Chapter 7 Statistical estimation

Last update on 2025-04-16 15:29:49+08:00

Table of contents

Parametric distribution estimation

Nonparametric distribution estimation

Optimal detector design

Chebyshev and Chernoff bounds

Experiment design

distribution estimation

estimate probability density $p(y)$ of a random variable from observed

parametric distribution estimation

choose from a family of densities $p_x(y)$ indexed by a parameter x

Maximum likelihood estimation

$$\text{maximize (over } x\text{)} \qquad \log p_x(y)$$

- y is observed value
- $l(x) = \log p_x(y)$ is called log-likelihood function
- this is maximum likelihood (ML) estimation
- can add constraints $x \in C$ explicitly or define $p_x(y) = 0$ for $x \notin C$
- convex optimization problem if $\log p_x(y)$ is concave in x for fixed y

Linear measurements with IID noise

linear measurement model

$$y_i = a_i^T x + v_i, \qquad i = 1, \ldots, m$$

- $x \in \mathbb{R}^n$ is vector of unknown parameters
- v_i is IID measurement noise, with density $p(x)$
- y_i is measurement: $y \in \mathbb{R}^m$ has density $p_x(y) = \prod_{i=1}^m p(y_i - a_i^T x)$

maximum likelihood estimate any solution x of

$$\text{maximize} \qquad l(x) = \sum_{i=1}^m \log p(y_i - a_i^T x)$$

examples

- Gaussian noise $\mathcal{N}(0, \sigma^2)$ with $p(z) = (2\pi\sigma^2)^{-1/2} e^{-z^2/(2\sigma^2)}$

$$l(x) = -\frac{m}{2} \log\left(2\pi\sigma^2\right) - \frac{1}{2\sigma^2} \sum_{i=1}^{m} \left(a_i^T x - y_i\right)^2$$

  ML estimate is LS solution

- Laplacian noise with $p(z) = (1/(2a)) e^{-|z|/a}$

$$l(x) = -m \log(2a) - \frac{1}{a} \sum_{i=1}^{m} \left|a_i^T x - y_i\right|$$

  ML estimate is ℓ_1-norm solution

- uniform noise on $[-a, a]$

$$l(x) = \begin{cases} -m \log(2a), & \left|a_i^T x - y_i\right| \leq a, \;\; i = 1, \ldots, m \\ -\infty, & \text{otherwise} \end{cases}$$

  ML estimate is any x with $\left|a_i^T x - y_i\right| \leq a$, i.e., ℓ_∞-norm solution with $\|Ax - y\|_\infty \leq a$

Logistic regression

random variable $y \in \{0, 1\}$ with distribution

$$p = \mathbf{prob}(y = 1) = \frac{e^{a^T u + b}}{1 + e^{a^T u + b}}$$

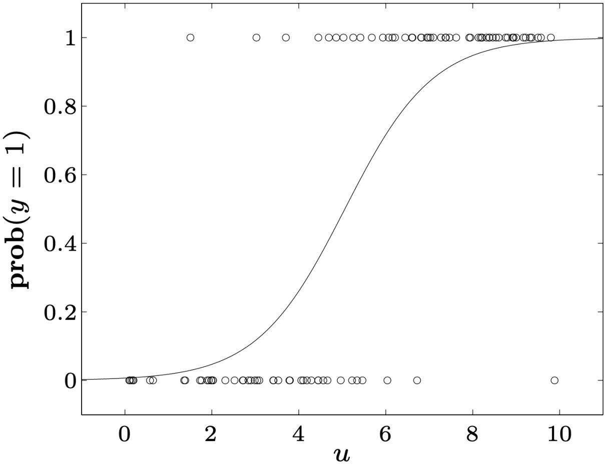
- a, b are parameters; $u \in \mathbb{R}^n$ are (observable) explanatory variables
- estimation problem: estimate a, b from m observations (u_i, y_i)

log-likelihood function for $y_1 = \cdots = y_k = 1$ and $y_{k+1} = \cdots = y_m = 0$

$$\begin{aligned} l(a, b) &= \log \left(\prod_{i=1}^{k} \frac{e^{a^T u_i + b}}{1 + e^{a^T u_i + b}} \prod_{i=k+1}^{m} \frac{1}{1 + e^{a^T u_i + b}} \right) \\ &= \sum_{i=1}^{k} (a^T u_i + b) - \sum_{i=1}^{m} \log \left(1 + e^{a^T u_i + b} \right) \end{aligned}$$

concave in a, b

example $(n = 1,\ m = 50$ measurements)



- circle shows 50 points (u_i, y_i)
- solid curve is ML estimate of $p = e^{au+b}/(1 + e^{au+b})$
- logistic regression is intrinsically a regression technique but can be used as classification

covariance estimation for Gaussian variable

Suppose $y \in \mathbb{R}^n$ is a Gaussian random variable with zero mean and covariance matrix $R = \mathbf{E}yy^T \in \mathbb{S}^n_{++}$,

$$p_R(y) = (2\pi)^{-n/2} \det(R)^{-1/2} \exp(-y^T R^{-1} y/2)$$

we estimate R based on N observables $y_1, \ldots, y_N \in \mathbb{R}^n$

let $Y = 1/N \sum_{k=1}^N y_k y_k^T$, then the log-likelihood function

$$l(R) = \log p_R(y_1, \ldots, y_N) = -(N/2) \log\det R - (N/2)\,\mathbf{tr}(R^{-1}Y) + C.$$

However, this log-likelihood function is not concave of R

Let $S = R^{-1}$ be the inverse of the covariance matrix, called information matrix or precision matrix. Then

$$l(S) = (N/2) \log\det S - (N/2)\,\mathbf{tr}(SY) + C.$$

is concave of S

the ML estimate of S is found by

$$\begin{array}{ll} \text{minimize} & \log\det S - \mathbf{tr}(SY) \\ \text{subject to} & S \in \mathcal{S} \subseteq \mathbb{S}^n_{++} \end{array}$$

- lower and upper matrix bounds $L \preceq R \preceq U$

$$U^{-1} \preceq R^{-1} \preceq L^{-1}$$

- a condition number constraint on R,

$$\lambda_{\max}(R) \leq \kappa_{\max}\lambda_{\min}(R)$$

can be expressed as

$$\lambda_{\max}(S) \leq \kappa_{\max}\lambda_{\min}(S)$$

maximum a posterior probability estimation

the condition density of x, given y, is given by Bayes' formula

$$p_{x|y}(x,y) = \frac{p(x,y)}{p_y(y)} = p_{y|x}(x,y)\frac{p_x(x)}{p_y(y)},$$

where prior density is $p_x(x)$

In the MAP estimation method, our estimate of x, given the observation y, is given by

$$\hat{x}_{\mathrm{map}} = \mathrm{argmax}_x p_{x|y}(x,y) = \mathrm{argmax}_x p_{y|x}(x,y)p_x(x) = \mathrm{argmax}_x p(x,y)$$

taking logarithms,

$$\hat{x}_{\mathrm{map}} = \mathrm{argmax}_x [\log p_{y|x}(x,y) + \log p_x(x)],$$

where the first term is the log-likelihood function and the second term penalizes choices of x

Parametric distribution estimation

Nonparametric distribution estimation

Optimal detector design

Chebyshev and Chernoff bounds

Experiment design

prior information

a random variable X in the finite set $\{\alpha_1, \dots, \alpha_n\} \subseteq \mathbb{R}$, the probability simplex is $\{p \in \mathbb{R}^n | p \succeq 0, \mathbf{1}^T p = 1\}$

- expectation $\mathbf{E}X = \sum_{i=1}^n \alpha_i p_i = \alpha$
- moment $\mathbf{E}X^2 = \sum_{i=1}^n \alpha_i^2 p_i = \beta$
- probability $\mathbf{prob}(X \geq 0) = \sum_{\alpha_i \geq 0} p_i \leq 0.3$
- the entropy of X, $-\sum_{i=1}^n p_i \log p_i$
- the Kullback-Leibler divergence $\sum_{i=1}^n p_i \log(p_i/q_i)$

bounding probabilities and expected values

$$\begin{array}{ll} \text{minimize} & \sum_{i=1}^{n} f(\alpha_i) p_i \\ \text{subject to} & p \in \mathcal{P} \end{array}$$

maximum likelihood estimation

$$\begin{array}{ll} \text{minimize} & \sum_{i=1}^{n} k_i \log p_i \\ \text{subject to} & p \in \mathcal{P} \end{array}$$

minimum KL divergence

$$\begin{array}{ll} \text{minimize} & \sum_{i=1}^{n} p_i \log(p_i/q_i) \\ \text{subject to} & p \in \mathcal{P} \end{array}$$

example (a probability distribution on 100 equidistance points in $[-1,1]$)

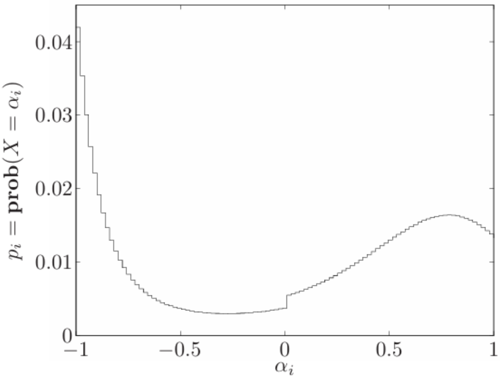


Figure 7.2 Maximum entropy distribution that satisfies the constraints (7.8).

Parametric distribution estimation

Nonparametric distribution estimation

Optimal detector design

Chebyshev and Chernoff bounds

Experiment design

Binary hypothesis testing

detection problem (hypothesis testing problem)

given observations of a random variable $X \in \{1, 2, \ldots, n\}$, choose between

- hypothesis 1: X was generated by distribution $p = (p_1, \ldots, p_n)$
- hypothesis 2: X was generated by distribution $q = (q_1, \ldots, q_n)$

deterministic and randomized detectors

- randomized detector: a nonnegative matrix $T \in \mathbb{R}^{2\times n}$ with $\mathbf{1}^T T = \mathbf{1}^T$
- deterministic detector: if all elements of T are 0 or 1
- if we observe $X = k$, we choose hypothesis 1 with probability t_{1k}, hypothesis 2 with probability t_{2k}

detection probability matrix

$$D = \begin{bmatrix} Tp & Tq \end{bmatrix} = \begin{bmatrix} 1 - P_{\mathsf{fp}} & P_{\mathsf{fn}} \\ P_{\mathsf{fp}} & 1 - P_{\mathsf{fn}} \end{bmatrix}$$

- P_{fp} is probability of selecting hypothesis 2 if X is generated by distribution 1 (false positive)
- P_{fn} is probability of selecting hypothesis 1 if X is generated by distribution 2 (false negative)

multicriterion formulation of detector design

$$
\begin{array}{ll}
\text{minimize (with respect to } \mathbb{R}^2_+) & (P_{\mathsf{fp}}, P_{\mathsf{fn}}) = ((Tp)_2, (Tq)_1) \\
\text{subject to} & t_{1k} + t_{2k} = 1, \qquad k = 1, \dots, n \\
& t_{ik} \geq 0, \qquad i = 1, 2, \qquad k = 1, \dots, n
\end{array}
$$

variables are entries of $T \in \mathbb{R}^{2 \times n}$

scalarization (with weight $\lambda > 0$)

$$\begin{array}{ll} \text{minimize} & (Tp)_2 + \lambda(Tq)_1 \\ \text{subject to} & t_{1k} + t_{2k} = 1, \qquad k = 1, \ldots, n \\ & t_{ik} \geq 0, \qquad i = 1, 2, \qquad k = 1, \ldots, n \end{array}$$

an LP with a simple analytical solution

$$(t_{1k}, t_{2k}) = \begin{cases} (1, 0) & p_k \geq \lambda q_k \\ (0, 1) & p_k < \lambda q_k \end{cases}$$

- a deterministic detector, given by a likelihood ratio test
- if $p_k = \lambda q_k$ for some k, any value $0 \leq t_{1k} \leq 1$, $t_{1k} = 1 - t_{2k}$ is optimal (i.e. Pareto-optimal detectors include non-deterministic detectors)

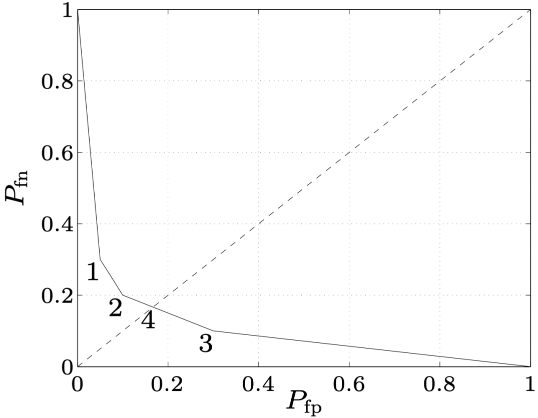
minimax detector

$$\begin{array}{ll} \text{minimize} & \mathbf{max}\{P_{\mathsf{fp}}, P_{\mathsf{fn}}\} = \mathbf{max}\{(Tp)_2, (Tq)_1\} \\ \text{subject to} & t_{1k} + t_{2k} = 1, \quad k = 1, \ldots, n \\ & t_{ik} \geq 0, \quad i = 1, 2, \quad k = 1, \ldots, n \end{array}$$

an LP; solution is usually not deterministic

example

$$P = \begin{bmatrix} 0.70 & 0.10 \\ 0.20 & 0.10 \\ 0.05 & 0.70 \\ 0.05 & 0.10 \end{bmatrix}$$



solutions 1,2,3 (and endpoints) are deterministic; 4 is minimax detector

Parametric distribution estimation

Nonparametric distribution estimation

Optimal detector design

Chebyshev and Chernoff bounds

Experiment design

We only discuss the Chebyshev bound

- If X is a random variable on $\mathbb{R}_+$ with $\mathbf{E}X = \mu$, then we have $\mathbf{prob}(X \geq 1) \leq \mu$, no matter what the distribution of X is.
- If X is a random variable on $\mathbb{R}$ with $\mathbf{E}X = \mu$ and $\mathbf{E}(X-\mu)^2 = \sigma^2$, then we have $\mathbf{prob}(|X-\mu| \geq a) \leq \sigma^2/a^2$, again no matter what the distribution of X is

The generalization

- Let X be random on $S \subseteq \mathbb{R}^m$, and $C \subseteq S$ be the set for which we want to bound $\mathbf{prob}(X \in C)$
- let $1_C(z) = 1$ if $z \in C$ and $1_C(z) = 0$ if $z \notin C$
- prior knowledge is known expected values of some functions

$$\mathbf{E}f_i(X) = a_i, \qquad i = 1, \ldots, n$$

- consider a linear combination of f_i,

$$f(z) = \sum_{i=1}^{n} x_i f_i(z),$$

  from which we have $\mathbf{E}f(X) = a^T x$

- suppose $f(z) \geq 1_C(z)$ for all $z \in S$, then we can upper bound $\mathbf{prob}(X \in C)$

$$a^T x = \mathbf{E} f(X) \geq \mathbf{E} 1_C(X) = \mathbf{prob}(X \in C)$$

- we search for the best such upper bound,

$$\begin{array}{ll} \text{minimize} & a_1 x_1 + \cdots + a_n x_n \\ \text{subject to} & f(z) = \sum_{i=1}^{n} x_i f_i(z) \geq 1 \text{ for } z \in C \\ & f(z) = \sum_{i=1}^{n} x_i f_i(z) \geq 0 \text{ for } z \in S,\ z \notin C \end{array}$$

- the formal dual

$$\begin{array}{ll} \text{maximize}_{\ p(z)} & \int_C p(z) dz \\ \text{subject to} & \int_S f_i(z) p(z) dz = a_i, \quad i = 1, \ldots, n \\ & \int_S p(z) dz = 1, \qquad p(z) \geq 0, \text{ for all } z \in S \end{array}$$

Parametric distribution estimation

Nonparametric distribution estimation

Optimal detector design

Chebyshev and Chernoff bounds

Experiment design

m linear measurements $y_i = a_i^T x + w_i$, $i = 1, \ldots, m$ of unknown $x \in \mathbb{R}^n$

- measurement errors w_i are IID $\mathcal{N}(0, 1)$
- ML (least-square) estimate is

$$\hat{x} = \left(\sum_{i=1}^{m} a_i a_i^T \right)^{-1} \sum_{i=1}^{m} y_i a_i$$

- error $e = \hat{x} - x$ has zero mean and covariance

$$E = \mathbf{E} e e^T = \left(\sum_{i=1}^{m} a_i a_i^T \right)^{-1}$$

  confidence ellipsoids are given by $\{x \mid (x - \hat{x})^T E^{-1} (x - \hat{x}) \leq \beta\}$

experiment design

choose $a_i \in \{v_1, \ldots, v_p\}$ (a set of possible test vectors) to make E 'small'

vector optimization formulation

$$\begin{array}{ll} \text{minimize (with respect to } \mathbb{S}^n_+) & E = \left(\sum_{k=1}^p m_k v_k v_k^T\right)^{-1} \\ \text{subject to} & m_1 + \cdots + m_p = m \\ & m_k \geq 0, \qquad m_k \in \mathbb{Z} \end{array}$$

- variables are m_k (number of vectors a_i which are equal to v_k)
- difficult in general, due to integer constraint

relaxed experiment design

assume $m \gg p$, use $\lambda = m_k/m$ as (continuous) real variable

$$\begin{array}{ll} \text{minimize (with respect to } \mathbb{S}^n_+) & E = \left(\sum_{k=1}^{p} m_k v_k v_k^T \right)^{-1} \\ \text{subject to} & m_1 + \cdots + m_p = m \\ & m_k \geq 0, \qquad m_k \in \mathbb{Z} \end{array}$$

ignoring the integer constraint, we arrive at

$$\begin{array}{ll} \text{minimize (with respect to } \mathbb{S}^n_+) & E = (1/m) \left(\sum_{k=1}^{p} \lambda_k v_k v_k^T \right)^{-1} \\ \text{subject to} & \lambda_1 + \cdots + \lambda_p = 1 \\ & \lambda_k \geq 0, \quad k = 1, \ldots, p \end{array}$$

- common scalarizations: minimize $\log\det E$, $\mathbf{tr}\, E$, $\lambda_{\text{max}}(E)$, ...
- can add other convex constraints, e.g. bound experiment cost $c^T\lambda \leq B$

D-optimal design

$$
\begin{array}{ll}
\text{minimize} & \log\det\left(\sum_{k=1}^{p} \lambda_k v_k v_k^T\right)^{-1} \\
\text{subject to} & \lambda \succeq 0 \\
& \mathbf{1}^T \lambda = 1
\end{array}
$$

interpretation: minimizes volume of confidence ellipsoids

dual problem

$$\begin{array}{ll} \text{maximize} & \log\det W + n\log n \\ \text{subject to} & v_k^T W v_k \leq 1, \qquad k = 1,\ldots,p \end{array}$$

interpretation: $\{x \mid x^T W x \leq 1\}$ is minimum volume ellipsoid centered at origin, that includes all test vectors v_k

complementary slackness for λ, W primal and dual optimal

$$\lambda_k \left(1 - v_k^T W v_k\right) = 0, \qquad k = 1,\ldots,p$$

optimal experiment uses vectors v_k on boundary of ellipsoid defined by W

computation reformulate primal problem with new variable X

$$\begin{array}{ll} \text{minimize} & \log\det X^{-1} \\ \text{subject to} & X = \sum_{k=1}^{p} \lambda_k v_k v_k^T, \qquad \lambda \succeq 0, \qquad \mathbf{1}^T\lambda = 1 \end{array}$$

$$L(X,\lambda,Z,z,\nu) = \log\det X^{-1} + \mathbf{tr}\left(Z\left(X - \sum_{k=1}^{p} \lambda_k v_k v_k^T\right)\right) - z^T\lambda + \nu\left(\mathbf{1}^T\lambda - 1\right)$$

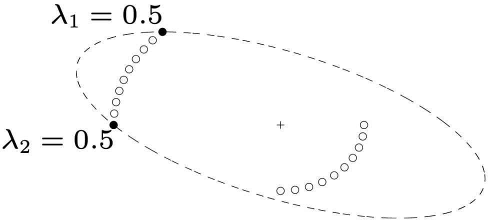
- minimize over X by setting gradient to zero to obtain $-X^{-1} + Z = 0$
- minimum over λ_k is $-\infty$ unless $-v_k^T Z v_k - z_k + \nu = 0$

dual problem

$$\begin{array}{ll} \text{maximize} & n + \log\det Z - \nu \\ \text{subject to} & v_k^T Z v_k \leq \nu, \qquad k = 1,\ldots,p \end{array}$$

change variable $W = Z/\nu$ and optimize over ν to get the above formulation

example $p = 20$



design uses two vectors, on boundary of ellipse defined by optimal W