Exponential families

Kaie Kubjas, 07.10.2020

Agenda

- Exponential families
- Discrete exponential families
- Gaussian exponential families
- Group projects

Exponential families

- An exponential family is a parametric statistical model with probability distributions of a certain form.
- General enough to include many of the most common families of probability distributions:
  - multivariate normal
  - exponential
  - Poisson
  - binomial (with fixed number of trials)
- Specific enough to have nice properties:
  - likelihood function is strictly concave [lecture 6]
  - exponential families have conjugate priors

Goals

- What is an exponential family?
- How to find the vanishing ideal of an exponential family?
  - Discrete exponential models: Hypothesis testing [lecture 7]
  - Gaussian exponential submodels: Conditional independence implications [previous lecture]

Statistical models

- A statistical model is a collection of probability distributions.
- A parametric statistical model is a collection of probability distributions indexed by a finite dimensional parameter space $\Theta \subseteq \mathbb{R}^d$:

$$\mathcal{M}_\Theta = \{P_\theta : \theta \in \Theta\}.$$

- In the case of continuous random variables, it is often specified in terms of corresponding probability density functions:

$$\mathcal{M}_\Theta = \{f_\theta : \theta \in \Theta\}.$$

- In the case of discrete random variables, it is often specified in terms of corresponding probability mass functions:

$$\mathcal{M}_\Theta = \{p_\theta : \theta \in \Theta\}.$$

[A probability distribution assigns a value $P(X \in S)$ to a set $S \subseteq R$; a probability mass function assigns a value $p(x) = P(X = x)$ to a number $x \in \mathbb{R}$.]

Statistic

Def: Let X be a random vector taking values in a set $\mathcal{X}$. A statistic is a function from $\mathcal{X}$ to $\mathbb{R}^k$ for some $k \in \mathbb{N}$.

Example:

- Let $X_1, \ldots, X_n$ be independent Bernoulli distributed random variables with expected value p. Denote $X = (X_1, \ldots, X_n)$.
- Then $T : \{0,1\}^n \to \mathbb{R}$ given by $T(X) = X_1 + \ldots + X_n$ is a statistic for X.

Sufficient statistic

Def: For a parametric statistical model $\mathscr{M}_\Theta$, a statistic T is sufficient if the probability density function or probability mass function factorizes as $f_\theta(x) = h(x)g(T(x), \theta)$.

Example:

- Let $X_1, \ldots, X_n$ be independent Bernoulli distributed random variables with expected value p. Denote $X = (X_1, \ldots, X_n)$.
- Then $T : \{0,1\}^n \to \mathbb{R}$ given by $T(X) = X_1 + \ldots + X_n$ is a statistic for X.
- $P(X_1 = x_1, \ldots, X_n = x_n) = p^{x_1}(1-p)^{1-x_1} \cdots p^{x_n}(1-p)^{1-x_n} = p^{T(x)}(1-p)^{1-T(x)}$
- $h(x) = 1$ and $g(T(x), p) = p^{T(x)}(1-p)^{1-T(x)} \Longrightarrow T$ is a sufficient statistic

Exponential families

- Let X be a random variable taking values in a set $\mathcal{X}$.
- An exponential family is the set of probability distributions whose probability mass function or density function can be expressed as

$$f_\theta(x) = h(x)e^{\eta(\theta)^t T(x) - A(\theta)}$$

for a given statistic $T : \mathcal{X} \to \mathbb{R}^k$, natural parameter $\eta : \Theta \to \mathbb{R}^k$, and functions $h : \mathcal{X} \to \mathbb{R}_{>0}$ and $A : \Theta \to \mathbb{R}$.

Exponential families

- Three equivalent forms:
  - $f_\theta(x) = h(x)e^{\eta(\theta)^t T(x) - A(\theta)}$
  - $f_\theta(x) = h(x)g(\theta)e^{\eta(\theta)^t T(x)}$
  - $f_\theta(x) = e^{\eta(\theta)^t T(x) - A(\theta) + B(x)}$

Binomial distribution

$X \sim \mathrm{Bin}(m, \theta)$, $\mathscr{X} = \{0, 1, \ldots, m\}$

$$p(x) = \binom{m}{x} \theta^x (1-\theta)^{m-x} = \binom{m}{x} \exp\left[\left(\log \frac{\theta}{1-\theta}\right) x + m \log(1-\theta)\right]$$

Binomial distribution

- $f_\theta(x) = h(x)e^{\eta(\theta)^t T(x) - A(\theta)}$
- Statistic $T : \mathcal{X} \to \mathbb{R}^k$, natural parameter $\eta : \Theta \to \mathbb{R}^k$, functions $h : \mathcal{X} \to \mathbb{R}_{>0}$ and $A : \Theta \to \mathbb{R}$
- Binomial distribution: $p(x) = \binom{m}{x} \theta^x (1-\theta)^{m-x} = \binom{m}{x} \exp\left[\left(\log \frac{\theta}{1-\theta}\right) x + m \log(1-\theta)\right]$
- Poll: What are k, T, η, h, A in this example?

1. $k = 1, T(x) = \log \frac{\theta}{1-\theta}, \eta = x, h = \binom{m}{x}, A = -m \log(1-\theta)$ - Wrong, because T should not depend on parameters
2. $k = 1, T(x) = x, \eta = \log \frac{\theta}{1-\theta}, h = \binom{m}{x}, A = -m \log(1-\theta)$ - Correct
3. $k = 2, T(x) = (x, m-x), \eta = (\theta, 1-\theta), h = \binom{m}{x}, A = 0$ - Wrong

Canonical form

- $f_\theta(x) = h(x)e^{\eta(\theta)^t T(x) - A(\theta)}$
- If $\eta(\theta) = \theta$, then the exponential family is said to be in canonical form.
- By defining a transformed parameter $\eta = \eta(\theta)$, it is always possible to convert an exponential family to canonical form.
- The function A is determined by the other functions: It makes the pdf (pmf) to integrate (sum) to one. Thus it can be written as a function of η.
- The canonical form is $f_\eta(x) = h(x)e^{\eta^t T(x) - A(\eta)}$.

Discrete exponential families

- Let X be a discrete random variable taking values in $\mathscr{X} = [r]$.
- Denote
  - $T(x) = a_x$ where $a_x = (a_{1x}, \ldots, a_{kx})^t$
  - $h(x) = h_x$, so $h = (h_1, \ldots, h_r) \in \mathbb{R}^r_{>0}$
  - $\eta = (\eta_1, \ldots, \eta_k)^t$ and $\theta_i = \exp(\eta_i)$
- Then $p_\eta(x) = h(x)e^{\eta^t T(x) - A(\eta)} = h_x e^{\sum_i \eta_i a_{ix} - A(\eta)} = h_x \prod_i e^{\eta_i a_{ix} - A(\eta)} = h_x \prod_i (e^{\eta_i})^{a_{ix}} e^{-A(\eta)} = h_x \prod_i \theta_i^{a_{ix}} \dfrac{1}{Z(\theta)}$

  where $Z(\theta) = \displaystyle\sum_{x \in \mathscr{X}} h_x \prod_j \theta_j^{a_{jx}}$.

Discrete exponential families

$$p_\theta(x) = \frac{1}{Z(\theta)} h_x \prod_j \theta_j^{a_{jx}} \text{ where } Z(\theta) = \sum_{x \in \mathcal{X}} h_x \prod_j \theta_j^{a_{jx}}$$

- If a_{jx} are integers for all j and x, then the parametrizing functions are rational functions.
- The entries a_{jx} can be recorded in the matrix $A = (a_{jx})_{j \in [k], x \in [r]} \in \mathbb{Z}^{k \times r}$.
- For $x \in \mathcal{X} = [r]$, the monomials $\prod_j \theta_j^{a_{jx}}$ correspond to a column of the matrix A.

Example: Let $A = \begin{pmatrix} 0 & 1 & 2 & 3 \\ 3 & 2 & 1 & 0 \end{pmatrix}$ and $h = \mathbf{1}$. Then

$$p_\theta = \frac{1}{Z(\theta)} \left(\theta_2^3, \theta_1\theta_2^2, \theta_1^2\theta_2, \theta_1^3\right) \text{ where } Z(\theta) = \theta_2^3 + \theta_1\theta_2^2 + \theta_1^2\theta_2 + \theta_1^3.$$

Discrete exponential families

- Let $A = (a_{jx})_{j\in[k],x\in[r]} \in \mathbb{Z}^{k\times r}$.
- The logarithm of the exponential family representation $p_\theta(x) = \frac{1}{Z(\theta)} h_x \prod_j \theta_j^{a_{jx}}$ gives

$$\log p_\theta(x) = \log h_x + \sum_j a_{jx} \log \theta_j - \log Z(\theta).$$

- If we assume that the matrix A contains the vector $\mathbf{1} = (1,1,\ldots,1)$ in its row span, then this is equivalent to requiring that $\log p$ belongs belongs to the affine space $\log(h) + \text{rowspan}(A)$.

Log-affine model

Def: Let $A \in \mathbb{Z}^{k\times r}$ be a matrix of integers such that $\mathbf{1} \in \text{rowspan}(A)$ and let $h \in \mathbb{R}^r_{>0}$. The log-affine model associated to A and h is the set of probability distributions

$$\mathscr{M}_{A,h} := \{p \in \text{int}(\Delta_{r-1}) : \log p \in \log h + \text{rowspan}(A)\}.$$

If $h = \mathbf{1}$, then $\mathscr{M}_A := \mathscr{M}_{A,\mathbf{1}}$ is called a log-linear model.

Log-affine model

Def: Let $A \in \mathbb{Z}^{k\times r}$ be a matrix of integers such that $\mathbf{1} \in \text{rowspan}(A)$ and let $h \in \mathbb{R}^r_{>0}$.
The monomial map associated to this data is the rational map

$$\phi^{A,h} : \mathbb{R}^k \to \mathbb{R}^r, \text{ where } \phi_j^{A,h} = h_j \prod_{i=1}^{k} \theta_i^{a_{ij}}.$$

NB! The normalizing constant $Z(\theta)$ is removed.

Example: Let $A = \begin{pmatrix} 0 & 1 & 2 & 3 \\ 3 & 2 & 1 & 0 \end{pmatrix}$. The monomial map is $\phi^A : \mathbb{R}^2 \to \mathbb{R}^4$ is given by

$$(\theta_1, \theta_2) \mapsto (\theta_2^3, \theta_1\theta_2^2, \theta_1^2\theta_2, \theta_1^3).$$

Discrete independent random variables

- Consider the parametrization

$$p_{ij} = \alpha_i \beta_j,$$

  where $i \in [2], j \in [2]$ and α_i, β_j are independent parameters.

- This is the parametrization of two discrete independent random variables.

- Poll 1: What are the matrix A and vector h representing the above parametrization? Answer: $A = \begin{pmatrix} 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 \\ 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 \end{pmatrix}$ and $h = \begin{pmatrix} 1 \\ 1 \\ 1 \\ 1 \end{pmatrix}$. Rows of A correspond to $\alpha_1, \alpha_2, \beta_1, \beta_2$ and columns of A to $p_{11}, p_{12}, p_{21}, p_{22}$.

- Poll 2: What is the size of the matrix of A if $i \in [r_1]$ and $j \in [r_2]$? Answer: $(r_1 + r_2) \times (r_1 r_2)$

Log-affine model

<u>Def:</u> Let $A \in \mathbb{Z}^{k \times r}$ and $h \in \mathbb{R}^r_{>0}$. The ideal

$$I_{A,h} := I(\phi^{A,h}(\mathbb{R}^k)) \subseteq \mathbb{R}[p]$$

is called the toric ideal associated to the pair A and h.

- If $h = \mathbf{1}$, then we denote $I_A := I_{A,1}$.
- Generators for the ideal $I_{A,h}$ are obtained from generators of the ideal I_A.

Log-affine model

Prop: Let $A \in \mathbb{Z}^{k\times r}$ and $h \in \mathbb{R}^r_{>0}$. Then

$$I_A = \langle p^u - p^v : u, v \in \mathbb{N}^r \text{ and } Au = Av\rangle.$$

Example: Let $A = \begin{pmatrix} 0 & 1 & 2 & 3 \\ 3 & 2 & 1 & 0 \end{pmatrix}$. The monomial map is $\phi^A : \mathbb{R}^2 \to \mathbb{R}^4$ is given by

$$(\theta_1, \theta_2) \mapsto (\theta_2^3, \theta_1\theta_2^2, \theta_1^2, \theta_2, \theta_1^3).$$

The toric ideal is

$$I_A = \langle p_1p_3 - p_2^2, p_1p_4 - p_2p_3, p_2p_4 - p_3^2\rangle.$$ [Poll]

toricMarkov -- calculates a generating set of the toric ideal I_A, given A; invokes "markov" from 4ti2

Synopsis

- Usage:
  `toricMarkov(A) or toricMarkov(A, InputType => "lattice") or toricMarkov(A,R)`
- Inputs:
  - `A`, a matrix, whose columns parametrize the toric variety; the toric ideal is the kernel of the map defined by `A`. Otherwise, if InputType is set to "lattice", the rows of `A` are a lattice basis and the toric ideal is the saturation of the lattice basis ideal.
  - `InputType => s`, default value null, which is the string "lattice" if rows of `A` specify a lattice basis
  - `R`, a ring, polynomial ring in which the toric ideal I_A should live
- Optional inputs:
  - InputType => ...,
- Outputs:
  - `B`, a matrix, whose rows form a Markov Basis of the lattice *{z integral : A z = 0}* or the lattice spanned by the rows of `A` if the option `InputType => "lattice"` is used

Description

Suppose we would like to comput the toric ideal defining the variety parametrized by the following matrix:

```
i1 : A = matrix"1,1,1,1;0,1,2,3"

o1 = | 1 1 1 1 |
     | 0 1 2 3 |

              2        4
o1 : Matrix ZZ  <--- ZZ
```

Since there are 4 columns, the ideal will live in the polynomial ring with 4 variables.

```
i2 : R = QQ[a..d]

o2 = R

o2 : PolynomialRing
```

```
i3 : M = toricMarkov(A)

o3 = | 0 1  -2 1 |
     | 1 -2 1  0 |
     | 1 -1 -1 1 |

              3        4
o3 : Matrix ZZ  <--- ZZ
```

Note that rows of M are the exponents of minimal generators of I_A. To get the ideal, we can do the following:

```
i4 : I = toBinomial(M,R)

               2           2
o4 = ideal (- c  + b*d, - b  + a*c, - b*c + a*d)

o4 : Ideal of R
```

Multivariate normal distribution

Let PD_m be the set of $m \times m$ symmetric positive definite matrices.

Def: Suppose $\mu \in \mathbb{R}^m$ and $\Sigma \in PD_m$. Then a random vector $X = (X_1, \ldots, X_m)$ is distributed according to the multivariate normal distribution $\mathcal{N}_m(\mu, \Sigma)$ if it has the density function

$$f_{\mu,\Sigma}(x) = \frac{1}{(2\pi)^{m/2} |\Sigma|^{1/2}} \exp\left\{ -\frac{1}{2}(x-\mu)^t \Sigma^{-1} (x-\mu) \right\}.$$

Normal distribution

- $f_{\mu,\sigma^2}(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left\{ -\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2 \right\} = \frac{1}{\sigma\sqrt{2\pi}} \exp\left\{ -\frac{1}{2\sigma}x^2 + \frac{\mu}{\sigma^2}x - \frac{\mu^2}{2\sigma^2} \right\}$
- Poll: What are T, η, h, A?
  1. $T = (x, x^2), \eta = \left(\frac{\mu}{\sigma^2}, -\frac{1}{2\sigma^2}\right), h = \frac{1}{\sigma\sqrt{2\pi}}, A = \frac{\mu^2}{2\sigma^2}$ - Wrong, because h depends on a parameter
  2. $T = (x, x^2), \eta = \left(\frac{\mu}{\sigma^2}, -\frac{1}{2\sigma^2}\right), h = \frac{1}{\sqrt{2\pi}}, A = \log\sigma + \frac{\mu^2}{2\sigma^2}$ - Correct
  3. $T = (x, -x^2/2), \eta = \left(\frac{\mu}{\sigma^2}, \frac{1}{\sigma^2}\right), h = \frac{1}{\sqrt{2\pi}}, A = \log\sigma + \frac{\mu^2}{2\sigma^2}$ - Correct

Multivariate normal distribution

- $f_{\mu,\Sigma}(x) = \dfrac{1}{(2\pi)^{m/2}|\Sigma|^{1/2}} \exp\left\{-\dfrac{1}{2}(x-\mu)^t\Sigma^{-1}(x-\mu)\right\}$

- $T : \mathcal{X} \to \mathbb{R}^m \times \mathbb{R}^{m(m+1)/2}$ given by

$$T(x) = (x_1, \ldots, x_m, -x_1^2/2, \ldots, -x_m^2/2, -x_1x_2, \ldots, -x_{m-1}x_m)^t$$

- $h(x) = (2\pi)^{m/2}$ for all $x \in \mathbb{R}^m$

- $\eta(\theta) = (\Sigma^{-1}\mu, \Sigma^{-1})$

- $A(\theta) = \dfrac{1}{2}\mu^t\Sigma^{-1}\mu + \dfrac{1}{2}\log|\Sigma|$

Concentration matrix

- The inverse of the covariance matrix plays an important role for Gaussian models as a natural parameter of the exponential family.
- It is called the concentration matrix or the precision matrix.
- It is often denoted $K = \Sigma^{-1}$.

Gaussian exponential families

- Choose a statistic $T(x)$ that maps $x \in \mathbb{R}^m$ to a vector of degree 2 polynomials with no constant term.

- This gives a subfamily of a regular multivariate Gaussian statistical model.

- Equivalently take a linear subspace L of the parameter space $\mathbb{R}^m \times PD_m$ of the regular exponential family.

- Commonly $L = L_1 \times L_2$ where $L_1 \subseteq \mathbb{R}^m$ and $L_2 \subseteq \mathbb{R}^{(m+1)m/2}$.

- Often L_1 is either $\{0\}$ or $\mathbb{R}^m$.

Gaussian exponential families

- Vanishing ideal is a subset of $\mathbb{R}[\mu, \sigma] := \mathbb{R}[\mu_i, \sigma_{ij} : 1 \leq i \leq j \leq m]$.
- If $L_1 = 0$, then the vanishing ideal has the form $\langle \mu_1, \ldots, \mu_m \rangle + I_2$, where I_2 is an ideal in $\mathbb{R}[\sigma]$.
- If $L_1 = \mathbb{R}^m$, then the vanishing ideal is generated by polynomials in $\mathbb{R}[\sigma]$.

Inverse linear space

- Exponential subfamily is a linear space in the space of concentration matrices.

- One is often interested in describing Gaussian models in the space of covariance matrices.

Def: Let $L \subseteq \mathbb{R}^{(m+1)m/2}$ be a linear space such that $L \cap PD_m$ is nonempty. The inverse linear space L^{-1} is the set of positive definite matrices

$$L^{-1} = \{K^{-1} : K \in L \cap PD_m\}.$$

- Gaussian exponential families have interesting ideals in $\mathbb{R}[\sigma]$.

Gaussian exponential families

Prop: If K is a concentration matrix for a Gaussian random vector, a zero entry $k_{ij} = 0$ is equivalent to a conditional independence statement $i \perp\!\!\!\perp j \,|\, [m] \backslash \{i, j\}$.

- The CI ideals that arise from zeros in the concentration matrix might not be primary.
- The linear space L in the concentration coordinators is irreducible and this allows us to parametrize the main component of the CI ideal.

Gaussian exponential families

- Let $m = 3$. Consider the Gaussian exponential family defined by the linear space of concentration matrices $L = \{K \in PD_3 : k_{12} = 0, k_{13} = 0\}$.

- This corresponds to CI statements $1 \perp\!\!\!\perp 2 \mid 3$ and $1 \perp\!\!\!\perp 3 \mid 2$.

- $J_{\mathscr{C}} = \langle \sigma_{12}\sigma_{33} - \sigma_{13}\sigma_{23}, \sigma_{13}\sigma_{22} - \sigma_{12}\sigma_{23} \rangle$

- The intersection axiom implies $1 \perp\!\!\!\perp \{2,3\}$, but no linear polynomials in $J_{\mathscr{C}}$. One option is to compute a primary decomposition of $J_{\mathscr{C}}$.

- Alternatively, we can use the parametrization of the model to compute the vanishing ideal.

lecture5.m2

```
restart
R = QQ[k11,k22,k23,k33,s11,s12,s13,s22,s23,s33]
K = matrix {{k11,0,0},{0,k22,k23},{0,k23,k33}}
S = matrix {{s11,s12,s13},{s12,s22,s23},{s13,s23,s33}}
I = ideal (K*S - identity(1))
J = eliminate ({k11,k22,k23,k33},I)
```

```
-:---  lecture5.m2     All L7     (Macaulay2)
```

```
i1 : R = QQ[k11,k22,k23,k33,s11,s12,s13,s22,s23,s33]

o1 = R

o1 : PolynomialRing

i2 : K = matrix {{k11,0,0},{0,k22,k23},{0,k23,k33}}

o2 = | k11 0   0   |
     | 0   k22 k23 |
     | 0   k23 k33 |

             3       3
o2 : Matrix R  <--- R

i3 : S = matrix {{s11,s12,s13},{s12,s22,s23},{s13,s23,s33}}

o3 = | s11 s12 s13 |
     | s12 s22 s23 |
     | s13 s23 s33 |

             3       3
o3 : Matrix R  <--- R

i4 : I = ideal (K*S - identity(1))

o4 = ideal (k11*s11 - 1, k22*s12 + k23*s13, k23*s12 + k33*s13, k11*s12, k22*s22
     --------------------------------------------------------------------------
     + k23*s23 - 1, k23*s22 + k33*s23, k11*s13, k22*s23 + k23*s33, k23*s23 +
     --------------------------------------------------------------------------
     k33*s33 - 1)

o4 : Ideal of R

i5 : J = eliminate ({k11,k22,k23,k33},I)

o5 = ideal (s13, s12)

o5 : Ideal of R

i6 :
```

```
U:**-  *M2*        Bot L58    (Macaulay2 Interaction:run)
```

Group projects

- Group projects take place instead of lectures 8-10
- Each group is assigned a different topic, in total four groups
- Each group meets 2-3 times during the first two weeks and prepares a presentation
- During the third week, everyone presents the topic they studied to three students who have studied a different topic
- Goal: Learn one topic in depth and basics about other topics

The method of moments

- “Algebraic Statistics of Gaussian Mixtures” by Carlos Amendola (Chapters 1 and 4)
- A method for estimating parameters of a model based on a dataset
- Focus on Gaussian mixtures
- Dataset: Naples’ crabs (1894 Karl Pearson)

The cone of sufficient statistics

- Chapter 8 of “Algebraic Statistics”
- A description of the cone of sufficient statistics
- Existence of the maximum likelihood estimate depends on whether a sufficient statistic is in the interior or on the boundary of the cone
- Polyhedral geometry
- Discrete exponential families
- Gaussian exponential families: matrix completion of positive semidefinite matrices

Exponential random graph models

- Chapter 11 of “Algebraic Statistics”
- Statistical models frequently used in the study of social networks
- Erdös-Renyi random graphs, stochastic block model, beta model
- Samples are typically not i.i.d.
- Hypothesis testing using Fisher’s exact test [Lecture 7]

Phylogenetic models

- Chapter 15 of “Algebraic Statistics” (some subchapters)
- Phylogenetics studies the evolutionary history of a collection of species
- A binary tree and a type of phylogenetic models gives a statistical model
- Goal: Reconstruct the evolutionary tree from data (e.g. aligned DNA sequences)

Next time

- Email me (kaie.kubjas@aalto.fi) your preferred topics
- Plan for the rest of the course:
  - Next time: Likelihood inference
  - Lecture 7: Fisher’s exact test
  - Instead of lectures 8-10: Group work
  - Lectures 11-12: Graphical models