Outliers.

Dato un insieme di N coppie di valori (x_i, y_i) con $i = 1, \ldots, N$, si vuole determinare la retta di equazione $y = mx + q$ che meglio li interpola, tramite il metodo dei minimi quadrati: si vuole cioè minimizzare la cifra di merito f data dalla somma dei quadrati delle differenze tra i valori osservati (y_i) e i valori calcolati sulla retta $(mx_i + q)$.

Alcuni dei valori y utilizzati sono affetti da errori di misura e pertanto non sono significativi (*outliers*), anzi distorcono la stima della retta ottimale. Si vuole quindi determinare per diversi valori di K quali sono i K dati da trascurare, qual è la retta ottimale che si ottiene trascurandoli e qual è il valore $f(K)$ corrispondente. Si considerano accettabili solo valori di K compresi tra 0 e $N/6$.

Naturalmente, più alto è il numero K di dati che si trascurano e più basso è il valore di $f(K)$ che si può ottenere. Si vuole quindi determinare quale sia un valore ragionevole per K, considerando il problema a due obiettivi conflittuali: minimizzare K e minimizzare $f(K)$. Quale criterio usereste? Perché? Con quale esito?

Formulare il problema, classificarlo e risolvere l'esempio con i dati riportati qui sotto (file *Outliers.txt*).

Dati.

Sono date N = 40 coppie di dati.

Soluzione. Il problema si può formulare sia con un modello di PNL con funzione obiettivo quadratica, sia con un modello di PL, imponendo che siano nulle tutte le N derivate parziali del primo ordine. Poiché il modello PNL è convesso, non c'è il rischio di finire in minimi locali e quindi l'ottimalità è comunque garantita.

Il modello non-lineare è

$$\begin{aligned} \text{minimize } z = & \sum_{i=1}^{N} (y_i - (mx_i + q))^2 \\ & m, q \text{ unrestricted} \end{aligned}$$

Per consentire che K dati vengano trascurati, si possono introdurre variabili binarie w_i, col significato "0 = considerato" e "1 = outlier". Il risultante modello PNLI è

$$\begin{aligned} \text{minimize } z = & \sum_{i=1}^{N} (1 - w_i)(y_i - (mx_i + q))^2 \\ \text{s.t. } & \sum_{i=1}^{N} w_i \leq K \\ & w_i \text{ binary } \forall i = 1, \ldots, N \\ & m, q \text{ unrestricted} \end{aligned}$$

x	y
12	324
16	309
21	342
25	337
28	385
29	431
32	427
35	407
36	395
38	578
43	502
44	431
45	462
49	486
50	515
51	642
52	515
53	540
54	547
56	547
59	554
62	587
67	556
68	612
69	579
71	641
72	509
73	622
74	620
75	511
76	635
77	659
78	632
79	632
86	661
87	698
91	572
92	702
95	717
98	753

Table 1: Dati osservati.

Il modello di PNLI garantisce l'ottimalità della soluzione, poiché il problema non-lineare è convesso per ogni data scelta delle variabili discrete.

Per valori di K tra 0 e 6, si ottengono i risultati seguenti

K	$f(K)$
0	90264
1	71733
2	53001
3	38265
4	25968
5	15513
6	13917

Table 2: Valori di K e di $f(K)$.

Questi valori possono essere visti come ascisse e ordinate di punti paretiani in due dimensioni.

Poiché le unità di misura dei due obiettivi conflittuali non sono direttamente confrontabili, è possibile ricorrere, ad esempio, al criterio della massima curvatura, che dipende solo dalla forma della regione paretiana. Si può notare che $f(K)$ diminuisce all'incirca linearmente per incrementi di K da 0 a 5, mentre il passaggio da $K = 5$ a $K = 6$ non porta ad una diminuzione di $f(K)$ altrettanto significativa. Può essere quindi una scelta ragionevole $K = 5$ (in effetti i dati dell'esempio sono stati costruiti proprio inserendo a mano 5 outliers in una serie generata a caso nell'intorno di una retta).