Statistical Estimation in Optimization

Andrea Bongco, Rachael Beckner,
Ixavier Higgins, & Emily Newman

WAKE FOREST
UNIVERSITY

Department of Mathematics

Winston Salem, NC

April 23, 2013

1. Parametric Distribution
   - Maximum Likelihood Estimation
   - Maximum A Posteriori Probability Estimation
2. Nonparametric Distribution
3. Optimal Detector Design and Hypothesis Testing
   - Deterministic and Randomized Detectors
   - Optimal Detector Design
   - Binary Hypothesis Testing
4. Experimental Designs
   - Relaxed Experiment Design
   - Scalarizations
5. Flower Exercise

Wake Forest University
Department of Mathematics

Maximum Likelihood Estimation (MLE)

In parametric distribution estimation, we consider a family of probability distributions on $\mathbb{R}^m$ indexed by a parameter vector $x \in \mathbb{R}^n$ with densities $p_x(\cdot)$. Take $y \in \mathbb{R}^n$ to be the set of observations. In Maximum Likelihood Estimation, we search for the parameter vector x which maximizes the likelihood of observing y. The log likelihood function is

$$\hat{x}_{mle} = \arg\max_x l(x) = \arg\max_x logp_x(y).$$

Maximum A Posteriori Probability Estimation (MAP)

An alternative to Maximum likelihood estimation is the Bayesian approach Maximum a posteriori probability estimation (MAP). Here prior information on the parameter vector, x, is implemented. Assume that x and y have a joint density $p(x, y) - x$ and y are considered random variables. A priori information is provided by the following marginal distributions.

$$p_x(x) = \int p(x, y)dy,$$

$$p_y(y) = \int p(x, y)dx.$$

and a posteriori information on the parameter vector can be described by the conditional distribution

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

MAP continued

The objective here is to maximize likelihood function.

$$\begin{aligned} \hat{x}_{map} &= \arg\max_x p_{x|y}(x,y) \\ &= \arg\max_x p_{y|x}(x,y)p_x(x) \\ &= \arg\max_x p(x,y). \end{aligned}$$

Notice that we are essentially maximizing the a similar function as in the MLE case. However, we have introduced a priori information into the computation.

Nonparametric Distribution

An alternate approach to statistical estimation are cases in which the family of distributions is unknown. Consider X with values in $\alpha_1, ..., \alpha_n \subseteq \mathbb{R}$. The distribution of X is characterized by $p \in \mathbb{R}^n$ where $\mathbf{prob}(\mathbf{X_k} = \alpha_\mathbf{k}) = \mathbf{p_k}$. (Note that p is semi positive definite and $1^T p = 1$).

Nonparametric Distribution continued

ML estimation can also be applied in the nonparametric setting given observations from that particular distribution. Let $X_1, ..., X_N$ be N independent samples from the distribution. Define k_i to be the number of observations with the value α_i such that $\sum_{i=1}^{n} k_i = N$. The log likelihood function is thus

$$l(p) = \sum_{i=1}^{n} k_i log(p_i).$$

Nonparametric Distribution continued

Typically, one uses prior information and use it as a means of finding eligible functions, p. For example, a set of function may have to satisfy the criteria that $E[X] = 2$, $var(X) = \alpha^2$, etc. Once the set of candidate distribution functions has been established, the next question is which function best fits the data. By solving the following convex function we can find the distribution p that has minimum Kullback-Leibler divergence from a given prior distribution q

$$\text{minimize } \sum_{i=1}^{n} p_i log\left(\frac{p_i}{q_i}\right), \text{ subject to } p \in P.$$

Deterministic and Randomized Detectors

Definition

estimator or *dectector* is given by the function $\psi : \{1, \ldots, n\} \to \{1, \ldots, m\}$ where an observed k is denoted $\psi(k) = \hat{\theta}$

Deterministic and Randomized Detectors continued

Definition

The *detection* probability matrix is defined as follows, $D = TP$ with elements given by $D_{ij} = (TP)_{ij} = \textbf{prob}(\hat{\theta} = i | \theta = j)$.

- Main diagonal of D gives the probability that $\hat{\theta} = i$ when $\theta = i$
- Off-diagonal represents the probability of mistaking $\theta = i$ for $\theta = j$
- When $D = I$ we call our detector perfect, so no matter our hypothesis we are correct

Optimal Detector Design

$$P^d = D_{ii} = \textbf{prob}(\hat{\theta} = i | \theta = i) \;\&\; P^e = 1 - D_{ii} = \textbf{prob}(\hat{\theta} \neq i | \theta = i)$$

The *minimax detector* goal is to minimize the maximum probability of error for a given value of θ.

- *Bias*: $\mathbf{E}_i(\hat{\theta} - \theta) = \sum_{j=1}^{m} (\theta_j - \theta_i) D_{ji}$.
- *Mean square error*: $\mathbf{E}_i(\hat{\theta} - \theta)^2 = \sum_{j=1}^{m} (\theta_j - \theta_i)^2 D_{ji}$.
- *Average absolute error*: $\mathbf{E}_i|\hat{\theta} - \theta| = \sum_{j=1}^{m} |\theta_j - \theta_i| D_{ji}$.

Binary Hypothesis Testing

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

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

Binary Hypothesis Testing continued

detection probability matrix:

$$D = [T_p \quad T_q] = \begin{bmatrix} 1 - P_{fp} & P_{fn} \\ P_{fp} & 1 - P_{fn} \end{bmatrix}$$

- 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)

Experimental Designs

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

- $y_i = a_i^T x + w_i, i = 1, ..., m$ where w_i is the noise resulting from measurements
- $\hat{x} = \left(\sum_{i=1}^{m} a_i a_i^T \right)^{-1} \sum_{i=1}^{m} y_i a_i$
- $e = \hat{x} - x$

Experimental Designs continued

- **Goal**: The goal of experiment design is to choose vectors a_i from the p possible test vectors $v_1, ..., v_p \in \mathbb{R}^n$ to minimize the covariance error E.
- **Set up**:

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

Relaxed Experiment Design

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

The optimal solution of this problem provides us with a lower bound to the experimental design problem.

Scalarizations

Frequently scalarizations are needed in experiment design. These include:

- *D-optimal design*: minimization of the determinate of the error covariance matrix E,
- *E-optimal design*: minimization of the norm of the error covariance matrix E,
- *A-optimal design*: minimization of the trace of the covariance matrix with an objective to mean of the norm of the error squared.

Flower Exercise