Chase Joyner

885 Homework 1 – M-estimation

September 15, 2015

The Calculus of M-Estimation by Stefanski and Boos gives a basic introduction and application to M-Estimation. An M-estimator $\hat{\boldsymbol{\theta}}$ is an estimator of $\boldsymbol{\theta}$ that satisfies

$$\sum_{i=1}^{n} \boldsymbol{\psi}(\mathbf{Y_i}, \hat{\boldsymbol{\theta}}) = \mathbf{0}, \tag{1}$$

where $\mathbf{Y}_1, ..., \mathbf{Y}_n$ are independent random vectors, $\boldsymbol{\theta}$ is a p-dimensional parameter, and $\boldsymbol{\psi}$ is a known $(p \times 1)$-functional that does not depend on i nor n. While many common estimators are not M-estimators, they can be written in the form of an M-estimator called *partial M-estimators*. A *partial M-estimator* is an estimator that alone is not an M-estimator, but is a component of an M-estimator. For example, the mean deviation from the sample mean, $\hat{\theta}_1 = n^{-1} \sum_{i=1}^{n} |Y_i - \bar{Y}|$ because there is no suitable $\boldsymbol{\psi}$ function such that $\sum_{i=1}^{n} \boldsymbol{\psi}(Y_i, \theta) = 0$ yields $\hat{\theta}_1$. However, this is a partial M-estimator since when combined with $\hat{\theta}_2 = \overline{Y}$ and the two functionals $\psi_1(y, \theta_1, \theta_2) = |y - \theta_2| - \theta_1$ and $\psi_2(y, \theta_1, \theta_2) = y - \theta_2$, we have

$$\sum_{i=1}^{n} \boldsymbol{\psi}(\mathbf{Y}_i, \hat{\theta}_1, \hat{\theta}_2) = \begin{pmatrix} \sum_{i=1}^{n} \Big(|Y_i - \hat{\theta}_2| - \hat{\theta}_1 \Big) \\ \sum_{i=1}^{n} \Big(Y_i - \hat{\theta}_2 \Big) \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}.$$

In addition to M-estimators, we will see later that we can add ψ functions to handle delta-method asymptotics for transformations of parameters, i.e. M-estimators are robust.

The basic approach of M-estimation is derived in section 2 of the paper. Assume, for now, that $Y_1, ..., Y_n$ are iid with distribution function F. The true parameter $\boldsymbol{\theta}_0$ is defined by

$$\mathrm{E}_F\big(\boldsymbol{\psi}(Y_1, \boldsymbol{\theta}_0)\big) = \int \boldsymbol{\psi}(y, \boldsymbol{\theta}_0) f(y) dy = \mathbf{0}. \tag{2}$$

If (2) above determines $\boldsymbol{\theta}_0$ uniquely, then there exists a sequence of M-estimators $\{\hat{\boldsymbol{\theta}}_n\}$ such that $\hat{\boldsymbol{\theta}}_n \stackrel{p}{\to} \boldsymbol{\theta}_0$. Define the function $\mathbf{G}_n(\boldsymbol{\theta}) = n^{-1} \sum_{i=1}^{n} \boldsymbol{\psi}(Y_i, \boldsymbol{\theta})$. A Taylor expansion about the true parameter $\boldsymbol{\theta}_0$ gives

$$\mathbf{0} = \mathbf{G}_n(\hat{\boldsymbol{\theta}}) = \mathbf{G}_n(\boldsymbol{\theta}_0) + \mathbf{G}_n'(\boldsymbol{\theta}_0)(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) + \mathbf{R}_n,$$

where $\mathbf{G}_n'(\boldsymbol{\theta}_0) = \frac{\partial \mathbf{G}_n(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}^T}\Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0}$. Rearranging the expansion, we arrive to

$$\sqrt{n}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) = \Big[-\mathbf{G}_n'(\boldsymbol{\theta}_0) \Big]^{-1} \sqrt{n} \mathbf{G}_n(\boldsymbol{\theta}_0) + \sqrt{n} \mathbf{R}_n^\star.$$

Define $\boldsymbol{\psi}'(y,\boldsymbol{\theta}) = \frac{\partial \boldsymbol{\psi}(y,\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}$. By the WLLN, as $n \to \infty$, we have

$$-\mathbf{G}_n'(\boldsymbol{\theta}_0) = \frac{1}{n}\sum_{i=1}^{n}\Big[-\boldsymbol{\psi}'(Y_i,\boldsymbol{\theta}_0)\Big] \xrightarrow{p} \mathrm{E}\Big[-\boldsymbol{\psi}'(Y_1,\boldsymbol{\theta}_0)\Big] = \mathbf{A}(\boldsymbol{\theta}_0).$$

Therefore, by CLT, we have

$$\sqrt{n}\mathbf{G}_n(\boldsymbol{\theta}_0) \xrightarrow{d} \mathrm{MVN}(0, \mathbf{B}(\boldsymbol{\theta}_0)),$$

where $\mathbf{B}(\boldsymbol{\theta}_0) = \mathrm{E}[\boldsymbol{\psi}(Y_1,\boldsymbol{\theta}_0)\boldsymbol{\psi}(Y_1,\boldsymbol{\theta}_0)^T]$. Note that $\sqrt{n}\mathbf{R}_n^{\star} \xrightarrow{p} \mathbf{0}$, which is difficult to prove, but holds under general assumptions. Combining the above results and appealing to Slutsky's Theorem, we conclude that

$$\hat{\boldsymbol{\theta}} \sim \mathrm{AMN}\left(\boldsymbol{\theta}_0, \frac{\mathbf{V}(\boldsymbol{\theta}_0)}{n}\right)$$

as $n \to \infty$, where $\mathbf{V}(\boldsymbol{\theta}_0) = \mathbf{A}(\boldsymbol{\theta}_0)^{-1}\mathbf{B}(\boldsymbol{\theta}_0)\big\{\mathbf{A}(\boldsymbol{\theta}_0)^{-1}\big\}^T$. Lastly, this work can be extended beyond equation (1). Suppose that we have an estimator, $\hat{\boldsymbol{\theta}}$, that satisfies

$$\sum_{i=1}^{n}\boldsymbol{\psi}(Y_i,\hat{\boldsymbol{\theta}}) = \mathbf{c}_n,$$

where $\mathbf{c}_n/\sqrt{n} \xrightarrow{p} \mathbf{0}$. Repeating the derivations as before with the sole change that $\mathbf{c}_n/\sqrt{n}$ is absorbed into the remainder quantity $\sqrt{n}\mathbf{R}_n^{\star}$, we arrive to the empirical estimators of $\mathbf{A}(\boldsymbol{\theta}_0)$ and $\mathbf{B}(\boldsymbol{\theta}_0)$,

$$\mathbf{A}_n(\mathbf{Y},\hat{\boldsymbol{\theta}}) = -\mathbf{G}_n'(\hat{\boldsymbol{\theta}}) = \frac{1}{n}\sum_{i=1}^{n}\Big[-\boldsymbol{\psi}'(Y_i,\hat{\boldsymbol{\theta}})\Big],$$

and

$$\mathbf{B}_n(\mathbf{Y},\hat{\boldsymbol{\theta}}) = \frac{1}{n}\sum_{i=1}^{n}\boldsymbol{\psi}(Y_i,\hat{\boldsymbol{\theta}})\boldsymbol{\psi}(Y_i,\hat{\boldsymbol{\theta}})^T$$

and

$$\mathbf{V}_n(\mathbf{Y},\hat{\boldsymbol{\theta}}) = \mathbf{A}_n(\mathbf{Y},\hat{\boldsymbol{\theta}})^{-1}\mathbf{B}_n(\mathbf{Y},\hat{\boldsymbol{\theta}})\big\{\mathbf{A}_n(\mathbf{Y},\hat{\boldsymbol{\theta}})^{-1}\big\}^T.$$

These ideas are now illustrated with examples.

The first example to illustrate M-estimation is to estimate the population mean and variance, μ, and σ^2. Let $\hat{\boldsymbol{\theta}} = (\overline{Y}, s_n^2)^T$ be the M-estimator defined by

$$\boldsymbol{\psi}(Y_i,\boldsymbol{\theta}) = \begin{pmatrix} Y_i - \theta_1 \\ (Y_i-\theta_1)^2 - \theta_2 \end{pmatrix}.$$

Denoting the true parameter values by $\boldsymbol{\theta}_0 = (\theta_{10}, \theta_{20})$, we have

$$\mathbf{A}(\boldsymbol{\theta}_0) = \mathrm{E}\Big[-\boldsymbol{\psi}'(Y_1,\boldsymbol{\theta}_0)\Big] = \mathrm{E}\begin{pmatrix} 1 & 0 \\ 2(Y_1-\theta_{10}) & 1 \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$$

since $\mathrm{E}(Y_1) = \theta_{10}$. Also, the matrix

$$\mathbf{B}(\boldsymbol{\theta}_0) = \mathrm{E}\Big[\boldsymbol{\psi}(Y_1,\boldsymbol{\theta}_0)\boldsymbol{\psi}(Y_1,\boldsymbol{\theta}_0)^T\Big]$$

has elements

$$
\begin{aligned}
\mathbf{B}(\boldsymbol{\theta}_0)_{11} &= \mathrm{E}\big[(Y_1-\theta_{10})^2\big] = \theta_{20} = \sigma^2 \\
\mathbf{B}(\boldsymbol{\theta}_0)_{12} = \mathbf{B}(\boldsymbol{\theta}_0)_{21} &= \mathrm{E}\big[(Y_1-\theta_{10})\big((Y_1-\theta_{10})^2-\theta_{20}\big)\big] = \mu_3 - 0 = \mu_3 \\
\mathbf{B}(\boldsymbol{\theta}_0)_{22} &= \mathrm{E}\Big[\big((Y_1-\theta_{10})^2-\theta_{20}\big)^2\Big] = \mu_4 - \sigma^4,
\end{aligned}
$$

which are estimated by

$$
\begin{aligned}
\mathbf{B}_n(\mathbf{Y},\hat{\boldsymbol{\theta}})_{11} &= \frac{1}{n}\sum_{i=1}^{n}(Y_i-\overline{Y})^2 = s_n^2 \\
\mathbf{B}_n(\mathbf{Y},\hat{\boldsymbol{\theta}})_{12} = \mathbf{B}_n(\mathbf{Y},\hat{\boldsymbol{\theta}})_{21} &= \frac{1}{n}\sum_{i=1}^{n}(Y_i-\overline{Y})\big((Y_i-\overline{Y})^2-s_n^2\big) = m_3 \\
\mathbf{B}_n(\mathbf{Y},\hat{\boldsymbol{\theta}})_{22} &== \frac{1}{n}\sum_{i=1}^{n}\big((Y_i-\overline{Y})^2-s_n^2\big)^2 = m_4 - s_n^4,
\end{aligned}
$$

where m_k denotes the kth sample moment. Since the matrix $\mathbf{A}(\boldsymbol{\theta}_0)$ need not be estimated, we simply have $\mathbf{V}_n(\mathbf{Y},\hat{\boldsymbol{\theta}}) = \mathbf{B}_n(\mathbf{Y},\hat{\boldsymbol{\theta}})$. Attached in the appendix is a simulation illustrating this example.

The second example of the paper shows how we can estimate the ratio of two population means. Let $\hat{\theta} = \overline{Y}/\overline{X}$, where $(Y_1, X_1), ..., (Y_n, X_n)$ are iid samples with $\mathrm{E}(Y_1) = \mu_Y$ and $\mathrm{E}(X_1) = \mu_X \neq 0$, $\mathrm{Var}(Y_1) = \sigma_Y^2$ and $\mathrm{Var}(X_1) = \sigma_X^2$, and $\mathrm{Cov}(Y_1, X_1) = \sigma_{YX}$. Also let $\psi(Y_i, X_i, \theta) = Y_i - \theta X_i$ be the functional for $\hat{\theta}$. This M-estimator leads to

$$\mathbf{A}(\boldsymbol{\theta}_0) = \mu_X,$$

and

$$\mathbf{B}(\boldsymbol{\theta}_0) = \mathrm{E}\big((Y_1-\theta_0 X_1)^2\big),$$

Therefore, we have

$$\mathbf{V}(\boldsymbol{\theta}_0) = \mathrm{E}\big((Y_1-\theta_0 X_1)^2\big)/\mu_X^2.$$

Notice that each of these must be estimated, which can be done by

$$
\begin{aligned}
\mathbf{A}_n(\mathbf{Y},\hat{\theta}) &= \overline{X} \\
\mathbf{B}_n(\mathbf{Y},\hat{\theta}) &= \frac{1}{n}\sum_{i=1}^{n}\left(Y_i - \frac{\overline{Y}}{\overline{X}}X_i\right)^2 \\
\mathbf{V}_n(\mathbf{Y},\hat{\theta}) &= \frac{1}{\overline{X}^2}\frac{1}{n}\sum_{i=1}^{n}\left(Y_i - \frac{\overline{Y}}{\overline{X}}X_i\right)^2.
\end{aligned}
$$

Attached in the appendix is a simulation illustrating this example.

The third example illustrates how the delta method can be implemented with M-estimation. Referring to example 1, suppose we are interested also in $s_n = \sqrt{s_n^2}$ and $\log(s_n^2)$. We can define $\psi_3(Y_i, \boldsymbol{\theta}) = \sqrt{\theta_2} - \theta_3$ and $\psi_4(Y_i, \boldsymbol{\theta}) = \log(\theta_2) - \theta_4$. These functions give the following matrices

$$\mathbf{A}(\boldsymbol{\theta}_0) = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & -\frac{1}{2\sqrt{\theta_{20}}} & 1 & 0 \\ 0 & -\frac{1}{\theta_{20}} & 0 & 1 \end{pmatrix},$$

$$\mathbf{B}(\boldsymbol{\theta}_0) = \begin{pmatrix} \frac{1}{\theta_{20}} & \frac{\mu_3}{2\theta_{20}^3} & 0 & 0 \\ \frac{\mu_3}{2\theta_{20}^3} & \frac{\mu_4 - \theta_{20}^2}{4\theta_{20}^4} & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix},$$

and

$$\mathbf{V}(\boldsymbol{\theta}_0) = \begin{pmatrix} \theta_{20} & \mu_3 & \frac{\mu_3}{2\sqrt{\theta_{20}}} & \frac{\mu_3}{\theta_{20}} \\ \mu_3 & \mu_4 - \theta_{20}^2 & \frac{\mu_4 - \theta_{20}^2}{2\sqrt{\theta_2 0}} & \frac{\mu_4 - \theta_{20}^2}{\theta_2 0} \\ \frac{\mu_3}{2\sqrt{\theta_{20}}} & \frac{\mu_4 - \theta_{20}^2}{2\sqrt{\theta_{20}}} & \frac{\mu_4 - \theta_{20}^2}{44\theta_{20}} & \frac{\mu_4 - \theta_{20}^2}{2\theta_{20}^{3/2}} \\ \frac{\mu_4}{\theta_{20}} & \frac{\mu_4 - \theta_{20}^2}{\theta_{20}} & \frac{\mu_4 - \theta_{20}^2}{2\theta_{20}^{3/2}} & \frac{\mu_4 - \theta_{20}^2}{\theta_{20}^2} \end{pmatrix}.$$

This idea is illustrated via simulation in the appendix.

Continuing in the paper, we jump to section 4 to discuss the situations when $\boldsymbol{\psi}$ is a nonsmooth function. More specifically, if $\boldsymbol{\psi}$ is not differentiable everywhere, we calculate the matrix $\mathbf{A}(\boldsymbol{\theta}_0)$ as

$$\mathbf{A}(\boldsymbol{\theta}_0) = -\frac{\partial}{\partial \boldsymbol{\theta}^T} \mathrm{E}_F \left[\boldsymbol{\psi}(Y_1, \boldsymbol{\theta}) \right] \Big|_{\boldsymbol{\theta} = \boldsymbol{\theta}_0}.$$

Notice that here the expectation is taken with respect to the true distribution F. We quickly discuss example 6 of the paper to illustrate this idea. Suppose we are interested in $\hat{\theta}$ that satisfies $\sum \psi_k(Y_i - \hat{\theta}) = 0$, where

$$\psi_k(x) = \begin{cases} x & |x| \le k, \\ \mathrm{sgn}(x)k & |x| > k. \end{cases}$$

Observe that this $\boldsymbol{\psi}$ is continuous everywhere, however not differentiable at $\pm k$. By the new definition of $\mathbf{A}(\boldsymbol{\theta}_0)$, we have

$$\begin{aligned} A(\theta_0) &= -\frac{\partial}{\partial \theta} \mathrm{E}_F \left[\psi_k(Y_1 - \theta) \right] \Big|_{\theta = \theta_0} \\ &= -\frac{\partial}{\partial \theta} \int \psi_k(y - \theta) f(y) dy \,\Big|_{\theta = \theta_0} \\ &= \int -\frac{\partial}{\partial \theta} \psi_k(y - \theta) \Big|_{\theta = \theta_0} f(y) dy \\ &= \int \psi_k'(y - \theta_0) f(y) dy. \end{aligned}$$

Also, we find that

$$B(\theta_0) = \mathrm{E}\big[\psi_k^2(Y_1 - \theta_0)\big] = \int \psi_k^2(y - \theta_0) f(y) dy$$

and thus

$$\mathbf{V}(\theta_0) = \frac{\int \psi_k^2(y - \theta_0) f(y) dy}{\big[\int \psi_k'(y - \theta_0) f(y) dy\big]^2}.$$

These can be estimated by $A_n(\mathbf{Y}, \hat{\theta}) = n^{-1} \sum_{i=1}^n \Big[-\psi_k'(Y_i - \hat{\theta})\Big]$ and $B_n(\mathbf{Y}, \hat{\theta}) = n^{-1} \sum_{i=1}^n \psi_k^2(Y_i - \hat{\theta})$.

The last section that we explore of the paper is section 5, which discusses Regression M-estimators. Consider the nonlinear model

$$Y_i = g(\mathbf{x}_i, \boldsymbol{\beta}) + e_i, \qquad i = 1, ..., n, \tag{3}$$

where g is a known differentiable function in $\boldsymbol{\beta}$, the errors $e_1, ..., e_n$ are independent with mean 0 and $\mathrm{Var}(e_i) = \sigma_i^2,\ i = 1, ..., n$, and $\mathbf{x}_1, ..., \mathbf{x}_n$, are known constant vectors. Under this setting, define the design matrix $\mathbf{X} = (\mathbf{x}_1, ..., \mathbf{x}_n)^T$. The least squares estimator of $\boldsymbol{\beta}$ satisfies

$$\sum_{i=1}^n (Y_i - g(\mathbf{x}_i, \hat{\boldsymbol{\beta}})) g'(\mathbf{x}_i, \hat{\boldsymbol{\beta}}) = \mathbf{0}$$

where the derivative of g is the partial derivative with respect to $\boldsymbol{\beta}$ evaluated at $\hat{\boldsymbol{\beta}}$. As in section 2, we expand the equation above about the true value $\boldsymbol{\beta}_0$ to obtain

$$\sqrt{n}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0) = \left[\frac{1}{n} \sum_{i=1}^n -\boldsymbol{\psi}'(Y_i, \mathbf{x}, \boldsymbol{\beta}_0)\right]^{-1} \frac{1}{\sqrt{n}} \sum_{i=1}^n \boldsymbol{\psi}(Y_i, \mathbf{x}_i, \boldsymbol{\beta}_0) + \sqrt{n} R_n^\star,$$

where $\boldsymbol{\psi}(Y_i, \mathbf{x}_i, \boldsymbol{\beta}_0) = (Y_i - g(\mathbf{x}_i, \boldsymbol{\beta}_0)) g'(\mathbf{x}_i, \boldsymbol{\beta}_0)$. From this, we obtain the estimate

$$\begin{aligned} A_n(\mathbf{X}, \mathbf{Y}, \boldsymbol{\beta}_0) &= \frac{1}{n} \sum_{i=1}^n \Big[-\boldsymbol{\psi}'(Y_i, \mathbf{x}_i, \boldsymbol{\beta}_0)\Big] \\ &= \frac{1}{n} \sum_{i=1}^n \Big[g'(\mathbf{x}_i, \boldsymbol{\beta}_0) g'(\mathbf{x}_i, \boldsymbol{\beta}_0)^T - (Y_i - g(\mathbf{x}_i \boldsymbol{\beta}_0)) g''(\mathbf{x}_i, \boldsymbol{\beta}_0)\Big]. \end{aligned}$$

Now, taking the expectation with respect to the true model, $\mathbf{Y}$ gets dropped and we're left with

$$A_n(\mathbf{X}, \boldsymbol{\beta}_0) = \frac{1}{n} \sum_{i=1}^n \mathrm{E}\Big[-\boldsymbol{\psi}'(Y_i, \mathbf{x}_i, \boldsymbol{\beta}_0)\Big] = \frac{1}{n} \sum_{i=1}^n g'(\mathbf{x}_i, \boldsymbol{\beta}_0) g'(\mathbf{x}_i, \boldsymbol{\beta}_0)^T. \tag{4}$$

Next, if we assume that the limit exists, then define

$$\begin{aligned} \mathbf{A}(\boldsymbol{\beta}_0) &= \lim_{n \to \infty} \frac{1}{n} \sum_{i=1}^n \mathrm{E}\Big[-\boldsymbol{\psi}'(Y_i, \mathbf{x}_i, \boldsymbol{\beta}_0)\Big] \\ &= \lim_{n \to \infty} \frac{1}{n} \sum_{i=1}^n g'(\mathbf{x}_i, \boldsymbol{\beta}_0) g'(\mathbf{x}_i, \boldsymbol{\beta}_0)^T. \end{aligned}$$

Evaluating equation (4) at $\hat{\boldsymbol{\beta}}$ will yield the estimator of $\mathbf{A}(\boldsymbol{\beta}_0)$, which is

$$\begin{aligned} A_n(\mathbf{X}, \hat{\boldsymbol{\beta}}) &= \frac{1}{n}\sum_{i=1}^{n} \mathrm{E}\Big[-\boldsymbol{\psi}'(Y_i, \mathbf{x}_i, \boldsymbol{\beta}_0)\Big]\Big|_{\boldsymbol{\beta}=\hat{\boldsymbol{\beta}}} \\ &= \frac{1}{n}\sum_{i=1}^{n} g'(\mathbf{x}_i, \hat{\boldsymbol{\beta}}) g'(\mathbf{x}_i, \hat{\boldsymbol{\beta}})^T. \end{aligned}$$

As for the $\mathbf{B}$ matrix, we have

$$\begin{aligned} \mathbf{B}_n(\mathbf{X}, \mathbf{Y}, \boldsymbol{\beta}_0) &= \frac{1}{n}\sum_{i=1}^{n} \mathrm{E}\Big[\boldsymbol{\psi}(Y_i, \mathbf{x}_i, \boldsymbol{\beta}_0)\boldsymbol{\psi}(Y_i, \mathbf{x}_i, \boldsymbol{\beta}_0)^T\Big] \\ &= \frac{1}{n}\sum_{i=1}^{n} \sigma_i^2 g'(\mathbf{x}_i, \boldsymbol{\beta}_0) g'(\mathbf{x}_i, \boldsymbol{\beta}_0)^T, \end{aligned}$$

which can be estimated by the mean-squared error

$$\begin{aligned} \mathbf{B}_n(\mathbf{X}, \mathbf{Y}, \hat{\boldsymbol{\beta}}) &= \frac{1}{n-p}\sum_{i=1}^{n} \boldsymbol{\psi}(Y_i, \mathbf{x}_i, \hat{\boldsymbol{\beta}})\boldsymbol{\psi}(Y_i, \mathbf{x}_i, \hat{\boldsymbol{\beta}})^T \\ &= \frac{1}{n-p}\sum_{i=1}^{n} (Y_i - g(\mathbf{x}_i, \hat{\boldsymbol{\beta}}))^2 g'(\mathbf{x}_i, \hat{\boldsymbol{\beta}}) g'(\mathbf{x}_i, \hat{\boldsymbol{\beta}})^T. \end{aligned}$$

Next, we look at example 8 of the paper to illustrate these ideas relating to regression M-estimators.

Consider the nonlinear regression model in equation (3), where $g(\mathbf{x}_i, \boldsymbol{\beta}) = \mathbf{x}_i^T\boldsymbol{\beta}$ and the least squares estimator of $\boldsymbol{\beta}$ satisfying

$$\sum_{i=1}^{n} \psi_k(Y_i - \mathbf{x}_i^T\hat{\boldsymbol{\beta}})\mathbf{x}_i = 0,$$

where ψ_k is the function defined in example 6. Then, we have

$$\boldsymbol{\psi}(Y_i, \mathbf{x}_i, \boldsymbol{\beta}) = \psi_k(Y_i - \mathbf{x}_i^T\boldsymbol{\beta})\mathbf{x}_i.$$

Then, we obtain the matrices

$$\mathbf{A}_n(\mathbf{X}, \mathbf{Y}, \boldsymbol{\beta}_0) = \frac{1}{n}\sum_{i=1}^{n} \psi_k'(e_i)\mathbf{x}_i\mathbf{x}_i^T,$$

and so

$$\mathbf{A}_n(\mathbf{X}, \boldsymbol{\beta}_0) = \frac{1}{n}\sum_{i=1}^{n} \mathrm{E}\Big[\psi_k'(e_i)\mathbf{x}_i\mathbf{x}_i^T\Big].$$

Also, we have

$$\mathbf{B}_n(\mathbf{X}, \boldsymbol{\beta}_0) = \frac{1}{n}\sum_{i=1}^{n} \mathrm{E}\Big[\psi_k(e_i)^2\mathbf{x}_i\mathbf{x}_i^T\Big].$$

Lastly, if the errors are identically distributed and by using the natural estimators for $\mathbf{A}_n$ and $\mathbf{B}_n$ above, we have

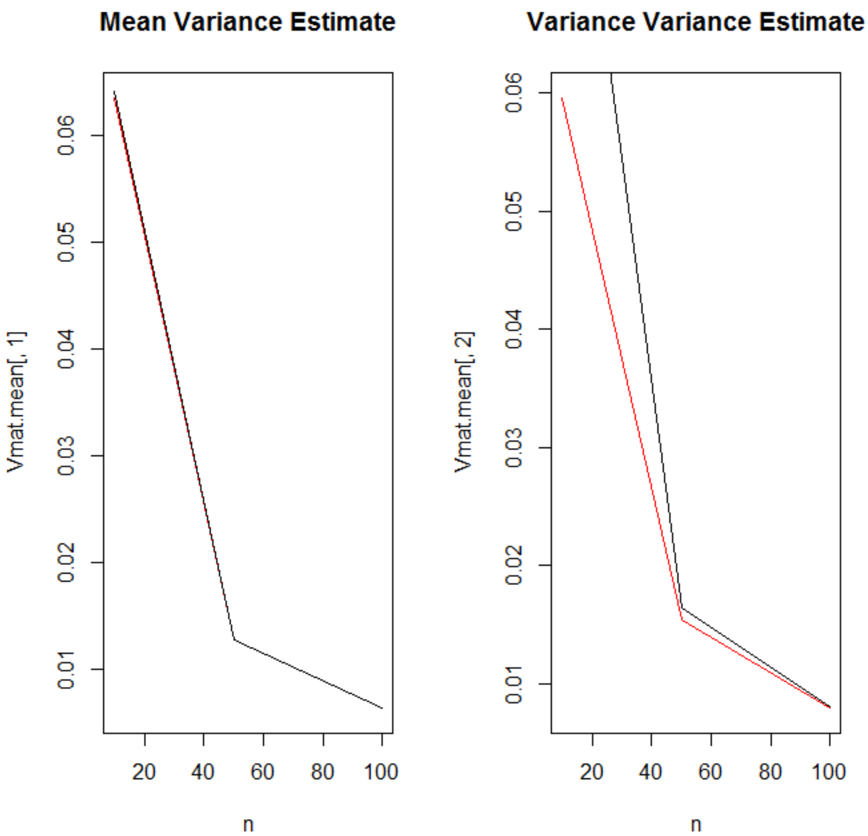
$$\mathbf{V}_n(\mathbf{X}, \boldsymbol{\beta}_0) = \left(\mathbf{X}^T\mathbf{X}/n\right)^{-1} \mathrm{E}[\psi_k(e_1)^2]/\mathrm{E}[\psi_k'(e_1)]^2.$$

In summary, M-estimators represent a large class of statistics, i.e. maximum likelihood estimators, sample moments, and even touches Bayesian estimators. Please refer to the Appendix section below for simulations to examples 1 through 3.

Appendix

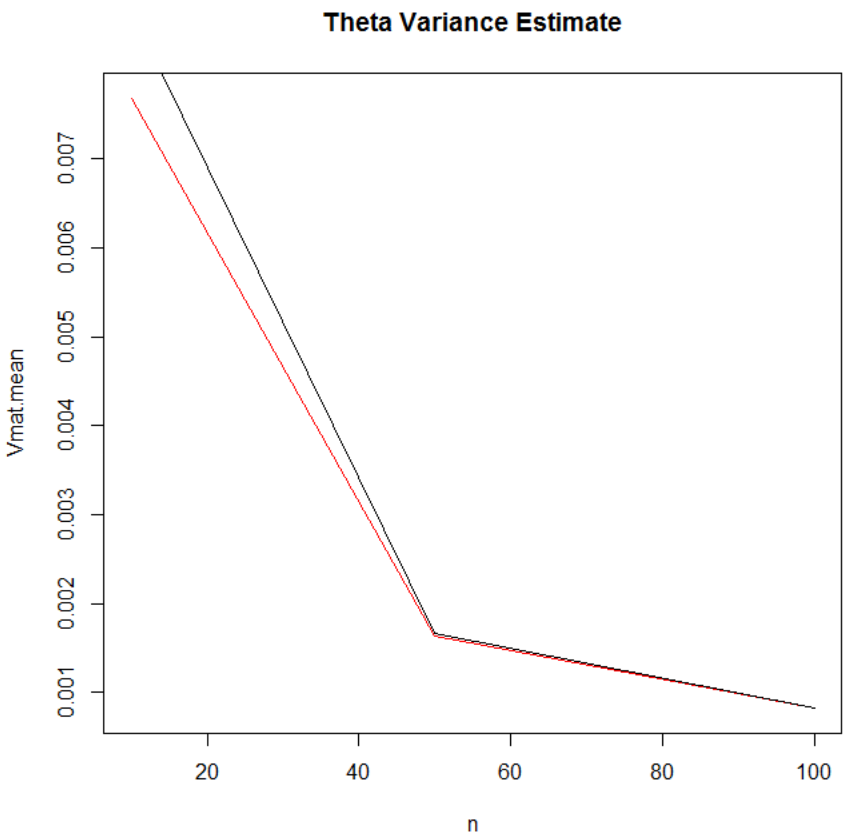
Example 1

The sample sizes are 10, 50, and 100. The samples are iid samples from N$(1.3, 0.8)$. Below is a graph representing the asymptotical properties of M-estimation. We see that the estimated mean variance (red line) is almost indistinguishable from the true mean variance (black line) and around a sample size of 50 that the estimated variances (red line) become very close to the true values (black line).



Example 2

The sample sizes are 10, 50, and 100. The samples of Y are iid from $\mathrm{N}(1.3, 0.8)$ and the samples of X are iid from $\mathrm{N}(3, 0.75)$. Here, we again see the asymptotics kick in relatively quickly around a sample size of $n = 50$.



Example 3

The sample sizes are 10, 50, and 100. The data are generated from $N(3, 0.5)$. As expected, we see the asymptotics kick in and conclude that M-estimation is valid.