Algebraic Statistics

Based on lectures by Carlos Améndola

Winter 2023

Updated on February 22, 2024
These notes are not endorsed by the lecturer.

Contents

1 Introduction **3**
1.1 Discrete Markov chain . . . 3
1.2 Maximum likelihood . . . 5
1.3 Hypothesis testing . . . 6

2 Probability Primer **8**
2.1 Covariance and covariance matrix . . . 8
2.2 Normal distribution . . . 9

3 Algebra Primer **11**
3.1 Elimination . . . 13
3.2 Implicitization . . . 15
3.3 Primary decomposition . . . 16

4 Conditional Independence **19**
4.1 Conditional independence axioms . . . 20
4.2 Discrete random variables . . . 22
4.3 Gaussian random variables . . . 24
4.4 Primary decomposition . . . 25

5 Statistics Primer **27**
5.1 4.12 . . . 29
5.2 Maximum likelihood estimation . . . 30

6 Exponential families **31**
6.1 Canonical form . . . 32
6.2 Discrete regular exponential families . . . 35
6.3 Gaussian regular exponential families . . . 39

7 Likelihood inference **41**
7.1 Maximum likelihood degree 41
7.2 Discrete models with constraints 48
7.3 Log-affine linear models 51
7.4 Gaussian linear concentration models 53

8 Fisher's exact test **56**
8.1 Asymptotic test 56
8.2 Independence model 57
8.3 Log-affine linear models 61
8.4 Markov bases 62

9 Review **64**

1 Introduction

1.1 Discrete Markov chain

Let $X_1, X_2, ..., X_m$ be a sequence of random variables on the same finite state space Ω. The joint probability distribution is $P(X_1 = x_1, ..., X_m = x_m)$ with $x_i \in \Omega$ such that $\sum P(X_1 = x_1, ..., X_n = x_n) = 1$. Lec01

Definition 1.1 (Markov chain). A sequence $X_1, ..., X_n$ is called a **Markov chain** if the probability of the next state depends only on the current state; that is,

$$\mathbb{P}(X_{n+1} = x_{n+1} \mid X_1 = x_1, ..., X_n = x_n) = \mathbb{P}(X_{n+1} = x_{n+1} \mid X_n = x_n).$$

Motto in algebraic statistics: **Statistical models are semi-algebraic sets**.

Here is an example of a statistical model that can be represented as a solution set of system of polynomial equations and inequalities.

Proposition 1.2 (Implicit Markov chain model). A vector $p \in \mathbb{R}^8$ is the probability distribution from a Markov chain model $\mathcal{M}$ if and only if

- $p \in \Delta_7$,
- $p_{010}p_{111} - p_{011}p_{110}$,
- $p_{000}p_{101} - p_{001}p_{100}$.

The second and third equality is obtained by taking the determinant of the matrix slices $j = 0$ and $j = 1$:

$$\begin{bmatrix} p_{000} & p_{001} \\ p_{100} & p_{101} \end{bmatrix} \quad \text{and} \quad \begin{bmatrix} p_{010} & p_{011} \\ p_{110} & p_{111} \end{bmatrix}.$$

Proof. $\Longrightarrow$: Assume we are given random variable X_1, X_2 and X_3 taking values in $\Omega = \{0, 1\}$ and satisfying the Markov chain property. A joint probability distribution can be given as an $\mathbb{R}^8$ vector

$$\mathbf{p} = \begin{bmatrix} p_{000} \\ p_{001} \\ \vdots \\ p_{111} \end{bmatrix} \in \Delta_7.$$

By definition of conditional probability, we compute

$$P(X_3 = k \mid X_1 = i, X_2 = j) = \frac{p_{ijk}}{p_{ij+}}$$
$$P(X_2 = j \mid X_1 = i) = \frac{p_{ij+}}{p_{i++}}.$$

Using the Markov chain property, we obtain the equality

$$\frac{p_{ijk}}{p_{ij+}} = \frac{p_{ij+}}{p_{i++}}.$$

Similarly, if for some different $i \neq i'$ we obtain $\frac{p_{i'jk}}{p_{i'j+}} = \frac{p_{ij+}}{p_{i++}}$, and thus

$$\frac{p_{ijk}}{p_{ij+}} = \frac{p_{i'jk}}{p_{i'j+}}.$$

This yields $p_{ijk}(p_{i'j0} + p_{i'j1}) = p_{i'jk}(p_{ij0} + p_{ij1})$.

- Substitute $i = j = k = 1, i' = 0$: $p_{111}(p_{010} + p_{011}) = p_{011}(p_{110} + p_{111})$. Thus, $p_{111}p_{010} - p_{011}p_{110} = 0$.
- Substitute $i = j = k = 0, i' = 1$: $p_{000}(p_{100} + p_{101}) = p_{100}(p_{000} + p_{001})$. Thus, $p_{000}p_{101} - p_{100}p_{001} = 0$.
- Substitute $i = j = 1, k = 0, i' = 0$: $p_{110}(p_{010} + p_{011}) = p_{010}(p_{110} + p_{111})$. Thus, $p_{110}p_{011} - p_{010}p_{111} = 0$.
- And so on (we obtain the same equations)...

□

Remark 1.3. A Markov chain model is an example of a **conditional independence model**.

Proposition 1.4 (Parametrization of the Markov chain model). A parametrization of the Markov chain model is given by

$$\varphi(\pi, \boldsymbol{\alpha}) = \begin{bmatrix} \vdots \\ \pi_{j_i} \prod_{i=2}^{m} \alpha_{i,j_{i-1},j_i} \\ \vdots \end{bmatrix} = \begin{bmatrix} p_{000} \\ p_{001} \\ \vdots \\ p_{111} \end{bmatrix} = \mathbf{p} \in \Delta_7$$

where $\pi_{j_i} = P(X_1 = j_i)$ and $\alpha_{i,j_{i-1},j_i} = P(X_i = j_i \mid X_{i-1} = j_{i-1})$.

Example 1.5. Let us consider the case with three random variables X_1, X_2 and X_3. We define $\alpha_{ij} = \alpha_{1ij}$ and $\beta_{jk} = \alpha_{2jk}$. Here are all the parameters

$$\boldsymbol{\pi} = (\pi_0, \pi_1), \boldsymbol{\alpha} = \begin{bmatrix} \alpha_{00} & \alpha_{01} \\ \alpha_{10} & \alpha_{11} \end{bmatrix}, \boldsymbol{\beta} = \begin{bmatrix} \beta_{00} & \beta_{01} \\ \beta_{10} & \beta_{11} \end{bmatrix}.$$

Note that the components of π sum to one as well as the rows of α and β. Thus, we have five *independent variables* $\pi_0, \alpha_{00}, \alpha_{10}, \beta_{00}$ and β_{10}. We later say that $\dim(\mathcal{M}) = 5$.

1.2 Maximum likelihood

Assume we observe the following data D:

$$000, 010, 110, 000, 101, 110, 100, 010, 110, 111, 000, 000, 010.$$

The count vector reads $\mathbf{u} = (4, 0, 3, 0, 1, 1, 3, 1)$. The likelihood function is

$$L_u(p) = \binom{n}{\mathbf{u}} \prod_{i,j,k} p_{ijk}^{u_{ijk}}$$

where $\binom{n}{\mathbf{u}} = \frac{n!}{\prod_{i,j,k} u_{ijk}!}$ is the multinomial coefficient. We want to maximize L_u. When we use the implicit model, we obtain a *constrained optimization problem*

$$\arg\max_{\mathbf{p}\in\Delta_7} \left\{ \prod_{ijk} {p_{ijk}}^{u_{ijk}} \right\} \text{ subject to } \begin{cases} p_{010}p_{111} - p_{011}p_{110} \\ p_{000}p_{101} - p_{001}p_{100} \end{cases}$$

An *unconstrained optimization problem* is given by the parametrized Markov model (although it is not quite unconstrained): optimize

$$\arg\max_{\boldsymbol{\pi},\boldsymbol{\alpha},\boldsymbol{\beta}} \left\{ \prod_{ijk} (\pi_i \alpha_{ij} \beta_{jk})^{u_{ijk}} \right\}$$

with constraints that $\boldsymbol{\pi}, \boldsymbol{\alpha}, \boldsymbol{\beta}$ are probability distributions.

A typical strategy is to compute the logarithm of L_D, equal the partial derivatives of that function to zero and solve the equations. The equations that we set to zero are called **score equations** or **critical equations**.

We define the log-likelihood function

$$l(\pi, \alpha, \beta \mid u) = \sum_{ijk} u_{ijk}(\log(\pi_i) + \log(\alpha_{ij}) + \log(\beta_{jk}))$$

Computing the derivative

$$\frac{\partial l}{\partial \pi_0} = \frac{u_{0++}(1-\pi_0)}{1} = u_{1++}\pi_0$$

Hence $\hat{\pi}_0 = \frac{u_{0++}}{u_{+++}}$ where $u_{+++} = N$ the total sample size. Other solutions $\hat{\alpha}_{10} = \frac{u_{10+}}{u_{1++}}$, $\hat{\beta}_{01} = \frac{u_{+01}}{u_{+0+}}$.

To make a prediction we compute $\hat{p}_{101} = \hat{\pi}_1 \hat{\alpha}_{10} \hat{\beta}_{01}$ where $\hat{\pi}_1 = \frac{u_{1++}}{u_{+++}}$. The estimator is then $\hat{p}_{101} = \frac{u_{10+}u_{+01}}{u_{+++}u_{+0+}}$. In this way

$$\hat{\pi}_i = \frac{u_{i++}}{u_{+++}}, \hat{\alpha}_{ij} = \frac{u_{ij+}}{u_{i++}}, \hat{\beta}_{ij} = \frac{u_{+jk}}{u_{+j+}}, \hat{p}_{ijk} = \frac{u_{ij+}u_{+jk}}{u_{+++}u_{+j+}}.$$

A closed form rational solution $\hat{\mathbf{p}}$ exists for our Markov chain model

$$\hat{p}_{ijk} = \frac{u_{ij+}u_{+jk}}{u_{+++}u_{+j+}}.$$

In general, score equations rarely have simple closed solutions; for instance the solution need not be unique.

Lec02, 04/17/23

Definition 1.6 (Homogeneous Markov chain). A **homogeneous Markov chain** is a Markov chain where the transition matrices stay the same for each step, i.e. $\alpha = \beta$.

1.3 Hypothesis testing

Having observed data, we found a distribution $\hat{p} \in \mathcal{M}$ in the model that 'best' explains the data.

How well does the model fit the data? Do we think that the data was generated by some distribution in the model?

Let p be the true generating distribution of the count u. Hypothesis test: $H_0 : p \in \mathcal{M}$ vs $H_1 : p \notin \mathcal{M}$. (In practice, for example clinical tests, one often wants that H_0 does not occur. One wants to reject H_0. One wants to have enough evidence to reject H_0. Only take H_1 if there is enough evidence.).

Idea: suppose $p \in \mathcal{M}$ and we observe u. Consider other outcomes V that could have been generated by p. The question is: what proportion of such data is more likely to have occurred than u? If this proportion is large, then either

1. hypothesis H_0 is false and H_1 holds
2. we were unlucky (a rare event happened)

For example, we throw a dive and we assume that the dice is fair. Assume we observe that nine out of ten times it was heads. If we assume that H_0 :the dice is fair, that count would be very rare.

No way to know if we were unlucky or H_0 is wrong. In practive, if probability is small enough $< 5\%$, then one says there is enough significant evidence to reject H_0. Otherwise, one says there is not enough evidence to reject.

Problem: probability of observing data v depends on (unknown) p.

Key insight from Fisher: Fisher's exact test: for some models the likelihood function does not depend on exact data but only through a (non-injective) function of the data. This function is called sufficient statistic. Back to $m = 3$ with parameters π, α, β.

$$\begin{aligned} P(\mathbf{u} \mid \boldsymbol{\pi}, \boldsymbol{\alpha}, \boldsymbol{\beta}) &= \binom{n}{u} \prod_{ijk} (\pi_i \alpha_{ij} \beta_{jk})^{u_{ijk}} \\ &= \binom{n}{u} \prod_{ijk} (\pi_i \alpha_{ij})^{u_{ij+}} \prod \beta_{jk}^{u_{+jk}} \end{aligned}$$

The quantities u_{ij+}, u_{+jk} are the *sufficient statistics*.

Fishers ida: compare observed data u with other data v. The definition of a Markov chain is

2 Probability Primer

30.10.2023 — Lecture 5

The covariance matrix is $\mathrm{Var}[\mathbb{X}] \in \mathbb{R}^{m\times m}$ with entries $\mathrm{Cor}(X_i, X_j)$ for $i, j \in \{1, ..., m\}$. The key property is that the covariance matrix is a symmetri positive semidefinite matrix.

Definition 2.1. Correlation

$$\mathrm{Corr}[X, Y] = \frac{\mathrm{Cov}(X, Y)}{\sqrt{\mathrm{Var}[X]}\sqrt{\mathrm{Var}[Y]}}$$

If X and Y are independent, then the correlation is zero.
Cauchy-Schwarz:

$$|\mathrm{Cov}(X, Y)| \leq \sqrt{\mathrm{Var}(X)}\sqrt{\mathrm{Var}(Y)}.$$

$\mathrm{Corr}[X, Y] \in [-1, 1]$
$\mathrm{Corr}[a_0 + a_1X, b_0 + b_1Y] = \mathrm{Corr}[X, Y]$ if $a_1, b_1 > 0$
Correlation matrix

$$\begin{bmatrix} 1 & ... & \mathrm{Corr}(X_1, X_m) \\ \vdots & \ddots & \vdots \\ \mathrm{Corr}(X_m, X_1) & ... & 1 \end{bmatrix}$$

Recall Gaussian: $f_{\mu,\sigma} = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$
Multivariate Gaussian

$$f_{\mu,\sigma} = \frac{1}{\sqrt{(2\pi)^m \det(\Sigma)}} e^{-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)}$$

with $\mathbb{E}[X] = \mu$ and $\mathrm{Var}[X] = \Sigma$. We write $X \sim N_m(\mu, \Sigma)$.

Theorem 2.2. Let $X \sim N_m(\mu, \Sigma)$. Let $I, J \subset [m]$ disjoint.

1. $X_J \sim N(\mu_J, \Sigma_{JJ})$
2. The random variable $X_U \mid X_j = x_j \sim N_{\#I}(\mu_I + \Sigma_{IJ}\Sigma_{JJ}^{-1}(x_J - \mu_J), \Sigma_{II} - \Sigma_{IJ}\Sigma_{JJ}^{-1}\Sigma_{JI})$. Note that $\Sigma_{II} - \Sigma_{IJ}\Sigma_{JJ}^{-1}\Sigma_{JI}$ is called the Schur component.
3. If $A \in \mathbb{R}^{k\times m}$ full rank, $k \leq m$, $b \in \mathbb{R}^k$, then $AX + b \sim N(A\mu + b, A\Sigma A^T)$

2.1 Covariance and covariance matrix

Definition 2.3 (Covariance). Let X, Y be random variables with finite second moments. The covariance is defined as

$$\mathrm{cov}(X, Y) = \mathbb{E}[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])].$$

The covariance is also an indicator for the strength of a linear relationship between X and Y.

Definition 2.4 (Correlation). The correlation ρ between X and Y is defined as

$$\rho_{X,Y} = \frac{\mathrm{cov(X, Y)}}{\sqrt{\mathrm{Var}(X)}\sqrt{\mathrm{Var}(Y)}}.$$

Let $\mathbf{X}$ be a random vector. The covariance matrix is a square matrix giving the covariance between each pair of elements of $\mathbf{X}$.

Proposition 2.5 (Properties). A covariance matrix is symmetric, positive semi-definite and its main diagonals contains variances. It is positive definite, if all the components of $\mathbf{X}$ are linearly independent.

2.2 Normal distribution

The **standard *univariate* normal distribution** $\mathcal{N}(0, 1)$ has as a density function the *Gaussian bell curve*

$$\phi(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2}.$$

In general, the ***univariate*** **normal distribution** $\mathcal{N}(\mu, \sigma)$ has as a density function

$$\phi(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2\sigma^2}(x-\mu)^2}.$$

Suppose $\mathbf{X} = (X_1, ..., X_m)$ is a random vector and each of its components $X_1, ..., X_k$ are independent and identically distributed as $X_i \sim \mathcal{N}(0, 1)$. Then, we say that $\mathbf{X}$ is distributed according to a **standard *multivariate* normal distribution** $\mathcal{N}_m(\mathbf{0}, \mathbf{I}_m)$; here $\mathbf{I}_m$ is the $m \times m$-identity matrix. The density function is

$$\phi_{\mathbf{0},\mathbf{I}}(\mathbf{x}) = \prod_{i=1}^{m} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x_i^2} = \frac{1}{(2\pi)^{\frac{m}{2}}} \cdot e^{-\frac{1}{2}\mathbf{x}^T\mathbf{x}}.$$

Let $\mu \in \mathbb{R}^m$ and $\Lambda \in \mathbb{R}^{m\times m}$ be of full rank m (so it is invertible). Define a transformation $\mathbf{Y} = \Lambda\mathbf{X} + \mu$. The new density function reads

$$\phi_{\mathbf{Y}}(\mathbf{y}) = \frac{1}{(2\pi)^{\frac{m}{2}}} e^{-\frac{1}{2}(\mathbf{y}-\mu)^T\Lambda^{-T}\Lambda^T(\mathbf{y}-\mu)} \cdot \det(\Lambda^{-1})$$

Define $\Sigma = \Lambda\Lambda^T$; then $\mathbf{Y}$ depends only on μ and some positive definite matrix Σ. We say $\mathbf{Y}$ is a **multivariate normal random vector**. So, any multivariate normal random vector is a non-singular affine transformation of a standard multivariate normal random vector.

Let $\Sigma \in \mathrm{PD}(m)$ be a positive definite matrix. The density function of the ***multivariate* normal distribution** $\mathcal{N}_m(\mu, \Sigma)$ is

$$\phi_{\mu,\Sigma}(\mathbf{y}) = \frac{1}{(2\pi)^{\frac{m}{2}}\sqrt{\det(\Sigma)}} e^{-\frac{1}{2}(\mathbf{y}-\mu)^T\Sigma^{-1}(\mathbf{y}-\mu)}.$$

The expectation of $\mathcal{N}_m(\mu, \Sigma)$ is μ and the covariance matrix is Σ.

Theorem 2.6 (Marginal distribution is normal). Let $B \subset [m]$, and $\mathbf{X} \sim \mathcal{N}(\mu, \Sigma)$. Then, the marginal distribution $\mathbf{X}_B$ of $\mathbf{X}$ over B is multivariate normal:

$$\mathbf{X}_B \sim \mathcal{N}(\mu_B, \Sigma_{B,B}).$$

Proof. Define $A = [m] \setminus B$. We marginalize over x_A (that is integrating out x_A). The covariance matrix can be partitioned into

$$\Sigma = \begin{bmatrix} \Sigma_{A,A} & \Sigma_{A,B} \\ \Sigma_{B,A} & \Sigma_{B,B} \end{bmatrix}.$$

Then,

$$\Sigma^{-1} = \begin{bmatrix} I & 0 \\ -\Sigma_{B,B}^{-1}\Sigma_{B,A} & I \end{bmatrix} \begin{bmatrix} & 0 \\ -\Sigma_{B,B}^{-1}\Sigma_{B,A} & I \end{bmatrix} \begin{bmatrix} I & 0 \\ -\Sigma_{B,B}^{-1}\Sigma_{B,A} & I \end{bmatrix}$$

... □

Theorem 2.7 (Conditional distribution is normal). Let $A, B \subset [m]$ be disjoint. Fix some $x_B \in \mathbb{R}^B$. The conditional distribution of X_A given $X_B = x_B$ is the multivariate normal distribution

$$\mathcal{N}(\mu_A + \Sigma_{A,B}\Sigma_{B,B}^{-1}(x_B - \mu_B), \Sigma_{A,A} - \Sigma_{A,B}\Sigma_{B,B}^{-1}\Sigma_{B,A}).$$

The conditional covariance matrix $\Sigma_{AA.B}$ is defined as $\Sigma_{A,A} - \Sigma_{A,B}\Sigma_{B,B}^{-1}\Sigma_{B,A}$.

Two random variables are uncorrelated if they are independent. The converse direction may not be true; however for normal distributed random variables it is true!

Proposition 2.8 (Independence equals uncorrelation). Fix a dimension $m \in \mathbb{N}$. Let $\mathbf{X} \sim \mathcal{N}_m(\mu, \Sigma)$. Let $A, B \subset [m]$ be disjoint. Then, $X_A \perp\!\!\!\perp X_B$ if and only if $\Sigma_{A,B} = \mathbf{0}$.

3 Algebra Primer

Here $\mathbb{N} = \{0, 1, 2, 3, ...\}$.

Definition 3.1 (Monomial). An expression of the form $\mathbf{x}^{\mathbf{u}} = x_1^{u_1} x_2^{u_2} \dots x_n^{u_n}$ is called a **monomial** where $\mathbf{u} \in \mathbb{N}^n$.

Definition 3.2 (Variety). Let $S \subset k[\mathbf{x}]$ be a subset of polynomials. The **variety** of S is defined as

$$V(S) = \{\mathbf{a} \in k \mid f(\mathbf{a}) = 0,\ \forall f \in S\}$$

Definition 3.3 (Vanishing ideal). Let $W \subset k^n$. We define

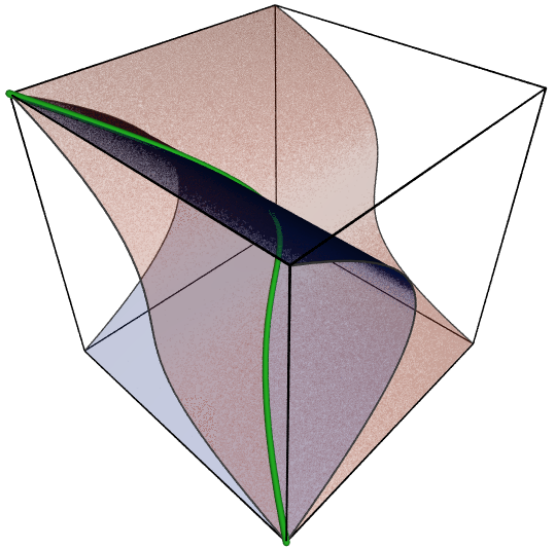
$$I(W) = \{f \in k[\mathbf{x}] \mid f(\mathbf{w}) = 0\ \forall \mathbf{w} \in W\}.$$

In algebraic statistics we often encounter varieties presented as ***parametric sets***

$$V(S) = \left\{\phi(t) : t \in k^d\right\} = \phi(k^d).$$

Example 3.4 (Twisted cube). Consider $\phi : \mathbb{C} \to \mathbb{C}^3 : t \mapsto (t, t^2, t^3)$. We have

$$V(\left\{x^2 - y, x^3 - z\right\}) = \phi(k^d).$$



Example 3.5 (Binomial model). Consider the map $\phi : \mathbb{C} \to \mathbb{C}^{n+1}$ given by $\phi_i(t) = \binom{n}{i} t^i (1-t)^{n-i}$ or alternatively

$$\phi : t \mapsto \begin{bmatrix} \binom{n}{0}(1-t)^n \\ \binom{n}{1} t(1-t)^{n-1} \\ \vdots \\ \binom{n}{n-1} t^{n-1}(1-t) \\ \binom{n}{n} t^n \end{bmatrix}.$$

The map $\phi_i(t)$ computes the probability to observe i-many successes where a success occurs with probability t. In other words, $\phi(t)$ denotes the probability distribution of a binomial random variable $X \sim \mathrm{Bin}(n, t)$.

The image $\phi([0, 1])$ consists of all possible probability distributions of a binomial random variable; it is a curve inside the probability simplex $\Delta_n \subset \mathbb{R}^{n+1}$.

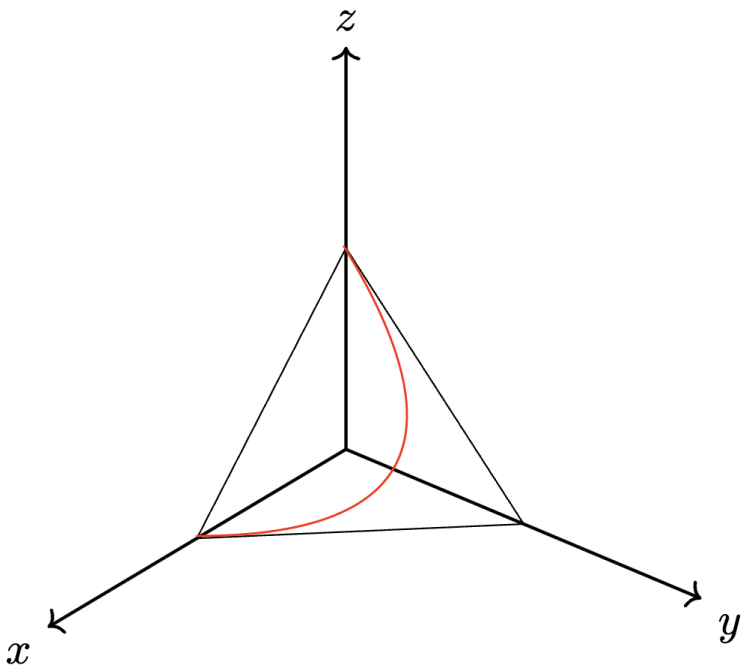


Figure 1: The red curves shows all binomial distributions in the binomial model.

The collection of all probability distributions $\phi(t)$ for $t \in [0, 1]$ will later be called a **parametric statistical model**.

Theorem 3.6 (Hilbert basis theorem). Every ideal $I \subset k[\mathbf{x}]$ is finitely generated.

Remark 3.7 (Implicitization problem). Given a rational map $\phi : k^d \to k^n$, we want to find a generating set of the ideal $I(\phi(k^d))$. This problem is called the **implicitization problem**.

Proposition 3.8 (Implicit binomial statistical model). The binomial model $\mathcal{M}_t = \{\mathbf{p} \mid \mathbf{p} \text{ is a binomial distribution}\}$ consists of points $\mathbf{x} \in \mathbb{R}^{n+1}$ inside Δ_n that satisfy all 2×2 minors of the matrix

$$\begin{bmatrix} \frac{x_0}{\binom{n}{0}} & \frac{x_1}{\binom{n}{1}} & \frac{x_2}{\binom{n}{2}} & \cdots & \frac{x_{n-1}}{\binom{n}{n-1}} \\ \frac{x_1}{\binom{n}{1}} & \frac{x_2}{\binom{n}{2}} & \frac{x_3}{\binom{n}{3}} & \cdots & \frac{x_n}{\binom{n}{n}} \end{bmatrix}.$$

Proof. We will solve it later using implicitization techniques. □

Example 3.9 (Binomial model). For $\phi_i(t) = \binom{2}{i} t^i (1-t)^{2-i}$ we have that

$$I(\phi(\mathbb{C})) = (x + y + z - 1, 4xz - y^2).$$

So the two polynomials $x + y + z - 1$ and $4xz - y^2$ solve the implicitization problem.

Theorem 3.10 (Nullstellensatz). If $k = \bar{k}$, then $I(V(I)) = \sqrt{I}$ for all ideals $I \subset k[\mathbf{x}]$.

Definition 3.11 (Zariski closure). Let $W \subset k^n$ be a subset. The set $V(I(W))$ is called the **Zariski closure** of W. It is the smallest algebraic variety that contains W.

3.1 Elimination

Lec, 11/06/23

Definition 3.12 (Elimination ideal). Let I be an ideal in $k[x_1, \ldots, x_n]$ and $j \in \{1, \ldots, n\}$. The j-th elimination ideal of I is

$$I_j = I \cap k[x_{j+1}, x_{j+2}, \ldots, x_n] \subset k[x_{j+1}, \ldots, x_n].$$

Proposition 3.13 (Vanishing ideal of projection). Let $V \subset k^{r+s}$ be a variety. Let $\pi : k^{r+s} \to k^r$ be the projection. Then, the vanishing ideal of $\pi(V)$ is the ideal $I(\pi(V)) = I \cap k[x_1, \ldots, x_r]$.

Proof. No proof. □

Lemma 3.14 (Image of a variety under projection). Let I_j be the j-th elimination ideal of $I \subset k[x_1, \ldots, x_n]$ and $\pi_j : k^n \to k^{n-j}$. Then $\pi_j(V(I)) \subset V(I_j)$.

Proof. Let $(a_{j+1}, \ldots, a_n) \in \pi_j(V(I))$; hence there exists $\mathbf{a}$ such that $\mathbf{a} = (a_1, \ldots, a_n) \in V(I)$. Let $f \in I_j \subset k[x_{j+1}, \ldots, x_n]$. This f is also in $I \subset k[x_1, \ldots, x_n]$. So $f(a_1, ..., a_n) = 0$ where $f \in I$. Hence, $\pi_j(V(I)) \subset V(I_j)$. □

Lemma 3.15. If $k = \bar{k}$, then $I(\pi_j(V(I))) \subset \sqrt{I_j}$.

Proof. Let Hilbert's Nullstellensatz (HNS) do all the work! Take some $f \in I(\pi_j(V(I))) \subset k[x_{j+1}, \ldots, n]$. It means $f(a_{j+1}, \ldots, a_n) = 0$ for all $a = (a_1, \ldots, a_n) \in V(I)$ or in other words $f(a_1, \ldots, a_n) = 0$ in $k[x_1, \ldots, x_n]$ for all $a \in V(I)$. Hence $f \in I(V(I)) = \sqrt{I}$ by HNS. Hence $f^m \in I$ for some m. Hence f^m still cannot contain variables $x_1, \ldots, x_j$. Hence $f^m \in k[x_{j+1}, \ldots, x_n] = I_j$. Thus, $f \in \sqrt{I_j}$. □

Theorem 3.16 (Closure Theorem). If $k = \bar{k}$, then $\overline{\pi_j(V(I))} = V(I_j)$.

Proof. Use Lemma 3.14: $\overline{\pi_j(V(I))} \subset \overline{V(I_j)} = V(I_j)$.

Use Lemma 3.15 and applying $V(\cdot)$ yields $V(\sqrt{I_j}) = V(I_j) \subset V(I(\pi_j(V(I))) = \overline{\pi_j(V(I))}$. □

Example 3.17. Let same $I = (xy - 1) \subset k[x, y]$. Then $V(I) = \{(x, \frac{1}{x}) \mid x \neq 0\}$. Then $\pi_1(V(I)) = k \setminus \{0\} \subset k = V(I_1)$ where $I_1 = \{0\}$.

Example 3.18. Let $k = \mathbb{R}$. Let $I = (x^2 + y^2) \subset \mathbb{R}[x, y]$. Then $V(I) = \{(0, 0)\}$. Then $\pi_1(V(I)) = \{0\} \subset \mathbb{R} = V(\{0\})$ with $I_1 = \{0\}$.

Lec, 11/07/23

Definition 3.19. A **j-th elimination ordering** is a term ordering where any monomial with $x_i, i \leq j$ indeterminate appearing to any nonzero power is larger than every monomial only in $x_i, i > j$ indeterminates.

Theorem 3.20 (Elimination Theorem). Let I be an ideal in $k[x_1, ..., x_n]$ with Gröbner basis G with respect to a j-th elimination ordering. Then $G_j = G \cap k[x_{j+1}, ..., x_n]$ is a Gröbner basis of I_j with respect to $\prec$.

Proof. To show that G_j is a Gröbner basis of I_j we need to show

- $G_j \subset I_j$: this follows because $G \subset I$;
- $\text{in}_\prec(G_j) \subset \text{in}_\prec(I_j)$: this is clear;
- $\text{in}_\prec(G_j) \supset \text{in}_\prec(I_j)$: Let $f \in I_j$. Then $f \in I$. Therefore, there exist some $g \in G$ such that $\text{in}_\prec(g)$ divides $\text{in}_\prec(f)$. Note that $\text{in}_\prec(f)$ only contains variables $x_{j+1}, ..., x_n$. Hence, $\text{in}_\prec(g)$ does not contain variables $x_1, ..., x_j$. Since we have an elimination order, all the other terms of g do not contain $x_1, ..., x_j$, as well. So $g \in G_j$. Thus, f is generated by some element in $\text{in}_\prec(G_j)$, namely g.

□

Corollary 3.21 (Lexicographic ordering). An ideal $I \subset k[x_1, ..., x_n]$, G a Gröbner basis of I with respect to lex $x_1 > x_2 > ... > x_n$ implies that $G_j = G \cap k[x_{j+1}, ..., x_n]$ is a Gröbner basis of I_j for all $j = 1, ..., n$.

Example 3.22 (Compute the closed image of a variety under projection). Let $k = \mathbb{C}$. Define $f(x, y) = xy^3 - x^2$ and $g(x, y) = x^3y^2 - y$. Consider the ideal $I = (f, g)$ and its variety $V(I)$. What is $\overline{\pi_1(V(I))}$?

- Compute a Gröbner basis of I with respect to lexicographic ordering $x \prec y$. We have $S(f, g) = y^2 - x^4$. The standard residue is $h(x, y) = y^2 - x^4$. So, we add it to G. Computing the S-polynomial between f, h and g, h yields: $k = f - xy(y^2 - x^4) = x^5y - x^2$ and$l = g - x^3(y^2 - x^4) = x^7 - y$. Next, we need to compute the standard residue of k and l. Repeating this process yields the unique reduced Gröbner basis $G = \{y - x^7, x^5y - x^2\}$. Now if we plug $x^7 = y$ into $x^5y - x^2$ we obtain $G = \{y - x^7, x^{12} - x^2\}$.
- Compute $G_1 = \{x^{12} - x^2\}$.
- Compute $\overline{\pi_1(V(I))} = V(I_1) = V(G_1) = V(x^{12} - x^2) = \{0\} \cup \{\xi \mid \xi^{10} = 1\} \subset \mathbb{C}$.

3.2 Implicitization

The *implicitization problem* is concerned with the following problem: Find the generators of the vanishing ideal of an image of a polynomial map; more precisely, given $\varphi : k^d \to k^r$ we want to compute the generators of $I(\varphi(k^d))$.

Example 3.23 (Twisted cube). Let $\varphi : k \to k^3, t \mapsto (t, t^2, t^3)$. Then $I(\varphi(k)) = (y - x^2, z - x^3)$.

Example 3.24 (Binomial model with two trials). Let $X \sim \text{Bin}(2, \theta)$. Then, the probability distribution is $\phi(\theta) = p_\theta = \begin{bmatrix}(1-\theta)^2 & 2\theta(1-\theta) & \theta^2\end{bmatrix}$ for $\theta \in [0, 1]$. What relations are satisfied by the components of p_θ? This is answered by implicitization. We will later see that $p_0 p_2 - \frac{1}{4} p_1^2$ holds.

Example 3.25 (Warning: Image of a polynomial function need not be a variety). Consider $\phi : \mathbb{C}^2 \to \mathbb{C}^2, (\theta_1, \theta_2) \mapsto (\theta_1, \theta_1\theta_2)$. Then, $\text{Im}(\phi) = \{(0, 0)\} \cup \big\{(a, b) \in \mathbb{C}^2 \mid a \neq 0\big\}$, which is not a variety. Note that the image is dense, i.e. $\overline{\text{Im}\phi} = \mathbb{C}^2$.

Remark 3.26 (Constructible set). It can be shown that if we have $k = \bar{k}$ and a polynomial function $\phi : k^d \to k^r$, then there exist a finite sequence of varieties $W_1 \supset W_2 \supset ... \supset W_s$ in k^r such that $\phi(k^d) = W_1 \setminus (W_2 \setminus ... \setminus W_s)$. This is called a *constructible* set.

In the example above, $\phi(\mathbb{C}^2) = \mathbb{C}^2 \setminus (\{(0, a) \mid a \in \mathbb{C}\} \setminus \{(0, 0)\})$.

Note if $k \neq \bar{k}$, then $\phi : \mathbb{R} \to \mathbb{R}, \phi(\theta) = \theta^2$ and $\text{Im}(\theta) = [0, \infty)$. However, it is true that $\phi(\mathbb{R}^d)$ is always a semialgebraic set (given by polynomial inequalities).

Theorem 3.27 (Implicitization theorem). Fix $k = \bar{k}$. Let $\phi = (\phi_1, \ldots, \phi_r) : k^d \to k^r$ be a polynomial. Define the ideal $I = (y_1 - \phi_1, y_2 - \phi_2, \ldots, y_r - \phi_r) \subset k[x_1, ..., x_d, y_1, ..., y_r]$. Then, $\overline{\text{image}(\phi)} = V(I_d)$.

Proof. We have $V(I) = \{(x_1, ..., x_d, y_1, ..., y_r) \mid \phi(x_1, ..., x_d) = (y_1, ..., xy_r)\}$. This is the graph of ϕ. Then

$$\begin{aligned}\pi_d(V) &= \{(x_1, ..., x_r) \mid \phi_1(t_1, ..., t_d) = x_1, ..., \phi_r(t_1, ..., t_d) = x_r\} \\ &= \Big\{(\phi_1(t_1, ..., t_d), ..., \phi_r(t_1, ..., t_d)) \mid (t_1, ..., t_d) \in k^d\Big\} = \phi(k^d).\end{aligned}$$

In words, the image of the parametrization is the projection of its graph. By closure theorem, we can compute the image of a projection by the elimination ideal; hence $V(I_d) = \overline{\pi_d(V(I))} = \overline{\phi(k^d)}$. □

Example 3.28 (Binomial model with two trials). Let $X \sim B(2, \theta)$ whose distribution is parametrized by $\phi : k \to k^3$, $\phi(\theta) = \begin{bmatrix}(1-\theta)^2 & 2\theta(1-\theta) & \theta^2\end{bmatrix}$. What relations are satisfied by the components of $\phi(\theta)$?

Define the ideal $I = (p_0 - (1-\theta)^2, p_1 - 2\theta(1-\theta), p_2 - \theta^2) \subset \mathbb{C}[\theta, p_0, p_1, p_2]$. By the Implicitization theorem, we want to compute $I \cap k[p_0, p_1, p_2]$. Using the Elimination

theorem, we can compute this elimination ideal with Gröbner basis. Consider the lexicographic ordering $y_3 \prec y_2 \prec y_1 \prec \theta$. Using Macaulay2 we obtain the reduced Gröbner basis

$$G = \{2\theta - p_1 - 2p_2, p_0 + p_1 + p_2 - 1, p_1^2 + 4p_1p_2 + 4p_2^2 - 4p_2\}$$
$$G_{\mathbf{p}} = \{p_0 + p_1 + p_2 - 1, p_1^2 + 4p_1p_2 + 4p_2^2 - 4p_2\}$$

Thus, by the Elimination theorem we obtain

$$V(I_1) = V(G_1) = V(p_0 + p_1 + p_2 - 1, p_1^2 + 4p_1p_2 + 4p_2^2 - 4p_2).$$

Note that the first equation implies $-p_1 - p_2 + 1 = p_0$ and the second implies $p_1^2 + 4p_1p_2 + 4p_2^2 - 4p_2 = p_1^2 - 4p_2(-p_1 - p_2 + 1)$; hence we plug the first into the second equation to obtain the *Hardy Weinberg equations*

$$V(p_0 + p_1 + p_2 - 1, p_0p_2 - \frac{1}{4}p_1^2).$$

Example 3.29 (General binomial model). In general for $X \sim \text{Bin}(n, \theta)$ this model is parametrized by $\phi = (\phi_0, \phi_1, \ldots, \phi_n)$, where $\phi_i = P(X = i) = \binom{n}{i}\theta^i(1-\theta)^{n-i}$. The implicit equations of $\phi(k)$ are $p_0 + p_1 + ... + p_n = 1$ and the 2×2-minors of the Hankel matrix

$$\begin{bmatrix} p_0/\binom{n}{0} & p_1/\binom{n}{1} & ... & p_{n-1}/\binom{n}{n-1} \\ p_1/\binom{n}{1} & p_2/\binom{n}{2} & ... & p_n/\binom{n}{n} \end{bmatrix}$$

For $n = 2$, we obtain $\det \begin{bmatrix} p_0 & p_1/2 \\ p_1/2 & p_2 \end{bmatrix} = p_0p_2 - \frac{1}{4}p_1^2$.

3.3 Primary decomposition

Definition 3.30 (Prime and reducible ideals). An ideal I is **prime** if $ab \in I$ implies that $a \in I$ or $b \in I$. An ideal I is called **reducible** if $I = I_1 \cap I_2$ for proper ideals $I_1, I_2 \supsetneq I$.

Definition 3.31 (Reducible variety). A variety V is called **reducible** if $V' \cup V'' = V$ for proper varieties $V, V' \subsetneq V$.

Proposition 3.32. A variety is irreducible if its associated ideal is prime.

Proposition 3.33. Every variety can be decomposed *uniquely* into finitely many irreducible varieties: $V = V_1 \cup ... \cup V_m$.

Proposition 3.34. Every ideal in a Noetherian ring can be decomposed into finitely many irreducible ideals: $I = I_1 \cap ... \cap I_m$.

Definition 3.35 (Primary ideal and associated prime ideal). An ideal I is called J-**primary** if $xy \in I$ and $x \notin I$ implies that $y \in J := \sqrt{I}$. We call $J = \sqrt{I}$ the **associated prime** of the primary ideal I.

Proposition 3.36 (Some facts in a Noetherian ring).

- Every prime ideal is irreducible.
- Not every irreducible ideal is prime.
- Every irreducible ideal is primary.
- The intersection of prime ideals need not be prime (e.g. take (2) and (3)).
- The intersection of J-primary ideals is J-primary.

Definition 3.37 (Primary decomposition). Let $I \subset k[x]$ be an ideal. A **primary decomposition** of I is a representation of $I = \bigcap_{i=1}^{r} Q_i$ as an intersection of finitely many primary ideals Q_i. The representation is **irredundant** if all $\sqrt{Q_i}$ are distinct and no $Q_i \supset \bigcap_{j \neq i} Q_j$ exists.

Corollary 3.38. A primary decomposition always exists in a Noetherian ring.

Definition 3.39 (Associated primes). Let I be decomposable such that $I = Q_1 \cap ... \cap Q_m$ is irredundant. Then $\sqrt{Q_i}$ are called the **associated primes** $\mathrm{Ass}(I)$ of I.

Proposition 3.40 (1st uniqueness). Let $I \subset k[x]$. Then, the associated primes of I are uniquely determined by I; thus the associated primes are independent of the specific primary decomposition of I. In other words, different irredundant primary decomposition yield the same associated primes.

Proposition 3.41 (2nd uniqueness). The minimal ideals in $\mathrm{Ass}(I)$ are called **minimal primes**. Primary ideals associated to the minimal ideals are uniquely determined by I. In other words,

$$I = \bigcap Q_i = \bigcap Q'_j, \quad \sqrt{Q}_k = \sqrt{Q'_{k'}} \text{ is minimal for some } k \in \mathbb{N} \implies Q_k = Q'_{k'}.$$

Proposition 3.42. Minimal primes are the smallest prime ideals containing I.

Definition 3.43. Embedded primes are associated primes that are not minimal.

Associated primes are unique. The primary ideals associated to minimal primes are unique. The primary ideals associated to embedded primes need *not* be unique.

Proposition 3.44 (Irreducible composition and minimal primes). The irreducible composition of a variety $V(I)$ is the union of $\bigcup_i V(P_i)$ where P_i are the minimal primes of I.

Definition 3.45 (Quotient ideal and saturation). Let I, J be ideals in $K[x]$. The **quotient ideal** is the ideal $I : J = \{f \in K[x] : fg \in I \ \forall g \in J\}$. The **saturation** of I by J is defined as $I : J^{\infty} = \bigcup_{k=1}^{\infty} I : J^k$.

Proposition 3.46. We have

$$I(V \setminus W) = I(V) : I(W).$$

Moreover,

$$\overline{V(I) \setminus V(J)} = V(I : J^\infty).$$

Definition 3.47 (Binomial ideal). A **binomial** is a polynomial of the form $\mathbf{x}^\alpha - \lambda \mathbf{x}^\beta$ in $k[\mathbf{x}]$. A **binomial ideal** is an ideal generated by binomials.

4 Conditional Independence

Conditional independence constraints are simple constraints on probability densities; they express the following: Lec09, 13.11.2023

Two sets of random variables are ***unrelated*** given a third set of random variables.

We assume the following

- $\mathbf{X} = (X_1, ..., X_m) \in \mathcal{X} = \prod_{i=1}^m \mathcal{X}_i$ is an m-dimensional random vectors.
- $f(\mathbf{x})$ is the density function of the joint probability distribution of $\mathbf{X}$ with respect to a product measure ν on $\mathcal{X}$
- given $A \subset [m]$, we write $\mathbf{X}_A = (X_a)_{a \in A}$, similarly $\mathcal{X}_A$
- given a partition $A_1 \mid A_2 \mid ... \mid A_k$ of $[m]$ we write $f(x_{A_1}, ..., x_{A_k})$ to group variables together.

We recall the density definitions. The **marginal density** of X_A is defined by

$$f_A(x_A) = \int_{\mathcal{X}_{A^c}} f(x_A, x_{A^c}) \, \mathrm{d}\nu_{A^c}(x_{A^c}), \quad \forall x_A \in \mathcal{X}_A$$

for $A \subset [m]$. The **conditional density** is

$$f_{A|B}(x_A \mid x_B) = \begin{cases} \frac{f_{A \cup B}(x_A, x_B)}{f_B(x_B)} & \text{if } f_B(x_B) > 0, \\ 0 & \text{otherwise} \end{cases}$$

given some $x_B \in \mathcal{X}_B$ and disjoint $A, B \subset [m]$.

Definition 4.1 (Conditional independence of random variables).
Let $A, B, C \subset [m]$ pairwise disjoint. We define $X_A \perp\!\!\!\perp X_B \mid X_C$ if

$$f_{A \cup B|C}(x_A, x_B \mid x_C) = f_{A|C}(x_A \mid x_C) \cdot f_{B|C}(x_B \mid x_C), \quad \forall x_A, x_B, x_C$$

We say that X_A is **conditionally independent of X_B given X_C**. We also write $A \perp\!\!\!\perp B \mid C$.

In plain English, this means that

Given X_C, knowing X_B does not give any information about X_A.

Proposition 4.2 (In plain English). If $f_{B|C}(X_B \mid X_C) > 0$, then $f_{A|B \cup C}(X_A \mid X_B, X_C) = f_{A|C}(X_A \mid X_C)$.

Proof. We have

$$\begin{aligned} f_{A|B\cup C}(X_A \mid X_B, X_C) &= \frac{f_{A\cup B\cup C}(X_A, X_B, X_C)}{f_{B\cup C}(X_B, X_C)} \cdot \frac{f_C(X_C)}{f_C(X_C)} \\ &= \frac{f_{A\cup B|C}(X_A, X_B \mid X_C)}{f_{B|C}(X_B \mid X_C)} \\ &= f_{A|C}(X_A \mid X_C). \end{aligned}$$

□

Example 4.3 (Independence of random variables). Set $C = \emptyset$. Then

$$X_A \perp\!\!\!\perp X_B \mid X_\emptyset = X_A \perp\!\!\!\perp X_B.$$

This corresponds to $f_{A\cup B}(x_A, x_B) = f_A(x_A) f_B(x_B)$.

4.1 Conditional independence axioms

We want to examine the following problem:

Question: Given a list of CI statements for some random vector $\mathbf{X}$, what other CI statements are also satisfied?

Here, we assume that the density f is unknown. Otherwise, we can just all the CI statements from the density.

Idea: Find some *easy* conditional independence implications which follow directly from definition. We call these inference rules **conditional independence axioms** or **conditional independence inference rules**.

Pearl (1988) proposes a list of four axioms from which he conjectured all CI relations can be deduced (something we call a **sound and complete system**). Soundness means everything stays true if deducing; completeness means everything that is true can be deduced.

Proposition 4.4 (Pearl's four axioms). Let $A, B, C, D \subset [m]$ be pairwise disjoint. Then, the following holds true:

- Symmetry: If $X_A \perp\!\!\!\perp X_B \mid X_C$ then $X_B \perp\!\!\!\perp X_A \mid X_C$;
- Decomposition: $X_A \perp\!\!\!\perp X_{B\cup D} \mid X_C \implies X_A \perp\!\!\!\perp X_B \mid X_C$;
- Weak union: $X_A \perp\!\!\!\perp X_{B\cup D} \mid X_C \implies X_A \perp\!\!\!\perp X_B \mid X_{C\cup D}$;
- Contraction: $X_A \perp\!\!\!\perp X_B \mid X_{C\cup D}$ and $X_A \perp\!\!\!\perp X_D \mid X_C \implies X_A \perp\!\!\!\perp X_{B\cup D} \mid X_C$.

One says that CI is a semigraphoid. Note that these four CI axioms are valid for *any* distribution.

Proof. • Symmetry follows easily from commutativity as we see

$$\begin{aligned} f_{A\cup B|C}(x_A, x_B \mid x_C) &= f_{A|C}(x_A \mid x_C) f_{B|C}(x_B \mid x_C) \\ &= f_{B|C}(x_B \mid x_C) f_{A|C}(x_A \mid x_C) \\ &= f_{B\cup A|C}(x_B, x_A \mid x_C). \end{aligned}$$

- Decomposition follows from marginalizing x_D. The statement $X_A \perp\!\!\!\perp X_{B\cup D} \mid X_C$ means $f_{A\cup B\cup D|C}(x_A, x_B, x_D \mid x_C) = f_{A|C}(x_A \mid x_C) f_{B\cup D|C}(x_B, x_D \mid x_C)$. Then, we marginalize (that is we integrate both sides over $\mathcal{X}_D$), and we get

$$\begin{aligned} &\int_{\mathcal{X}_D} \frac{f_{A\cup B\cup D\cup C}(x_A, x_B, x_D, x_C)}{f_C(x_C)} \nu(x_D) \\ &= \int_{\mathcal{X}_D} \frac{f_{A\cup C}(x_A, x_C)}{f_C(x_C)} \frac{f_{B\cup D\cup C}(x_B, x_D, x_C)}{f_C(x_C)} \nu(x_D). \end{aligned}$$

  By definition of marginal density, the above equation is the same as

$$\frac{f_{A\cup B\cup C}(x_A, x_B, x_C)}{f_C(x_C)} = \frac{f_{A\cup C}(x_A, x_C)}{f_C(x_C)} \frac{f_{B\cup C}(x_B, x_C)}{f_C(x_C)}.$$

  Hence, $f_{A\cup B|C}(x_A, x_B \mid x_C) = f_{A|C}(x_A \mid x_C) f_{B|C}(x_B \mid x_C)$; thus $X_A \perp\!\!\!\perp X_B \mid X_C$.

- Weak union follows easily from condition on x_D. We take $f_{A\cup B\cup D|C}(x_A, x_B, x_D \mid x_C) = f_{A|C}(x_A \mid x_C) f_{B\cup D|C}(x_B, x_D \mid x_C)$, and divide it by $f_{D|C}(x_D \mid x_C)$:

$$\frac{f_{A\cup B\cup C\cup D}(x_A, x_B, x_C, x_D)}{f_{C\cup D}(x_C, x_D)} = \frac{f_{A\cup C}(x_A, x_C)}{f_C(x_C)} \frac{f_{B\cup C\cup D}(x_B, x_C)}{f_{C\cup D}(x_C, x_D)}.$$

  Note that by the decomposition axiom, we have $X_A \perp\!\!\!\perp X_D \mid X_C$. Thus,

$$f_{A\cup C}(x_A, x_C)/f_C(x_C) = f_{A\cup C\cup D}(x_A, x_C, x_D)/f_{C\cup D}(x_C, x_D).$$

  Hence, $f_{A\cup B|C\cup D}(x_A, x_B \mid x_C, x_D) = f_{A|C\cup D}(x_A \mid x_C, x_D) f_{B|C\cup D}(x_B \mid x_C, x_D)$ which says that $X_A \perp\!\!\!\perp X_B \mid X_{C\cup D}$.

- We start with $f_{A\cup B|C\cup D}(x_A, x_B \mid x_C, x_D) = f_{A|C\cup D}(x_A \mid x_C, x_D) f_{B|C\cup D}(x_B \mid x_C, x_D)$. Multiply by $f_{C\cup D}(x_C, x_D)$ and then divide by $f_C(x_C) > 0$ to get

$$f_{A\cup B\cup D|C}(x_A, x_B, x_D \mid x_C) = f_{A\cup D|C}(x_A, x_D \mid x_C) f_{B|C\cup D}(x_B \mid x_C, x_D).$$

  With $X_A \perp\!\!\!\perp X_D \mid X_C$ we obtain

$$f_{A\cup B\cup D|C}(x_A, x_B, x_D \mid x_C) = f_{A|C}(x_A \mid x_C) f_{D|C}(x_D \mid x_C) f_{B|C\cup D}(x_B \mid x_C, x_D).$$

  The right-hand side is the same as $f_{A|C}(x_A \mid x_C) f_{B\cup D|C}(x_B, x_D \mid x_C)$. Hence, $X_A \perp\!\!\!\perp X_{B\cup D} \mid X_C$.

□

Studeny (1992) disproved Pearl's conjecture: it is impossible to find a finite set of axioms from which all CI relations can be deduced.

We present a further fifth axiom that does not hold in general, but holds only for densities f such that $f(x) > 0$ for every $x \in \mathcal{X}$.

Proposition 4.5 (Intersection axiom). If $f(x) > 0$ for all $x \in \mathcal{X}$ then $X_A \perp\!\!\!\perp X_B \mid X_{C \cup D}$ and $X_A \perp\!\!\!\perp X_C \mid X_{B \cup D} \implies X_A \perp\!\!\!\perp X_{B \cup C} \mid X_D$.

Proof. ... □

A simple corollary follows for $D = \emptyset$: if X_A is independent of X_B given X_C, and X_A independent of X_C given X_B, then X_A is independent of $X_{B \cup C}$.

4.2 Discrete random variables

If $\mathbf{X}$ is a discrete random variable, then CI statements are algebraic constraints on the distribution, namely they are quadratic. This is where the algebra comes into perspective.

Let $\mathbf{X} = (X_1, ..., X_m) \in \mathcal{R} = \prod_{j=1}^{m} [r_j]$ be a vector of discrete random variables. The joint density function $f : (i_1, ..., i_m) \mapsto \mathbb{P}(X_1 = i_1, ..., X_m = i_m) = p_{i_1 ... i_m}$ of $\mathbf{X}$ can be represented by a $r_1 \times r_2 \times ... \times r_m$ tensor $\mathbf{p} = (p_{i_1 ... i_m})_{(i_1, ..., i_m) \in \mathcal{R}}$.

Example 4.6 (Representation of the joint density function by a matrix). For $m = 2$ we obtain a matrix

$$\mathbf{p} = \begin{bmatrix} p_{11} & \cdots & p_{1r_2} \\ \vdots & \ddots & \vdots \\ p_{r_1 1} & \cdots & p_{r_1 r_2} \end{bmatrix} \in \mathbb{R}^{[r_1] \times [r_2]}.$$

Lemma 4.7 (Rank criterion for independence). Let $\mathbf{X} = (X_1, X_2)$ be a vector of discrete random variables. Then, $X_1 \perp\!\!\!\perp X_2$ if and only if $\operatorname{rank}(\mathbf{p}) = 1$.

Proof. Note that $\mathbf{p} \neq \mathbf{0}$ since $\mathbf{p}$ is a probability mass function.

- $\implies$: Since $X_1 \perp\!\!\!\perp X_2$, we have $p_{ij} = \mathbb{P}(X_1 = i, X_2 = j) = \mathbb{P}(X_1 = i) P(X_2 = j)$ for all $i \in [r_1], j \in [r_2]$. By marginalizing we know that $\mathbb{P}(X_1 = i) = p_{i+}$ and $\mathbb{P}(X_2 = j) = p_{+j}$; so $p_{ij} = p_{i+} p_{+j}$. Hence

$$\mathbf{p} = \begin{bmatrix} p_{1+} \\ p_{2+} \\ \cdots \\ p_{r_1 +} \end{bmatrix} \cdot \begin{bmatrix} p_{+1} & p_{+2} & \cdots & p_{+r_2} \end{bmatrix}.$$

  We see that $\mathbf{p}$ is of rank 1 since $\mathbf{p} \neq \mathbf{0}$.

- $\Longleftarrow$: assume that $\mathbf{p} = \mathbf{a}\mathbf{b}^T$. Without loss of generality, we assume that $\mathbf{a} \geq \mathbf{0}$, $\mathbf{b} \geq \mathbf{0}$ and $\|\mathbf{a}\|_1 = \|\mathbf{b}\|_1 = 1$. Note that $p_{i+} = p_{i1} + ... + p_{ir_2} = a_i b_1 + ... + a_i b_{r_2} = a_i b_+$, $p_{+j} = b_j a_+$ and $1 = p_{++} = a_+ b_+$. Then

$$p_{ij} = a_i b_j = a_i \underbrace{b_+ a_+}_{=1} b_j = p_{i+} p_{+j}.$$

This proves that $X_1 \perp\!\!\!\perp X_2$.

□

Corollary 4.8 (Minor criterion for independence). Let $\mathbf{X} = (X_1, X_2)$ be a vector of discrete random variables. Then, $X_1 \perp\!\!\!\perp X_2$ if and only if $\det \begin{bmatrix} p_{i_1 j_1} p_{i j_2} \\ p_{i_2 j_1} p_{i_2 j_2} \end{bmatrix} = 0$ for all $i_1, i_2 \in [r_1]$ and $j_1, j_2 \in [r_2]$.

Proof. Follows from the linear algebra fact that a matrix has rank r if and only if there exists a $r \times r$ minor that is non-zero and all $(r+1) \times (r+1)$-minors are zero. □

Proposition 4.9 (Algebraic characterization of conditional independence). Let $\mathbf{X}$ be a discrete random vector and $A, B, C \subset [m]$ be pairwise disjoint. Then, $X_A \perp\!\!\!\perp X_B \mid X_C$ if and only if

$$p_{i_A i_B i_C +} \cdot p_{j_A j_B i_C +} - p_{i_A j_B i_C +} \cdot p_{j_A i_B i_C +} = 0$$

for all $i_A, j_A \in \mathcal{R}_A, i_B, j_B \in \mathcal{R}_B, i_C \in \mathcal{R}_C$.

Alternatively, we can write this as

$$\begin{aligned} &\mathbb{P}(X_A = i_A, X_B = i_B, X_C = i_C)\mathbb{P}(X_A = j_A, X_B = j_B, X_C = i_C) \\ &= \mathbb{P}(X_A = i_A, X_B = j_B, X_C = i_C)\mathbb{P}(X_A = j_A, X_B = i_B, X_C = i_C). \end{aligned}$$

Proof. Proof skipped. Reduce this to the lemma by first marginalizing (so that $A \cup B \cup C = [m]$) and then conditioning on X_C (so that we may assume $C = \emptyset$). Apply the lemma. □

Lec10, 14.11.2023

The above proposition allows us to do algebra since we deal with polynomials. We will see that probability distributions satisfying the CI statements form a *semi-algebraic* set.

Definition 4.10 (Conditional independence ideal). Let $\mathbf{X}$ be a discrete random vector and $A, B, C \subset [m]$ be pairwise disjoint. The **conditional independence ideal** $I_{A \perp\!\!\!\perp B|C} \subset \mathbb{C}[\mathbf{p}]$ is defined as

$$(p_{i_A i_B i_C +} p_{j_A j_B i_C +} - p_{i_A j_B i_C +} p_{j_A i_B i_C +} : i_A, j_A \in \mathcal{R}_A, i_B, j_B \in \mathcal{R}_B, i_C \in \mathcal{R}_C).$$

$I_{A \perp\!\!\!\perp B|C} \subset \mathbb{C}[\mathbf{p}]$ is the ideal generated by all the *quadratic polynomials* in Proposition 4.9.

In general, if $\mathcal{C} = \{A_i \perp\!\!\!\perp B_i \mid C_i\}$ is a set of conditional independence statements, then we define

$$I_{\mathcal{C}} := \sum_{A \perp\!\!\!\perp B|C \in \mathcal{C}} I_{A \perp\!\!\!\perp B|C}.$$

Definition 4.11 (Discrete conditional independence model). The set of all probability distributions $\mathcal{M}_{\mathcal{C}}$ that satisfy the quadratic polynomial equations is called the **discrete conditional independence model**; in mathematical terms, $\mathcal{M}_{\mathcal{C}}$ is defined as

$$\mathcal{M}_{\mathcal{C}} := V_{\mathbb{R}}(I_{\mathcal{C}}) \cap \Delta_{\mathcal{R}},$$

where $\Delta_{\mathcal{R}} = \left\{\mathbf{p} \in \mathbb{R}^{\mathcal{R}}_{\geq 0} : \sum p_{i_1,\ldots,i_m} = 1\right\}$. The CI model $\mathcal{M}_{\Xi}$ is a *semi-algebraic* set.

Example 4.12. Consider $\mathbf{X} = (X_1, X_2)$ and $1 \perp\!\!\!\perp 2$. Then the CI ideal is $I_{1 \perp\!\!\!\perp 2} = \{p_{i_1 j_1} p_{i_2 j_2} - p_{i_2 j_1} p_{i_1 j_2} : i_1, i_2 \in [r_1], j_1, j_2 \in [r_2]\}$. These are all 2×2-minors of $\mathbf{p}$.

4.3 Gaussian random variables

Now, we assume $\mathbf{X} \sim \mathcal{N}_m(\mu, \Sigma)$. We will see that $X_A \perp\!\!\!\perp X_B \mid X_C$ corresponds to an *algebraic* constraint.

Proposition 4.13 (Rank criterion for conditional independence). Fix $\mu \in \mathbb{R}^m$ and a positive definite matrix $\Sigma \in \mathrm{PD}(m)$. Let $\mathbf{X} \sim \mathcal{N}_m(\mu, \Sigma)$, and $A, B, C \subset [m]$ pairwise disjoint. Then $X_A \perp\!\!\!\perp X_B \mid X_C$ if and only if

$$\operatorname{rank}(\Sigma_{A \cup C, B \cup C}) = \#C.$$

Proof. The conditional distribution of $X_{A \cup B}$ given X_C is normal with expected value $\mu_{A \cup B} + \Sigma_{A \cup B, A \cup B} \Sigma^{-1}_{C,C}(x_C - \mu_C)$ and covariance matrix $\Sigma_{A \cup B, A \cup B} - \Sigma_{A \cup B, C} \Sigma^{-1}_{C,C} \Sigma_{C, A \cup B}$. For a normal distribution we know that $A \perp\!\!\!\perp B$ if the covariance matrix with rows A and columns B is zero. Hence, we have

$$X_A \perp\!\!\!\perp X_B \mid X_C \iff \Sigma_{A,B} - \Sigma_{A,C} \Sigma^{-1}_{C,C} \Sigma_{C,B} = 0.$$

The expression on the right hand is known as the *Schur complement* of $\Sigma_{C,C}$ in $\Sigma_{A \cup C, B \cup C}$:

$$\Sigma_{A \cup C, B \cup C} = \begin{bmatrix} \Sigma_{A,B} & \Sigma_{A,C} \\ \Sigma_{C,B} & \Sigma_{C,C} \end{bmatrix}.$$

One can prove that $\Sigma_{A,B} = \Sigma_{A,C} \Sigma^{-1}_{C,C} \Sigma_{C,B}$ if and only if rank of $\Sigma_{A \cup C, B \cup C}$ is $\#C$. □

Constraining the rank of a matrix to be k is algebraically equivalent to constraining all $(k+1) \times (k+1)$-minors to be zero.

Definition 4.14 (Gaussian conditional independence ideal). Let $A, B, C \subset [m]$ be pairwise disjoint. The **Gaussian conditional independence ideal** $I_{A \perp\!\!\!\perp B|C} \subset \mathbb{R}[\sigma_{ij} : 1 \leq i, j \leq m]$ is defined as

$$I_{A \perp\!\!\!\perp B|C} = ((\#C + 1) \times (\#C + 1) \text{ minors of } \Sigma_{A \cup C, B \cup C}).$$

Example 4.15. Let $\mathcal{C} = \{1 \perp\!\!\!\perp 3, 1 \perp\!\!\!\perp 3 \mid 2\}$. What is the Gaussian conditional independence ideal $I_{\mathcal{C}}$?

$$I_{\mathcal{C}} = (\sigma_{1,3}, \det \Sigma_{\{1,2\},\{3,2\}}) = (\sigma_{1,3}, \det \begin{bmatrix} \sigma_{1,3} & \sigma_{1,2} \\ \sigma_{2,3} & \sigma_{2,2} \end{bmatrix})$$

So, the conditional independence model consists of all positive definite matrices Σ such that

$$\sigma_{1,3} = 0 \text{ and } \sigma_{1,2}\sigma_{2,3} = \sigma_{1,3}\sigma_{2,2}.$$

This is equivalent to

$$\sigma_{1,3} = 0 \text{ and } \sigma_{1,2}\sigma_{2,3} = 0.$$

The solution set splits into two linear spaces

$$L_1 = \{\Sigma : \sigma_{1,3} = \sigma_{1,2} = 0\} \text{ and } L_2 = \{\Sigma : \sigma_{1,3} = \sigma_{2,3} = 0\}.$$

So $1 \perp\!\!\!\perp (2,3)$ and $3 \perp\!\!\!\perp (1,2)$.

Thus, for multi-variate Gaussians, we have $1 \perp\!\!\!\perp 3, 1 \perp\!\!\!\perp 3 \mid 2 \implies 1 \perp\!\!\!\perp (2,3) \vee 3 \perp\!\!\!\perp (1,2)$.

4.4 Primary decomposition

Example 4.16 (Gaussian conditional independence). Assume $\mathbf{X} \sim \mathcal{N}_m(\mu, \Sigma)$. Given the list of conditional independence $\mathcal{C} = \{1 \perp\!\!\!\perp 2 \mid 3, 2 \perp\!\!\!\perp 3\}$, can we infer that $2 \perp\!\!\!\perp \{1, 3\}$?

You may recognize this as the *contraction axiom*, which we have already proven using probability theory. This time we would like to use primary decomposition to prove this axiom.

The Gaussian conditional independence ideal $I_{\mathcal{C}}$ corresponding to $\mathcal{C}$ contains all Gaussian probability distributions that satisfy $\mathcal{C}$; it is of the form

$$I_{\mathcal{C}} = \left(\sigma_{2,3}, \det \begin{bmatrix} \sigma_{1,2} & \sigma_{1,3} \\ \sigma_{3,2} & \sigma_{3,3} \end{bmatrix}\right) = (\sigma_{2,3}, \sigma_{1,2}\sigma_{3,3} - \sigma_{1,3}\sigma_{3,2}).$$

Since $\sigma_{23} = \sigma_{32}$ we have $I_{\mathcal{C}} = (\sigma_{23}, \sigma_{12}\sigma_{33}) = (\sigma_{23}, \sigma_{12}) \cap (\sigma_{23}, \sigma_{33})$. The vanishing set of $I_{\mathcal{C}\mathcal{C}}$ consists of two components $V(\sigma_{23}, \sigma_{12})$ and $V(\sigma_{23}, \sigma_{33})$; the latter is the set of all matrices vanishing entries in σ_{33} and σ_{23}. Since positive definite matrices have positive entries on the diagonal, we have $V(\sigma_{23}, \sigma_{33}) \cap \mathrm{PD}_3 = \emptyset$. Thus, only the first component is relevant for us, and we see that $2 \perp\!\!\!\perp \{1, 3\}$.

Example 4.17 (Binary random variable). Now assume $\mathbf{X} = (X_1, X_2, X_3)$ is a binary random vector, that is $X_i \in \{1, 2\}$ for all $i = 1, 2, 3$. The associated conditional independence ideal $I_{\mathcal{C}}$ of $\mathcal{C} = \{1 \perp\!\!\!\perp 2 \mid 3, 2 \perp\!\!\!\perp 3\}$ is

$$\begin{aligned} I_{\mathcal{C}} = ((p_{111} + p_{211})(p_{122} + p_{222}) - (p_{112} + p_{212})(p_{121} + p_{221}), \\ p_{111}p_{221} - p_{121}p_{211}, \\ p_{112}p_{222} - p_{122}p_{212}). \end{aligned}$$

Here is the sketch of the computation. For $\mathcal{C}_2 = \{1 \perp\!\!\!\perp 2 \mid 3\}$ we must compute $I_{\mathcal{C}_2} = (p_{i_1 i_2 i_3} p_{j_1 j_2 i_3} - p_{i_1 j_2 i_3} p_{j_1 i_2 i_3})$ for all $i_1, i_2, j_1, j_2, i_3 \in \{1, 2\}$. For $i_1 = i_2 = j_1 = j_2 = i_3 = 0$ we see that the quadratic polynomial vanishes; also for $i_1 = i_2 = j_1 = j_2 = 0$ and $i_3 = 0$. Similarly, for other combinations except for $i_1 = i_2 = i_3 = 1$ and $j_1 = j_2 = 2$ as well as $i_1 = i_2 = 1$ and $i_3 = j_1 = j_2 = 2$. Further computation yields that these are the only two quadratic polynomials; hence $I_{\mathcal{C}_2}$ is generated by $p_{111}p_{221} - p_{121}p_{211}$ and $p_{112}p_{222} - p_{122}p_{212}$. For $\mathcal{C}_1 = \{2 \perp\!\!\!\perp 3\}$ we must compute the generators

$$(p_{1 i_2 i_3} + p_{2 i_2 i_3})(p_{1 j_2 j_3} + p_{2 j_2 j_3}) - (p_{1 i_2 j_3} + p_{2 i_2 j_3})(p_{1 i_3 j_2} + p_{2 i_3 j_2})$$

for all $i_2, i_3, j_2, j_3 \in \{1, 2\}$. We see that the only non-zero quadratic polynomial is $(p_{111} + p_{221})(p_{122} + p_{222}) - (p_{112} + p_{212})(p_{121} + p_{221})$.

...

5 Statistics Primer

Lec13, 27.11.2023

Definition 5.1 (Statistical model). A **statistical model** $\mathcal{M}$ is a collection of probability distributions or density functions. A **parametric statistical model** $\mathcal{M}_\Theta$ is a statistical model of the form $\mathcal{M}_\Theta = \{p_\theta : \theta \in \Theta\}$, where $\Theta \subset \mathbb{R}^d$ is a finite-dimensional parameter space.

Definition 5.2 (Identifiable Model). A statistical model is called **identifiable** if $p_\theta = p_{\theta'}$ implies $\theta = \theta'$.

Example 5.3 (Binomial random variable model). Let $n \in \mathbb{N}$. For $\theta \in \Theta = [0,1]$ we define $p_\theta(k) = \binom{n}{k}\theta^k(1-\theta)^{n-k}$. A statistical model is given by $M_\theta = \{p_\theta : \theta \in \Theta\}$.

Example 5.4 (Multivariate normal random vector). Let $X \in \mathbb{R}^m$ be a real random vector. Let $\Theta = \mathbb{R}^m \times \mathrm{PD}_m$. Define $\mathcal{M}_\Theta = \{p_\theta(x) : \theta \in \Theta\}$ where

$$p_\theta = (2\pi)^{-\frac{m}{2}}|\Sigma|^{-\frac{1}{2}} \exp\left(-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)\right).$$

Definition 5.5 (Implicit statistical models). Implicit statistical models are given by constraints on the probability distributions or densities.

Example 5.6 (Implicit binomial statistical model). The binomial model can be implicitly given by the intersection of Δ_n with the 2×2-minors of

$$\begin{bmatrix} \frac{x_0}{\binom{n}{0}} & \frac{x_1}{\binom{n}{1}} & \frac{x_2}{\binom{n}{2}} & \cdots & \frac{x_{n-1}}{\binom{n}{n-1}} \\ \frac{x_1}{\binom{n}{1}} & \frac{x_2}{\binom{n}{2}} & \frac{x_3}{\binom{n}{3}} & \cdots & \frac{x_n}{\binom{n}{n}} \end{bmatrix}.$$

Here is another implicit model.

Example 5.7 (Independent random variables). Let X, Y be two discrete random variables with $X \in [r]$ and $Y \in [c]$. Let $\mathcal{R} = [r] \times [c]$ and $\Delta_\mathcal{R} = \Delta_n$ where $n = |\mathcal{R}| - 1$. The **model of independence** $\mathcal{M}_{X \perp\!\!\!\perp Y}$ is defined by the constraints

$$\begin{aligned} \mathcal{M}_{X \perp\!\!\!\perp Y} &= \Delta_\mathcal{R} \cap \{\mathbf{p} \in \mathbb{R}^{r \times c} : \mathbb{P}(X = i, Y = j) = \mathbb{P}(X = i)\mathbb{P}(Y = j),\ \forall i \in [r], j \in [c]\} \\ &= \Delta_\mathcal{R} \cap \{\mathbf{p} \in \mathbb{R}^{r \times c} : \mathrm{rank}(\mathbf{p}) = 1\}. \end{aligned}$$

This is an *implicit* representation of $\mathcal{M}_{X \perp\!\!\!\perp Y}$. For a *parametric* representation we define $\theta = (\alpha, \beta) \in \Delta_{r-1} \times \Delta_{c-1}$ with $\mathbb{P}_\theta(X_1 = i, X_2 = j) = \alpha_i\beta_j$.

Example 5.8 (Main example of implicit statistical models). Consider a set of conditional independence statements $\mathcal{C} = \{A_1 \perp\!\!\!\perp B_1 \mid C_1, A_2 \perp\!\!\!\perp B_2 \mid C_2, \dots\}$. If $\mathbf{X}$ is a discrete random vector, then the discrete conditional independence model is implicit:

$$\mathcal{M}_\mathcal{C} = V\left(\sum_\mathcal{C} I_{A \perp\!\!\!\perp B \mid C}\right) \cap I_{\Delta_\mathcal{R}}.$$

If $\mathbf{X} \sim \mathcal{N}_m(\mu, \Sigma)$ is a Gaussian vector, then the Gaussian conditional independence model is also implicit

$$\mathcal{M}_{\mathcal{C}} = V\left(\sum_{\mathcal{C}} I_{A \perp\!\!\!\perp B|C}\right) \cap \mathrm{PD}_m.$$

Definition 5.9 (Statistic). Let $\mathbf{X} \in \mathcal{X}$ be a random vector with state space $\mathcal{X}$. A **statistic** is a function from the state space $\mathcal{X}$ to some set.

Examples are

- $\bar{X}_n = \frac{1}{n}\sum_{i=1}^n X_i$,
- $S_n^2 = \frac{1}{n-1}\sum_{i=1}^n (X_i - \bar{X}_n)^2$,
- $T = \max\{X_1, X_2, ..., X_n\}$,
- $T' = 5$.

We informally describe what a sufficient statistic is. We say T is a **sufficient statistic** if the statistician who knows the value of T can do just as good a job of estimating the unknown parameter θ as the statistician who knows the entire random sample.

Definition 5.10 (Sufficient statistic). A statistic T is called **sufficient** if $\mathbb{P}(X = x \mid T(X) = t, \theta) = \mathbb{P}(X = x \mid T(X) = t)$.

Theorem 5.11 (Factorization theorem). Let $\mathbf{X}$ be a random vector with joint density $f(\mathbf{x} \mid \theta)$. A statistic T is sufficient if and only if the joint density f can be factored as follows

$$f(\mathbf{x} \mid \theta) = u(\mathbf{x})v(T(\mathbf{x}), \theta)$$

for non-negative functions u, v.

Example 5.12 (Sufficient statistic of the likelihood function of a multivariate normal distribution). Given samples $X^{(1)}, \ldots, X^{(N)}$ i.i.d. normal, the joint distribution is given by

$$p_{\mu,\Sigma}(X) = \prod_{i=1}^N \frac{1}{(2\pi)^{\frac{m}{2}}|\Sigma|^{\frac{1}{2}}} \exp\left(-\frac{1}{2}(X^{(i)} - \mu)^T\Sigma^{-1}(X^{(i)} - \mu)\right).$$

A sufficient statistic is given by

$$T(X) = \begin{bmatrix} \bar{X} \\ \frac{1}{N}\sum_{i=1}^N (X^{(i)} - \bar{X})(X^{(i)} - \bar{X})^T \end{bmatrix}.$$

5.1 4.12

Last topic: Parameter estimation. Given a sample $X^{(1)}, X^{(2)}, ..., X^{(N)}$ iid $\sim P_\theta \in \mathcal{M}_\Theta$ estimate the value of $\theta \in \Theta$. We have an estimator $\hat{\theta}(X^{(1)}, ..., X^{(N)})$.

Example 5.13. Assume $X^{(1)}, ..., X^{(N)} \sim \text{Ber}(\theta)$ with $\mathbb{E}[\bar{X}_N] = \theta$ unbiased, $\hat{\theta}(X^{(1)}, ..., X^{(N)}) \coloneqq \bar{X}_N$.

Methods of moment: let $X = (X_1, ..., X_m)$ be a random vector. Let $\alpha \in \mathbb{N}^m$, the α-th moment is defined to be $\mu_\alpha = \mathbb{E}[X_1^{\alpha_1} X_2^{\alpha_2} ... X_m^{\alpha_m}]$. The order of moment is $|\alpha| = \sum \alpha_i$.

Another example: Binomial models

Example 5.14 (Multidimensional gaussian). Assume $X^{(1)}, ..., X^{(N)} \sim N_m(\mu, \Sigma)$.

- $M_1 \coloneqq \mathbb{E}[X] = \mu \in \mathbb{R}^m$.
- $M_2 \coloneqq \mathbb{E}[XX^T] = \Sigma + \mu\mu^T$
- $M_3 \coloneqq \mathbb{E}[X \otimes X \otimes X]$

The first two order moment suffice to solve the system.

$$\hat{\mu}_1 = \mu, \hat{\mu}_2 = \Sigma + \mu\mu^T$$

and

$$\begin{aligned}\hat{\Sigma} = \hat{M}_2 - \hat{M}_1\hat{M}_1^T &= \frac{1}{N}\sum_{i=1}^N X^{(i)}(X^{(i)})^T - \bar{X}_N\bar{X}_N^T \\ &= \frac{1}{N}\sum_{i=1}^N (X^{(i)} - \bar{X})(X^{(i)} - \bar{X})^T = \frac{N-1}{N}S_N^2 = \tilde{S}_N^2\end{aligned}$$

Sample covariance: $S_N^2 = \frac{1}{N-1}\sum_{i=1}^N (X^{(i)} - \bar{X})^2$.

Given $X^{(1)}, ..., X^{(N)}$ iid, $\mathbb{E}[X^{(i)}] = \mu$ and $\text{Var}[X^{(i)}] = \sigma^2$. Is the sample covariance unbiased, i.e. $\mathbb{E}[S_N^2] = \sigma^2$ and $\mathbb{E}[\tilde{S}_N^2] = (1 - \frac{1}{N})\sigma^2$.

Proof.

$$\begin{aligned}\mathbb{E}[\sum_{i=1}^N (X^{(i)} - \bar{x})^2] = \sum_{i=1}^N \mathbb{E}[(X^{(i)} - \bar{x})^2] = \sum_{i=1}^N \text{Var}[X^{(i)} - \bar{x}] &= \sum_{i=1}^N \text{Var}[(1 - \frac{1}{N})X^{(i)} - \frac{1}{N}\sum_j^N X^{(j)}] \\ &= \sum_{i=1}^N ((1 - \frac{1}{N})^2 \text{Var}[X^{(i)}] + \frac{1}{N^2}\sum \text{Var}[\text{X}^{(\text{j})}]) \\ &= \sum (\frac{N-1}{N})^2\sigma^2 + \frac{N-1}{N^2}\sigma^2 \\ &...\end{aligned}$$

□

Gaussian mixture: $f(x) = \lambda f_{\mu_1,\sigma_1^2}(x) + (1-\lambda) f_{\mu_2,\sigma_2^2}(x)$. Assume $X_1 \sim N(\mu, \sigma_1^2)$ and $X_2 \sim N(\mu_2, \sigma_2^2)$.

Assume $X^{(1)}, ..., X^{(N)}$ iid $\sim f$. Then $m_1 = \lambda\mu_1 + (1-\lambda)\mu_2$ and $m_2 = \lambda(\mu_1^2 + \sigma_1^2) + (1-\lambda)(\mu_2^2 + \sigma_2^2)$, $m_3 = \lambda(\mu_1^3 + 3\mu_1\sigma_1^2) + ...$, $m_4 = \lambda(\mu_1^4 + 6\mu_1^2\sigma_1^2 + 3\sigma_1^4) + (1-\lambda)...$ and $m_5 = \lambda(\mu_1^5 + 10\mu_1^3\sigma_1^2 + 15\mu_1\sigma_1^4) + (1-\lambda)$.

Solve system for $\lambda, \mu_1, \mu_2, \sigma_1^2, \sigma_2^2$ and $\hat{m}_1 = m_1, ..., \hat{m}_5 = m_5$. Reduces to monis polynomial in $p = \mu_1\mu_2$.

Example 5.15. Assume $X^{(1)}, ..., X^{(N)} \sim \text{Unif}(-\theta, \theta)$ with $\theta > 0$. What is $\hat{\theta}_{MOM}$? So $\bar{x}_\mu = \hat{\mu}_1 = \mu_1(\theta) = 0$. This does not work.

$\hat{\theta}_{MOM} = \sqrt{3\hat{\mu}_2}$.

Example 5.16. $(x^{(1)}, ..., x^{(5)}) = (0, 2, -1, -1, 0)$, $\hat{\mu}_1 = 0$ and $\hat{\mu}_2 = \frac{1}{5}(0^2+2^2+1+1+0^2) = \frac{6}{5}$. So $\hat{\theta}_{MOM}(x^{(1)}, ..., x^{(5)} = \sqrt{\frac{18}{5}})$. THis is smaller than $\sqrt{\frac{20}{5}} = 2$. Not a good estimator.

5.2 Maximum likelihood estimation

We recall that $\frac{\partial \mathbf{x}^T\mathbf{A}\mathbf{x}}{\partial \mathbf{x}} = \mathbf{x}^T(\mathbf{A} + \mathbf{A}^T)$, $\frac{\partial \text{tr}(\mathbf{AB})}{\partial \mathbf{A}} = \mathbf{B}^T$, and $\frac{\partial \log(\det \mathbf{A})}{\partial \mathbf{A}} = (\mathbf{A}^{-1})^T$.

Proposition 5.17 (Multivariate normal model). For a multivariate normal random vector, the maximum likelihood estimator of μ and Σ are given by

$$\hat{\mu} = \frac{1}{N}\sum_{i=1}^{N} X^{(i)}, \quad \text{and} \quad \hat{\Sigma} = \frac{1}{N}(X^{(i)} - \hat{\mu})(X^{(i)} - \hat{\mu})^T.$$

Proof. The log-likelihood function is given by

$$\ell_X(\mu, \Sigma) = -\frac{1}{2}\sum_{i=1}^{N}\left(m\log(2\pi) + \log(\det(\Sigma)) + (X^{(i)} - \mu)^T\Sigma^{-1}(X^{(i)} - \mu)\right).$$

Write $X^{(i)} - \bar{X} + \bar{X} - \mu$ for $X^{(i)} - \mu$, and we see that the log-likelihood function becomes

$$\ell_X(\mu, \Sigma) = -\frac{mN}{2}\log(2\pi) - \frac{N}{2}\log\det(\Sigma) - \frac{N}{2}\text{tr}(\Sigma^{-1}S) - \frac{N}{2}(\bar{X} - \mu)^T\Sigma^{-1}(\bar{X} - \mu).$$

Using the derivative rules we see that

$$\hat{\mu} = \bar{X} \quad \text{and} \quad \hat{\Sigma} = S = \frac{1}{N}\sum(X^{(i)} - \hat{\mu})(X^{(i)} - \hat{\mu})^T.$$

□

6 Exponential families

Definition 6.1 (Exponential family). A statistical model $\mathcal{M} = \{p_\theta\}_{\theta\in\Theta}$ on $\mathcal{X}$ is called a k-dimensional **exponential family** if p is of the form

$$p_\theta(x) = h(x) \cdot \exp\left(\langle \eta(\theta), T(x)\rangle - B(\theta)\right) \quad \forall x \in \mathcal{X},$$

where

- $h : \mathcal{X} \to (0, \infty)$,
- $\eta : \Theta \to \mathbb{R}^k$ is called the *natural parameter*,
- $T : \mathcal{X} \to \mathbb{R}^k$ is a sufficient statistic (see factorization theorem by Fisher-Neyman),
- $B : \Theta \to \mathbb{R}$ is measurable.

The function $B(\theta)$ is a normalizing constant; its explicit form is

$$B(\theta) = \log \int_{\mathcal{X}} h(x) \cdot \exp\left(\langle \eta(\theta), T(x)\rangle\right) \, \mathrm{d}x.$$

Example 6.2 (Binomial random variable). Fix $n \in \mathbb{N}$. Let $X \sim \mathrm{Bin}(n, \theta)$, $\mathcal{X} = \{0, 1, \ldots, n\}$. We have

$$p_\theta(x) = \binom{n}{x}\theta^x(1-\theta)^{n-x} = \binom{n}{x}\exp\left\{x\log(\theta) + (n-x)\log(1-\theta)\right\}.$$

This is a 2-dimensional exponential family with sufficient statistic $T(x) = (x, n-x)$ and natural parameter $\eta(\theta) = (\log(\theta), \log(1-\theta))$.

We can also write the binomial distribution as a one-dimensional exponential family

$$p_\theta(x) = \binom{n}{x}\exp\left\{x\log\left(\frac{\theta}{1-\theta}\right) + n\log(1-\theta)\right\}$$

with $T(x) = x$, $\eta(\theta) = \frac{\theta}{1-\theta}$ and $B(\theta) = -n\log(1-\theta)$.

Definition 6.3 (Minimal exponential family). A k-dimensional exponential family is called **minimal** if the sufficient statistic $T(\mathcal{X})$ does not lie on a proper affine subspace of $\mathbb{R}^k$, i.e. there exists no $c \in \mathbb{R}$ and $\mathbf{v} \in \mathbb{R}^k \setminus \{0\}$ such that $\langle \mathbf{v}, T(x)\rangle = c$ for all $x \in \mathcal{X}$.

Example 6.4 (Nonminimal binomial random variable). In the previous example, the two-dimensional exponential family of the binomial random variable is not minimal since $\langle \mathbf{1}, T(x)\rangle = n$ is a constant.

Proposition 6.5 (Minimal implies identifiable natural parameter). If an exponential family is minimal, then the natural parameter $\eta(\theta)$ is identifiable, i.e. it is not overparametrized.

Proof. Assume $\langle \eta(\theta), T(x) \rangle = \langle \eta(\theta'), T(x) \rangle$. Then $\langle \eta(\theta) - \eta(\theta'), T(x) \rangle = 0$. Since the exponential family is minimal, we must have $\eta(\theta) = \eta(\theta')$. □

Example 6.6 (Univariate normal distribution). Let $X \sim \mathcal{N}(\mu, \sigma^2)$. Define $\mathcal{X} = \mathbb{R}$ and $\theta = (\mu, \sigma^2) \in \mathbb{R} \times (0, \infty)$. The density function is

$$p_\theta(x) = \left(2\pi\sigma^2\right)^{-\frac{1}{2}} \exp\left\{-\frac{1}{2\sigma^2}(x-\mu)^2\right\} = \left(2\pi\sigma^2\right)^{-\frac{1}{2}} \exp\left\{-\frac{x^2}{2\sigma^2} + \frac{x\mu}{\sigma^2} - \frac{\mu^2}{2\sigma^2}\right\}.$$

This is a two-dimensional exponential family with sufficient statistic $T(x) = (x, -\frac{x^2}{2})$, natural parameter $\eta(\theta) = (\frac{\mu}{\sigma^2}, \sigma^{-2})$ and $B(\theta) = \frac{\mu^2}{2\sigma^2}$.

Proposition 6.7 (Joint distribution of i.i.d. exponential family is exponential family). Let $X^{(1)}, \ldots, X^{(N)}$ be random variables i.i.d. from a k-dimensional exponential family $\mathcal{M} = \{p_\theta\}_{\theta \in \Theta}$. Then the joint distribution of $X^{(1)}, \ldots, X^{(N)}$ is a k-dimensional exponential family with sufficient statistic $\sum_{i=1}^N T(X^{(i)})$.

Proof. We have

$$p_\theta(x^{(1)}, \ldots, x^{(N)}) = \prod_{i=1}^N p_\theta(x^{(i)}) = \prod_{i=1}^N h(x^{(i)}) \cdot \exp\left(\langle \eta(\theta), \sum_{i=1}^N T(x^{(i)}) \rangle - NB(\theta)\right).$$

□

6.1 Canonical form

We can reparametrize the exponential family in terms of natural parameters

$$p_\eta(x) = h(x) \cdot \exp\left(\langle \eta, T(x) \rangle - A(\eta)\right) \quad \forall x \in \mathcal{X}$$

with log *partition function* (or *cumulant generating function*)

$$A(\eta) = \log \int_{\mathcal{X}} h(x) \cdot \exp\left(\langle \eta, T(x) \rangle\right) \mathrm{d}x,$$

provided that the natural parameter is taken from

$$\eta \in N = \left\{\eta \in \mathbb{R}^k : \int_{\mathcal{X}} h(x) e^{\langle \eta, T(x) \rangle} \mathrm{d}x < \infty\right\}.$$

Theorem 6.8. Let $\mathcal{M}$ be a k-dimensional exponential family. Then

- N is a convex set,
- $A : N \to \mathbb{R}$ is convex (and strictly convex if $\mathcal{M}$ is minimal),
- the log-likelihood function $\ell_X(\eta)$ is concave on N

Theorem 6.9. Let $\mathcal{M}$ be a k-dimensional exponential family with sufficient statistic $T(x)$ and log partition function $A(\eta)$. Then

$$\nabla A(\eta) = \mathbb{E}_\eta[T(X)] \quad \text{and} \quad \nabla^2 A(\eta) = \mathrm{Var}_\eta[T(X)].$$

Corollary 6.10. Let $\mathcal{M}$ be a regular exponential family with suffcient statistic $T(X)$. Given data X, the maximum likelihood estimator, if it exists, is the unique solution to the equation

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

Proof. The likelihood function is given by

$$p_\eta(x) = h(x)e^{\langle \eta, T(x)\rangle - A(\eta)}$$

Taking the log gives

$$\log p_\eta(x) = \log h(x) + \langle \eta, T(x)\rangle - A(\eta).$$

The gradient of the log-likelihood function is

$$\nabla_\eta \ell_x(\eta) = T(x) - \nabla A(\eta).$$

So, the critical equation reads $T(x) = \mathbb{E}_\eta[T(X)]$. □

Definition 6.11.

Example 6.12 (Binomial random variable in canonical form). We know

$$p_\theta(x) = \binom{n}{x} \exp\left\{ x \log\left(\frac{\theta}{1-\theta}\right) + n\log(1-\theta)\right\}.$$

Substitute $\eta = \log\left(\frac{\theta}{1-\theta}\right)$. Then $e^\eta = \frac{\theta}{1-\theta}$, so $\theta = \frac{e^\eta}{1+e^\eta}$, and $1-\theta = \frac{1}{1+e^\eta}$. Therefore,

$$p_\eta(x) = \binom{n}{x} \exp\left\{\eta x - A(\eta)\right\}$$

with $A(\eta) = n\log\left(1+e^\eta\right)$.

Example 6.13 (Discrete random variable). Let X be a discrete random variable taking values in $\mathcal{X} = [r]$. Define the sufficient statistic $T : \mathcal{X} \to \mathbb{N}^{r-1}$ by

$$T : x \mapsto \begin{bmatrix} \mathbb{1}_{\{1\}} \\ \mathbb{1}_{\{2\}} \\ \vdots \\ \mathbb{1}_{\{r-1\}} \end{bmatrix}.$$

Note that $T(x) = \mathbf{0}$ if and only if $x = r$. Moreover, define $h(x) = 1$ for all $x \in \mathcal{X}$. By formula, we have

$$A(\eta) = \log(1 + \sum_{x=1}^{r-1} \exp(\eta_x)),$$

where we integrate with respect to the count measure. Hence, for all $x \in [r-1]$ we get

$$p_\eta(x) = \frac{e^{\eta_x}}{1 + \sum_{k=1}^{r-1} e^{\eta_k}},$$

and for $x = r$ we have

$$p_\eta(x) = 1 - \sum_{k=1}^{r-1} p_\eta(k) = \frac{1}{1 + \sum_{k=1}^{r-1} e^{\eta_k}}.$$

Since $1 + \sum_{x=1}^{r-1} \exp(\eta_x)$ is finite for all $\eta \in \mathbb{R}^{r-1}$, the natural parameter space is $\mathbb{R}^{r-1}$. We obtain an exponential family $\{P_\eta : \eta \in \mathbb{R}^{r-1}\}$, where the natural parameter η can be interpreted as the log-odds ratio; p_η is equal to some probability distribution $(p_1, \ldots, p_r)$ if and only if $\eta_x = \log(\frac{p_x}{p_r})$ for all $x = 1, ..., r-1$. So we obtain a correspondence between the natural parameter space $\mathbb{R}^{r-1}$ and the probability simplex Δ_{r-1}.

Remark 6.14 (Trace trick). Let's recall the *trace trick*. For any vector $\mathbf{x}$ and matrix $\mathbf{A}$ we have $\mathbf{x}^T\mathbf{A}\mathbf{x} = \mathrm{tr}(\mathbf{x}^T\mathbf{A}\mathbf{x})$ since $x^T\mathbf{A}\mathbf{x}$ is a scalar. Under trace we have $\mathbf{x}^T\mathbf{A}\mathbf{x} = \mathbf{x}\mathbf{x}^T\mathbf{A}$ and $\mathbf{x}^T\mathbf{A}\mathbf{x} = \mathbf{A}\mathbf{x}\mathbf{x}^T$.

Example 6.15 (Multivariate normal random variables). Define $\mathcal{X} = \mathbb{R}^m$ and $h(x) = 1$ for all $x \in \mathcal{X}$. Consider the matrix

$$\begin{bmatrix} x_1^2 & x_1x_2 & \cdots & x_1x_m \\ \vdots & x_2^2 & \cdots & x_2x_m \\ \vdots & \cdots & \ddots & \vdots \\ \cdots & \cdots & & x_m^2 \end{bmatrix} \in \mathbb{R}^{m \times m}.$$

Define a statistic $T : \mathbb{R}^m \to \mathbb{R}^m \times \mathbb{R}^{m(m+1)/2}$ by $\mathbf{x} \mapsto \begin{bmatrix} \mathbf{x} \\ \mathbf{d} \\ \mathbf{o} \end{bmatrix}$ where $\mathbf{d} = -(x_i^2/2)_{i \in [m]}$ and

$$\mathbf{o} = -\begin{bmatrix} x_1x_2 \\ \vdots \\ x_1x_m \\ x_2x_3 \\ \vdots \\ x_{m-1}x_m \end{bmatrix};$$

the vector $\mathbf{o}$ contains all the off-diagonal elements of the matrix, the vector $\mathbf{d}$ contains the diagonal elements of the matrix. Hence, the dimension of $[\mathbf{d}, \mathbf{o}]$ is $m(m+1)/2$.

Given a natural parameter $\eta \in \mathbb{R}^{m+m(m+1)/2}$, we write

$$\eta_{[m]} \coloneqq \begin{bmatrix} \eta_1 \\ \vdots \\ \eta_m \end{bmatrix} \in \mathbb{R}^m \quad \text{and} \quad \eta_{[m\times m]} = \begin{bmatrix} \eta_{m+1} & \cdots & \\ & \ddots & \vdots \\ & & \eta_{2m} \end{bmatrix} \in Symm(m).$$

The natural parameter space is $N = \mathbb{R}^m \times \mathrm{PD}_m$. For $\eta \in N$ we have

$$A(\eta) = -\frac{1}{2}\left(\log\left(\det(\eta_{[m\times m]})\right) - \eta_{[m]}^T \eta_{[m\times m]}^{-1} \eta_{[m]} - m\log(2\pi)\right).$$

So, the Lebesgue density p_η can be written as

$$\begin{aligned} p_\eta(\mathbf{x}) &= \exp\left\{\langle \eta, T(\mathbf{x})\rangle - A(\eta)\right\} \\ &= \exp\left\{\langle \eta, T(\mathbf{x})\rangle - \frac{1}{2}\eta_{[m]}^T \eta_{[m\times m]} \eta_{[m]}\right\} \cdot \frac{1}{\sqrt{(2\pi)^m \det\left(\eta_{[m\times m]}^{-1}\right)}}. \end{aligned}$$

Write

$$\exp\left\{\langle \eta, T(\mathbf{x})\rangle\right\} = \exp\left\{\eta_{[m]}^T \mathbf{x} - \frac{1}{2}\mathbf{x}^T \eta_{[m\times m]} \mathbf{x}\right\} = \exp\left\{\eta_{[m]}^T \mathbf{x} - \frac{1}{2}\mathrm{tr}\left(\eta_{[m\times m]} \mathbf{x}\mathbf{x}^T\right)\right\}.$$

Now substitute

$$\Sigma = \eta_{[m\times m]}^{-1} \quad \text{and} \quad \mu = \Sigma \eta_{[m]}.$$

We obtain

$$\begin{aligned} \exp\left\{-\frac{1}{2}(\mathbf{x}-\mu)^T \Sigma^{-1}(\mathbf{x}-\mu)\right\} &= \exp\left\{-\frac{\left(\mathbf{x}^T\Sigma^{-1}\mathbf{x} - \mathbf{x}^T\Sigma^{-1}\mu - \mu^T\Sigma^{-1}\mathbf{x} + \mu^T\Sigma^{-1}\mu\right)}{2}\right\} \\ &= \exp\left\{\frac{-\mathbf{x}^T\Sigma^{-1}\mathbf{x} + 2\mathbf{x}^T\Sigma^{-1}\mu - \mu^T\Sigma^{-1}\mu}{2}\right\} \\ &= \exp\left\{-\frac{1}{2}\mathbf{x}^T\eta_{[m\times m]}\mathbf{x} + \mathbf{x}^T\eta_{[m]} - \frac{1}{2}\eta_{[m]}^T\eta_{[m\times m]}^{-1}\eta_{[m]}\right\} \end{aligned}$$

Definition 6.16. The inverse of the covariance matrix is called the **concentration matrix** or **precision matrix**.

6.2 Discrete regular exponential families

Let X be a discrete random variable with state space $\mathcal{X} = [c]$. Assume it is an exponential family of order k with statistic $T : \mathcal{X} \to \mathbb{R}^k$, i.e.

Canonical form

$$p_\eta(x) = h(x) \cdot \exp\left\{\langle \eta, T(x)\rangle - A(\eta)\right\}.$$

We can also rewrite this as

Parametrization by rational functions in θ_j

$$p_\theta(x) = \frac{1}{Z(\theta)} h_x \prod_{j=1}^{k} \theta_j^{t_{jx}}$$

for $\mathbf{h} = [h(1), \ldots, h(c)]$, $\theta_j = \exp(\eta_j)$ and $T(x)_j = t_{jx}$, i.e.

$$\mathbf{T} = (t_{ij}) = \begin{bmatrix} | & ... & | \\ T(1) & ... & T(c) \\ | & ... & | \end{bmatrix} \in \mathbb{R}^{k \times c}.$$

To normalize p_θ to one we set $Z(\theta) = \sum_x h_x \prod_j \theta_j^{t_{jx}}$. If all the exponents $\mathbf{T}$ are integers, then we see that the discrete exponential family is parametrized by rational functions.

Definition 6.17 (Log-affine model). Let $\mathbf{A} \in \mathbb{Z}^{r \times c}$ such that $\mathbf{1} \in \text{rowspan}(\mathbf{A})$ and $\mathbf{h} \in \mathbb{R}^c_{>0}$. The **log-affine model** associated to $\mathbf{A}$ and $\mathbf{h}$ is the set of all probability distributions

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

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

To $\mathbf{A}$ and $\mathbf{h}$ we can associate a **monomial map** $\phi^{\mathbf{A},\mathbf{h}} : \mathbb{R}^r \to \mathbb{R}^c$ defined by the *rational* map

$$\theta \mapsto \begin{bmatrix} h_1 \prod_{j=1}^r \theta_j^{a_{j1}} \\ h_2 \prod_{j=1}^r \theta_j^{a_{j2}} \\ \vdots \\ h_c \prod_{j=1}^r \theta_j^{a_{jc}} \end{bmatrix} = \left[h_i \theta^{\mathbf{a}_{\cdot i}}\right]_{i=1,\ldots,c} \in \mathbb{R}^c.$$

Equipped with the monomial map we can associate a **toric ideal** to $\mathbf{A}$ and $\mathbf{h}$ by

$$I_{\mathbf{A},\mathbf{h}} = I\left(\phi^{\mathbf{A},\mathbf{h}}(\mathbb{R}^r)\right) \subset \mathbb{R}[\mathbf{p}].$$

Toric ideal = vanishing ideal of the image of a monomial map.

Question: What are the generators of the toric ideal $I_{\mathbf{A},\mathbf{h}}$?

It suffices to consider the case $\mathbf{h} = \mathbf{1}$ since we can always divide by $\mathbf{h}$ to obtain the general case; i.e. we substitute $p_j \mapsto \frac{p_j}{h_j}$.

In case of $I_{\mathbf{A}}$ (that is $\mathbf{h} = \mathbf{1}$) we deal with binomial ideals.

Proposition 6.18 (Toric ideals of log-linear models are binomial). Let $\mathbf{A} \in \mathbb{Z}^{r \times c}$. Then the toric ideal $I_{\mathbf{A}}$ is a binomial ideal and

$$I_{\mathbf{A}} = (\mathbf{x}^{\alpha} - \mathbf{x}^{\beta} \mid \alpha, \beta \in \mathbb{N}^c \text{ with } A\alpha = A\beta).$$

If $\mathbf{1} \in \text{rowspan}(\mathbf{A})$, then the toric ideal is homogeneous.

Proof. We prove something stronger: I_A is a k-linear combination of binomials $x^{\alpha} - x^{\beta}$ with $A\alpha = A\beta$ for $\alpha, \beta \in \mathbb{N}^c$. The direction $\supset$ is clear. ... □

Example 6.19. Let $\mathbf{A} = \begin{bmatrix} 0 & 1 & 2 \\ 2 & 1 & 0 \end{bmatrix}$ and $\mathbf{h} = \begin{bmatrix} 1 & 2 & 1 \end{bmatrix}$ then

$$p_0 \mapsto \theta_1^2, \quad p_1 \mapsto 2\theta_1\theta_2, \quad p_2 \mapsto \theta_2^2.$$

The toric ideal is $I_{\mathbf{A},\mathbf{h}} = (4p_0p_2 - p_1^2)$.

Example 6.20 (Twisted cube). Let $\mathbf{A} = \begin{bmatrix} 0 & 1 & 2 & 3 \\ 3 & 2 & 1 & 0 \end{bmatrix}$. The toric ideal is the vanishing ideal of the parametrization of

$$p_1 \mapsto \theta_2^3, \quad p_2 \mapsto \theta_1\theta_2^2, \quad p_3 \mapsto \theta_1^2\theta_2, \quad p_4 \mapsto \theta_1^3.$$

The toric ideal is generated by three quadratic binomials

$$p_1p_3 - p_2^2, p_2p_4 - p_3^2, p_1p_4 - p_2p_3.$$

Example 6.21 (Binomial random variable with three trials). Now assume $\mathbf{h} = [1, 3, 3, 1]$. Then the toric ideal is the vanishing ideal of the parametrization of

$$p_1 \mapsto \theta_2^3, \quad p_2 \mapsto 3\theta_1\theta_2^2, \quad p_3 \mapsto 3\theta_1^2\theta_2, \quad p_4 \mapsto \theta_1^3.$$

The toric ideal $I_{\mathbf{A},\mathbf{h}}$ is generated by three quadratic binomials

$$3p_1p_3 - p_2^2, 3p_2p_4 - p_3^2, 9p_1p_4 - p_2p_3.$$

Intersecting the toric ideal with the probability simplex yields the binomial probability distribution with three trials. To see this note that the probability distribution is given by $p_k = \binom{3}{k}\theta^k(1-\theta)^{3-k}$; setting $\theta_1 = \theta$ and $\theta_2 = 1 - \theta$ yields the above parametrization.

Example 6.22 (Discrete independent random variables). Assume we have two independent variables X and Y with state space $\mathcal{X} = [r_1]$ and $\mathcal{Y} = [r_2]$. We parametrize the joint distribution by

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

Question: What is the matrix $\mathbf{A}$ associated to the toric ideal?

Assume $r_1 = 3$ and $r_2 = 4$. Then

$$\mathbf{A} = \begin{bmatrix} 1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 \\ 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 1 \end{bmatrix} \in \mathbb{Z}^{7 \times 12}.$$

Question: What does $\mathbf{A}$ do?

Given count vector $\mathbf{u} = [u_{11}, u_{12}, ..., u_{34}]$, we have that $\mathbf{Au} = \begin{bmatrix} \text{row sums} \\ \text{column sums} \end{bmatrix}$ computes the row and column sums of the count vector. Since $\mathbf{A}$ represents a sufficient statistic, $T(\mathbf{x})$ computes the row and column sums of observed data.

Question: How does the toric ideal look like?

It is generated by all binomials $\mathbf{p^u} - \mathbf{p^{u'}}$ such that $\mathbf{Au} = \mathbf{Au'}$; in other words the counts $\mathbf{u}$ and $\mathbf{u'}$ must have the same row and column sum.

Note that $I_\mathbf{A}$ equals $I_{X \perp\!\!\!\perp Y} = (p_{i_1 j_1} p_{i_2 j_2} - p_{i_1 j_2} p_{i_2 j_1})$. *The independence ideal is a toric ideal!*

Example 6.23 (Complete independence model). Let X, Y, Z be discrete random variables with state spaces $\mathcal{X} = \mathcal{Y} = \mathcal{Z} = \{0, 1\}$. The associated matrix $\mathbf{A}$ of the log-linear model is

$$\mathbf{A} = \begin{bmatrix} 1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 \\ 1 & 1 & 0 & 0 & 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 & 0 & 0 & 1 & 1 \\ 1 & 0 & 1 & 0 & 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 & 0 & 1 & 0 & 1 \end{bmatrix} \in \mathbb{Z}^{(2+2+2) \times (2^3)} = \mathbb{Z}^{6 \times 8}.$$

It is the vanishing ideal of the following parametrization

$$p_{111} = \alpha_1 \beta_1 \gamma_1, \quad p_{112} = \alpha_1 \beta_1 \gamma_2, \quad p_{121} = \alpha_1 \beta_2 \gamma_1, \quad p_{122} = \alpha_1 \beta_2 \gamma_2,$$
$$\dots$$

To compute the complete independence model $\mathcal{M} = \{\mathbf{p} \mid \log(\mathbf{p}) \in \text{rowspan}(\mathbf{A})\}$ we compute the toric ideal $I_\mathbf{A}$. Using Macaulay2 or Singular we compute the toric ideal

$$\begin{aligned} I_\mathbf{A} = (&p_6 p_7 - p_5 p_8, p_2 p_7 - p_1 p_8, p_2 p_3 - p_1 p_4, \\ &p_1 p_6 - p_2 p_5, p_2 p_8 - p_4 p_6, p_3 p_6 - p_2 p_7, \\ &p_4 p_5 - p_2 p_7, p_2 p_7 - p_4 p_5, p_1 p_7 - p_3 p_5). \end{aligned}$$

If we label the columns of $\mathbf{A}$ as follows

$$\mathbf{A} = \begin{bmatrix} | & | & | & | & | & | & | & | \\ p_1 & p_2 & p_3 & p_4 & p_5 & p_6 & p_7 & p_8 \\ | & | & | & | & | & | & | & | \end{bmatrix}$$
$$= \begin{bmatrix} | & | & | & | & | & | & | & | \\ p_{111} & p_{112} & p_{121} & p_{122} & p_{211} & p_{212} & p_{221} & p_{222} \\ | & | & | & | & | & | & | & | \end{bmatrix},$$

we obtain that

$$I_{\mathbf{A}} = (p_{21,2}p_{22,1} - p_{21,1}p_{22,2},\ p_{11,2}p_{12,1} - p_{11,1}p_{12,2},\ p_{11,2}p_{22,1} - p_{11,1}p_{22,2},$$
$$p_{1,11}p_{2,12} - p_{1,12}p_{2,11},\ p_{1,12}p_{2,22} - p_{1,22}p_{2,12},\ p_{1,21}p_{2,12} - p_{1,12}p_{2,21},$$
$$p_{112}p_{221} - p_{122}p_{211},\ p_{211}p_{122} - p_{221}p_{112},\ p_{221}p_{111} - p_{211}p_{121},),$$

which is $I_{\{1\perp\!\!\!\perp\{2,3\},2\perp\!\!\!\perp\{1,3\},3\perp\!\!\!\perp\{2,3\}\}}$. The toric ideal is the complete independence model of three variables.

6.3 Gaussian regular exponential families

Let $\mathbf{X} \sim \mathcal{N}_m(\mu, \Sigma)$ and $\theta = (\mu, \Sigma) \in \mathbb{R}^m \times \mathrm{PD}_m$. The density function is

$$\begin{aligned} p_\theta(\mathbf{x}) &= (2\pi)^{-\frac{m}{2}} \det(\Sigma)^{-\frac{1}{2}} \cdot \exp\left(-\frac{1}{2}(\mathbf{x}-\mu)^T\Sigma^{-1}(\mathbf{x}-\mu)\right) \\ &\propto \exp\left(-\frac{1}{2}\mathbf{x}^T\Sigma^{-1}\mathbf{x} + \mathbf{x}^T\Sigma^{-1}\mu - \frac{1}{2}\mu^T\Sigma^{-1}\mu\right) \\ &\propto \exp\left(-\frac{1}{2}\mathbf{x}^T\Sigma^{-1}\mathbf{x} + \mathbf{x}^T\Sigma^{-1}\mu\right) \\ &= \exp\left(\langle\Sigma^{-1}, -\frac{1}{2}\mathbf{x}\mathbf{x}^T\rangle + \langle\Sigma^{-1}\mu, \mathbf{x}\rangle\right). \end{aligned}$$

We get the following *sufficient statistic*, *natural parameter* and *natural parameter space*:

$$T(\mathbf{x}) = \begin{bmatrix} \mathbf{x} \\ -\frac{1}{2}\mathbf{x}\mathbf{x}^T \end{bmatrix} \quad \text{and} \quad \eta = \begin{bmatrix} \Sigma^{-1}\mu \\ \Sigma^{-1} \end{bmatrix} \in N = \mathbb{R}^m \times \mathrm{PD}_m.$$

We now consider submodels of the Gaussian model: we explore the natural parameter space $N' = L \cap (\mathbb{R}^m \times \mathrm{PD}_m)$, where L is a linear subspace. In applications $L = L_1 \times L_2$ where $L_1 \subset \mathbb{R}^m$ and $L_2 \subset \mathrm{PD}_m$ are linear subspaces; often, $L_1 = \{\mathbf{0}\}$ or $L_1 = \mathbb{R}^m$. The associated vanishing ideal lives in $\mathbb{R}[\mu_i, \sigma_{ij} \mid 1 \le i \le j \le m]$.

- If $L_1 = \{\mathbf{0}\}$, then the vanishing ideal is of the form $I = (\mu_1, \ldots, \mu_m) + I'$ for $I' \subset \mathbb{R}[\sigma_{ij}]$ since $\mu = \mathbf{0}$.
- If $L_1 = \mathbb{R}^m$, then there is no restriction on μ; hence I only depends on σ_{ij}.

In either case we can focus on the vanishing ideal associated to L_2.

Definition 6.24 (Inverse linear space). Let $L \subset \mathbb{R}^{\frac{(m+1)m}{2}}$ be a linear space such that $L \cap \mathrm{PD}_m \neq \emptyset$. The **inverse linear space** is the set of all positive definite matrices

$$L^{-1} = \left\{ K^{-1} \mid K \in L \cap \mathrm{PD}_m \right\}.$$

Definition 6.25 (Gaussian linear concentration model). The **Gaussian linear concentration model** defined by L consists of all distributions $\mathcal{N}_m(\cdot, \Sigma)$ such that $\Sigma \in L^{-1}$, i.e. $\Sigma^{-1} \in L$. In other words, the concentration matrix $K = \Sigma^{-1}$ is in L.

We can derive a special meaning if L constraints some off-diagonal elements K_{ij} to zero. It corresponds to saturated conditional independence statements.

Proposition 6.26 (Concentration model and conditional independence). Let $K = (K_{ij})$ be a concentration matrix of a multivariate random normal vector. For $i \neq j \in [m]$, we have

$$K_{ij} = 0 \iff X_i \perp\!\!\!\perp X_j \mid X_{[m]\setminus\{i,j\}}.$$

Proof. We have

$$K = \Sigma^{-1} = \frac{1}{\det(\Sigma)} \mathrm{adj}(\Sigma),$$

where the adjugate is given by $\mathrm{adj}(\Sigma)_{ij} = (-1)^{i+j} \det(\tilde{\Sigma}_{ji})$ and

$$\tilde{\Sigma}_{ji} = \Sigma_{[m]\setminus\{j\},[m]\setminus\{i\}} = \Sigma_{\mathcal{C}\cup\{i\},\mathcal{C}\cup\{j\}}$$

for $C = [m] \setminus \{i, k\}$. Thus,

$$k_{ij} = 0 \iff \det(\Sigma_{C\cup\{i\},\mathcal{C}\cup\{j\}}) = 0 \iff X_i \perp\!\!\!\perp X_j \mid X_C.$$

The last equivalence follows from $A \perp\!\!\!\perp B \mid C$ if and only if $\mathrm{rank}(\Sigma_{A\cup C, B\cup C}) = \#C$ if and only if every $(\#C + 1) \times (\#C + 1)$ minor of $\Sigma_{A\cup C, B\cup C}$ vanishes. □

Example 6.27 (Gaussian independence model). Consider the Gaussian exponential family with $m = 3$ random variables and

$$L = \{K \in \mathrm{PD}_3 \mid k_{12} = k_{13} = 0\}.$$

This means that $1 \perp\!\!\!\perp 2 \mid 3$ and $1 \perp\!\!\!\perp 3 \mid 2$. The intersection axiom implies that $1 \perp\!\!\!\perp \{2, 3\}$. If we look at the conditional independence ideal

$$I_{\mathcal{C}} \cap \mathrm{PD}_3 = (\sigma_{12}\sigma_{33} - \sigma_{13}\sigma_{23}, \sigma_{13}\sigma_{22} - \sigma_{12}\sigma_{23}) \cap \mathrm{PD}_3,$$

the intersection axiom is not immediately visible; do we know that $\sigma_{1,2}$ and $\sigma_{1,3}$ are in $I_{\mathcal{C}} \cap \mathrm{PD}_3$?

- We could use primary decomposition and extract the component that is in PD_3.
- We could use the parametrization of the Gaussian exponential family and compute the vanishing ideal of that parametrization. If we do this, then we see that the vanishing ideal is generated by σ_{12} and σ_{13}.

7 Likelihood inference

Given a sequence of samples $\mathbf{X} = (X^{(1)}, \ldots, X^{(N)})$ and model $\mathcal{M}_{\theta \in \Theta}$, $\Theta \subset \mathbb{R}^d$ maximize the log-likelihood function $\operatorname{argmax}_{\theta \in \Theta} \ell_{\mathbf{X}}(\theta)$, where $\ell_{\mathbf{X}}(\theta) = \log\left(\prod_{i=1}^{N} f_\theta(X^{(i)})\right)$. The **score equations** are

$$\begin{cases} \frac{\partial \ell_{\mathbf{X}}}{\partial \theta_1}(\theta) = 0 \\ \frac{\partial \ell_{\mathbf{X}}}{\partial \theta_2}(\theta) = 0 \\ \vdots \\ \frac{\partial \ell_{\mathbf{X}}}{\partial \theta_d}(\theta) = 0 \end{cases}$$

Question: What is the algebraic structure of the score equations?

7.1 Maximum likelihood degree

Assume a discrete model $\mathcal{M}_\Theta \subset \Delta_{r-1}$ with state space $[r]$. The log-likelihood function is

$$\ell_{\mathbf{X}}(\theta) = \sum_{j=1}^{r} u_j \log(p_j(\theta)).$$

Observe that if the p_js are rational functions, then so are the score equations

$$\frac{\partial \ell_{\mathbf{X}}}{\partial \theta_k}(\theta) = \sum_{j=1}^{r} \frac{u_j}{p_j} \frac{\partial p_j}{\partial \theta_k}(\theta) = 0.$$

Definition 7.1 (Maximum likelihood degree). The number of complex solutions (if finite) to the score equations for generic u is called the **maximum likelihood degree** of the parametric discrete model.

The ML degree is an algebraic measure of *complexity*. The best case if the ML degree is one, then there is a unique solution.

We need to check that the ML degree is well-defined.

Proposition 7.2 (ML degree is well-defined). For generic data $u \in \mathbb{N}^r$, the number of complex solutions is independent of u.

Proof. Consider the coefficient field $\mathbb{C}(\mathbf{u})$. Assume $p_j(\theta) = \frac{f_j}{g_j}(\theta)$ for all $j = 1, \ldots, r$. The log likelihood function becomes $\sum_{j=1}^{r} u_j \cdot (\log(f_j(\theta)) - \log(g_j(\theta)))$. The score functions are $\sum_{j=1}^{r} u_j (\frac{1}{f_j} \frac{\partial f_j}{\partial \theta_k}(\theta) - \frac{1}{g_j} \frac{\partial g_j}{\partial \theta_k}(\theta)) = 0$ for all $k = 1, \ldots, d$. Multiply by $\prod_{j=1}^{r} f_j g_j$ and we define an ideal

$$I = \left(\prod_{j=1}^{r} f_j g_j \sum_{j=1}^{r} u_j \left(\frac{1}{f_j} \frac{\partial f_j}{\partial \theta_k}(\theta) - \frac{1}{g_j} \frac{\partial g_j}{\partial \theta_k}(\theta) \right) : k = 1, \ldots, d \right).$$

Define $J = I : \left(\prod_{j=1}^{r} f_j g_j\right)^{\infty}$ and compute $V(J)$. The ML degree is the number of complex solutions θ in $V(J)$. □

Example 7.3 (What does generic mean?). Suppose we have the following score equation for some model

$$u_1\theta^2 + u_2\theta + u_3 = 0.$$

Here, generic $u = (u_1, u_2, u_3)$ means that $u_1 \neq 0$ and $u_2^2 - 4u_1u_3 \neq 0$.

Example 7.4 (Twisted cube). Suppose have the following model

$$\mathcal{M}_{\theta\in[0,1]} = \left\{ \frac{1}{1+\theta+\theta^2+\theta^3} \begin{bmatrix} 1 \\ \theta \\ \theta^2 \\ \theta^3 \end{bmatrix} : \theta \in [0,1] \right\} \subset \Delta_3.$$

Note that this is the log-linear model for the twisted cube with

$$\mathbf{A} = \begin{bmatrix} 1 & 1 & 1 & 1 \\ 0 & 1 & 2 & 3 \end{bmatrix}, \quad p_1 \mapsto \theta_1, \quad p_2 \mapsto \theta_1\theta_2, \quad p_3 \mapsto \theta_1\theta_2^2, \quad p_4 \mapsto \theta_1\theta_2^3.$$

Define $s = \frac{1}{1+\theta+\theta^2+\theta^3}$. The likelihood function is

$$L_u(\theta) = s^{u_0}(s\theta)^{u_1}(s\theta^2)^{u_2}(s\theta^3)^{u_3} = s^{u_0+u_1+u_2+u_3}\theta^{u_1+2u_2+3u_3}.$$

The log-likelihood function is

$$\ell_u(\theta) = u_+ \log(s) + (u_1 + 2u_2 + 3u_3)\log(\theta).$$

Calculating the score equations

$$3(u_+ - u_3)\theta^3 + 2(u_+ - u_2)\theta^2 + (u_+ - u_1)\theta - (u_1 + u_2^2 + u_3^3) = 0,$$

we get that the ML degree is three.

Example 7.5 (Binomial random variable). Let X be a binomial random variable with two trials. We have the parametrization

$$\theta \mapsto \begin{bmatrix} (1-\theta)^2 \\ 2\theta(1-\theta) \\ \theta^2 \end{bmatrix} \in \Delta_2.$$

The log-likelihood function is

$$\begin{aligned} \ell_u(\theta) &= 2u_0\log(1-\theta) + u_1\log(2\theta(1-\theta)) + 2u_2\log(\theta) \\ &= u_1\log(2) + (u_1 + 2u_2)\log(\theta) + (2u_0 + u_1)\log(1-\theta). \end{aligned}$$

Take the derivative, and we see that the ML degree is one.

Example 7.6 (Independence model). Let $X \perp\!\!\!\perp Y$, so the joint distribution is parametrized by

$$p_{ij} = \theta_i \theta_j.$$

Let's say $\mathcal{X} = \mathcal{Y} = [2]$. Then

$$(\theta_1, \theta_2) \mapsto \begin{bmatrix} \theta_1\theta_2 & \theta_1(1-\theta_2) \\ (1-\theta_1)\theta_2 & (1-\theta_1)(1-\theta_2) \end{bmatrix} = \mathbf{p}.$$

The log-likelihood function $\ell_u(\theta_1, \theta_2)$ is

$$\begin{aligned} u_{11}\log(\theta_1\theta_2) + u_{12}\log(\theta_1(1-\theta_2)) + u_{21}\log((1-\theta_1)\theta_2) + u_{22}\log((1-\theta_1)(1-\theta_2)) \\ = \log(\theta_1)(u_{11}+u_{12}) + \log(\theta_2)(u_{11}+u_{21}) + \log(1-\theta_1)(u_{21}+u_{22}) \\ + \log(1-\theta_2)(u_{12}+u_{22}). \end{aligned}$$

Compute the score equations

$$\begin{aligned} \frac{\partial}{\partial\theta_1}\ell_u(\theta_1,\theta_2) &= \frac{u_{11}+u_{12}}{\theta_1} - \frac{u_{21}+u_{22}}{1-\theta_1} = 0 \\ \frac{\partial}{\partial\theta_2}\ell_u(\theta_1,\theta_2) &= \frac{u_{11}+u_{21}}{\theta_2} - \frac{u_{12}+u_{22}}{1-\theta_2} = 0 \end{aligned}$$

Then

$$\begin{aligned} (1-\theta_1)(u_{11}+u_{12}) &= \theta_1(u_{21}+u_{22}) \\ (1-\theta_2)(u_{11}+u_{21}) &= \theta_2(u_{12}+u_{22}), \end{aligned}$$

and

$$\theta_1 = \frac{u_{11}+u_{12}}{u_{++}} \quad \text{and} \quad \theta_2 = \frac{u_{11}+u_{21}}{u_{++}}.$$

Multivariate normal model

We see in this section that in case of multivariate normal random vectors, the score equations are rational in $\theta = (\mu, \Sigma) \in \mathbb{R}^m \times \mathrm{PD}_m$. The log-likelihood function is (see Proposition 5.17)

$$\begin{aligned} \ell_X(\mu,\Sigma) &= -\frac{N}{2}(m\log(2\pi) + \log(\det(\Sigma))) - \frac{1}{2}\sum_{i=1}^{N}(X^{(i)}-\mu)^T\Sigma^{-1}(X^{(i)}-\mu) \\ &= -\frac{N}{2}\left(m\log(2\pi) + \log\det(\Sigma) + \mathrm{tr}(\Sigma^{-1}S) + (\bar{X}-\mu)^T\Sigma^{-1}(\bar{X}-\mu)\right). \end{aligned}$$

We write

$$\begin{aligned}\sum_{i=1}^{N}(X^{(i)}-\mu)^T\Sigma^{-1}(X^{(i)}-\mu) &= \sum_{i=1}^{N}(X^{(i)}-\bar{X}+\bar{X}-\mu)^T\Sigma^{-1}(X^{(i)}-\bar{X}+\bar{X}-\mu)\\ &= \sum_{i=1}^{N}(X^{(i)}-\bar{X})^T\Sigma^{-1}(X^{(i)}-\bar{X})+(\bar{X}-\mu)^T\Sigma^{-1}(\bar{X}-\mu)\\ &= N\mathrm{tr}(S\Sigma^{-1})+\sum_{i=1}^{N}(\bar{X}-\mu)^T\Sigma^{-1}(\bar{X}-\mu)\\ &= N\left(\mathrm{tr}(S\Sigma^{-1})+(\bar{X}-\mu)^T\Sigma^{-1}(\bar{X}-\mu)\right).\end{aligned}$$

So, the log-likelihood function can be expressed in terms of the sufficient statistic $T(X) = (\bar{X}, S)$:

$$\ell_X(\mu,\Sigma) = -\frac{N}{2}(m\log(2\pi)+\log(\det(\Sigma)))-\frac{N}{2}\mathrm{tr}(S\Sigma^{-1})-\frac{N}{2}(\bar{X}-\mu)^T\Sigma^{-1}(\bar{X}-\mu).$$

Corollary 7.7 (Full gaussian model). The score equations of the full Gaussian model $\Theta = \mathbb{R}^m \times \mathrm{PD}_m$ are rational in μ and Σ; unique solutions are

$$\hat{\mu} = \bar{X} \quad \text{and} \quad \hat{\Sigma} = S;$$

hence the ML degree is one.

MLE reduces to a simpler problem if we fix $\Sigma = \mathrm{ID}_m$ and only estimate μ from some parameter space $\Theta_1 \subset \mathbb{R}^m$.

Proposition 7.8 (Restricted mean and identitiy covariance matrix). Let $X \in \mathbb{R}^m$ be a normal random vector. Assume we want to estimate the parameter of the model $\mathcal{M}_{\theta\in\Theta}$ for $\Theta = \Theta_1 \times \{\mathrm{ID}_m\}$ and $\Theta \subset \mathbb{R}^m$. Then the maximum likelihood is equivalent to the least squares method.

Proof. We need to estimate $\mu \in \Theta_1$; using maximum likelihood, we have

$$\hat{\mu} = \mathrm{argmax}_{\mu\in\Theta_1}(\bar{X}-\mu)^T(\bar{X}-\mu) = \mathrm{argmax}_{\mu\in\Theta_1}\|\bar{X}-\mu\|^2.$$

This means we search for the μ in Θ_1 that is closest to the sample mean. □

We define the **ED degree** (Euclidean distance degree) to be the number of critical points of the Euclidean distance $\|\bar{X}-\mu\|^2$ function for generic $\bar{X}$. Critical points are the solutions to the score equations $\frac{\partial}{\partial\mu}\|\bar{X}-\mu\|^2 = 0$.

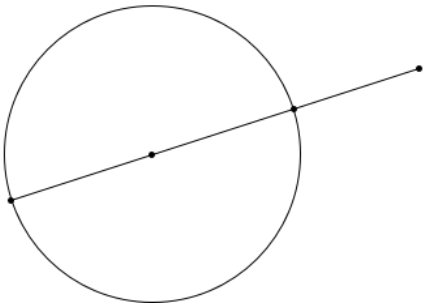
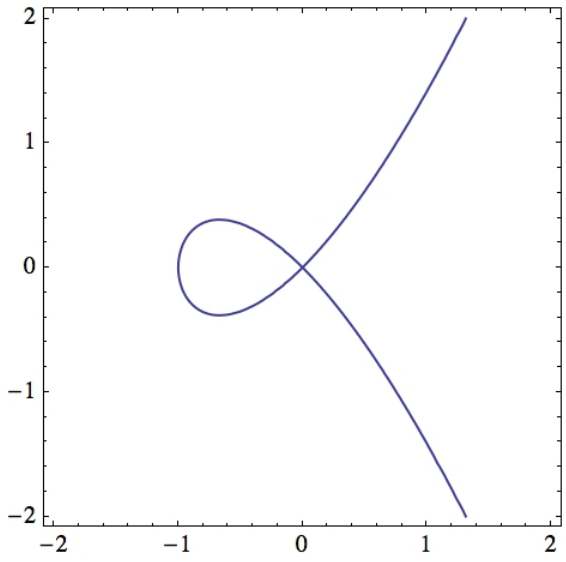


Figure 2: The ED degree of a line is one, of a circle two, of a parabola three and of an ellipsoid four.

Corollary 7.9 (ML Degree equals ED degree). The ML degree equals the ED degree for Gaussian models with unit covariance.

Corollary 7.10 (Rationality of score equations for Gaussian models with unit covariance). The score equations for Gaussian models with identity covariance matrix are rational in μ.

Example 7.11 (Nodal cubic). Assume $\Theta_1 = \left\{ \begin{bmatrix} t^2-1 \\ t(t^2-1) \end{bmatrix} \mid t \in \mathbb{R} \right\}$; this is a nodal cubic.



Assume the covariance is fixed to be the identity matrix. Then the ML degree equals the ED degree; find $\mu \in \mathbb{R}^2$ in Θ_1 such that it minimizes

$$\|\bar{X} - \mu\|^2 = (\bar{X}_1 - t^2 + 1)^2 + (\bar{X}_2 - t(t^2-1))^2.$$

Computing the derivative yields

$$\begin{aligned}&\frac{\partial}{\partial t}\left((\bar{X}_1 - t^2 + 1)^2 + (\bar{X}_2 - t(t^2-1))^2\right)\\&= 2(\bar{X}_1 - t^2 + 1)\cdot(-2t) + 2(\bar{X}_2 - t(t^2-1))\cdot(-3t^2+1).\end{aligned}$$

Setting this to zero, we get a polynomial score equation in t of degree 5. The ED degree equals 5.

Proposition 7.12 (Unrestricted mean and restricted covariance matrix). Let $\Theta = \mathbb{R}^m \times \Theta_1$, where $\Theta_1 \subset \mathrm{PD}_m$. Then the maximum likelihood estimate is

$$\hat{\mu} = \bar{X} \quad \text{and} \quad \hat{\Sigma} = \operatorname{argmin}_{\Sigma\in\Theta_1} \log\det(\Sigma) + \operatorname{tr}(\Sigma^{-1}S).$$

Alternatively, in terms of the precision matrix $K = \Sigma^{-1}$, we have

$$\hat{\mu} = \bar{X} \quad \text{and} \quad \hat{K} = \operatorname{argmax}_{K\in\Theta_1^{-1}} \log\det(K) - \operatorname{tr}(KS).$$

Proposition 7.13 (Centered model and restricted covariance matrix). Another special case is given if we have a centered model, that is $\Theta = \{0\} \times \Theta_1$. Then we have the same result as in the previous proposition but now with $S = \frac{1}{N}\sum_{i=1}^N x^{(i)}x^{(i)^T}$.

Example 7.14 (Full independence and centered model). Let $\mathbf{X} = (X_1, X_2, X_3)$ be a random normal vector. Assume Θ is centered and given by full independence $X_1 \perp\!\!\!\perp (X_2, X_3), X_2 \perp\!\!\!\perp (X_1, X_3)$ and $X_3 \perp\!\!\!\perp (X_1, X_2)$. Then

$$\Theta_1 = \left\{ \begin{bmatrix} \sigma_{11} & 0 & 0 \\ 0 & \sigma_{22} & 0 \\ 0 & 0 & \sigma_{33} \end{bmatrix} \right\} \subset \mathrm{PD}_3 \quad \text{and} \quad \Theta_1^{-1} = \left\{ \begin{bmatrix} \frac{1}{\sigma_{11}} & 0 & 0 \\ 0 & \frac{1}{\sigma_{22}} & 0 \\ 0 & 0 & \frac{1}{\sigma_{33}} \end{bmatrix} \right\} \subset \mathrm{PD}_3.$$

Further, assume that we are given the sample covariance matrix

$$S = \begin{bmatrix} s_{11} & s_{12} & s_{13} \\ s_{21} & s_{22} & s_{23} \\ s_{31} & s_{32} & s_{33} \end{bmatrix}.$$

Let $a = \sigma_{11}^{-1}$, $b = \sigma_{22}^{-1}$ and $c = \sigma_{33}^{-1}$. The likelihood function if K is

$$\log\det K - \operatorname{tr}(KS) = \log(abc) - (s_{11}a + s_{22}b + s_{33}c).$$

Take the derivative with respect to a, b and c and set to zero. We get that the ML degree is one and

$$\hat{\Sigma} = \begin{bmatrix} s_{11} & 0 & 0 \\ 0 & s_{22} & 0 \\ 0 & 0 & s_{33} \end{bmatrix}.$$

Example 7.15 (Bivariate centered model and restricted covariance matrix). Let $\mathbf{X} = (X_1, X_2)$ be a random normal vector. Assume it is centered and

$$\Theta_1 = \left\{ \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix} \mid \rho \in (-1, 1) \right\}.$$

The parameter ρ is the correlation between X_1 and X_2, i.e. $\rho = \mathrm{Corr}(X_1, X_2)$. Assume we have sample covariance matrix

$$S = \begin{bmatrix} s_{11} & s_{12} \\ s_{12} & s_{22} \end{bmatrix};$$

note that this matrix is symmetric. Also, we compute the precision matrix

$$\Sigma^{-1} = K = \frac{1}{1-\rho^2} \begin{bmatrix} 1 & -\rho \\ -\rho & 1 \end{bmatrix}$$

The log-likelihood is

$$\ell_{\mathbf{X}}(\theta) \propto \log\det\Sigma + \mathrm{tr}(\Sigma^{-1}S) = \log(1-\rho^2) + \frac{1}{1-\rho^2}(s_{11} + s_{22} - 2\rho s_{12}).$$

Take the derivative with respect to ρ and setting this to zero gives

$$\frac{\partial}{\partial\rho}\ell_{\mathbf{X}}(\theta) \propto \frac{-2\rho}{1-\rho^2} + \frac{-2s_{12}(1-\rho^2) - (s_{11} + s_{22} - 2\rho s_{12}) \cdot (-2\rho)}{(1-\rho^2)^2} = 0.$$

The score equation is

$$\rho^3 - s_{12}\rho^2 + (s_{11} + s_{22} - 1)\rho - s_{12} = 0.$$

The ML degree is three.

Example 7.16 (Ungeneric data). Continuing the last example, suppose we have "ungeneric" data S with $s_{12} = 0$. Then, the score equation becomes

$$\rho(\rho^2 + s_{11} + s_{22} - 1) = 0.$$

One MLE solution is $\hat{\rho} = 0$. If $s_{11} + s_{22} > 1$, then $\hat{\rho} = 0$ is the unique solution, and we have $\hat{\Sigma} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$. For the other case $s_{11} + s_{22} < 1$, we get two further solutions on top of $\hat{\rho} = 0$; this is the generic case.

Example 7.17. Assume a centered normal model and $\Theta_1 = \{\Sigma \in \mathrm{PD}_4 \mid \sigma_{12} = \sigma_{34} = 0\}$; this means that $X_1 \perp\!\!\!\perp X_2$ and $X_3 \perp\!\!\!\perp X_4$. We have

$$\Sigma = \begin{bmatrix} \sigma_{11} & 0 & \sigma_{13} & \sigma_{14} \\ 0 & \sigma_{22} & \sigma_{23} & \sigma_{24} \\ \sigma_{13} & \sigma_{23} & \sigma_{33} & 0 \\ \sigma_{14} & \sigma_{24} & 0 & \sigma_{44} \end{bmatrix}$$

Assume we observed this random (and generic) sample covariance matrix

$$S = \begin{bmatrix} 19 & 3 & 5 & 7 \\ 3 & 23 & 11 & 13 \\ 5 & 11 & 31 & -1 \\ 7 & 13 & -1 & 37 \end{bmatrix};$$

note that this matrix is positive definite. We want to use Macaulay2 to compute the critical solutions of

$$\log\det\Sigma + \operatorname{tr}(\Sigma^{-1}S).$$

To compute Σ^{-1} we use $\Sigma^{-1} = \frac{1}{\det\Sigma}\operatorname{adj}(\Sigma)$. If we compute the critical solutions of $\log\det\Sigma + \operatorname{tr}(\Sigma^{-1}S)$ with Macaulay2, we need to first turn the equations into a polynomial. Define $f = \det\Sigma$ and $L = \operatorname{tr}(\operatorname{adj}(\Sigma)S)$. Then the log-likelihood becomes

$$\ell = \log f + \frac{1}{f}L.$$

So the score equations are

$$\frac{1}{f}\frac{\partial f}{\partial\sigma_{ij}} + \frac{\frac{\partial L}{\sigma_{ij}}f - L\frac{\partial f}{\sigma_{ij}}}{f^2} = 0.$$

Multiply by f^2, and we get an ideal in the polynomial ring. Compute the vanishing set of the saturated ideal with respect to (f); we need the saturation to remove unwanted zeroes that were introduced by f. Then, we see that the ML degree is 17.

7.2 Discrete models with constraints

Let $\mathcal{M} \subset \Delta_{r-1}$ be a family of discrete models.

Definition 7.18 (Likelihood function in projective space). Let $V \subset \mathbb{P}^{r-1}$ be an irreducible projective variety over $\mathbb{C}$, and let $\mathbf{u} \in \mathbb{N}^r$ be a vector of counts. The likelihood function $L_{\mathbf{u}} : V \to \mathbb{C}$ is defined as

$$L_{\mathbf{u}}(\mathbf{p}) = \frac{p_1^{u_1}\cdots p_r^{u_r}}{(p_1+\cdots+p_r)^{u_1+\cdots+u_r}}.$$

Note that if $\mathbf{p} \in \Delta_{r-1}$, we recover the usual likelihood function. Moreover, $L_{\mathbf{u}}$ is a rational function in $\mathbf{p}$ of degree zero (the numerator is of degree u_+ and the denominator has also degree u_+).

Previously, we assigned the ML degree to statistical models; now we assign the ML degree to projective varieties.

Definition 7.19 (ML degree of projective varieties). The ML degree of V is the number of critical points $\mathbf{p}$ of $L_{\mathbf{u}}$ for generic $\mathbf{u}$ on $\mathbf{p} \in V_{\text{reg}} \setminus \mathcal{H}$ where $\mathcal{H} = V(p_1 ... p_r(p_1 + ... + p_r))$.

Since $\mathbf{p}$ is constrained, we use Lagrange multipliers to find the critical points of $L_{\mathbf{u}}$.

Remark 7.20 (Lagrange multipliers). Assume $V = (f_1, \ldots, f_k)$. These are our k constraints. We optimize the log of $L_{\mathbf{u}}$ plus the Lagrange multipliers:

$$\mathcal{L}_{\mathbf{u}}(\mathbf{p}, \lambda) = \left(\sum_{i=1}^{r} u_i \log(p_i) \right) - u_+ \log(p_+) + \sum_{j=1}^{k} \lambda_j f_j.$$

So, the score equations are

$$\begin{aligned} \frac{\partial \mathcal{L}_{\mathbf{u}}}{\partial p_i}(\mathbf{p}) &= \frac{u_i}{p_i} - \frac{u_+}{p_+} + \sum_{j=1}^{k} \lambda_j \frac{\partial f_j}{\partial p_i}(\mathbf{p}) = 0 \\ \frac{\partial \mathcal{L}_{\mathbf{u}}}{\partial \lambda_j}(\mathbf{p}) &= f_j(\mathbf{p}) = 0. \end{aligned}$$

Example 7.21 (Binomial model with two trials). Let X be a discrete random variable. Assume V is given by $p_1^2 - 4p_0p_2$. Given counts $\mathbf{u} = (u_0, u_1, u_2)$ find θ such that $\mathbf{p}_\theta \in V$ that maximizes $L_{\mathbf{u}}$. Note that the implicit equation characterizes the binomial model with two trials. So in the end, we should end up with the ML estimator for the binomial model. Let's see if this is indeed the case.

The Lagrange function is

$$\mathcal{L}_{\mathbf{u}}(\mathbf{p}, \lambda) = u_0 \log(p_0) + u_1 \log(p_1) + u_2 \log(p_2) - \lambda p_1^2 + \lambda 4 p_0 p_2.$$

We have

$$\begin{aligned} \frac{\partial}{\partial p_0} : \frac{u_0}{p_0} - \frac{u_+}{p_+} + 4\lambda p_2 = 0 \\ \cdots \end{aligned}$$

Solve for p_0, p_1, p_2 and λ in terms of u_0, u_1 and u_2 gives

$$\begin{bmatrix} \hat{p}_0 \\ \hat{p}_1 \\ \hat{p}_2 \end{bmatrix} = \begin{bmatrix} \frac{(2u_0+u_1)^2}{4(u_0+u_1+u_2)^2} \\ \frac{(2u_0+u_1)(u_1+2u_2)}{2(u_0+u_1+u_2)^2} \\ \frac{(u_1+2u_2)^2}{4(u_0+u_1+u_2)^2} \end{bmatrix}$$

This is consistent with $\text{Bin}(2, \theta)$ where

$$\hat{\theta} = \frac{u_1 + 2u_2}{2(u_0 + u_1 + u_2)}.$$

Note that if by chance the count $\mathbf{u}$ lies in $V(p_1^2 - 4p_0p_2)$ then $\lambda = 0$ and $\hat{\mathbf{p}} = \frac{1}{\mathbf{u}_+}(u_0, u_1, u_2)$.

Example 7.22 (Independence model). Assume the statistical model

$$\mathcal{M} = \left\{ \begin{bmatrix} p_{11} & p_{12} \\ p_{21} & p_{22} \end{bmatrix} \in \Delta_3 \mid p_{11}p_{22} - p_{12}p_{21} = 0 \right\}$$

What is this model? What is the MLE? We know that this is the independence model $X \perp\!\!\!\perp Y$ for $X, Y \in [2]$. The MLE is given by

$$\hat{p}_{ij} = \frac{u_{i+}u_{+j}}{u_{++}^2}.$$

If we had computed critical solutions to the Lagrange equation we would have obtained the exact same result as above.

June Huh proved in 2014 that if the ML degree of a variety is one, then $\hat{\mathbf{p}}$ is always a rational function in $\mathbf{u}$ of degree zero where numerator and denominator are products of linear form.

Definition 7.23 (Algebraic torus). The **algebraic torus** is $(\mathbb{C}^*)^r = (\mathbb{C} \setminus \{0\})^r$.

Definition 7.24 (Very affine variety). A **very affine variety** is a set of the form $V \cap (\mathbb{C}^*)^r$ where $V \subset \mathbb{C}^r$ is a variety.

Our statistical model live in the probability simplex; from an algebraic geometry point of view, we consider our model to live in a very affine variety.

Theorem 7.25 (June Huh, 2014). Let X be a discrete random variable with state space $\mathcal{X} = [r]$, and V a very affine variety. Let MLdegree$(V) = 1$. Then, there exists $\mathbf{h} \in \mathbb{C}^r$ and a matrix $\mathbf{B} \in \mathbb{Z}^{n \times r}$ where all columns sum to zero such that

$$\hat{p}_j(u_1, \ldots, u_r) = h_j \prod_{i=1}^{n} \left(\sum_{k=1}^{r} b_{ik}u_k \right)^{b_{ij}}$$

$(\mathbf{B}, \mathbf{h})$ is called a **Horn pair**.

Example 7.26 (Binomial model with two trials). Let $\mathbf{h} = \begin{bmatrix} 1 & 2 & 1 \end{bmatrix}$ and $\mathbf{B} = \begin{bmatrix} 2 & 1 & 0 \\ 0 & 1 & 2 \\ -2 & -2 & -2 \end{bmatrix}$. Then

$$\begin{aligned} \hat{p}_0 &= (2u_0 + 1u_1)^2(0u_0 + u_1 + 2u_2)^0(-2u_0 - 2u_1 - 2u_2)^{-2} \\ &= \frac{(2u_0 + u_1)^2}{(-2u_0 - 2u_1 - 2u_2)^2} \end{aligned}$$

Similarly,

$$\begin{aligned} \hat{p}_1 &= 2\left((2u_0 + u_1)(u_1 + 2u_2)(-2u_0 - 2u_1 - 2u_2)^{-2}\right) \\ \hat{p}_2 &= (2u_0 + u_1)^0(u_1 + 2u_2)^2(-2u_0 - 2u_1 - 2u_2)^{-2} \end{aligned}$$

Example 7.27 (Independence model). Assume $X \perp\!\!\!\perp Y$, $X, Y \in [2]$. We know $\hat{p}_{ij} = \frac{p_{i+}p_{+j}}{p_{++}^2}$. What is $\mathbf{h}$ and $\mathbf{B}$? We have

$$\begin{aligned}
\hat{p}_{11} &= \frac{(u_{11}+u_{12})(u_{11}+u_{21})}{u_{++}^2} \\
\hat{p}_{12} &= \frac{(u_{11}+u_{12})(u_{12}+u_{22})}{u_{++}^2} \\
\hat{p}_{21} &= \frac{(u_{21}+u_{22})(u_{11}+u_{21})}{u_{++}^2} \\
\hat{p}_{22} &= \frac{(u_{21}+u_{22})(u_{12}+u_{22})}{u_{++}^2}.
\end{aligned}$$

Then

$$\mathbf{h} = \begin{bmatrix} 4 & 4 & 4 & 4 \end{bmatrix} \quad \text{and} \quad \mathbf{B} = \begin{bmatrix} 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 \\ 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 \\ -2 & -2 & -2 & -2 \end{bmatrix}.$$

Note that $\mathbf{B}$ contains $\mathbf{A}$ as a submatrix, where $\mathbf{A}$ is the matrix from the log-linear model.

Theorem 7.28 (June Huh, Euler characteristic). Let $V \subset (\mathbb{C}^*)^r$ be a very affine variety of dimension d. If V is smooth (i.e. it contains no singular points), then $\text{MLdegree}(V) = (-1)^d \mathcal{X}(V)$, where $\mathcal{X}(V)$ is the Euler characteristic of V.

7.3 Log-affine linear models

We now consider the MLE of log-affine linear models.

Theorem 7.29 (Birch's Theorem). Let $\mathcal{M}$ be a *discrete* regular exponential family. Let $A \in \mathbb{Z}^{k \times r}$ such that $\mathbf{1} \in \text{rowspan}(A)$, $\mathbf{h} \in \mathbb{R}^r_{>0}$ and $\mathbf{u}$ vector of counts from N i.i.d. samples. Then the MLE, if it exists, is the unique solution to the equation

$$A\bar{\mathbf{u}} = A\mathbf{p}$$

with $\bar{\mathbf{u}} = \frac{1}{N}\mathbf{u}$ is the empirical distribution and $\mathbf{p} \in \mathcal{M}_{A,h}$. Note that if we do not restrict $\mathbf{p}$ to lie in the probability simplex, then there may exist multiple solutions;.

Proof. This follows from Corollary 6.10, which concerned regular, not necessarily discrete, exponential families. Note that $\mathbf{u}$ is always a sufficient statistic for a discrete model. $A\mathbf{u}$ is the minimal sufficient statistic for a log-linear model.

Let $\mathbf{X} = (X^{(1)}, \ldots, X^{(N)})$ be an i.i.d. sample with state space $\mathcal{X} = [r]^N$. Then,

$$T(\mathbf{X}) = \sum_{i=1}^{N} T(X^{(i)}).$$

Recall that for log-affine linear models, the columns $\mathbf{a}_{\cdot j}$ of A consists of $T(1), \ldots, T(r)$. So, we get

$$T(\mathbf{X}) = \sum_{i=1}^{r} u_i T(j) = \sum_{i=1}^{r} u_i \mathbf{a}_{\cdot i} = A\mathbf{u}.$$

Next, note that

$$\mathbb{E}[T(X)] = N \cdot \mathbb{E}[T(X^{(N)})] = N \sum_{i=1}^{r} p_i T(i) = NA\mathbf{p}.$$

Hence, $NA\mathbf{p} = \mathbb{E}[T(X)] = T(X) = A\mathbf{u}$. □

Let $A = \begin{bmatrix} | & \cdots & | \\ \mathbf{a}_{\cdot 1} & \cdots & \mathbf{a}_{\cdot r} \\ | & \cdots & | \end{bmatrix} \in \mathbb{Z}^{k \times r}$ such that $\mathbf{1} \in \text{rowspan}(A)$. Let $\mathbf{h} \in \mathbb{R}^r_{>0}$. We can define a lattice polytope by $P(A) = \text{conv}\{a_1, \ldots, a_r\} \subset \mathbb{R}^k$. A point p lies in the relative interior of a polytope $P = \text{conv}\{p_i\}$ if there exists a positive convex combination of the vertices that equals p; note that not every convex combination for p need be positive.

Theorem 7.30 (Existence of MLE). Let $\mathcal{M}_{A,\mathbf{h}}$ be a discrete exponential family. The MLE exists for $\mathcal{M}_{A,\mathbf{h}}$ if and only if $A\bar{\mathbf{u}} \in \text{relint}(P(A))$.

Proof. $\Longrightarrow$: Assume the MLE exists for $\mathcal{M}_{A,\mathbf{h}}$. By Birch's Theorem there exists a $\mathbf{p} \in \mathcal{M}_{A,\mathbf{h}}$ such that $A\mathbf{p} = A\bar{\mathbf{u}}$. Recall that $\mathbf{p} \in \mathcal{M}_{A,\mathbf{h}}$ means that $\mathbf{p}$ lies in the interior of the probability simplex. Hence, $A\bar{\mathbf{u}}$ lies in the relative interior of $P(A)$ since all $p_i > 0$ for $i = 1, \ldots, r$.

If $\mathbf{b} = A\bar{\mathbf{u}} \in \text{relint}(A)$, then there exists positive $\mathbf{v} \in \mathbb{R}^r_{>0}$ such that $N\mathbf{b} = A\mathbf{v}$. Observe that the likelihood $\prod p_i^{u_i}$ can be replaced by $\prod p_i^{v_i}$ since $A\mathbf{u} = A\mathbf{v}$. Next, note that on the boundary of $\mathbf{p} \in \mathcal{M}_{A,\mathbf{h}}$, the likelihood is zero because all exponents $v_i > 0$ and some $p_i = 0$. Note that in the interior of $\mathcal{M}_{A,\mathbf{h}}$ the likelihood is positive. Since the likelihood function of an exponential family is concave, a maximum is attained in the interior of $\mathcal{M}_{A,\mathbf{h}}$. □

There are only two cases: if $\bar{\mathbf{u}}$ lies on the boundary of $P(A)$, then the MLE does not exist. If $\bar{\mathbf{u}}$ lies in the relative interior of $P(A)$, then the MLE exists and is unique.

Corollary 7.31. If all the components of $\mathbf{u}$ are positive, then the MLE exists and is unique.

Example 7.32 (Independence model). Consider $\mathcal{M}_{X \perp\!\!\!\perp Y}$ with $X, Y \in [2]$. Does the MLE exist? For which data does it exist?

$\mathcal{M}_{X \perp\!\!\!\perp Y}$ is a log-linear model with

$$A = \begin{bmatrix} | & | & | & | \\ \mathbf{a}_{\cdot 1} & \mathbf{a}_{\cdot 2} & \mathbf{a}_{\cdot 3} & \mathbf{a}_{\cdot 4} \\ | & | & | & | \end{bmatrix} = \begin{bmatrix} 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 \\ 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 \end{bmatrix}.$$

The polytope $P(A) \subset \mathbb{R}^4$ lives in $\mathbb{R}^2$. Assume we are given some data $\mathbf{u} \in \mathbb{N}^4$. If all $u_i > 0$, then the MLE exists. If some $u_i = 0$, then it can go wrong.

- Let $\mathbf{u} = (2, 0, 0, 3)$. Then

$$A\bar{\mathbf{u}} = \frac{1}{5}(2 \cdot \mathbf{a}_{\cdot 1} + 3 \cdot \mathbf{a}_{\cdot 4}) = \frac{1}{5} \begin{bmatrix} 2 \\ 3 \\ 2 \\ 3 \end{bmatrix}$$

  The point $\frac{1}{5} \begin{bmatrix} 2 & 3 & 2 & 3 \end{bmatrix}^T$ lives in the relative interior of $P(A)$ since we can find a positive convex combination of vertices of A such that

$$\frac{1}{5} \begin{bmatrix} 2 \\ 3 \\ 2 \\ 3 \end{bmatrix} = A \begin{bmatrix} 1 \\ 1 \\ 1 \\ 2 \end{bmatrix}.$$

- Let $\mathbf{u} = (1, 0, 1, 0)$. Then $A\bar{\mathbf{u}} = \frac{1}{2}(1, 1, 2, 0)$. This lives on the boundary of the polyope; hence no MLE exists.

7.4 Gaussian linear concentration models

Let $L \subset \mathbb{R}^{\frac{m(m+1)}{2}}$ be a linear space of matrices such that $L \cap \mathrm{PD}_m \neq \emptyset$. Recall the Gaussian linear concentration model $\mathcal{M}_L$; this model consists of all normal distributions in $\mathcal{N}(\cdot, \Sigma)$ parametrized by $\Sigma \in L^{-1} = \left\{\Sigma \in \mathrm{PD}_m \mid \Sigma^{-1} \in L\right\}$. The likelihood is given by

$$\ell_K(S) = \log\det K - \mathrm{tr}(KS).$$

A minimal sufficient statistic is $T(X) = \pi_L(S)$, the orthogonal projection of the sample covariance matrix S onto the linear space L.

Theorem 7.33 (MLE estimator for Gaussian linear concentration models). Let $\Theta = \mathbb{R}^m \times L^{-1}$ be the parameter space of $\mathcal{M}_L$. Let $X^{(1)}, \ldots, X^{(N)} \in \mathbb{R}^m$ i.i.d. sample. Then the MLE for $(\mu, \Sigma) \in \Theta$, if it exists, is $(\bar{X}, \hat{\Sigma})$ where $\hat{\Sigma}$ is the unique solution to

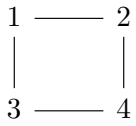
$$\pi_L(\hat{\Sigma}) = \pi_L(S), \quad \hat{\Sigma} \in L^{-1}.$$

Proposition 7.34 (Existence of MLE for Gaussian linear concentration models). The MLE exists for Gaussian linear concentration models if and only if $\pi_L^{-1}(S) \neq \emptyset$ if and only if $\pi_L(S) \in \text{relint}(\mathcal{C}_L)$; here $\mathcal{C}_L$ denotes the cone of matrices in L.

Example 7.35 (Gaussian graphical models). The linear space L can be given by a Gaussian graphical model G. It consists of matrices

$$L = \{(k_{ij})_{i,j} \mid k_{ij} = 0 \iff \text{there is no edge between } i \text{ and } j \text{ in } G\}\,.$$

For instance, consider the following Gaussian graphical model



There is no edge between $(2,3)$ and $(1,4)$; so L consists of matrices

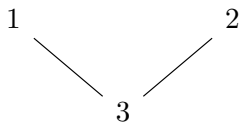
$$L = \left\{ \begin{bmatrix} * & * & * & 0 \\ * & * & 0 & * \\ * & 0 & * & * \\ 0 & * & * & * \end{bmatrix} \right\}.$$

Theorem 7.36 (MLE for Gaussian graphical models). Let G be an undirected graph. The MLE estimator $(\hat{\mu}, \hat{\Sigma})$ of the Gaussian graphical model $\mathcal{M}_G$ given sufficient statistic $T(X) = (\bar{X}, S)$ is

$$\bar{\mu} = \bar{X} \quad \text{and} \quad \hat{\Sigma}_{ij} = s_{ij} \text{ if } i \leftrightarrow j \in G \text{ or } i = j, \quad \hat{\Sigma}^{-1}_{ij} = 0 \text{ if } i \leftrightarrow j \notin G$$

provided the MLE exists. In words, whenever there is an edge between i and j in G, we set $\hat{\Sigma}_{ij} = s_{ij}$; when there is no edge, the (i,j)-th component of $\hat{\Sigma}$ is only implicitly given by the inverse.

Example 7.37. Consider the following Gaussian graphical model G:



The linear space is

$$L = \left\{ \begin{bmatrix} * & 0 & * \\ 0 & * & * \\ * & * & * \end{bmatrix} \right\}.$$

What is $\hat{\Sigma}$ of the graphical model $\mathcal{M}_G$? Assume we have data S. Then, we know by the above theorem that

$$\hat{\Sigma} = \begin{bmatrix} s_{11} & x & s_{13} \\ x & s_{22} & s_{23} \\ s_{13} & s_{23} & s_{33} \end{bmatrix}.$$

The entry $\hat{\Sigma}_{12}$ must be chosen in a way such that $\hat{\Sigma}_{12}^{-1} = 0$; using the formula $\hat{\Sigma}_{ij}^{-1} = (-1)^{i+j} \frac{1}{\det \hat{\Sigma}} \det(\hat{\Sigma}_{ji})$, we obtain

$$\hat{\Sigma}_{12}^{-1} = (-1)^{1+2} \frac{1}{\det \hat{\Sigma}} (x s_{33} - s_{13} s_{23}) = 0.$$

So $x = \frac{s_{13} s_{23}}{s_{33}}$.

8 Fisher's exact test

8.1 Asymptotic test

Question: Is the true generating distribution $\mathbf{p} \in \Delta_{rc-1}$ really part of $\mathcal{M}_{X \perp\!\!\!\perp Y}$?
We define

- H_0: $X \perp\!\!\!\perp Y$
- H_1: X and Y are no independent

Definition 8.1 (Chi-Square-statistic). It is defined by

$$\chi^2(\mathbf{u}) = \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{(u_{ij} - e_{ij})^2}{e_{ij}}$$

where $e_{ij} = \frac{u_{i+}u_{+j}}{u_{++}}$ is the expected count.

Remark 8.2. If X and Y were independent, then we would expect $(u_{ij} - e_{ij})^2$ to be small.

Decision rule: If $\chi^2(\mathbf{u})$ is small, we keep H_0. If $\chi^2(\mathbf{u})$ is large, we reject H_0.

Reasoning. Suppose now, $(u_{ij} - e_{ij})^2$ is large. What does that mean? Either, X and Y are independent, but our estimate is bad, or X and Y are dependent. Since the estimator is optimal, we infer that X and Y are dependent. □

Definition 8.3 (p-value). The ***p*-value** of the χ^2-test is the probability under H_0 to observe an equal or greater value of χ^2 than the observed one $\chi^2(\mathbf{u})$.

How do we compute this p-value? We can approximate it with the following theorem.

Theorem 8.4. Let $X \in [r]$ and $Y \in [c]$. Let $\mathbf{p} \in \mathcal{M}_{X \perp\!\!\!\perp Y} \subset \Delta_{rc-1}$ be the true joint distribution. If $\mathbf{u}_N$ is a contingency table with sample size N, then

$$\chi^2(\mathbf{u}_N) \overset{\text{distribution}}{\to} \chi^2_{(r-1)(c-1)}$$

as $N \to \infty$. We call the $\chi^2_{(r-1)(c-1)}$ the *Chi-squared distribution* with $(r-1)(c-1)$ degrees of freedom. It is the distribution of $\sum_{i=1}^{(r-1)(c-1)} z_i^2$ where all $z_i \sim \mathcal{N}(0,1)$ i.i.d. Note that this distribution does not depend on the true distribution $\mathbf{p}$.

Recall that convergence in distribution means

$$\lim_{N \to \infty} \mathbb{P}(\chi^2(\mathbf{u}_N) \geq t) = \mathbb{P}(\chi^2_{(r-1)(c-1)} \geq t) \quad \forall t.$$

Hence, to compute the p-value, that is to compute

$$\mathbb{P}(\chi^2(U) \geq \chi^2(\mathbf{u})),$$

we can instead compute the **asymptotic *p*-value**

$$\mathbb{P}(\chi^2_{(r-1)(c-1)} \geq \chi^2(\mathbf{u})).$$

Example 8.5 (Glasses and handedness). Assume $X, Y \in \{0, 1\}$ and

$$X \dots \text{wears glasses if } X = 1,$$
$$Y \dots \text{handedness if } Y = 1.$$

Let $N = 100$. Assume we have the following contingency table.

X Y	right	left	$\mathbf{u}_{\cdot +}$
glasses	43	9	52
no glasses	44	4	48
$\mathbf{u}_{+\cdot}$	87	13	100

We assert $H_0 : p \in \mathcal{M}_{X \perp\!\!\!\perp Y}$ and $H_1 : p \notin \mathcal{M}_{X \perp\!\!\!\perp Y}$. Can we reject H_0? Let's say our confidence interval is $1 - 0.05 = 95\%$.

We compute the expected counts.

$$\hat{p}_{11} = \frac{52 \cdot 87}{100 \cdot 100}$$
$$\hat{e}_{11} = \frac{52 \cdot 87}{100} = 45.24$$
$$\hat{e}_{12} = \frac{52 \cdot 13}{100} = 6.76$$
$$\hat{e}_{21} = \frac{48 \cdot 13}{100} = 41.76$$
$$\hat{e}_{22} = \frac{48 \cdot 13}{100} = 6.24$$

Then

$$\chi^2(\mathbf{u}) = \frac{(45.24 - 43)^2}{45.24} + \frac{(6.76 - 9)^2}{6.76} + \frac{(41.76 - 44)^2}{41.76} + \frac{(6.24 - 4)^2}{6.24} \approx 1.7774.$$

Is this a high or a low χ^2 value? To answer this we compute the asymptotic p-value

$$\mathbb{P}(\chi^2_1 \geq 1.7774) \approx 0.1825 > 0.05.$$

We do not reject H_0, in other words handedness and wearing glasses are independent.

Problem: Can we trust this asymptotic approximation? What if N is small?
The rescue is Fisher's exact test (1934).

8.2 Independence model

Let $X \in [r]$ and $Y \in [c]$. Assume $\mathbf{p} \in \Delta_{rc-1}$ is the joint distribution of X and Y. Consider a set of tables of size N

$$T(N) = \left\{ \mathbf{u} \in \mathbb{N}^{r \times c} \mid u_{++} = N \right\}.$$

Let $U \in T(N)$ be a random variable.

Question: What is the probability of observing a contingency table $\mathbf{u}$?

Proposition 8.6. It holds $U \sim \text{Multinomial}(N, \mathbf{p})$, i.e.

$$\mathbb{P}(U = \mathbf{u}) = \binom{N}{u_{11}u_{12}...u_{rc}} \prod_{i=1}^{r}\prod_{j=1}^{c} p_{ij}^{u_{ij}} \quad \forall \mathbf{u} \in T(N).$$

Here $\binom{N}{u_{11}u_{12}...u_{rc}} = \frac{N!}{u_{11}!u_{12}!...u_{rc}!}$.

Problem: This distribution depends on the true joint distribution $\mathbf{p}$.

Fisher's idea was to restrict to tables with the same row and column sums as the observed data $\mathbf{u}$. *What is the probability to observe* $\mathbf{u}$ *among all tables with the same row and column sum as* $\mathbf{u}$*?*

Definition 8.7 (Hypergeometric distribution). The **hypergeometric distribution** $X \sim \text{HypGeo}(N, K, n)$ is defined as

$$\mathbb{P}(X = k) = \frac{\binom{K}{k}\binom{N-K}{n-k}}{\binom{N}{n}}$$

for $\max\{0, n-(N-K)\} \leq k \leq \min\{K, n\}$.

We specialize to the case $r = c = 2$ before considering general r and c.

Proposition 8.8 (Conditional distribution of observing a contingengy table for the independence model). Let $r = c = 2$ and $\mathbf{p} \in \mathcal{M}_{X \perp\!\!\!\perp Y}$. Assume $\mathbf{u} \in T(N)$. Then

$$(U \mid U_{1+} = \mathbf{u}_{1+}, U_{+1} = \mathbf{u}_{+1}) \sim \text{HypGeo}(N, \mathbf{u}_{1+}, \mathbf{u}_{+1}).$$

That means

$$\mathbb{P}(U_{11} = u_{11} \mid U_{1+} = \mathbf{u}_{1+}, U_{+1} = \mathbf{u}_{1+}) = \frac{\binom{\mathbf{u}_{1+}}{u_{11}}\binom{N-\mathbf{u}_{1+}}{\mathbf{u}_{+1}-u_{11}}}{\binom{N}{\mathbf{u}_{+1}}}.$$

Remark 8.9. Note that $U_{11} = u_{11}$ given $U_{1+} = \mathbf{u}_{1+}, U_{+1} = \mathbf{u}_{1+}$ uniquely determines the contingency table since $\mathbf{u}_{2+} = N - \mathbf{u}_{1+}$ and $\mathbf{u}_{+2} = N - \mathbf{u}_{+1}$.

Remark 8.10. Note that this conditional distribution $\text{HypGeo}(N, \mathbf{u}_{1+}, \mathbf{u}_{+1})$ does not depend on $\mathbf{p}$.

Definition 8.11 (Conditional exact p-value of observing a contingency table). The **conditional exact p-value** is defined as

$$\mathbb{P}(\chi^2(U) \geq \chi^2(u) \mid U_{1+} = \mathbf{u}_{1+}, U_{+1} = \mathbf{u}_{+1}).$$

Example 8.12 (The tea test). Assume $X, Y \in \{\text{Milk}, \text{Tea}\}$ and

X ... ingredient added first (ground truth),
Y ... ingredient added first according to colleague.

X Y	Milk	Tea	$\mathbf{u}_{\cdot+}$
Milk	3	1	4
Tea	1	3	4
$\mathbf{u}_{+\cdot}$	4	4	8

As usual $H_0 : \mathbf{p} \in \mathcal{M}_{X \perp\!\!\!\perp Y}$. If H_0 is true, then the colleague is just guessing. Computation of the expected count yields $\mathbf{e} = \begin{bmatrix} 2 & 2 \\ 2 & 2 \end{bmatrix}$. Then,

$$\chi^2(\mathbf{u}) = \frac{(3-2)^2}{2} + \frac{(1-2)^2}{2} + \frac{(1-2)^2}{2} + \frac{(3-2)^2}{2} = 2.$$

Now we want to compute $\mathbb{P}(\chi^2(U) \geq 2 \mid U_{1+} = 4, U_{+1} = 4)$. For $u_{11} = 0$ we obtain

$$\mathbb{P}(U_{11} = 0 \mid U_{1+} = 4, U_{+1} = 4) = \frac{\binom{4}{0}\binom{4}{4}}{\binom{8}{4}} = \frac{1}{70} \approx 0.014.$$

Moreover,

$$\chi^2(\mathbf{u}) = \frac{(0-2)^2}{2} + \frac{(4-2)^2}{2} + \frac{(4-2)^2}{2} + \frac{(0-2)^2}{2} = 8.$$

Further computation of $u_{11} \in \{1, 2, 3, 4\}$ gives

u_{11}	$\mathbb{P}(U_{11} = u_{11})$	$\chi^2(\mathbf{u})$
0	0.014	8
1	0.229	2
2	0.514	0
3	0.229	2
4	0.014	8

Hence,

$$\mathbb{P}(\chi^2(U) \geq 2 \mid U_{1+} = 4, U_{+1} = 4) = 0.486.$$

We generalize to arbitrary $r, c \in \mathbb{N}$.

Definition 8.13 (Multivariate hypergeometric distribution). Let $\mathbf{X} = (X_i)_{i=1,\dots,k}$ be a random vector. Let $M_1, \dots, M_k, M, n \in \mathbb{N}$ such that $n \leq M = \sum_{i=1}^k M_i$. Then $\mathbf{X} \sim \text{MultiHypGeom}(M_1, \dots, M_k, n)$ if the probability mass function is

$$p_{\mathbf{X}}(\mathbf{x}) = p_{\mathbf{X}}(x_1, \dots, x_k) = \frac{\prod_{i=1}^k \binom{M_i}{x_i}}{\binom{M}{n}}, \quad \forall \mathbf{x} \in \mathbb{N}^k : \sum_{i=1}^k x_i = n.$$

M_i denotes the subpopulation of class i. The total population is $M = \sum_{i=1}^k M_i$.

Remark 8.14. Note that $\sum_{i=1}^k \binom{M_i}{x_i} = \binom{M}{n}$.

Proposition 8.15 (How many ways exist to obtain a column of a contingengy table?). Let $j \in [c]$. Let u_{+j} be fixed. The number of ways of obtaining columns with counts $\begin{bmatrix} u_{1j} & u_{2j} & \dots & u_{rj} \end{bmatrix}^T$ is

$$\binom{u_{+j}}{u_{1j} u_{2j} \dots u_{rj}}.$$

Moreover, the number of ways of ordering the class sizes is

$$\binom{N}{u_{1+} u_{2+} \dots u_{r+}}.$$

Theorem 8.16 (Conditional distribution of observing contingency table for the independence model (general case)). Let $\mathbf{p} \in \mathcal{M}_{X \perp\!\!\!\perp Y}$. Fix some data $\mathbf{u} \in T(N)$. Given marginal counts $u_{+\cdot}$ and $u_{\cdot +}$, it holds

$$\mathbb{P}(U = \mathbf{u} \mid U_{+\cdot} = u_{+\cdot}, U_{\cdot +} = u_{\cdot +}) = \frac{\binom{u_{+1}}{u_{11} u_{21} \dots u_{r1}} \cdots \binom{u_{+c}}{u_{1c} u_{2c} \dots u_{rc}}}{\binom{N}{u_{1+} \dots u_{r+}}}.$$

Using a more compact notation we write

$$\mathbb{P}(U = \mathbf{u} \mid U_{+\cdot} = u_{+\cdot}, U_{\cdot +} = u_{\cdot +}) = \frac{\binom{u_{+1}}{\mathbf{u}_{\cdot 1}} \cdots \binom{u_{+c}}{\mathbf{u}_{\cdot c}}}{\binom{N}{\mathbf{u}_{\cdot +}}}.$$

Definition 8.17 (Exact p-value). The exact p-value is defined as

$$\mathbb{P}(\chi^2(U) \geq \chi^2(\mathbf{u}) \mid U_{\cdot +} = u_{\cdot +}, U_{+\cdot} = u_{+\cdot}).$$

Proposition 8.18 (Explicit formula for the exact p-value). The explicit formula for the exact p-value is

$$\mathbb{P}(\chi^2(U) \geq \chi^2(\mathbf{u}) \mid U_{\cdot +} = u_{\cdot +}, U_{+\cdot} = u_{+\cdot}) = \frac{\sum_{\mathbf{v}} \left(\mathbb{1}_{\chi^2(\mathbf{v}) \geq \chi^2(\mathbf{u})} \cdot \frac{1}{\prod_{i=1}^{r} \prod_{j=1}^{c} v_{ij}!} \right)}{\sum_{\mathbf{v}} \frac{1}{\prod_{i=1}^{r} \prod_{j=1}^{c} v_{ij}!}},$$

where we sum $\mathbf{v}$ over $T(\mathbf{u}_{\cdot +}, \mathbf{u}_{+\cdot}) = \{\mathbf{w} \in T(N) \mid \mathbf{w}_{\cdot +} = \mathbf{u}_{\cdot +}, \mathbf{w}_{+\cdot} = \mathbf{u}_{+\cdot}\}$.

8.3 Log-affine linear models

We generalize to log-affine linear models $\mathcal{M}_{A,\mathbf{h}} \subset \Delta_{r-1}$. Assume we collect data $X^{(1)}, \ldots, X^{(n)} \in [r]$ i.i.d. according to some $\mathbf{p} \in \operatorname{int}(\mathcal{M}_{A,\mathbf{h}})$. We want to test

$$H_0 : \mathbf{p} \in \mathcal{M}_{A,\mathbf{h}} \quad \text{versus} \quad H_1 : \mathbf{p} \notin \mathcal{M}_{A,\mathbf{h}}.$$

If the true distribution is indeed in $\mathcal{M}_{A,\mathbf{h}}$, then we can write

$$\mathbf{p}_j = \frac{1}{Z(\theta)} h_j \theta^{\mathbf{a}_{\cdot j}}$$

for some positive vector $\mathbf{h} \in \mathbb{R}^r$, $\theta = (\theta_1, \ldots, \theta_k)$ and $A \in \mathbb{Z}^{k \times r}$. The probability of observing a contingency table $\mathbf{u}$ is given by

$$L(\mathbf{u} \mid \theta) = \frac{1}{Z(\theta)^n} \binom{n}{\mathbf{u}} \mathbf{h}^{\mathbf{u}} \theta^{A\mathbf{u}}.$$

We define the fiber of a table $\mathbf{u}$ to be the set of all tables $\mathbf{v}$ that have the same sufficient statistic as $\mathbf{u}$.

Definition 8.19 (Fiber of a contingency table). Let $\mathbf{u} \in \mathbb{N}^r$. The **fiber** is defined as $\mathcal{F}(\mathbf{u}) = \{\mathbf{v} \in \mathbb{N}^r \mid A\mathbf{v} = A\mathbf{u}\}$.

Proposition 8.20 (Conditional probability is independent on θ). We have $L(\mathbf{v} \mid \mathbf{v} \in \mathcal{F}(\mathbf{u}), \theta) = \mathbb{P}(U = \mathbf{v} \mid v \in \mathcal{F}(\mathbf{u}))$.

Proof. The key idea is that the conditional probability $L(\mathbf{v} \mid v \in \mathcal{F}(\mathbf{u}), \theta)$ does not depend on θ since we have

$$\begin{aligned} L(\mathbf{v} \mid \mathbf{v} \in \mathcal{F}(\mathbf{u}), \theta) &= \frac{Z(\theta)^{-n} \binom{n}{\mathbf{v}} \mathbf{h}^{\mathbf{v}} \theta^{A\mathbf{v}}}{\sum_{\mathbf{w} \in \mathcal{F}(\mathbf{u})} Z(\theta)^{-n} \binom{n}{\mathbf{w}} \mathbf{h}^{\mathbf{w}} \theta^{A\mathbf{w}}} \\ &= \frac{Z(\theta)^{-n} \binom{n}{\mathbf{v}} \mathbf{h}^{\mathbf{v}} \theta^{A\mathbf{v}}}{\theta^{A\mathbf{v}} Z(\theta)^{-n} \sum_{\mathbf{w} \in \mathcal{F}(\mathbf{u})} \binom{n}{\mathbf{w}} \mathbf{h}^{\mathbf{w}}} \\ &= \frac{\binom{n}{\mathbf{v}} \mathbf{h}^{\mathbf{v}}}{\sum_{\mathbf{w} \in \mathcal{F}(\mathbf{u})} \binom{n}{\mathbf{w}} \mathbf{h}^{\mathbf{w}}} \end{aligned}$$

Hence, we have $L(\mathbf{v} \mid \mathbf{v} \in \mathcal{F}(\mathbf{u}), \theta) = \mathbb{P}(U = \mathbf{v} \mid v \in \mathcal{F}(\mathbf{u}))$. □

Corollary 8.21. If we have a log-linear model, then we have

$$\mathbb{P}(U = \mathbf{v} \mid \mathbf{v} \in \mathcal{F}(\mathbf{u})) \propto \binom{N}{\mathbf{v}} \propto \frac{1}{\prod_{i=1}^{r} v_i!}.$$

Definition 8.22 (Exact conditional p-value). The exact conditional p-value is defined as

$$\frac{1}{\#\mathcal{F}(\mathbf{u})} \sum_{\mathbf{v} \in \mathcal{F}(\mathbf{u})} \mathbb{1}_{\chi^2(\mathbf{v}) \geq \chi^2(\mathbf{u})} \cdot \mathbb{P}(U = \mathbf{v} \mid \mathbf{v} \in \mathcal{F}(\mathbf{u})).$$

Problem: The fiber $\mathcal{F}(\mathbf{u})$ is too big to enumerate.
Solution: Approximate the p-value by generating N random samples $\mathbf{v_1}, \ldots, \mathbf{v}_n$ from the fiber $\mathcal{F}(\mathbf{u})$ and computing $\frac{1}{N} \sum_{\mathbf{v} \in \mathcal{F}(\mathbf{u})} \mathbb{1}_{\chi^2(\mathbf{v}) \geq \chi^2(\mathbf{u})} \cdot \mathbb{P}(U = \mathbf{v} \mid \mathbf{v} \in \mathcal{F}(\mathbf{u}))$.

How do we sample? We use Markov bases to sample from the fiber $\mathcal{F}(\mathbf{u})$.

8.4 Markov bases

Definition 8.23 (Markov basis). Let $A \in \mathbb{Z}^{k \times r}$ and let $\ker_{\mathbb{Z}}(A) = \{\mathbf{b} \in \mathbb{Z}^r \mid A\mathbf{b} = \mathbf{0}\}$. A finite set $B \subset \ker_{\mathbb{Z}}(A)$ is called a **Markov basis** if for all tables $\mathbf{u} \in \mathbb{N}^r$ and $\mathbf{v} \in \mathcal{F}(\mathbf{u})$ there exist $\mathbf{b}_1, \ldots, \mathbf{b}_s \in \pm B$ such that

- $\mathbf{v} = \mathbf{u} + \mathbf{b}_1 + \cdots + \mathbf{b}_s$ and,
- $\mathbf{u} + \mathbf{b}_1 + \cdots + \mathbf{b}_t \geq \mathbf{0}$ for all $t \leq s$.

The elements of B are called **moves**.

Remark 8.24 (Graphical interpretation). Fix $A \in \mathbb{Z}^{k \times r}$. For every table $\mathbf{u}$ we define a graph $G_{\mathbf{u}} = (V, E)$ in the following way: $V = \mathcal{F}(\mathbf{u})$ and $E = \{(\mathbf{v}, \mathbf{w}) \mid \mathbf{v} - \mathbf{w} \in B'\}$. Then B is a Markov basis if and only if $G_{\mathbf{u}}$ is connected for all $\mathbf{u} \in \mathbb{N}^r$.

Given a Markov basis, we can sample from the distribution of the fiber by the *Metropolis-Hastings* algorithm.

Theorem 8.25. The output sequence from the Metropolis-Hastings algorithm is an ergodic Markov chain on $\mathcal{F}(\mathbf{u})$ with stationary distribution $\mathbb{P}(U = \cdot \mid \cdot \in \mathcal{F}(\mathbf{u}))$.

How do we compute a Markov basis?
For log-affine linear models $\mathcal{M}_{A,\mathbf{h}}$ we can construct the ring homomorphism

$$\varphi : \mathbb{R}[p_1, \ldots, p_r] \to \mathbb{R}[\theta_1^{\pm}, \ldots, \theta_k^{\pm}], \quad p_i \mapsto h_i \theta^{\mathbf{a}_{\cdot i}},$$

whose kernel is the toric ideal $I_{A,\mathbf{h}}$. We can always substitute $p_i \mapsto \frac{p_i}{h_i}$ so that we can assume $\mathbf{h} = \mathbf{1}$ without loss of generality. Then, I_A is a binomial ideal of the form $I_A = (\mathbf{p}^{\alpha^+} - \mathbf{p}^{\alpha^-} \mid \alpha \in \ker_{\mathbb{Z}}(A))$.

Theorem 8.26 (Fundamental theorem of Markov bases (Diaconis, Sturmfels 1998)). A finite subset $B \subset \ker_{\mathbb{Z}}(A)$ is a Markov basis for $A \in \mathbb{Z}^{k\times r}$ if and only if $I_A = (\mathbf{p}^{\mathbf{b}^+} - \mathbf{p}^{\mathbf{b}^-} \mid \mathbf{b} \in B)$.

Proof. We define $I_B = (\mathbf{p}^{\mathbf{b}^+} - \mathbf{p}^{\mathbf{b}^-} \mid \mathbf{b} \in B)$.

$\Longrightarrow$: $\supset$ is easy. For $\subset$ let $\mathbf{p}^{\mathbf{u}} - \mathbf{p}^{\mathbf{v}} \in I_A$. By assumption there exist $b_1, \dots, b_s \in \pm B$ such that $v = u + \sum_{i=0}^{s} b_i$ and $u + \sum_{i=1}^{t} b_i \in \mathbb{N}^r$ for all $t \leq s$.

Induction on s. For $s = 1$ we have $v = u+b = u+b^+-b^-$. Then $v-b^+ = u-b^- \in \mathbb{N}^r$; it is in $\mathbb{N}^r$ because b^+ and b^- have disjoint support. Then $p^{v-b^+}(p^{b^+}-p^{b^-}) = p^u-p^v \in I_B$.

Assume $v = u+\sum_{i=0}^{s-1} b_i$ and $u+\sum_{i=1}^{t} b_i \in \mathbb{N}^r$ for all $t \leq s-1$. Then $u' = u+b_1 \in \mathbb{N}^r$; hence $p^v - p^{u'} \in I_B$. Moreover, $u' = u + b_1$; so $p^{u'} - p^u \in I_B$. Thus, $p^v - p^u \in I_B$.

$\Longleftarrow$: Assume $I_A = I_B$; hence for any $p^u - p^v$ we have $p^u - p^v = \sum_{l=1}^{s} p^{d_l}(p^{b_l^+} - p^{b_l^-})$. Induction on s.

For $s = 1$ we have $p^u - p^v = p^d(p^{b^+} - p^{b^-})$. Hence, $u = d + b^+$ and $v = d + b^-$. Then $u - b^+ = d = v - b^-$. Therefore, $v = u + b$.

Assume for $s-1$ the claim is true. We have $p^u - p^v = \sum_{l=1}^{s} p^{d_l}(p^{b_l^+} - p^{b_l^-})$. We see that for some $k \in [s]$ we must have $p^u = p^{d_k}p^{b_k^+}$; assume without loss of generality $k = 1$. Hence, $u = d_1 + b_1^+$ and $p^v + p^{d_1+b_1^-} = \sum_{l=2}^{s} p^{d_l}(p^{b_l^+} - p^{b_l^-})$. By induction claim we have that there exists a connected path between v and $d_1 + b_1^-$. Consider $u = d_1 + b_1^+ = d_1 + b_1^- + b_1^+ - b_1^- = v + \sum_j b'_j + b_1$. □

Example 8.27 (Independence model). Consider $\mathcal{M}_{X \perp\!\!\!\perp Y}$ with $X, Y \in [2]$. The sufficient statistic is $T(u) = T(u_{11}, u_{12}, u_{21}, u_{22}) = (u_{1+}, u_{2+}, u_{+1}, u_{+2})$; hence we have

$$A = \begin{bmatrix} 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 \\ 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 \end{bmatrix}.$$

We know from the independence model that $I_A = (p_{11}p_{22} - p_{12}p_{21}) = (\mathbf{p}^{(1,0,0,1)} - \mathbf{p}^{(0,1,1,0)})$. So a Markov basis is $B = \{(1, -1, -1, 1)\}$.

9 Review

- What is the saturation ideal?
- What is the marginal density?
- What is the conditional density?
- What is conditional independence?
- Pearl's conditional independence axioms
- What is the intersection axiom?
- Rank criterion for independence
- Algebraic criterion for conditional independence of discrete random vectors
- Conditional independence ideal of discrete random vectors
- Conditional independence model of discrete random vectors
- Rank criterion for conditional independence of Gaussian random vectors
- Algebraic criterion for conditional independence of Gaussian random vectors
- Conditional independence ideal of Gaussian random vectors
- Compute the conditional independence model of Gaussian random vectors for $\mathcal{C} = \{1 \perp\!\!\!\perp 3, 1 \perp\!\!\!\perp 3 \mid 2\}$. We should get $1 \perp\!\!\!\perp \{2, 3\}$ or $3 \perp\!\!\!\perp \{1, 2\}$.
- Prove that $(1 \perp\!\!\!\perp 2 \mid 3)$ and $(2 \perp\!\!\!\perp 3)$ imply $2 \perp\!\!\!\perp (1, 3)$.
- Prove the closure theorem
- How to compute the j-th elimination ideal? Use the Elimination Theorem, see Theorem 3.20.
- How do you compute the closed image of a variety under a projection? What is $\overline{\pi(V(I))}$?
- Implicitization of binomial model with two trials
- Implicitization of general binomial model
- What is a statistical model?
- What is a sufficient statistic?
- Factorization theorem
- Log likelihoo function of a multivariate normal distribution

- Sufficient statistic of the likelihood function of a multivariate normal distribution
- MLE of Gaussian model
- Exponential family definition
- Exponential form of binomial model
- Exponential form of univariate normal model
- Log-affine linear models
- What is a toric ideal?
- Toric ideals and binomial ideals
- Sufficient statistic and natural parameter of Gaussian model
- Inverse linear space
- Gaussian linear concentration model
- Concentration matrix and conditional independence
- Assume concentration matrix K has elements $k_{12} = k_{13} = 0$. What does this mean? What is the ideal?
- Maximum Likelihood degree definition
- Log-likelihood function of a discrete model
- Score equations of a discrete model
- Compute MLE for binomial model with r trials given samples $X^{(1)}, ..., X^{(n)}$
- What is the ML degree of the twisted cube?
- What is the ML degree of the binomial model with two trials?
- What is the ML degree of the independence model?
- What is the ML degree of the full Gaussian model?
- What is the ML degree when covariance is ID matrix and mean is from a nodal cubic $t^2 - 1, t(t^2 - 1)$?
- MLE of restricted covariance Gaussian model with centered / unrestricted mean
- MLE of full independence Gaussian model and centered
- Likelihood in projective space

- Likelihood degree of projective variety
- Lagrange multipliers
- Score equations with lagrange multipliers
- Example: discrete with $r = 3$ and $V = (p_1^2 - 4p_0p_2)$
- MLE of independence model 2x2
- MLE of binomial model (2 trials)?
- June Huh Theorem
- June Huh applied on binomial model with two trials
- June Huh and independence model
- Birchs Theorem
- When does the MLE exist for log affine linear models?
- Define the Gaussian linear concentration model
- MLE of Gaussian linear concentration model
- Gaussian graphical model
- MLE of graphical model
- Example: Compute the MLE of the Gaussian graphical model with $1 \perp\!\!\!\perp 2 \mid 3$
- Chi square statistic
- Asymptotic p-value, how does it work?
- Probability of observing a contingency table; what is the problem?
- Definition of hypergeometric distribution
- Conditional distribution of observing a contingency table for $r = c = 2$ and general case
- exact p-value
- explicit formula of the p-value
- Probability of observing a contingency table of a log-affine linear model
- What is a fiber?
- Conditional probability of observing a contingency table of a log-affine linear model + Independence Proof

- Exact conditional p-value of a log-affine linear model
- Markov basis
- Fundamental theorem of Markov basis
- Example: 2x2 independence model