Convex Optimization

Stephen Boyd   Lieven Vandenberghe

Revised slides by Stephen Boyd, Lieven Vandenberghe, and Parth Nobel

7. Statistical estimation

Outline

Maximum likelihood estimation

Hypothesis testing

Experiment design

Maximum likelihood estimation

- **parametric distribution estimation:** choose from a family of densities $p_x(y)$, indexed by a parameter x (often denoted θ)
- we take $p_x(y) = 0$ for invalid values of x
- $p_x(y)$, as a function of x, is called **likelihood function**
- $l(x) = \log p_x(y)$, as a function of x, is called **log-likelihood function**

- **maximum likelihood estimation (MLE):** choose x to maximize $p_x(y)$ (or $l(x)$)
- a convex optimization problem if $\log p_x(y)$ is concave in x for fixed y
- not the same as $\log p_x(y)$ concave in y for fixed x, *i.e.*, $p_x(y)$ is a family of log-concave densities

Linear measurements with IID noise

linear measurement model

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

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

maximum likelihood estimate: any solution x of

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

(y is observed value)

Examples

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

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

  ML estimate is least-squares solution

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

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

  ML estimate is ℓ_1-norm solution

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

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

  ML estimate is any x with $|a_i^T x - y_i| \leq a$

Logistic regression

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

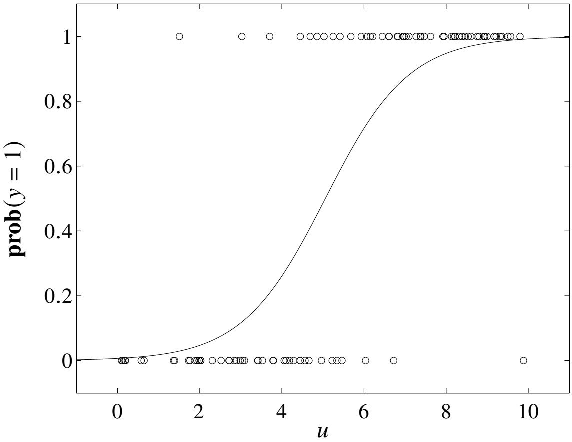
$$p = \mathbf{prob}(y = 1) = \frac{\exp(a^T u + b)}{1 + \exp(a^T u + b)}$$

- a, b are parameters; $u \in \mathbf{R}^n$ are (observable) explanatory variables
- estimation problem: estimate a, b from m observations (u_i, y_i)
- log-likelihood function (for $y_1 = \cdots = y_k = 1$, $y_{k+1} = \cdots = y_m = 0$):

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

concave in a, b

Example



- $n = 1$, $m = 50$ measurements; circles show points (u_i, y_i)
- solid curve is ML estimate of $p = \exp(au + b)/(1 + \exp(au + b))$

Gaussian covariance estimation

- fit Gaussian distribution $\mathcal{N}(0, \Sigma)$ to observed data $y_1, \ldots, y_N$
- log-likelihood is

$$
\begin{aligned}
l(\Sigma) &= \frac{1}{2} \sum_{k=1}^{N} \left(-2\pi n - \log\det \Sigma - y^T \Sigma^{-1} y \right) \\
&= \frac{N}{2} \left(-2\pi n - \log\det \Sigma - \mathbf{tr}\, \Sigma^{-1} Y \right)
\end{aligned}
$$

  with $Y = (1/N) \sum_{k=1}^{N} y_k y_k^T$, the empirical covariance
- l is **not** concave in Σ (the $\log\det \Sigma$ term has the wrong sign)
- with no constraints or regularization, MLE is empirical covariance $\Sigma^{\mathrm{ml}} = Y$

Change of variables

- change variables to $S = \Sigma^{-1}$
- recover original parameter via $\Sigma = S^{-1}$
- S is the **natural parameter** in an **exponential family** description of a Gaussian
- in terms of S, log-likelihood is

$$l(S) = \frac{N}{2}\left(-2\pi n + \log\det S - \mathbf{tr}\, SY\right)$$

  which is **concave**

- (a similar trick can be used to handle nonzero mean)

Fitting a sparse inverse covariance

- S is the **precision matrix** of the Gaussian
- $S_{ij} = 0$ means that y_i and y_j are independent, conditioned on y_k, $k \neq i, j$
- sparse S means
  - many pairs of components are conditionally independent, given the others
  - y is described by a sparse (Gaussian) Bayes network

- to fit data with S sparse, minimize convex function

$$-\log\det S + \mathbf{tr}\, SY + \lambda \sum_{i \neq j} |S_{ij}|$$

over $S \in \mathbf{S}^n$, with hyper-parameter $\lambda \geq 0$

Example

- example with $n = 4$, $N = 10$ samples generated from a sparse S^{true}

$$S^{\text{true}} = \begin{bmatrix} 1 & 0 & 0.5 & 0 \\ 0 & 1 & 0 & 0.1 \\ 0.5 & 0 & 1 & 0.3 \\ 0 & 0.1 & 0.3 & 1 \end{bmatrix}$$

- empirical and sparse estimate values of Σ^{-1} (with $\lambda = 0.2$)

$$Y^{-1} = \begin{bmatrix} 3 & 0.8 & 3.3 & 1.2 \\ 0.8 & 1.2 & 1.2 & 0.9 \\ 3.2 & 1.2 & 4.6 & 2.1 \\ 1.2 & 0.9 & 2.1 & 2.7 \end{bmatrix}, \qquad \hat{S} = \begin{bmatrix} 0.9 & 0 & 0.6 & 0 \\ 0 & 0.7 & 0 & 0.1 \\ 0.6 & 0 & 1.1 & 0.2 \\ 0 & 0.1 & 0.2 & 1.2 \end{bmatrix}.$$

- estimation errors: $\quad \|S^{\text{true}} - Y^{-1}\|_F^2 = 49.8, \quad \|S^{\text{true}} - \hat{S}\|_F^2 = 0.2$

Outline

Maximum likelihood estimation

Hypothesis testing

Experiment design

(Binary) hypothesis testing

detection (hypothesis testing) problem
given observation of a random variable $X \in \{1, \ldots, n\}$, choose between:

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

randomized detector

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

Detection probability matrix

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

- 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)

- **multi-objective formulation of detector design**

$$\begin{array}{ll} \text{minimize (w.r.t. } \mathbf{R}^2_+) & (P_{\mathrm{fp}}, P_{\mathrm{fn}}) = ((Tp)_2, (Tq)_1) \\ \text{subject to} & t_{1k} + t_{2k} = 1, \quad k = 1, \ldots, n \\ & t_{ik} \geq 0, \quad i = 1, 2, \quad k = 1, \ldots, n \end{array}$$

variable $T \in \mathbf{R}^{2 \times n}$

Scalarization

- scalarize with weight $\lambda > 0$ to obtain

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

- 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

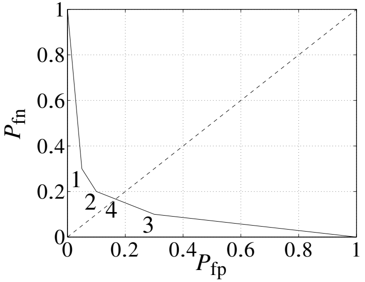
- minimize maximum of false positive and false negative probabilities

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

- an LP; solution is usually not deterministic

Example

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



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

Outline

Maximum likelihood estimation

Hypothesis testing

Experiment design

Experiment design

- m linear measurements $y_i = a_i^T x + w_i$, $i = 1, \ldots, m$ of unknown $x \in \mathbf{R}^n$
- measurement errors w_i are IID $\mathcal{N}(0, 1)$
- ML (least-squares) 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

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

- confidence ellipsoids are given by $\{x \mid (x - \hat{x})^T E^{-1} (x - \hat{x}) \leq \beta\}$
- **experiment design**: choose $a_i \in \{v_1, \ldots, v_p\}$ (set of possible test vectors) to make E 'small'

Vector optimization formulation

- formulate as vector optimization problem

$$
\begin{array}{ll}
\text{minimize (w.r.t. } \mathbf{S}_+^n) & E = \left(\sum_{k=1}^p m_k v_k v_k^T \right)^{-1} \\
\text{subject to} & m_k \geq 0, \quad m_1 + \cdots + m_p = m \\
& m_k \in \mathbf{Z}
\end{array}
$$

- variables are m_k, the number of vectors a_i equal to v_k
- difficult in general, due to integer constraint
- common scalarizations: minimize $\log\det E$, $\mathbf{tr}\, E$, $\lambda_{\max}(E)$, ...

Relaxed experiment design

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

$$\begin{array}{ll} \text{minimize (w.r.t. } \mathbf{S}^n_+) & E = (1/m)\left(\sum_{k=1}^p \lambda_k v_k v_k^T\right)^{-1} \\ \text{subject to} & \lambda \succeq 0, \quad \mathbf{1}^T\lambda = 1 \end{array}$$

- a convex relaxation, since we ignore constraint that $m\lambda_k \in \mathbf{Z}$
- optimal value is lower bound on optimal value of (integer) experiment design problem
- simple rounding of $\lambda_k m$ gives heuristic for experiment design problem

D-optimal design

- scalarize via log determinant

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

- interpretation: minimizes volume of confidence ellipsoids

Dual of D-optimal experiment design problem

dual problem

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

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

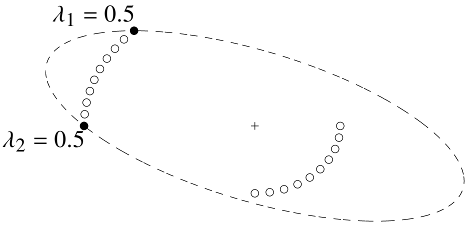
complementary slackness: for λ, W primal and dual optimal

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

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

Example

$(p = 20)$



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

Derivation of dual

first reformulate primal problem with new variable X:

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

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

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

dual problem

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

change variable $W = Z/\nu$, and optimize over ν to get dual of page 7.21