H.; Baraniuk, R. G. «Multiscale image segmentation using wavelet-domain hidden Markov models». *IEEE Trans. Image Process.*, 10 (9) (2001), 1309–1321.

[24] Choi, M. J.; Lim, J. J.; Torralba, A.; Willsky, A. S. «Exploiting hierarchical context on a large database of object categories». A: *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*. San Francisco, CA: IEEE, 2010, 129–136.

[25] Ciliberto, C.; Cueto, M. A.; Mella, M.; Ranestad, K.; Zwiernik, P. «Cremona linearizations of some classical varieties». A: *From Classical to Modern Algebraic Geometry*. Cham: Birkhäuser/Springer, 2016, 375–407. (Trends Hist. Sci.)

[26] Comon, P. «Independent component analysis, a new concept?». *Signal Process.*, 36 (3) (1994), 287–314.

[27] Comon, P.; Jutten, C. (ed.). *Handbook of Blind Source Separation. Independent Component Analysis and Applications*. Academic Press, 2010.

[28] Cong, F.; Lin, Q.-H.; Kuang, L.-D.; Gong, X.-F.; Astikainen, P.; Ristaniemi, T. «Tensor decomposition of EEG signals: A brief review». *Journal of Neuroscience Methods*, 248 (2015), 59–69.

[29] Crouse, M. S.; Nowak, R. D.; Baraniuk, R. G. «Wavelet-based statistical signal processing using hidden Markov models». *IEEE Trans. Signal Process.*, 46 (4) (1998), 886–902.

[30] Dempster, A. P.; Laird, N. M.; Rubin, D. B. «Maximum likelihood from incomplete data via the EM algorithm». Amb discussió. *J. Roy. Statist. Soc. Ser. B*, 39 (1) (1977), 1–38.

[31] Develin, M.; Sullivant, S. «Markov bases of binary graph models». *Ann. Comb.*, 7 (4) (2003), 441–466.

[32] Drton, M. «Algebraic problems in structural equation modeling». A: *The 50th Anniversary of Gröbner Bases*. Tòquio: Mathematical Society of Japan, 2018, 35–86. (Adv. Stud. Pure Math.; 77)

[33] Drton, M.; Foygel, R.; Sullivant, S. «Global identifiability of linear structural equation models». *Ann. Statist.*, 39 (2) (2011), 865–886.

[34] Drton, M.; Sturmfels, B.; Sullivant, S. *Lectures on Algebraic Statistics*. Basilea: Birkhäuser Verlag, 2009. (Oberwolfach Semin.; 39)

[35] Drton, M.; Weihs, L. «Generic identifiability of linear structural equation models by ancestor decomposition». *Scand. J. Stat.*, 43 (4) (2016), 1035–1045.

[36] Eriksson, B.; Dasarathy, G.; Barford, P.; Nowak, R. «Toward the practical use of network tomography for internet topology discovery». A: *Proceedings IEEE INFOCOM*. San Diego, CA: IEEE, 2010, 1–9.

[37] Fernández-Sánchez, J.; Casanellas, M. «Invariant versus classical quartet inference when evolution is heterogeneous across sites and lineages». *Systematic Biology*, 65 (2) (2016), 280–291.

[38] Foygel, R.; Draisma, J.; Drton, M. «Half-trek criterion for generic identifiability of linear structural equation models». *Ann. Statist.*, 40 (3) (2012), 1682–1713.

[39] Freund, Y.; Haussler, D. «Unsupervised learning of distributions on binary vectors using two layer networks». A: *Advances in Neural Information Processing Systems*. Vol. 4. Morgan-Kaufmann, 1991, 912–919.

[40] Garcia, L. D.; Stillman, M.; Sturmfels, B. «Algebraic geometry of Bayesian networks». *J. Symbolic Comput.*, 39 (3-4) (2005), 331–355.

[41] Harshman, R. A.; Lundy, M. E. «PARAFAC: Parallel factor analysis». *Comput. Statist. Data Anal.*, 18 (1) (1994), 39–72.

[42] Hartley, R.; Zisserman, A. *Multiple View Geometry in Computer Vision*. 2a ed. Amb un pròleg d'Olivier Faugeras. Cambridge: Cambridge University Press, 2003.

[43] Hitchcock, F. L. «The expression of a tensor or a polyadic as a sum of products». *J. Math. and Phys.*, 6 (1-4) (1927), 164–189.

[44] Janson, S. «Normal convergence by higher semi-invariants with applications to sums of dependent random variables and random graphs». *Ann. Probab.*, 16 (1) (1988), 305–312.

[45] Janzamin, M.; Ge, R.; Kossaifi, J.; Anandkumar, A. «Spectral learning on matrices and tensors». *Foundations and Trends® in Machine Learning*, 12 (5-6) (2019), 393–536.

[46] Ji, Y.; Wang, Q.; Li, X.; Liu, J. «A survey on tensor techniques and applications in machine learning». *IEEE Access*, 7 (2019), 162950–162990.

[47] Kruskal, J. B. «Three-way arrays: rank and uniqueness of trilinear decompositions, with application to arithmetic complexity and statistics». *Linear Algebra Appl.*, 18 (2) (1977), 95–138.

[48] Landsberg, J. M. *Tensors: Geometry and Applications*. Providence, RI: American Mathematical Society, 2012. (Grad. Stud. Math.; 128)

[49] Lauritzen, S. L. *Graphical Models*. Oxford Sci. Publ. Nova York: The Clarendon Press, Oxford University Press, 1996. (Oxford Statist. Sci. Ser.; 17)

[50] Lawrence, N. D. «Gaussian process latent variable models for visualisation of high dimensional data». A: *NIPS'03: Proceedings of the 16th International Conference on Neural Information Processing Systems*. Cambridge, MA: MIT Press, 2004, 329–336.

[51] Lukacs, E. «Some extensions of a theorem of Marcinkiewicz». *Pacific J. Math.*, 8 (1958), 487–501.

[52] Makhijani, M. K.; Balu, N.; Yamada, K.; Yuan, C.; Nayak, K. S. «Accelerated 3D MERGE carotid imaging using compressed sensing with a hidden Markov tree model». *Journal of Magnetic Resonance Imaging*, 36 (5) (2012), 1194–1202.

[53] Marcinkiewicz, J. «Sur une propriété de la loi de Gauß». *Math. Z.*, 44 (1) (1939), 612–618.

[54] McCullagh, P. *Tensor Methods in Statistics. Monographs on Statistics and Applied Probability*. Chapman and Hall/CRC, 2018.

[55] Mesters, G.; Zwiernik, P. «Nonndependent components analysis». *Ann. Statist.*, 52 (6) (2024), 2506–2528.

[56] Michałek, M.; Sturmfels, B. *Invitation to Nonlinear Algebra*. Providence, RI: American Mathematical Society, 2021. (Grad. Stud. Math.; 211)

[57] Montúfar, G. «Restricted Boltzmann machines: introduction and review». A: *Information Geometry and its Applications*. Cham: Springer, 2018, 75–115. (Springer Proc. Math. Stat.; 252)

[58] Montúfar, G.; Morton, J. «Discrete restricted Boltzmann machines». *J. Mach. Learn. Res.*, 16 (2015), 653–672.

[59] Mourad, R.; Sinoquet, C.; Zhang, N. L.; Liu, T.; Leray, P. «A survey on latent tree models and applications». *J. Artificial Intelligence Res.*, 47 (2013), 157–203.

[60] Nafees, S.; Rice, S. H.; Wakeman, C. A. «Analyzing genomic data using tensor-based orthogonal polynomials with application to synthetic RNAs». *NAR Genomics and Bioinformatics*, 2 (4) (2020), lqaa101.

[61] Pachter, L.; Sturmfels, B. (ed.). *Algebraic Statistics for Computational Biology*. Nova York: Cambridge University Press, 2005.

[62] Pearl, J. *Causality. Models, Reasoning, and Inference*. Cambridge: Cambridge University Press, 2000.

[63] Pearson, K. «Contributions to the mathematical theory of evolution». *Philos. Trans. Roy. Soc. London Ser. A*, 185 (1894), 71–110.

[64] Pearson, K. «Method of moments and method of maximum likelihood». *Biometrika*, 28 (1/2) (1936), 34–59.

[65] Pfeiffer, M.; Betizeau, M.; Waltispurger, J.; Pfister, S. S.; Douglas, R. J.; Kennedy, H.; Dehay, C. «Unsupervised lineage-based characterization of primate precursors reveals high proliferative and morphological diversity in the OSVZ». *Journal of Comparative Neurology*, 524 (3) (2016), 535–563.

[66] Pistone, G.; Riccomagno, E.; Wynn, H. P. *Algebraic Statistics. Computational Commutative Algebra in Statistics*. Boca Raton, FL: Chapman & Hall/CRC, 2001. (Monogr. Statist. Appl. Probab.; 89)

[67] Qi, Y.; Comon, P.; Lim, L.-H. «Semialgebraic geometry of nonnegative tensor rank». *SIAM J. Matrix Anal. Appl.*, 37 (4) (2016), 1556–1580.

[68] Rabiner, L. R. «A tutorial on hidden Markov models and selected applications in speech recognition». A: *Proceedings of the IEEE*, 77 (2) (1989), 257–286.

[69] Ricci, M. M. G.; Levi-Civita, T. «Méthodes de calcul différentiel absolu et leurs applications». *Math. Ann.*, 54 (1-2) (1900), 125–201.

[70] Ringe, D.; Warnow, T.; Taylor, A. «Indo-European and computational cladistics». *Transactions of the Philological Society*, 100 (1) (2002), 59–129.

[71] Robeva, E.; Seigal, A. «Duality of graphical models and tensor networks». *Inf. Inference*, 8 (2) (2019), 273–288.

[72] Romberg, J. K.; Choi, H.; Baraniuk, R. G. «Bayesian tree-structured image modeling using wavelet-domain hidden Markov models». *IEEE Trans. Image Process.*, 10 (7) (2001), 1056–1068.

[73] Ruffini, M.; Casanellas, M.; Gavaldà, R. «A new method of moments for latent variable models». *Mach. Learn.*, 107 (8-10) (2018), 1431–1455.

[74] Schreiber, J.; Durham, T.; Bilmes, J.; Noble, W. S. «Avocado: a multi-scale deep tensor factorization method learns a latent representation of the human epigenome». *Genome Biology*, 21, article núm. 81 (2020).

[75] Seigal, A.; Montúfar, G. «Mixtures and products in two graphical models». *J. Algebr. Stat.*, 9 (1) (2018), 1–20.

[76] Shiers, N.; Aston, J. A. D.; Smith, J. Q.; Coleman, J. S. «Gaussian tree constraints applied to acoustic linguistic functional data». *J. Multivariate Anal.*, 154 (2017), 199–215.

[77] Shimizu, S.; Inazumi, T.; Sogawa, Y.; Hyvärinen, A.; Kawahara, Y.; Washio, T.; Hoyer, P. O.; Bollen, K. «DirectLiNGAM: a direct method for learning a linear non-Gaussian structural equation model». *J. Mach. Learn. Res.*, 12 (2011), 1225–1248.

[78] Smolensky, P. «Information processing in dynamical systems: Foundations of harmony theory». *Parallel Distributed Process*, 1 (1986).

[79] Speed, T. P. «Cumulants and partition lattices». *Austral. J. Statist.*, 25 (2) (1983), 378–388.

[80] Steel, M. *Phylogeny—Discrete and Random Processes in Evolution*. Filadèlfia, PA: Society for Industrial and Applied Mathematics (SIAM), 2016. (CBMS-NSF Regional Conf. Ser. in Appl. Math.; 89)

[81] Sturmfels, B.; Timme, S.; Zwiernik, P. «Estimating linear covariance models with numerical nonlinear algebra». *Algebr. Stat.*, 11 (1) (2020), 31–52.

[82] Sturmfels, B.; Uhler, C. «Multivariate Gaussian, semidefinite matrix completion, and convex algebraic geometry». *Ann. Inst. Statist. Math.*, 62 (4) (2010), 603–638.

[83] Sturmfels, B.; Zwiernik, P. «Binary cumulant varieties». *Ann. Comb.*, 17 (1) (2013), 229–250.

[84] Sullivant, S. «Algebraic geometry of Gaussian Bayesian networks». *Adv. in Appl. Math.*, 40 (4) (2008), 482–513.

[85] Sullivant, S. *Algebraic Statistics*. Providence, RI: American Mathematical Society, 2018. (Grad. Stud. Math.; 194)

[86] Tramontano, D.; Drton, M.; Etesami, J. «Parameter identification in linear non-Gaussian causal models under general confounding». Preprint (2024). [Disponible en línia a: https://arxiv.org/abs/2405.20856]

[87] Uhler, C. «Geometry of maximum likelihood estimation in Gaussian graphical models». *Ann. Statist.*, 40 (1) (2012), 238–261.

[88] Vavasis, S. A. «On the complexity of nonnegative matrix factorization». *SIAM J. Optim.*, 20 (3) (2009), 1364–1377.

[89] Verma, T.; Pearl, J. «Equivalence and synthesis of causal models». A: *Proceedings of the Sixth Conference on Uncertainty in Artificial Intelligence*. Nova York: Elsevier Science Inc., 1990, 220–227.

[90] Wainwright, M. J.; Jordan, M. I. «Graphical models, exponential families, and variational inference». *Foundations and Trends® in Machine Learning*, 1 (1-2) (2008), 1–305.

[91] Wang, K.; Seigal, A. «Identifiability of overcomplete independent component analysis». Preprint (2024). [Disponible en línia a: https://arxiv.org/abs/2401.14709]

[92] Watanabe, S. *Algebraic Geometry and Statistical Learning Theory*. Cambridge: Cambridge University Press, 2009. (Cambridge Monogr. Appl. Comput. Math.; 25)

[93] Willsky, A. S. «Multiresolution Markov models for signal and image processing». A: *Proceedings of the IEEE*, 90 (8) (2002), 1396–1458.

[94] Yang, K. D.; Katcoff, A.; Uhler, C. «Characterizing and learning equivalence classes of causal DAGs under interventions». *Proceedings of Machine Learning Research*, 80 (2018), 5537–5546.

[95] Žabokrtskỳ, Z.; Popel, M. «Hidden Markov tree model in dependency-based machine translation». A: *Proceedings of the ACL-IJCNLP 2009 Conference Short Papers*, 2009, 145–148.

[96] Zhang, N. L. «Hierarchical latent class models for cluster analysis». *J. Mach. Learn. Res.*, 5 (2003/04), 697–723.

[97] Zhang, Y.; Chen, X.; Zhou, D.; Jordan, M. I. «Spectral methods meet EM: a provably optimal algorithm for crowdsourcing». *J. Mach. Learn. Res.*, 17 (2016), article núm. 102, 44 p.

[98] Zwiernik, P. «*L*-cumulants, *L*-cumulant embeddings and algebraic statistics». *J. Algebr. Stat.*, 3 (1) (2012), 11–43.

[99] Zwiernik, P. *Semialgebraic Statistics and Latent Tree Models*. Boca Raton, FL: Chapman & Hall/CRC, 2016. (Monogr. Statist. Appl. Probab.; 146)

[100] Zwiernik, P.; Uhler, C.; Richards, D. «Maximum likelihood estimation for linear Gaussian covariance models». *J. R. Stat. Soc. Ser. B. Stat. Methodol.*, 79 (4) (2017), 1269–1292.

Luis Sierra
Department of Statistical Sciences
University of Toronto
luis.sierra@mail.utoronto.ca

Marta Casanellas
Departament de Matemàtiques
Universitat Politècnica de Catalunya
marta.casanellas@upc.edu

Piotr Zwiernik
Department of Economics and Business
Universitat Pompeu Fabra
piotr.zwiernik@upf.edu