STAT 603 : MONOTONE REGRESSION

1. The set-up

We are considering the nonparametric model

$$y_{in} = f_o(x_{in}) + d_{in} \ , \quad i = 1, 2, \cdots, n \ , \tag{1}$$

where the x_{in} are design points, lying in the interval $[\,0\,,\,1\,]$. We assume deterministic design points. (Alternatively, if the design is random, then we condition on the design.) For convenience, assume that $x_{in} < x_{i+1,n}$ for all i. The noise vector $d_n = (\, d_{1,n}, d_{2,n}, \cdots, d_{n,n}\,)^T$ follows a normal distribution

$$d_n \sim \mathcal{N}(\,0\,,\,\sigma^2 I\,) \ . \tag{2}$$

(Here, I is the $n \times n$ identity matrix.) Finally, we assume that the unknown function f_o is decreasing and positive. The positivity is really only for convenience. By decreasing we mean non-increasing.

The goal is to estimate f_o from the data $y_n = (\, y_{1,n}, y_{2,n}, \cdots, y_{n,n}\,)^T$. The normality of the data suggests that we estimate by least-squares, so

$$\begin{array}{ll} \text{minimize} & \sum_{i=1}^{n} |\, f(x_{in}) - y_{in}\,|^2 \\ \text{subject to} & f \text{ is decreasing} \ . \end{array} \tag{3}$$

It is clear (?) that the estimated values $f(x_{in})$ might be unique, but that in between, the estimator is quite arbitrary. By way of example, if

$$b_i = f(x_{in}) \ , \quad i = 1, 2, \cdots, n \ , \tag{4}$$

are given and f is decreasing then the choices

$$f(\,t\,) = b_i \, \frac{x_{i+1,n} - t}{x_{i+1,n} - x_{in}} + b_{i+1,n} \, \frac{t - x_{in}}{x_{i+1,n} - x_{in}} \ , \quad x_{in} \leqslant t \leqslant x_{i+1,n} \ , \tag{5}$$

and

$$f(\,t\,) = b_i \ , \quad x_{in} \leqslant t < x_{i+1,n} \ , \tag{6}$$

(note the (in)equality signs !) are both reasonable choices. Thus, we concentrate on estimating the b_i. In applied regression language, we are interested in estimating $\widehat{Y}$. The minimization problem then becomes

$$\begin{aligned} &\text{minimize} && \| \, b - y_n \, \|^2 \\ &\text{subject to} && b_1 \geqslant b_2 \geqslant \cdots \geqslant b_n \ . \end{aligned} \tag{7}$$

which we may clean up as (throwing in the factor $\frac{1}{2}$)

$$\begin{aligned} &\text{minimize} && \tfrac{1}{2} \, \| \, b - y_n \, \|^2 \\ &\text{subject to} && D\, b \leqslant 0 \text{ (component wise)} \, , \end{aligned} \tag{8}$$

where the difference matrix $D \in \mathbb{R}^{(n-1)\times n}$ is defined as

$$D = \begin{bmatrix} -1 & 1 & & & \\ & -1 & 1 & & \\ & & \ddots & \ddots & \\ & & & -1 & 1 \end{bmatrix} . \tag{9}$$

It is worthwhile noting that the design has disappeared form (7) and (8)-(9).

(10) Exercise. Show that the dual problem is

$$\begin{aligned} &\text{maximize} && - \tfrac{1}{2} \, \| \, D^T \lambda - y_n \, \|^2 + \tfrac{1}{2} \, \| \, y_n \, \|^2 \\ &\text{subject to} && \lambda \geqslant 0 \ , \end{aligned} \tag{11}$$

and if λ^* solves (11), then $b^* = y - D^T \lambda^*$ solves (8), and vice versa.

2. Least Concave Majorants

The primal and the dual problems (8) and (11) provide the generic way to the solution of the monotone regression problem. Thus, not much attention is paid to the precise form of the matrix D. However, a special theory of some importance is possible.

To investigate the possibilities, the Lagrange Multiplier Theorem tells us that b solves (8) if and only if there exists a $\lambda \in \mathbb{R}^{n-2}$ such that

$$\begin{aligned} & b - y + D^T \lambda = 0 \ , \\ & \lambda \geqslant 0 \ , \ \ Db \leqslant 0 \ , \ \ \langle \, \lambda \, , \, Db \, \rangle = 0 \ . \end{aligned} \tag{12}$$

Written component by component, the equation $b - y + D^T\lambda = 0$ reads

$$
\begin{aligned}
&b_1 - y_{1,n} - \lambda_1 = 0 \ , \\
&b_2 - y_{2,n} + \lambda_1 - \lambda_2 = 0 \ , \\
&\qquad\quad \vdots \\
&b_{n-1} - y_{n-1,n} - \lambda_{n-1} = 0 \ , \\
&b_n - y_{n,n} + \lambda_{n-1} = 0 \ .
\end{aligned} \tag{13}
$$

The idea of adding these equations suggests itself. Thus, define

$$
B_i = \sum_{j=1}^{i} b_j \ , \quad Y_i = \sum_{j=1}^{i} y_{jn} \ , \tag{14}
$$

and $B_0 = Y_0 = 0$. Then, by adding consecutive equations in (13), we get

$$
\begin{aligned}
&B_1 - Y_1 - \lambda_1 = 0 \ , \\
&B_2 - Y_2 - \lambda_2 = 0 \ , \\
&\qquad\quad \vdots \\
&B_{n-1} - Y_{n-1} - \lambda_{n-1} = 0 \ , \\
&B_n - Y_n = 0 \ .
\end{aligned} \tag{15}
$$

Since the λ_i are nonnegative, we conclude that

$$
\begin{aligned}
&B_0 = Y_0 \ , \quad B_n = Y_n \ , \\
&B_i \geqslant Y_i \ , \quad i = 1, 2, \cdots, n-1.
\end{aligned} \tag{16}
$$

COMMENT. Pretend for a moment that the b_i and y_{in} are nonnegative, and add up to 1. (In other words, we are talking about discrete probability distributions.) Then, (16) says that B majorizes Y. See the four B's, and MARSHALL and OLKIN (1979).

Obviously, the cumulative distribution B is increasing and concave, in the sense that

$$
B_{i-1} - 2\,B_i + B_{i-1} - 2 \leqslant 0 \ , \quad i = 1, 2, \cdots, n-1 \ , \tag{17}
$$

i.e., the graph obtained by connecting the dots (i, B_i), $i = 0, 1, 2, \cdots, n$, is concave.

To summarize, if b is the solution of the monotone regression problem, then B majorizes Y, and is concave.

We finish by showing that B is the smallest such function. So, let $C = (C_0, C_1, \cdots, C_n)^T$ satisfy

$$\begin{aligned} &C_0 = 0 , \quad C_n = Y_n , \\ &C_i \geqslant Y_i , \quad i = 1, 2, \cdots, n-1 . \end{aligned} \tag{18}$$

We are going to show that then

$$B_i \leqslant C_i , \quad i = 0, 1, 2, \cdots, n . \tag{19}$$

So, here it goes.

Obviously, if $B_i = Y_i$ for some i, then $B_i \leqslant C_i$. So, suppose that $B_i > Y_i$. In fact, assume that there are indices i and $i + k$ $(k \geqslant 1)$ with

$$\begin{aligned} &B_i = Y_i , \quad B_{i+k} = Y_{i+k} , \\ &B_j > Y_j , \quad j = i+1, \cdots, i+k-1 . \end{aligned} \tag{20}$$

From (15), we glean that

$$\begin{aligned} &\lambda_i = 0 , \quad \lambda_{i+k} = 0 , \\ &\lambda_j > 0 , \quad j = i+1, \cdots, i+k-1 . \end{aligned}$$

Now, the complementarity condition (12), $\langle \lambda , D b \rangle = 0$, actually says that

$$\lambda_j [Db]_j = 0 , \quad j = 1, 2, \cdots, n-1 .$$

Since $\lambda_j > 0$ for $j = i+1, \cdots, i+k-1$, we get

$$[Db]_j = 0 , \quad j = i+1, \cdots, i+k-1 ,$$

which translates to

$$b_{i+1} = b_{i+2} = \cdots = b_{i+k} .$$

Writing it in yet another way, we get

$$B_{j-1} - 2\,B_j + B_{j+1} = 0 \;, \quad j = i, i+1, \cdots, i+k-1 \;.$$

The picture (!) that emerges is that the points

$$(\, j \,,\, B_j \,) \;, \quad j = i+1, \cdots, i+k-1 \;,$$

lie on the straight line connecting the points $(\, i \,,\, Y_i \,)$ and $(\, i+k \,,\, Y_{i+k} \,)$.

Now, $C_i \geqslant Y_i = B_i$ and $C_{i+k} \geqslant Y_{i+k} = B_{i+k}$, and C is concave, so the points $(\, j \,,\, C_j \,)$, $j = i+1, \cdots, i+k$ must lie above the line connecting the points $(\, i \,,\, C_i \,)$ and $(\, i+k \,,\, C_{i+k} \,)$. which obviously lies above the line connecting the points $(\, i \,,\, B_i \,)$ and $(\, i+k \,,\, B_{i+k} \,)$. (Draw the digram.) This says that

$$B_j \leqslant C_j \;, \quad j = i, i+1, \cdots, i+k \;. \tag{22}$$

So, the optimality of B (or, rather, b) and (18)-(20) imply (22).

Thus, (19) follows and

$$B \text{ is the Least Concave Majorant of } Y \text{ !!!!} \tag{23}$$

3. The pool-adjacent-violators algorithm gets you closer

The pool-adjacent-violators algorithm for computing the solution of (7) starts with the initial guess $b = y_n$, and introduces weights, which initially consists of the vector of all ones $w = \mathbb{1}$. The interpretation is that the estimator f is given by

$$f(\, t \,) = b_i \;, \quad W_{i-1} < \, t \, \leqslant W_i \;, \quad i = 1, 2, \cdots, n \;. \tag{24}$$

with W the cumulative distribution of the w. Thus, the above interval has length w_i.

The basic step of the algorithm is to locate a violation of the monotonicity constraints, i.e., it finds an index i such that

$$b_i < b_{i+1} \;, \tag{25}$$

and fixes it by averaging. The new b_i and the new weight are

$$\begin{aligned} \beta &= \frac{w_i\, b_i + w_{i+1}\, b_{i+1}}{w_i + w_{i+1}} \;, \\ \varpi &= w_i + w_{i+1} \;, \end{aligned} \tag{26}$$

and the two adjacent intervals are combined into one. Thus, in (24), the number of intervals is diminished by 1.

It is obvious that locating violators can succeed at most $n-1$ times. Inspection of the previous section reveals that this is indeed a way to compute the least concave majorant of Y. (Well, now. It does take some drawing of diagrams.)

What we wish to show is that if b solves (7), and f_o is indeed decreasing, then

$$\sum_{i=1}^{n} |\, b_i - f_o(x_{in})\,|^2 \leqslant \sum_{i=1}^{n} |\, y_{in} - f_o(x_{in})\,|^2 \;. \tag{27}$$

In fact, it holds for any decreasing function φ,

$$\sum_{i=1}^{n} |\, b_i - \varphi(x_{in})\,|^2 \leqslant \sum_{i=1}^{n} |\, y_{in} - \varphi(x_{in})\,|^2 \;. \tag{27$'$}$$

(Note that in (27) there is some "connection" between the data y_{in} and f_o. In (27$'$), most definitely there is not.)

Comment. The result (27) might be nice mathematics and fit into the theme of this course (it is and it does!), but shirley, it has no statistical significance!? Well, yes. However, you can replace the data y_n by any estimator $\widehat{Y}$, e.g., a smoothing spline or polynomial estimator. And then it says that we do better by making $\widehat{Y}$ monotone by computing the least concave majorant of its cumulative sums.

To prove (27$'$), it suffices to take any decreasing function φ and prove that for the basic step (26),

$$\int_{W_{i-1}}^{W_{i+1}} |\, \beta - \varphi(\,t\,)\,|^2\, dt \leqslant J_i(\,b_i, b_{i+1}) \;, \tag{28}$$

where

$$J_i(p,q) \stackrel{\text{def}}{=} \int_{W_{i-1}}^{W_i} |\, p - \varphi(t)\,|^2\, dt + \int_{W_i}^{W_{i+1}} |\, q - \varphi(t)\,|^2\, dt \ . \tag{29}$$

Note that $J_i(p,q)$ is a quadratic function of the real variables p and q.

To prove (28) consider the minimization problem

$$\begin{aligned} &\text{minimize} && J_i(p,q) \\ &\text{subject to} && w_i\, p + w_{i+1}\, q = w_i\, b_i + w_{i+1}\, b_{i+1} \ , \\ & && p \leqslant q \ . \end{aligned} \tag{30}$$

There are now two claims: If $(p,q) = (p^*, q^*)$ is the (unique!) solution to (30), then

$$J_i(p^*, q^*) \leqslant J_i(\, b_i, b_{i+1}) \ , \tag{31}$$

and

$$p^* = q^* = \frac{w_i\, b_i + w_{i+1}\, b_{i+1}}{w_i + w_{i+1}} \ , \tag{32}$$

The first claim says that (p^*, q^*) gives a smaller value of the objective function than $(\, b_i, b_{i+1})$, which is (another) feasible point for the minimization problem (30). The second claim is the crux of the matter. To prove it, we reason by way of contradiction: Suppose $p^* < q^*$. In fact, let

$$q^* - p^* = \delta > 0 \ .$$

Now, consider the points (p,q),

$$p = p^* + x\, w_{i+1} \ , \quad q = q^* - x\, w_i \quad \text{for} \quad x \leqslant \delta / (\, w_i + w_{i+1}\,) \ .$$

For these x, then

$$w_i\, p + w_{i+1}\, q = w_i\, p^* + w_{i+1}\, q^* = w_i\, b_i + w_{i+1}\, b_{i+1} \ ,$$

and

$$p - q = p^* - q^* + x\,(\, w_i + w_{i+1}\,) \leqslant p^* - q^* + \delta = 0 \ ,$$

so that these points (p, q) are feasible for the problem (30).

Now, consider

$$f(x) = J_i(\, p^* + x\, w_{i+1}\, ,\, q^* - x\, w_i\,)\ , \quad x \leqslant \delta/(\, w_i + w_{i+1}\,)\ .$$

Then, f is minimal at $x = 0$, and since that is an interior point of the domain of definition of f, then $f\,'(0) = 0$. This in fact constitutes a contradiction, as we now show.

First, we need to compute the partial derivatives of $J_i(p, q)$. To compute these, rewrite

$$\begin{aligned}\int_{W_{i-1}}^{W_i} |\, p - \varphi(\,t\,)\,|^2\, d\,t\ &= w_i\, p^2 - 2\, w_i\, p\, \Phi_i + \int_{W_{i-1}}^{W_i} |\, \varphi(\,t\,)\,|^2\, d\,t \\ &= w_i\,(\, p - \Phi_i\,)^2 + \Bigl\{\, \text{terms not depending on } p\,\Bigr\}\ .\end{aligned}$$

where

$$\Phi_i = \frac{1}{w_i} \int_{W_{i-1}}^{W_i} \varphi(\,t\,)\, d\,t\ \ , \quad i = 1, 2, \cdots, n\ .$$

Thus, Φ_i is the mean of φ over the interval $(\, W_{i-1}, W_i\,)$. Note that for decreasing φ,

$$\varphi(W_{i-1}) \geqslant \Phi_i \geqslant \varphi(W_i)\ . \tag{33}$$

(Draw the diagram!)

It follows that

$$\frac{\partial}{\partial p}\, J_i(p, q) = 2\, w_i\,(\, p - \Phi_i\,)\ , \quad \frac{\partial}{\partial q}\, J_i(p, q) = 2\, w_{i+1}\,(\, q - \Phi_{i+1}\,)\ .$$

Consequently, since

$$f\,'(x) = w_{i+1}\, \frac{\partial}{\partial p}\, J_i(p, q) - w_i\, \frac{\partial}{\partial q}\, J_i(p, q)\ ,$$

we have

$$\begin{aligned}f\,'(0) &= 2\, w_i\, w_{i+1}\,(\, p^* - \Phi_i\,) - 2\, w_i\, w_{i+1}\,\bigl(\, q^* - \Phi_{i+1}\,\bigr) \\ &= 2\, w_i\, w_{i+1}\,(\, p^* - \Phi_i - q^* - \Phi_{i+1}\,)\ .\end{aligned}$$

Now, $f\,'(0) = 0$ implies that

$$p^* - q^* = \Phi_i - \Phi_{i+1} \ ,$$

and this is the advertised contradiction : The left hand side is strictly negative by assumption, but the right hand side is positive (nonnegative, actually). In detail, since φ is decreasing, by (33),

$$\Phi_i \geqslant \varphi(W_i\,) \geqslant \Phi_{i+1} \ .$$

The conclusion is that the assumption $p^* < q^*$ is false, so that $p^* = q^*$. Of course, if $p^* = q^*$, then the second part of the claim (32) follows.

This concludes the proof of (27).

References

Marshall, A.W., Olkin, I. (1979), *Inequalities: Theory of majorization and its applications*, Academic Press, New York.

Barlow, R.E., Bartholomew, D.J., Bremner, J.M., Brunk, H.D. (1972), *Statistical inference under order restrictions*, John Wiley and Sons, New York.