Likelihood inference

Kaie Kubjas, 14.10.2020

- Send me preferences for group work topics today.
- From Homework 3, one can resubmit only problems for which there was an attempt at the original submission.
- Period II: Matrix Theory course by Vanni Noferini
- Hourly-based teacher positions for spring 2021 (deadline November 5)

Agenda

- Gaussian exponential families
- Parameter estimation
- Maximum likelihood estimation
    - Implicit models
    - Exponential families

Gaussian exponential families

Exponential families

- Let X be a random variable taking values in a set $\mathscr{X}$.
- 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 $T : \mathscr{X} \to \mathbb{R}^k$, natural parameter $\eta : \Theta \to \mathbb{R}^k$, and functions $h : \mathscr{X} \to \mathbb{R}_{>0}$ and $A : \Theta \to \mathbb{R}$.

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 $x \in \mathbb{R}^m$
- $\eta(\theta) = (\Sigma^{-1}\mu, \Sigma^{-1})$
- $A(\theta) = \dfrac{1}{2}\mu^t\Sigma^{-1}\mu + \dfrac{1}{2}\log|\Sigma|$

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.

- Example: Let $m = 3$ and $T(x) = (x_1, x_2, x_3, -x_1^2/2, -x_2^2/2, -x_3^2/2, -x_2x_3)^t$.

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

- Example: Let $m = 3$ and $L = \mathbb{R}^3 \times \{K \in PD_3 : k_{12} = 0, k_{13} = 0\}$.

Inverse linear space

- We focus on the cases $\{0\} \times L$ or $\mathbb{R}^m \times L$. Then exponential subfamily is determined by a linear space in the space of concentration matrices.

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

<u>Def:</u> 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

- The vanishing ideal of L^{-1} is a subset of $\mathbb{R}[\sigma] := \mathbb{R}[\sigma_{ij} : 1 \leq i \leq j \leq m]$.
- Gaussian exponential subfamilies 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 Gaussian exponential 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)

Parameter estimation

Parameter estimation

- A typical problem in statistics: Given a parametric model, estimate some or all parameters of the model based on data.
  - Maximum likelihood estimation [today]
  - Method of moments [one of the group projects]
- Do not assume that the model accurately fits the data -> hypothesis testing [next time for discrete exponential families]

Parameter estimation

Def: Let $\mathscr{M}_\Theta$ be a parametric statistical model. Suppose we want to estimate a fixed parameter θ. An estimator of θ is a function $\hat{\theta}$ from the state space to $\mathbb{R}$ that is used to infer the value of θ.

Example: Consider the family of binomial distributions Bin$(2,\theta)$

$$\left\{ \left(\theta^2, 2\theta(1-\theta), (1-\theta)^2\right) : \theta \in [0,1] \right\}.$$

Let $X^{(1)}, \ldots, X^{(n)}$ be i.i.d. samples from a distribution p_θ in this family. Let $u = (u_0, u_1, u_2)$ be the vector of counts, i.e. $u_j = \#\{i : X^{(i)} = j\}$. Then $\sqrt{\frac{u_0}{n}}$ is an estimator of the parameter θ.

Def: The estimator $\hat{\theta}$ is consistent if $\hat{\theta}$ converges to θ in probability as the sample size tends to infinity, i.e.

$$\lim_{n \to \infty} P(\|\hat{\theta}_n - \theta\|_2 > \epsilon) = 0 \text{ for all } \epsilon > 0.$$

Maximum likelihood estimation

- Let $D = \{X^{(1)}, X^{(2)}, \ldots, X^{(n)}\}$ be data from some model with parameter space Θ.

- Likelihood function (discrete case): $L(\theta \mid D) := p_\theta(D)$ - the probability of observing the data D given the parameter θ

- Likelihood function (continuous case): $L(\theta \mid D) := f_\theta(D)$ - the value of the density function evaluated at the data

- The maximum likelihood estimate $\hat{\theta}$ is the maximizer of the likelihood function:

$$\hat{\theta} = \text{argmax}_{\theta \in \Theta} L(\theta \mid D).$$

Maximum likelihood estimation

I.i.d. sampling: $L(\theta \mid D) = \prod_{i=1}^{n} L(\theta \mid X^{(i)})$

- Likelihood function (discrete case): $L(\theta \mid D) = \prod_{i=1}^{n} p_\theta(X^{(i)})$
- Let $u \in \mathbb{N}^r$ be the vector of counts, i.e. $u_j = \#\{i : X^{(i)} = j\}$: $L(\theta \mid D) = \prod_{i=1}^{n} p_\theta(X^{(i)}) = \prod_{j=1}^{r} p_\theta(j)^{u_j}$
- Example for $\left\{ \left(\theta^2, 2\theta(1-\theta), (1-\theta)^2\right) : \theta \in [0,1] \right\}$: $L(\theta \mid D) = (\theta^2)^{u_0} \cdot (2\theta(1-\theta))^{u_1} \cdot ((1-\theta)^2)^{u_2}$
- Likelihood function (continuous case): $L(\theta \mid D) = \prod_{i=1}^{n} f_\theta(X^{(i)})$

Log-likelihood function

- The log-likelihood function is

$$l(\theta \mid D) = \log L(\theta \mid D)$$

- I.i.d. data: turns a product into a sum
- Example:
  - $L(\theta \mid D) = (\theta^2)^{u_0} \cdot (2\theta(1-\theta))^{u_1} \cdot ((1-\theta)^2)^{u_2}$
  - $l(\theta \mid D) = u_0 \log(\theta^2) + u_1 \log(2\theta(1-\theta)) + u_2 \log((1-\theta)^2)$
- The likelihood and log-likelihood function have the same maximizer, because logarithm is a monotone function

Breakout rooms

Score equations

Let $\Theta \subseteq \mathbb{R}^d$ be an open full-dimensional parameter set.

<u>Def:</u> The score equations or critical equations of the model $\mathcal{M}_\Theta$ are the equations obtained by setting the gradient of the log-likelihood function to zero:

$$\frac{\partial}{\partial \theta_i} l(\theta \mid D) = 0, \quad i = 1, \ldots, d.$$

Score equations example

$\mathcal{M}_{X \perp\!\!\!\perp Y} = \{p = \begin{pmatrix} p_{11} & p_{12} \\ p_{21} & p_{22} \end{pmatrix} \in \Delta_3 : p_{ij} = \alpha_i \beta_j, (\alpha, \beta) \in \Delta_1 \times \Delta_1\}$ and $u = \begin{pmatrix} 19 & 141 \\ 17 & 149 \end{pmatrix}$

Log-likelihood function: $l(\alpha, \beta \,|\, u) = 160 \log \alpha_1 + 166 \log \alpha_2 + 36 \log \beta_1 + 290 \log \beta_2$

$$= 160 \log \alpha_1 + 166 \log(1 - \alpha_1) + 36 \log \beta_1 + 290 \log(1 - \beta_1)$$

Score equations:

$$\frac{\partial l(\alpha, \beta \,|\, u)}{\partial \alpha_1} = \frac{160}{\alpha_1} - \frac{166}{1 - \alpha_1} = 0$$

$$\frac{\partial l(\alpha, \beta \,|\, u)}{\partial \beta_1} = \frac{36}{\beta_1} - \frac{290}{1 - \beta_1} = 0$$

Score equations

- Since Θ is open, the maximum likelihood estimate might not exist.
- If Θ were closed, then the maximum likelihood estimate might not be a solution to the score equations.

Discrete setup

- A parametric model given by a rational map $p : \Theta \to \Delta_{r-1}$
- I.i.d. samples $X^{(1)}, \ldots, X^{(n)}$ such that each $X^{(i)} \sim p$ for some unknown distribution p
- The vector of counts $u \in \mathbb{N}^r$, given by $u_j = \#\{i : X^{(i)} = j\}$
- Log-likelihood function $l(\theta \,|\, u) = \sum_{j=1}^{r} u_j \log p_j$
- Score equations $\sum_{j=1}^{r} \frac{u_j}{p_j} \frac{\partial p_j}{\partial \theta_i} = 0$

ML degree

Theorem: Let $\mathscr{M}_\Theta \subseteq \Delta_{r-1}$ be a statistical model. For generic data, the number of solutions to the score equations is independent of u.

Generic = data is outside a variety

Def: The number of solutions to the score equations for generic u is called the maximum likelihood degree (ML degree) of the parametric discrete statistical model $\mathscr{M}_\Theta$.

Implicit models

Implicit models

- Implicit models are given as the intersection of the interior of the probability simplex $\text{int}(\Delta_{r-1})$ and the variety $V(I)$, where $I = \langle g_1, \ldots, g_k \rangle$.

- Let us denote it by $V_{\text{int}(\Delta)}(I)$. Given a vector of counts $u = (u_1, \ldots, u_r)$, we would like to maximize the log-likelihood function

$$l(p \mid u) = \sum_{i=1}^{r} u_i \log p_i$$

over $V_{\text{int}(\Delta)}(I)$.

Implicit models example

- $\mathscr{M}_{X \perp\!\!\!\perp Y} = \{P = \begin{pmatrix} p_{11} & p_{12} \\ p_{21} & p_{22} \end{pmatrix} \in \Delta_3 : p_{11}p_{22} - p_{12}p_{21} = 0\}$ and $u = \begin{pmatrix} 19 & 141 \\ 17 & 149 \end{pmatrix}$
- Want to maximize $l(p \mid u) = 19 \log p_{11} + 141 \log p_{12} + 17 \log p_{21} + 149 \log p_{22}$ over $\mathscr{M}_{X \perp\!\!\!\perp Y}$.
- The constraints are $p_{11} + p_{12} + p_{21} + p_{22} = 1$ and $p_{11}p_{22} - p_{12}p_{21} = 0$.

Lagrange multipliers

- Recall that the method of Lagrange multipliers is used to solve the following constrained optimization problem:

$$\max f(x)$$

$$\text{subject to } g_i(x) = 0 \text{ for } i = 1,\ldots,k$$

- The Lagrangian of this optimization problem is

$$L(x, \lambda) = f(x) - \sum_{i=1}^{k} \lambda_i g_i(x).$$

- Example: $L(x, \lambda) = l(p \mid u) - \lambda_1(p_{11} + p_{12} + p_{21} + p_{22} - 1) - \lambda_2(p_{11}p_{22} - p_{12}p_{21})$

Lagrange multipliers

The constrained critical points of f are among the unconstrained critical points of L. Hence one has to solve

$$g_1 = 0, \ldots, g_k = 0,$$

$$\frac{\partial f}{\partial x_1} - \sum_{i=1}^{k} \lambda_i \frac{\partial g_i}{\partial x_1} = 0, \ldots, \frac{\partial f}{\partial x_m} - \sum_{i=1}^{k} \lambda_i \frac{\partial g_i}{\partial x_r} = 0$$

Lagrange multipliers

The gradient of the log-likelihood function is $\begin{pmatrix} \frac{u_1}{p_1} & \cdots & \frac{u_r}{p_r} \end{pmatrix}$. Hence:

$$g_1 = 0, \ldots, g_s = 0,$$

$$\frac{u_1}{p_1} - \sum_{i=1}^{k} \lambda_i \frac{\partial g_i}{\partial p_1} = 0, \ldots, \frac{u_r}{p_r} - \sum_{i=1}^{k} \lambda_i \frac{\partial g_i}{\partial p_r} = 0$$

Lagrange multipliers

- Clearing the denominators gives a system of polynomial equations:

$$g_1 = 0, \ldots, g_s = 0,$$

$$u_1 - p_1 \sum_{i=1}^{k} \lambda_i \frac{\partial g_i}{\partial p_1} = 0, \ldots, u_r - p_r \sum_{i=1}^{k} \lambda_i \frac{\partial g_i}{\partial p_r} = 0$$

- When clearing the denominators, one might introduce new solutions where one of the p_i is zero (but this happens only if one of u_i is zero)

Lagrange multipliers

- In the statistical setting, one constraint is $p_1 + \ldots + p_r = 1$. Set $g_0 = p_1 + \ldots + p_r - 1$.
- Then $u_1 - p_1 \sum_{i=0}^{k} \lambda_i \frac{\partial g_i}{\partial p_1} = 0, \ldots, u_r - p_r \sum_{i=0}^{k} \lambda_i \frac{\partial g_i}{\partial p_r} = 0$ is equivalent to u being in the row span of the augmented Jacobian matrix

$$J' = \begin{pmatrix} p_1 & p_2 & \cdots & p_r \\ p_1 \frac{\partial g_1}{\partial p_1} & p_2 \frac{\partial g_1}{\partial p_2} & \cdots & p_r \frac{\partial g_1}{\partial p_r} \\ \vdots & \vdots & \ddots & \vdots \\ p_1 \frac{\partial g_k}{\partial p_1} & p_2 \frac{\partial g_k}{\partial p_2} & \cdots & p_r \frac{\partial g_k}{\partial p_r} \end{pmatrix}.$$

Lagrange multipliers

- Example:
  $L(x, \lambda) = l(p \mid u) - \lambda_1(p_{11} + p_{12} + p_{21} + p_{22} - 1) - \lambda_2(p_{11}p_{22} - p_{12}p_{21})$

- $p \in V(I)$ is a critical point of $l(p \mid u)$ if u is in the row span of the matrix
  $$\begin{pmatrix} p_{11} & p_{12} & p_{21} & p_{22} \\ p_{11}p_{22} & -p_{12}p_{21} & -p_{12}p_{21} & p_{11}p_{22} \end{pmatrix}$$

Lagrange multipliers

- Consider the ideal I_l generated by: $g_1, \ldots, g_s$,
$$u_1 - p_1 \sum_{i=0}^{k} \lambda_i \frac{\partial g_i}{\partial p_1}, \ldots, u_r - p_r \sum_{i=0}^{k} \lambda_i \frac{\partial g_i}{\partial p_r}.$$

- Whether the variety of the ideal is finite, can be checked with the command dim(I_l): dim=0 means that the system has finitely many solutions.

- If there are finitely many solutions, then the number of solutions can be computed with degree(I_l).

- The solutions can be found for example with the solve command in Mathematica.

Exponential families

Concave functions

Def: A set $S \subseteq \mathbb{R}^d$ is convex if for all $x, y \in S$, also $(x + y)/2 \in S$.

Def: Let S be a convex set.

- A function $f : S \to \mathbb{R}$ is convex if $f((x + y)/2) \leq (f(x) + f(y))/2$ for all $x, y, \in S$.

- A function $f : S \to \mathbb{R}$ is concave if $f((x + y)/2) \geq (f(x) + f(y))/2$ for all $x, y, \in S$.

Concave functions

Prop: Let S be a closed convex set and $f : S \to \mathbb{R}$ be a concave function. Then the set $U \subseteq S$ where f attains its maximum value is a convex set. If f is strictly concave, i.e. $f((x+y)/2) > (f(x)+f(y))/2$ for all $x \neq y$, then f has a unique global maximum, if a maximum exists.

Exponential families

The canonical form of an exponential family is $f_\eta(x) = h(x)e^{\eta^t T(x) - A(\eta)}$

- statistic $T : \mathscr{X} \to \mathbb{R}^k$,
- function $h : \mathscr{X} \to \mathbb{R}_{>0}$, and
- function $A : H \to \mathbb{R}$.

Exponential families

Prop: Let $\mathscr{M}$ be an exponential family with minimal sufficient statistics $T(x)$ and natural parameter η, with density $f_\eta(x) = h(x)e^{\eta^t T(x) - A(\eta)}$. Then the likelihood function is strictly concave. Furthermore, the maximum likelihood estimate, if it exists, is the solution to

$$T(x) = \mathbb{E}_\eta[T(X)],$$

where x denotes the data vector.

I.i.d. samples

I.i.d. samples $X^{(1)}, \ldots, X^{(n)}$ yield a new exponential family with the same parameter η, the sufficient statistic

$$T_n(X^{(1)}, \ldots, X^{(n)}) = \sum_{i=1}^{n} T(X)^{(i)}$$

and with

$$h_n(X^{(1)}, \ldots, X^{(n)}) = \prod_{i=1}^{n} h(X^{(i)}).$$

Discrete exponential families

<u>Cor:</u> Let $A \subseteq \mathbb{Z}^{k\times r}$ such that $\mathbf{1} \in \text{rowspan}(\mathbf{A})$, let $h \in \mathbb{R}^r_{>0}$, and let u be the vector of counts from n i.i.d. samples. Then the maximum likelihood estimate in the log-linear model $\mathscr{M}_{A,h}$ given the data u is the unique solution, if it exists, to the equations

$$Au = nAp \text{ and } p \in \mathscr{M}_{A,h}.$$

Gaussian exponential families

Cor: Let L be a linear space in $\mathbb{R}^{m(m+1)/2}$ such that $L \cap PD_m$ is not empty, and let $\mathbb{R}^m \times \mathscr{M}_{L^{-1}}$ be the corresponding parameter space of the Gaussian exponential family. Let $X^{(1)}, \ldots, X^{(n)} \in \mathbb{R}^m$ be i.i.d. samples and let $\bar{X}$ and S be the corresponding sample mean and sample covariance matrix. Then the maximum likelihood estimate for $(\mu, \Sigma) \in \mathbb{R}^m \times \mathscr{M}_{L^{-1}}$ is $(\bar{X}, \hat{S})$, where $\hat{S}$ is the unique solution, if it exists, to the equations

$$\pi(S) = \pi(\hat{S}) \text{ and } \hat{S} \in \mathscr{M}_{L^{-1}},$$

where π denotes the orthogonal projection onto L.

Next time

- Hypothesis testing for discrete exponential families
- Reading task based on “Algebraic algorithms for sampling from conditional distributions” by Diaconis and Sturmfels