Statistical Estimation in Optimization

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

April 22, 2013

Introduction

In optimization theory, we are routinely concerned with trying to maximize (minimize) a function subject to a set of constraints. Statistical estimation of parameters is a perfect example of optimization theory in practice. As we shall see in the following sections, statistical estimation results in parameters maximizing the log-likelihood function.

Parametric distribution estimation

Statistical estimation of model parameters are usually handled in two fundamentally different manners. In the first situation, we assume that the distributional form is known. 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 $y \in \mathbb{R}^n$, then the log likelihood function is

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

Thus we are searching for the vector $\hat{x}$ that maximizes the likelihood of observing the fixed y (value, vector, or matrix). The maximum likelihood (ML) estimation problem is a convex optimization so long as the log-likelihood function $l(x)$ is concave for all y. As long as the set of constraints can be described by linear equality and convex inequality constraints, ML estimates can be computed via convex optimization.

The Bayesian version of maximum likelihood estimation is called Maximum a posteriori probability (MAP) estimation. The primary difference between this method and ML estimation is that prior information on the parameter vector, x, is implemented. We assume that x and y have a joint density $p(x, y) - x$ and y are considered random variables. A priori knowledge of x and y are given by the marginal density functions

$$\begin{aligned} p_x(x) &= \int p(x, y)dy, \\ p_y(y) &= \int p(x, y)dx. \end{aligned}$$

In the MAP estimates, the equation below represents our a posteriori estimate of the $\mathbf{x}$ when the observed values of y are input.

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

Thus, our convex optimization problem is to find the an estimate maximizing the conditional density of x given our data. We have

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

Fundamentally, ML estimator and MAP estimates are the same. However, the MAP estimates penalize for choosing $\mathbf{x}$ that are unlikely. Otherwise, we are maximizing essentially the same likelihood functions.

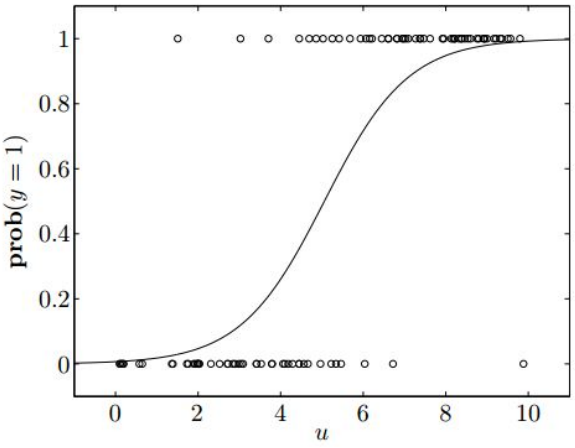
Assume we have $y = 0$ the event that a person does not have allergies and $y = 1$ the event that he/she does. The probabilities of the events are defined as $\mathbf{prob}(y = 1) = p$ and $\mathbf{prob}(y = 0) = 1 - p$. p is defined to depend on a vector of explanatory variables $u \in \mathbb{R}$. The variable u composes the explanatory variable, the number of times an individual sneezes in 1 hr when exposed to pollen. Note that u could be in $\mathbb{R}^n$, where u_i =(weight measurement, height, medical history,...) for person i.

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

Given some observations of the the explanatory variables and their outcomes, i.e. for u_i, $y_i = 0 or 1$. a and b in the function above are the model parameters we are attempting to estimate. The log-likelihood function to maximize is

$$\sum_{i=1}^{q} log p_i + \sum_{i=q+1}^{m} log(1 - p_i).$$

The figure below shows 50 points (u_i, y_i). Note that the curve shows $\mathbf{prob}(y = 1) = \frac{exp(au+b)}{1+exp(au+b)}$ for the ML parameters a,b. As we see, the probability that an individual has allergies $(y = 1)$ is greatest when $u \geq 8$ and is smallest when $u \leq 2$.



Nonparametric distribution estimation

Statistical estimation also can be accomplished by estimating distributions nonparametrically. 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 $1^T p = 1$.

Maximum likelihood estimation can also be applied in the nonparametric estimation of p based on 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

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

Thus, maximization of this function is equivalent to ML estimation as previous discussed. We will be finding the p that maximizes the likelihood of the observing data drawn from this unknown distribution.

Typically, one can implement prior information and use it as a means of finding eligible functions, p. 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

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

Optimal detector design and hypothesis

The goal is to determine the distribution that generates the sample X that has values in $\{1, \ldots, n\}$, where our distribution depends on our parameter $\theta \in \{1, \ldots, m\}$. This distribution of X can be represented by a matrix $P \in \mathbb{R}^{n \times m}$ where the elements are $p_{kj} = \mathbf{prob}(\mathbf{X} = \mathbf{k} | \theta = \mathbf{j})$. The different values for θ are our hypotheses that we will test.

Deterministic and randomized detectors

Our *estimator* or *dectector* is given by the function $\psi : \{1, \ldots, n\} \to \{1, \ldots, m\}$ where an observed k is denoted $\psi(k) = \hat{\theta}$. The *maximum likelihood detector* is represented by $\hat{\theta} = \psi_{ml}(k) = \arg\max_j p_{kj}$, so $\psi_{ml}(k)$ is the maximum value in the j^{th} column of our matrix P.

Our *randomized detector* is defined in terms of a matrix $T \in \mathbb{R}^{m \times n}$ with elements $t_{ik} = \mathbf{prob}(\hat{\theta} = i | X = k)$. This means that when $X = k, \hat{\theta} = i$ has probability t_{ik}. We need the matrix T to have each column vector be a unit vector, so $t_k \succeq 0$ and $1^T t_k = 1$. The detection probability matrix is defined as follows, $D = TP$ with elements given by $D_{ij} = (TP)_{ij} = \mathbf{prob}(\hat{\theta} = i | \theta = j)$. The main diagonal of D gives the probability that $\hat{\theta} = i$ when $\theta = i$, while the off-diagonal represents the probability of mistaking $\theta = i$ for $\theta = j$. Moreover, when $D = I$ we call our detector perfect, so no matter our hypothesis we are correct.

To obtain the *detection probabilities* we take the diagonal of D as a vector, denoted as P^d. When we take the compliment of the vector P^d we get the *error probabilities*, denoted P^e. Elements for P^d and P^e are given by the following equations:

$$P^d = D_{ii} = \mathbf{prob}(\hat{\theta} = i | \theta = i) \text{ and } P^e = 1 - D_{ii} = \mathbf{prob}(\hat{\theta} \neq i | \theta = i)$$

Optimal detector design

Some interesting values we want to look at are the bias, mean-square error, and average absolute error. The bias and mean-square error are equivalent to the expected value and variance of a basic random variable.

- *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*: $\mathbf{E}_i|\hat{\theta}-\theta| = \sum_{j=1}^m |\theta_j - \theta_i| D_{ji}$.

Now we want to find the minimum and maximum of our probabilities. The lower bound of our hypotheses will be given by $P_j^d = D_{jj} \geq L_j$. The maximum we can impose is the probability we will allow for mistaking $\theta = i$ for $\theta = j$. This is given by the following inequality $D_{ij} \leq U_{ij}$. Both of the inequalities associated with our bounds are linear constraints on T. The *minimax detector* goal is to minimize the maximum probability of error for a given value of θ.

Multicriterion formulation and scalarization

Lets examine Bayes detector design, where $q \in \mathbb{R}^m$ with elements $q_i = \mathbf{prob}(\theta = i)$. For this detector p_{ij} are conditional probabilities of X given θ then our goal is to minimize $q^T P^e$, where $q^T P^e = \sum_{j=1}^m q_j \sum_{i \neq j} D_{ij} = \sum_{i,j=1}^m W_{ij} D_{ij}$. The weighted matrix W is defined in this case as $W_{ij} = q_j$ when $i, j = 1, \ldots, m, i \neq j$ and $W_{ii} = 0$ for $i = 1, \ldots, m$. We finally get the optimal detector $\hat{\theta} = \arg\max_j (p_{kj} q_j)$ where $X = k$. This detector is called a maximum a posteriori probability (MAP) detector. We get maximum likelihood (ML) detector when $q = (1\backslash m)\mathbf{1}$ which minimizes the mean probability of error.

Binary hypothesis testing

When working with the special case of *binary hypothesis testing* we have $\theta = 1$, distribution p, denoting a normal even while $\theta = 2$, distribution q, is a abnormal event. For this case the detection probability matrix is a 2×2 matrix $D = \begin{bmatrix} 1 - P_{fp} & P_{fp} \\ f_{fp} & 1 - P_{fn} \end{bmatrix}$, where P_{fp} is the probability of a false positive and P_{fn} is the probability of a false negative. Now we can look at the *likelihood ratio threshold test* where the optimal detector is

$$\hat{\theta} = \begin{cases} 1 & W_{21} p_k > W_{12} q_k \\ 2 & W_{21} p_k \leq W_{12} q_k. \end{cases}$$

The test is negative if $\dfrac{p_k}{q_k} > \dfrac{W_{12}}{W_{21}}$ and positive otherwise.

Robust detectors

Instead of assuming we know all the possible distributions we will assume they are not known. Let $\mathcal{P}$ be the set of possible distributions, P, where the detection matrix D will depend on our distribution P. With the worse-case detection probaility matrix, D^{wc} defined as

$$D_{ij}^{wc} = \sup_{P \in \mathcal{P}} D_{ij} \text{ when } i \neq j \text{ and } D_{ii}^{wc} = \inf_{P \in \mathcal{P}} D_{ii} \text{ for } i, j \in \{1, \ldots, m\}.$$

The worse case probability error, $P_i^{wce} = 1 - D_{ii}^{wc}$, which is the maximum probability of error when $\theta = i$. Robust minimax detector is a good genearal example of this. In this case $\mathcal{P}$ is finite, and our goal is to minimize the worst-case probability of error.

$$\max_i P_i^{wce} = \max_{i=1,\ldots,m} \sup_{P \in \mathcal{P}} (1 - (TP)_{ii}) = 1 - \min_{i=1,\ldots,m} \inf_{P \in \mathcal{P}} (TP)_{ii}.$$

This is piecewise-linear and concave, which easily translates to be the polyhedron $\mathbf{conv}\mathcal{P}$ which has the same worse-case detection matrix and robust minimax detector.

Experiment design

Consider the problem of estimating the vector $x \in \mathbb{R}^n$ using measurements or experiments $y_i = a_i^T x + w_i, i = 1, ..., m$ where w_i is the noise resulting from measurements. The maximum likelihood estimate is denoted by $\hat{x} = \left(\sum_{i=1}^{m} a_i a_i^T\right)^{-1} \sum_{i=1}^{m} y_i a_i$. The error associated with this estimation is $e = \hat{x} - x$ with a zero mean and the covariance matrix

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

This matrix provides us with information on the accuracy of the estimation. The confidence level can be illustrated as an ellipsoid with the equation

$$\epsilon = \{z | (z - \hat{x})^T E^{-1} (z - \hat{x}) \leq \beta\}.$$

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. In other words, the objective is to find a set of m measurements or experiments from the p possible experiments to generate the maximum information. We denote m_j to be the number of vectors a_i equal to v_k and we can use this notation to show that the error covariance matrix E is only dependent on the number of experiments chosen.

To set up the basic experiment design problem we need choose the numbers of each type of experiment $m_1, ..., m_p \in \mathbb{Z}$ for the possible experiments $v_1, ..., v_p$ such that E is minimal. This gives us the vector optimization problem

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

over the positive semi definite cone.

The relaxed experiment design problem

When this problem gets difficult to solve, the relaxed experimental design given by

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

is used as a convex optimization problem. The optimal solution of this problem provides us with a lower bound to the experimental design problem.

Scalarizations

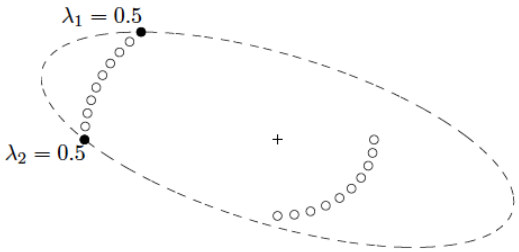
Frequently scalarizations are needed in experiment design. These include:

- *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.

The most commonly used scalarization is D-optimal design. Using this means that we minimize the volume of the confidence ellipsoid ignoring the contract factor $1/m$ in E and taking the logarithm of the function. This gives us the convex optimization problem modeled by

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

An example of D-optimal design can be represented graphically below.



Considering a problem with $X \in \mathbb{R}^2$ and $p = 20$. The measurement vectors a_i are represented as circles in the figure and the origin is marked with a cross. The ellipsoid is the minimum volume ellipsoid, that contains the points v_i.