m-Estimators and z-Estimators

Guy Lebanon

m-Estimators and z-estimators (also called estimating equations) are natural extensions of the MLE. They enjoy similar consistency and are asymptotically normal, although with sometimes higher asymptotic variance. There are several reasons for studying these estimators: (a) they may be more comptuationally efficient than the MLE, (b) they may be more robust (resistent to deviations from the assumptions) than MLE, and (c) they can be analyzed using techniques that do not assume the true model is within the assumed parametric family. We follow along lines similar to [2] where more details may be found.

We assume that we have n samples $X_1, \ldots, X_n \overset{\text{iid}}{\sim} P \in \mathcal{Q}$ and we consider a parametric family $\mathcal{P} = \{p_\theta : \theta \in \Theta\} \subset \mathcal{Q}$ for the purpose of approximating P. Note that P is not necessarily a member of $\mathcal{P}$. The m-estimator associated with a given a function $m_\theta(x)$ is

$$\arg\max_{\theta\in\Theta} M_n(\theta) \quad \text{where} \quad M_n(\theta) = \frac{1}{n}\sum_{i=1}^{n} m_\theta(X_i).$$

The z-estimator associated with a given vector valued function $\psi_\theta = (\psi_{\theta,1}, \ldots, \psi_{\theta,l}) : X \to \mathbb{R}^l$ is the value of θ in Θ satisfying the following l-equations

$$\Psi_n(\theta) = 0 \quad \text{where} \quad \Psi_n(\theta) = \frac{1}{n}\sum_{i=1}^{n} \psi_\theta(X_i).$$

The two estimators are equivalent if m_θ is concave and smooth in θ and $\psi_{\theta,i}(x) = \partial m_\theta(x)/\partial\theta_i$. The case $m_\theta(x) = \log p_\theta(x)$ or $\psi_{\theta,i}(x) = \partial \log p_\theta(x)/\partial\theta_i$ reduces m or z-estimators to the MLE. In some cases it is convenient to work with m-estimators and in other cases with z-estimators.

Consistency

Consistency, in this case, corresponds to the convergence of the m-estimator to $\theta_0 \overset{\text{def}}{=} \arg\max_{\theta\in\Theta} M(\theta)$ where $M(\theta) \overset{\text{def}}{=} \mathsf{E}_P m_\theta(x)$. Note that this does not mean convergence to the truth i.e., $M(\theta) \neq P$ since P may lie outside $\{p_\theta : \theta \in \Theta\}$. Rather we have convergence to θ_0 - the "projection" of P on $\{p_\theta : \theta \in \Theta\}$. Note that in the case of the MLE, the projection is in the KL-divergence sense: $\theta_0 = \arg\min_{\theta\in\Theta} D_{KL}(P \,||\, p_\theta)$. The proposition below is in terms of m-estimates but a similar one holds for z-estimates.

Proposition 1 ([2]). *Assume* $\sup_{\theta\in\Theta} |M_n(\theta) - M(\theta)| \xrightarrow{\text{p}} 0$ *(law or large numbers convergence* $M_n(\theta) \xrightarrow{\text{p}} M(\theta)$ *is uniform over* θ*), and for all* $\epsilon > 0$, $\sup_{\theta:d(\theta,\theta_0)\geq\epsilon} M(\theta) < M(\theta_0)$*. Then for any sequence of estimators* $\hat{\theta}_n$ *with* $M_n(\hat{\theta}_n) \geq M_n(\theta_0) - o_P(1)$ *we have* $\hat{\theta}_n \xrightarrow{\text{p}} \theta_0$

The first condition is satisfied by the uniform strong law of large numbers and is satisfied for example if Θ is compact, M_n is continuous in θ for all x and if $|M_n(x)| < K(x)$, $\forall x, \theta$ for some function K with $\mathsf{E}_P K(X) < \infty$ [1]. There are other less restrictive conditions. The second condition correspond to θ_0 being isolated from the rest of the function and may be easily verified by examining the shape of the function M. For example, it holds for concave and continuous M over a compact set Θ. The assertion $M_n(\hat{\theta}_n) \geq M_n(\theta_0) - o_P(1)$ is trivially satisfied if $\hat{\theta}_n$ is an m-estimator (maximizes M_n).

Proof. The uniform convergence of M_n to M implies $M_n(\theta_0) \xrightarrow{\text{p}} M(\theta_0)$ and since $M_n(\hat{\theta}_n) \geq M_n(\theta_0) - o_P(1)$ we have $M_n(\hat{\theta}_n) \geq M(\theta_0) - o_P(1)$ and $M(\theta_0) - M(\hat{\theta}_n) \leq M_n(\hat{\theta}_n) - M(\hat{\theta}_n) + o_P(1) \leq \sup_{\theta\in\Theta} |M_n(\theta) - M(\theta)| + o_P(1) \xrightarrow{\text{p}} 0$. By the second assumption, $\forall\epsilon > 0$ there exists $\eta > 0$ with $M(\theta) < M(\theta_0) - \eta$ for every θ for

which $d(\theta,\theta_0) \geq \epsilon$. Thus, the event $\{d(\hat{\theta}_n,\theta_0) \geq \epsilon\}$ is contained in $\{M(\hat{\theta}_n) < M(\theta_0) - \eta\}$ whose probability converges to 0. □

Asymptotic Normality

We prove below that z-estimators $\hat{\theta}_n$ (the zero of the vector valued $\Psi_n(\theta) = \frac{1}{n}\sum_{i=1}^n \psi_\theta(X_i)$) are asymptotically a Gaussian whose mean is the zero of $\mathsf{E}_P\psi_\theta(X)$ which we denote by θ_0. We denote the matrix of partial derivates of ψ_θ by $\dot{\psi}_\theta$. The result below reduces to the standard MLE asymptotic normality if $\psi_\theta = \nabla \log p_\theta$ and $P \in \{p_\theta : \theta \in \Theta\}$, but is more general since it applies to general z-estimates and does not assume $P \in \{p_\theta : \theta \in \Theta\}$. A similar result may be stated for m-estimators.

Proposition 2. *We assume that $\Theta \subset \mathbb{R}^l$ is open and convex, $\mathsf{E}_P\psi_{\theta_0}(X) = 0$, $\mathsf{E}_P\|\psi_{\theta_0}(X)\|^2 < \infty$, $\mathsf{E}_P\dot{\psi}_{\theta_0}(X)$ exists and is non-singular, and $|\ddot{\Psi}_{ij}| = |\partial^2\psi_\theta(x)/\partial\theta_i\theta_j| < g(x)$ for all i,j and θ in a neighborhood of θ_0 for some integrable g. Then every consistent estimator sequence $\hat{\theta}_n$ for which $\Psi_n(\hat{\theta}_n) = 0\ \forall n$ satisfies*

$$\hat{\theta}_n = \theta_0 - \frac{1}{n}(\mathsf{E}_P\dot{\psi}_{\theta_0})^{-1}\sum_{i=1}^n \psi_{\theta_0}(X_i) + o_P\left(\frac{1}{\sqrt{n}}\right), \quad \textit{and} \tag{1}$$

$$\sqrt{n}(\hat{\theta}_n - \theta_0) \rightsquigarrow N(0, (\mathsf{E}_P\dot{\psi}_{\theta_0})^{-1}(\mathsf{E}_P\psi_{\theta_0}\psi_{\theta_0}^\top)(\mathsf{E}_P\dot{\psi}_{\theta_0})^{-1}). \tag{2}$$

Proof. By Taylor's theorem there exists a random vector $\tilde{\theta}_n$ on the line segment between θ_0 and $\hat{\theta}_n$ for which

$$0 = \Psi_n(\hat{\theta}_n) = \Psi_n(\theta_0) + \dot{\Psi}_n(\theta_0)(\hat{\theta}_n - \theta_0) + \frac{1}{2}(\hat{\theta}_n - \theta_0)^\top \ddot{\Psi}_n(\tilde{\theta}_n)(\hat{\theta}_n - \theta_0)$$

which we re-arrange as

$$\sqrt{n}\dot{\Psi}_n(\theta_0)(\hat{\theta}_n - \theta_0) + \sqrt{n}\frac{1}{2}(\hat{\theta}_n - \theta_0)^\top \ddot{\Psi}_n(\tilde{\theta}_n)(\hat{\theta}_n - \theta_0) = -\sqrt{n}\Psi_n(\hat{\theta}_n) = -\sqrt{n}\Psi_n(\theta_0) + o_P(1) \tag{3}$$

where the second equality is due to the fact that $\hat{\theta}_n \xrightarrow{\text{P}} \theta_0$ and that continuous functions preserves limits. Since $\dot{\Psi}_n(\theta_0)$ converges by the law of large numbers to $\mathsf{E}_P\dot{\psi}_\theta(X)$ and $\ddot{\Psi}_n(\tilde{\theta}_n)$ converges to a matrix of bounded values in the neighborhood of θ_0 (for large n) Equation (3) becomes

$$\sqrt{n}(\mathsf{E}_P\dot{\psi}_\theta(X) + o_P(1) + \frac{1}{2}(\hat{\theta}_n - \theta_0)O_P(1))(\hat{\theta}_n - \theta_0) = \sqrt{n}(\mathsf{E}_P\dot{\psi}_\theta(X) + o_P(1))(\hat{\theta}_n - \theta_0) = -\sqrt{n}\Psi_n(\theta_0) + o_P(1)$$

since $\hat{\theta}_n - \theta_0 = o_P(1)$ and $o_P(1)O_p(1) = o_P(1)$ (the notation $O_P(1)$ denotes stochastically bounded and it applies to $\ddot{\Psi}_n(\tilde{\theta}_n)$ as described above). The matrix $\mathsf{E}_P\dot{\psi}_\theta(X) + o_P(1)$ converges to a non-singular matrix and multiplying by its inverse proves (1).

Equation (2) follows from (1) by noticing that $n^{-1/2}\sum_{i=1}^n \psi_{\theta_0}(X_i)$ is an average of iid RVs with expectation 0. Applying Slutsky's theorem followed by the central limit theorem to the right hand side establishes normality while a simple calculation establishes the variance in (2). □

If we neglect the remainder in (1), the (asymptotic) influence function $\mathcal{I}$ is

$$\begin{aligned}\mathcal{I}_n(z) \stackrel{\text{def}}{=} \hat{\theta}_n(X_1,\ldots,X_{n-1},z) - \hat{\theta}_{n-1}(X_1,\ldots,X_{n-1}) &\approx \frac{1}{n}(\mathsf{E}_P\dot{\psi}_{\theta_0})^{-1}\psi_{\theta_0}(z) - \frac{1}{n(n-1)}\sum_{i=1}^{n-1}(\mathsf{E}_P\dot{\psi}_{\theta_0})^{-1}\psi_{\theta_0}(X_i) \\ &= \frac{1}{n}(\mathsf{E}_P\dot{\psi}_{\theta_0})^{-1}\psi_{\theta_0}(z) + o_P\left(\frac{1}{n}\right). \end{aligned} \tag{4}$$

References

[1] T. S. Ferguson. *A Course in Large Sample Theory.* Chapman & Hall, 1996.

[2] A. W. van der Vaart. *Asymptotic Statistics.* Cambridge University Press, 1998.