M-Estimation with Imprecise Data

Marco Cattaneo
Department of Mathematics, University of Hull

ISIPTA '15, Pescara, Italy
20 July 2015

precise case

data: $X_1, \dots, X_n \in \mathcal{X}$ i.i.d. with (unknown) distribution P_X

goal: estimating the value(s) θ_0 of $\theta \in \Theta$ that minimize(s)

$$\underbrace{L(P_X, \theta)}_{\text{loss/distance}} \overset{\text{e.g.}}{=} \underbrace{E_{P_X}[\rho(X, \theta)]}_{\text{risk}} \overset{\text{e.g.}}{=} \underbrace{E_{P_X}\left[(X-\theta)^2\right]}_{\text{mean squared error: } \theta_0 = E_{P_X}[X]}$$

ML estimate (nonparametric) of $L(P_X, \cdot)$: the function $L(\hat{P}_X, \cdot)$ obtained by plugging in the empirical distribution of the data $\hat{P}_X$

ML decision (Cattaneo, 2013): the estimate(s) $\hat{\theta}_0$ that minimize(s)

$$\underbrace{L(\hat{P}_X, \theta)}_{\hat{\theta}_0\text{: minimum distance estimator}} \overset{\text{e.g.}}{=} \underbrace{\tfrac{1}{n}\textstyle\sum_{i=1}^n \rho(X_i, \theta)}_{\hat{\theta}_0\text{: M-estimator}} \overset{\text{e.g.}}{=} \underbrace{\tfrac{1}{n}\textstyle\sum_{i=1}^n (X_i - \theta)^2}_{\hat{\theta}_0 = \frac{1}{n}\sum_{i=1}^n X_i\text{: least squares estimator}}$$

asymptotic consistency: under some regularity conditions (Wolfowitz, 1957; Huber, 1964),

$$\hat{\theta}_0 \xrightarrow[n \to \infty]{\text{a.s.}} \theta_0$$

references

Cattaneo, M. (2013). Likelihood decision functions. *Electron. J. Stat.* 7, 2924–2946.

Cattaneo, M., and Wiencierz, A. (2012). Likelihood-based Imprecise Regression. *Int. J. Approx. Reasoning* 53, 1137–1154.

Cattaneo, M., and Wiencierz, A. (2014). On the implementation of LIR: the case of simple linear regression with interval data. *Comput. Stat.* 29, 743–767.

Ferson, S., Kreinovich, V., Hajagos, J., Oberkampf, W., and Ginzburg, L. (2007). *Experimental Uncertainty Estimation and Statistics for Data Having Interval Uncertainty*. Technical Report SAND2007-0939. Sandia National Laboratories.

Huber, P. J. (1964). Robust estimation of a location parameter. *Ann. Math. Stat.* 35, 73–101.

Manski, C. F. (2003). *Partial Identification of Probability Distributions*. Springer.

Schollmeyer, G., and Augustin, T. (2015). Statistical modeling under partial identification: Distinguishing three types of identification regions in regression analysis with interval data. *Int. J. Approx. Reasoning* 56, 224–248.

Schuyler, Q., Hardesty, B. D., Wilcox, C., and Townsend, K. (2014). Global analysis of anthropogenic debris ingestion by sea turtles. *Conserv. Biol.* 28, 129–139.

Utkin, L. V., and Coolen, F. P. A. (2011). Interval-valued regression and classification models in the framework of machine learning. In *ISIPTA '11*, eds. F. Coolen, G. de Cooman, T. Fetz, and M. Oberguggenberger. SIPTA, 371–380.

Wiencierz, A., and Cattaneo, M. (2015). On the validity of minimin and minimax methods for Support Vector Regression with interval data. In *ISIPTA '15*, eds. T. Augustin, S. Doria, E. Miranda, and E. Quaeghebeur. Aracne, 325–332.

Wolfowitz, J. (1957). The minimum distance method. *Ann. Math. Stat.* 28, 75–88.

imprecise case

data: $S_1, \dots, S_n \subseteq \mathcal{X}$ i.i.d. with (unknown) distribution P_S, such that $X_i \in S_i$

- the distribution P_S of the imprecise data only partially determines the distribution P_X of the (unobservable) precise data: let $[P_S]$ be the set of all distributions P_X compatible with P_S (in the sense that $X_i \in S_i$ is possible)
- assumptions reducing $[P_S]$ are also possible (e.g., all "beta distributions" on interval data) for the ML decision, but not for the black-box approach to estimation (see definitions below)

black-box approach (e.g., Ferson et al., 2007): since $X_1, \dots, X_n$ are only known to lie in $S_1, \dots, S_n$, replace the estimate $\hat{\theta}_0(X_1, \dots, X_n)$ with (the convex hull of) the set of estimates

$$\left\{\hat{\theta}_0(X_1, \dots, X_n) : X_i \in S_i\right\}$$

ML estimate (nonparametric) of $L(P_X, \cdot\,)$: usually not unique, it corresponds to the the set $\{L(P_X, \cdot\,) : P_X \in [\hat{P}_S]\}$ of all functions obtained by plugging in the distributions P_X compatible with the empirical distribution of the (imprecise) data $\hat{P}_S$

ML decision: the estimate(s) $\hat{\theta}_0$ that minimize(s)

$$\left\{L(P_X, \theta) : P_X \in [\hat{P}_S]\right\} \overset{\text{e.g.}}{=} \underbrace{\left\{E_{P_X}[\rho(X, \theta)] : P_X \in [\hat{P}_S]\right\}}_{= co\left\{\frac{1}{n}\sum_{i=1}^n \rho(X_i, \theta) \,:\, X_i \in S_i\right\}} \overset{\text{e.g.}}{=} \underbrace{\left\{E_{P_X}\left[(X - \theta)^2\right] : P_X \in [\hat{P}_S]\right\}}_{= co\left\{\frac{1}{n}\sum_{i=1}^n (X_i - \theta)^2 \,:\, X_i \in S_i\right\}}$$

asymptotic consistency, depending on the definition of minimum: under some regularity conditions (and possibly "smoothing corrections"),

pointwise dominance:

$$\hat{\theta}_0 \xrightarrow[n \to \infty]{\text{a.s.}} \{\arg\min_{\theta \in \Theta} L(P_X, \theta) : P_X \in [P_S]\}$$

- pointwise dominance ("maximality") and black-box approach ("E-admissibility") have the same limit, called sharp collection region by Schollmeyer and Augustin (2015)
- e.g., set of undominated regression functions of LIR approach (Cattaneo and Wiencierz, 2012, 2014), which uses interval dominance for computational reasons

minimax:

$$\hat{\theta}_0 \xrightarrow[n \to \infty]{\text{a.s.}} \arg\min_{\theta \in \Theta} \max_{P_X \in [P_S]} L(P_X, \theta)$$

- estimate and limit are usually unique, which greatly simplifies computation, description, and interpretation of the results: see logistic regression example below
- e.g., minimax SVR estimate (Utkin and Coolen, 2011; Wiencierz and Cattaneo, 2015), or LRM regression function of LIR approach (Cattaneo and Wiencierz, 2012, 2014)

minimin:

$$\hat{\theta}_0 \xrightarrow[n \to \infty]{\text{a.s.}} \{\theta \in \Theta : L(P_X, \theta) = 0,\ P_X \in [P_S]\}$$

- in parametric models the limit is the identification region (Manski, 2003) of the parameter θ (when L corresponds to a distance between distributions), called sharp marrow region by Schollmeyer and Augustin (2015): see parametric model example below
- e.g., minimin SVR estimate (Utkin and Coolen, 2011; Wiencierz and Cattaneo, 2015)

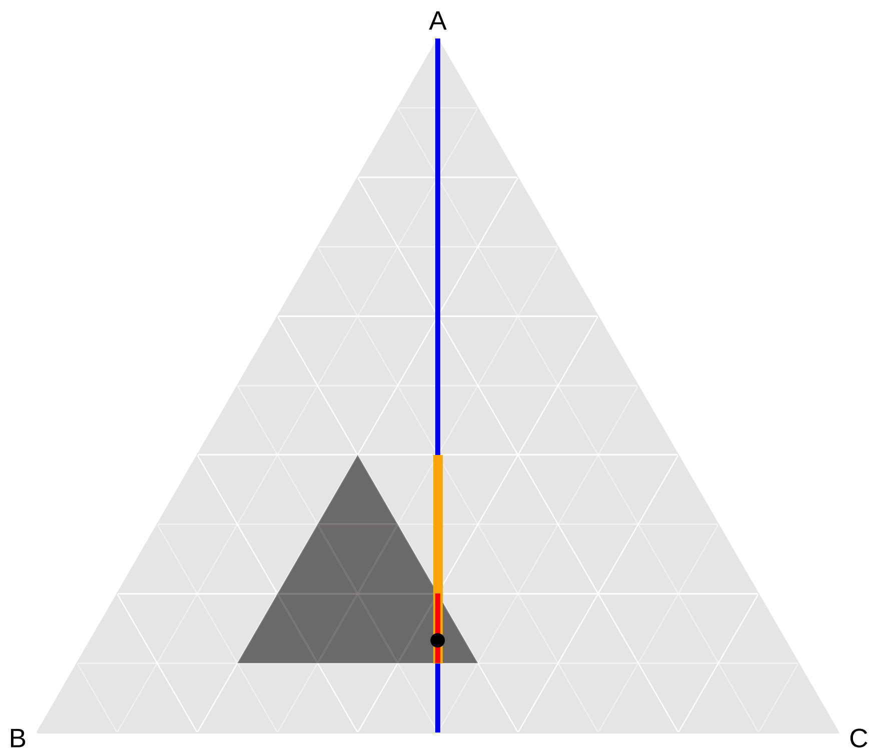
example: parametric model

precise data: $X_1, \dots, X_n \in \mathcal{X} = \{A, B, C\}$ i.i.d. with (unknown) distribution $P_X = (p_A, p_B, p_C)$

parametric model (represented by blue line): $p_B = p_C = \frac{1-\theta}{2}$ with $\theta = p_A \in \Theta = [0,1]$, i.e., $P_{X,\theta} = \left(\theta, \frac{1-\theta}{2}, \frac{1-\theta}{2}\right)$ with $\theta \in [0,1]$

loss $L(P_X, \theta)$: Euclidean distance between P_X and $P_{X,\theta}$

empirical distribution of precise data: $\hat{P}_X = \left(\frac{n_A}{n}, \frac{n_B}{n}, \frac{n_C}{n}\right)$, where n_A, n_B, n_C are the count data of A, B, C, respectively



ML decision with precise data: $\hat{\theta}_0 = \frac{n_A}{n}$

- asymptotic consistency: $\hat{\theta}_0 \xrightarrow[n\to\infty]{\text{a.s.}} \theta$
- $\hat{\theta}_0$ is also the parametric ML estimator: i.e., the M-estimator with the Kullback–Leibler divergence from P_X to $P_{X,\theta}$ as loss $L(P_X, \theta)$

imprecise data: $S_1, \dots, S_n \in \{\{A\}, \{B\}, \{C\}, \mathcal{X}\}$ i.i.d. with (unknown) distribution $P_S = (q_A, q_B, q_C, q_{na})$ (i.e., data are either precisely observed, or missing), such that $X_i \in S_i$

- $[P_S] = \{P_X : p_j \geq q_j \text{ for all } j \in \mathcal{X}\}$ is the set of all distributions P_X compatible with $P_S = (q_A, q_B, q_C, q_{na})$
- e.g., the gray area represents the set $[P_S]$ of all distributions P_X compatible with $P_S = (0.1, 0.4, 0.2, 0.3)$

empirical distribution of imprecise data: $\hat{P}_S = \left(\frac{n_A}{n}, \frac{n_B}{n}, \frac{n_C}{n}, \frac{n_{na}}{n}\right)$, where n_A, n_B, n_C, n_{na} are the count data of A, B, C, and missing, respectively

ML decision with imprecise data:

pointwise dominance: $\hat{\theta}_0 = \left[\frac{n_A}{n}, \frac{n_A + n_{na}}{n}\right]$

- asymptotic consistency: $\hat{\theta}_0 \xrightarrow[n\to\infty]{\text{a.s.}} \{p_A : P_X \in [P_S]\}$ (represented by orange segment)
- $\hat{\theta}_0$ is also the convex hull of the set of estimates $\left\{\hat{\theta}_0(X_1, \dots, X_n) : X_i \in S_i\right\}$ (black-box approach)

minimax: $\hat{\theta}_0 = \frac{2}{3}\frac{n_A}{n} + \frac{1}{3}\left(\left(1 - 2\,\frac{n_B \vee n_C}{n}\right) \vee \frac{n_A}{n}\right)$

- asymptotic consistency: $\hat{\theta}_0 \xrightarrow[n\to\infty]{\text{a.s.}} \frac{2}{3}\,q_A + \frac{1}{3}\left(1 - 2\,(q_B \vee q_C)\right)$ (represented by black point)
- $\hat{\theta}_0$ changes if the Euclidean distance $L(P_X, \theta)$ between P_X and $P_{X,\theta}$ is replaced by the Kullback–Leibler divergence from P_X to $P_{X,\theta}$ (while this is not the case for the other two definitions of minimum)

minimin: $\hat{\theta}_0 = \left[\frac{n_A}{n}, \left(1 - 2\,\frac{n_B \vee n_C}{n}\right) \vee \frac{n_A}{n}\right]$

- asymptotic consistency: $\hat{\theta}_0 \xrightarrow[n\to\infty]{\text{a.s.}} \{p_A : P_{X,\theta} \in [P_S]\}$ (represented by red segment)
- $\hat{\theta}_0$ estimates the set of all θ compatible with the distribution of the (imprecise) data: this is often the goal when the parametric model is assumed to be true

example: logistic regression

precise data: $(X_1, Y_1), \ldots, (X_{468}, Y_{468}) \in \mathbb{R} \times \{0,1\}$ i.i.d. with (unknown) distribution $P_{(X,Y)}$, describing the presence ($Y = 1$) or absence ($Y = 0$) of marine debris in the gastrointestinal system of a green turtle that died at time X

logistic regression: estimates $(\hat{\alpha}, \hat{\beta})$ of the regression parameters $(\alpha, \beta) = \theta \in \Theta = \mathbb{R}^2$ are obtained by minimizing

$$\begin{aligned} L(\hat{P}_{(X,Y)}, (\alpha, \beta)) &= \sum_{i=1}^{n} \left(Y_i \ln\left(1 + \exp(-\alpha - \beta X_i)\right) + (1 - Y_i) \ln\left(1 + \exp(\alpha + \beta X_i)\right)\right) \\ &= -\ln \prod_{i=1}^{n} \left(\frac{1}{1+\exp(-\alpha-\beta X_i)}\right)^{Y_i} \left(1 - \frac{1}{1+\exp(-\alpha-\beta X_i)}\right)^{1-Y_i} \end{aligned}$$

- $(\hat{\alpha}, \hat{\beta})$ are the parametric ML estimates when $P(Y = 1 \mid X) = \frac{1}{1+\exp(-\alpha-\beta X)}$ is assumed
- of particular interest is the question if the probability of debris ingestion increased over time ($\beta > 0$) or not ($\beta \leq 0$)

imprecise data: $[\underline{X}_1, \overline{X}_1] \times \{Y_1\}, \ldots, [\underline{X}_{468}, \overline{X}_{468}] \times \{Y_{468}\} \subset \mathbb{R} \times \{0,1\}$ i.i.d. with (unknown) distribution $P_{[\underline{X},\overline{X}] \times \{Y\}}$ (Schuyler et al., 2014)

minimax logistic regression: estimates $(\hat{\alpha}_m, \hat{\beta}_m)$ are obtained by minimizing

$$\max_{\hat{P}_{(X,Y)} \in [\hat{P}_{[\underline{X},\overline{X}] \times \{Y\}}]} L(\hat{P}_{(X,Y)}, (\alpha, \beta)) = \begin{cases} L(\hat{P}_{(Y\,\overline{X} + (1-Y)\,\underline{X},\, Y)}, (\alpha, \beta)) & \text{if } \beta \leq 0 \\ L(\hat{P}_{(Y\,\underline{X} + (1-Y)\,\overline{X},\, Y)}, (\alpha, \beta)) & \text{if } \beta \geq 0 \end{cases}$$

- computing the minimax logistic regression corresponds to computing two (standard) logistic regressions, with the two extreme cases for the precise X data: $Y\,\overline{X} + (1 - Y)\,\underline{X}$ and $Y\,\underline{X} + (1 - Y)\,\overline{X}$
- $(\hat{\alpha}_m, \hat{\beta}_m) \approx (-67, 0.033)$, and the significant positivity of $\hat{\beta}_m$ (with p-value ≈ 0.001) in the logistic regression with worst-case precise X data (i.e., $Y\,\underline{X} + (1 - Y)\,\overline{X}$) should imply also the significant positivity of $\hat{\beta}$ in the logistic regression with the true precise X data: that is, the ingestion of marine debris by green turtles increased over time