Lecture 13: Profile likelihoods, GEE, and GMM

Profile likelihoods

Let $\ell(\theta,\xi)$ be a likelihood (or empirical likelihood), where θ and ξ are not necessarily vector-valued.
It may be difficult to maximize the likelihood $\ell(\theta,\xi)$ simultaneously over θ and ξ.
For each fixed θ, let $\xi(\theta)$ satisfy

$$\ell(\theta,\xi(\theta)) = \sup_{\xi} \ell(\theta,\xi).$$

The function

$$\ell_P(\theta) = \ell(\theta,\xi(\theta))$$

is called a *profile likelihood* function for θ.
Suppose that $\widehat{\theta}_P$ maximizes $\ell_P(\theta)$.
Then $\widehat{\theta}_P$ is called a maximum profile likelihood estimator of θ.
$\widehat{\theta}_P$ may be different from an MLE of θ.
Although this idea can be applied to parametric models, it is more useful in semi-parametric models, especially when θ is a parametric component and ξ is a nonparametric component.

Example

Consider the empirical likelihood

$$\ell(G) = \prod_{i=1}^{n} P_G(\{x_i\}), \qquad G \in \mathscr{F}$$

subject to the constraints

$$p_i > 0, \quad i = 1, ..., n, \quad \sum_{i=1}^{n} p_i = 1, \quad \text{and} \quad \sum_{i=1}^{n} p_i \psi(x_i, \theta) = 0,$$

where $\theta \in \mathscr{R}^k$ is an unknown parameter vector ψ is a known function from $\mathscr{R}^d \times \mathscr{R}^k$ to $\mathscr{R}^s$, and $k \leq s$.

Maximizing this empirical likelihood is equivalent to maximizing

$$H(p_1, ..., p_n, \omega, \lambda, \theta) = \log\left(\prod_{i=1}^{n} p_i\right) + \omega\left(1 - \sum_{i=1}^{n} p_i\right) - n\sum_{i=1}^{n} p_i \lambda^\tau \psi(x_i, \theta),$$

where ω and λ are Lagrange multipliers.

$$\frac{\partial H}{\partial p_i} = \frac{1}{p_i} - \omega - n\lambda^\tau \psi(x_i, \theta) \quad i = 1, ..., n$$

Example (continued)

Setting $\partial H/\partial p_i = 0$ and multiplying it by p_i leads to

$$1 = \omega p_i + n\lambda^\tau \psi(x_i,\theta) \quad i = 1,...,n$$

Taking the sum over i on both sides of this expression gives $\omega = n$, since $\sum_{i=1}^n p_i = 1$ and $\sum_{i=1}^n p_i \psi(x_i,\theta) = 0$.
Then the solution is

$$p_i(\theta) = n^{-1}\{1+[\lambda_n(\theta)]^\tau \psi(x_i,\theta)\}^{-1}, \quad i = 1,...,n,$$

with a $\lambda_n(\theta)$ satisfying

$$\frac{1}{n}\sum_{i=1}^n \frac{\psi(x_i,\theta)}{1+[\lambda_n(\theta)]^\tau \psi(x_i,\theta)} = 0$$

Substituting $p_i(\theta)$ into $\ell(G)$ leads to the following profile empirical likelihood for θ:

$$\ell_P(\theta) = \prod_{i=1}^n \frac{1}{n\{1+[\lambda_n(\theta)]^\tau \psi(x_i,\theta)\}}.$$

Example (continued)

If $\widehat{\theta}$ is a maximum of $\ell_P(\theta)$, then $\widehat{\theta}$ is a maximum profile empirical likelihood estimator of θ and the corresponding estimator of p_i is $p_i(\widehat{\theta})$.
A result similar to Theorem 5.4 and a result on asymptotic normality of $\widehat{\theta}$ are established in Qin and Lawless (1994), under some conditions on ψ.

Missing data

Assume that $X_1,...,X_n$ are i.i.d. random variables from an unknown c.d.f. F and some X_i's are missing.
Let $\delta_i = 1$ if X_i is observed and $\delta_i = 0$ if X_i is missing.
Suppose that (X_i, δ_i) are i.i.d. and let

$$\pi(x) = P(\delta_i = 1 | X_i = x).$$

If X_i and δ_i are independent, i.e., $\pi(x) \equiv \pi$ does not depend on x, then the empirical c.d.f. based on observed data, i.e., the c.d.f. putting mass r^{-1} to each observed X_i, where r is the number of observed X_i's, is an unbiased and consistent estimator of F, provided that $\pi > 0$.

Missing data

On the other hand, if $\pi(x)$ depends on x (called nonignorable missingness), then the empirical c.d.f. based on observed data is a biased and inconsistent estimator of F.

In fact, the empirical c.d.f. based on observed data is an unbiased estimator of $P(X_i \leq x|\delta_i = 1)$, which is generally different from the unconditional probability $F(x) = P(X_i \leq x)$.

If both π and F are in parametric models, then we can apply the method of maximum likelihood.

For example, if $\pi(x) = \pi_\theta(x)$ and $F(x) = F_\vartheta(x)$ has a p.d.f. f_ϑ, where θ and ϑ are vectors of unknown parameters, then a parametric likelihood of (θ, ϑ) is

$$\ell(\theta, \vartheta) = \prod_{i=1}^{n} [\pi_\theta(x_i) f_\vartheta(x_i)]^{\delta_i} (1 - \pi)^{1-\delta_i},$$

where $\pi = \int \pi_\theta(x) f_\vartheta(x) dx$.

Computationally, it may be difficult to maximizing this likelihood, since π is an integral.

Missing data

Suppose now that $\pi(x) = \pi_\theta(x)$ is the parametric component and F is the nonparametric component.
Then an empirical likelihood can be defined as

$$\ell(\theta, G) = \prod_{i=1}^{n} [\pi_\theta(x_i) p_i]^{\delta_i} (1-\pi)^{1-\delta_i} \qquad p_i = P_G(\{x_i\})$$

subject to $p_i \geq 0$, $\sum_{i=1}^{n} \delta_i p_i = 1$, $\sum_{i=1}^{n} \delta_i p_i [\pi_\theta(x_i) - \pi] = 0$, $i = 1, ..., n$.
It can be shown (exercise) that the logarithm of the profile empirical likelihood for (θ, π) (with a Lagrange multiplier) is

$$\sum_{i=1}^{n} \left\{ \delta_i \log\left(\pi_\theta(x_i)\right) + (1-\delta_i)\log(1-\pi) - \delta_i \log\left(1 + \lambda[\pi_\theta(x_i) - \pi]\right) \right\}.$$

Under some conditions, it can be shown that the estimators $\widehat{\theta}$, $\widehat{\pi}$, and $\widehat{\lambda}$ obtained by maximizing this likelihood are consistent and asymptotically normal and that the empirical c.d.f. putting mass $\widehat{p}_i = r^{-1}\{1 + \widehat{\lambda}[\pi_{\widehat{\theta}}(X_i) - \widehat{\pi}]\}^{-1}$ to each observed X_i is consistent for F.
The result can be extended when there is an observed covariate.

Generalized estimating equation (GEE)

The method of GEE is a powerful and general method of deriving point estimators, which includes many previously described methods as special cases, such as the method of moments, the least squares, the maximum likelihood, M-estimators, quasi-likelihoods,etc.

Assume that $X_1,...,X_n$ are independent (not necessarily identically distributed) random vectors, where the dimension of X_i is d_i, $i = 1,...,n$ ($\sup_i d_i < \infty$), and that we are interested in estimating θ, a k-vector of unknown parameters related to the unknown population.

Let $\Theta \subset \mathscr{R}^k$ be the range of θ, ψ_i be a Borel function from $\mathscr{R}^{d_i} \times \Theta$ to $\mathscr{R}^k$, $i = 1,...,n$, and

$$s_n(\gamma) = \sum_{i=1}^{n} \psi_i(X_i,\gamma), \qquad \gamma \in \Theta.$$

If θ is estimated by $\widehat{\theta} \in \Theta$ satisfying $s_n(\widehat{\theta}) = 0$, then $\widehat{\theta}$ is called a GEE estimator.

The equation $s_n(\gamma) = 0$ is called a GEE.

Motivation

Usually GEE's are chosen so that

$$E[s_n(\theta)] = \sum_{i=1}^{n} E[\psi_i(X_i, \theta)] = 0,$$

where the expectation E may be replaced by an asymptotic expectation defined in §2.5.2 if the exact expectation does not exist.

If this is true, then $\widehat{\theta}$ is motivated by the fact that $s_n(\widehat{\theta}) = 0$ is a sample analogue of $E[s_n(\theta)] = 0$.

Example

- The LSE: under model $X_i = \beta^\tau Z_i + \varepsilon_i$, the LSE of β is a solution of the equation

$$\sum_{i=1}^{n} \psi(X_i, \gamma) = \sum_{i=1}^{n} (X_i - \gamma^\tau Z_i) Z_i = 0$$

- The MLE: $\psi(x, \theta) = \partial \log f_\theta(x) / \partial \theta$

Proposition 5.2. (Consistency of GEE estimators)

Suppose that $X_1, ..., X_n$ are i.i.d. from F and $\psi_i \equiv \psi$, a bounded and continuous function from $\mathscr{R}^d \times \Theta$ to $\mathscr{R}^k$. Let $g(t) = \int \psi(x,t)dF(x)$. Suppose that $g(\theta) = 0$ and $\partial g(t)/\partial t$ exists and is of full rank at $t = \theta$. Then $\widehat{\theta}_n \to_p \theta$.

Other results can be found in the textbook.

Asymptotic normality of GEE estimators

If a GEE estimator $\widehat{\theta}$ is consistent, then its asymptotic normality can be established using Taylor's expansion

$$s_n(\widehat{\theta}) - s_n(\theta) = -s_n(\theta) \approx \nabla s_n(\theta)(\widehat{\theta} - \theta)$$

Then

$$\sqrt{n}(\widehat{\theta} - \theta) \approx -[\nabla s_n(\theta)]^{-1} \sqrt{n} s_n(\theta)$$

Since s_n is a sum of independent random vectors, an application of the CLT leads to

$$\sqrt{n} V_n^{-1/2} (\widehat{\theta} - \theta) \to_d N(0, I_k)$$

where $V_n = [\nabla s_n(\theta)]^{-1} \mathrm{Var}(s_n(\theta)) [\nabla s_n(\theta)]^{-1}$

Generalized method of moments (GMM)

In some cases, the number of equations is larger than k, the dimension of θ.
That is, we have more than necessary equations.
For example, in a parametric problem where a k-dimenisonal θ and finite $E(X_1^m)$, $m > k$, how do we apply the method of moments?
Suppose that we have a set of $m \geq k$ functions

$$\psi_j(x, \theta), \quad j = 1, ..., m$$

such that $E_\theta[\psi_j(X_i, \theta)] = 0$ for all j and ψ_j's are not linearly dependent, i.e., the $m \times m$ matrix whose (j, j')th element is $E_\theta[\psi_j(X_i, \theta)\psi_{j'}(X_i, \theta)]$ is positive definite, which can usually be achieved by eliminating some redundant functions when ψ_j's are linearly dependent.
Let

$$G_n(\theta) = \left(\frac{1}{n} \sum_{i=1}^n \psi_1(x_i, \theta), ..., \frac{1}{n} \sum_{i=1}^n \psi_m(x_i, \theta) \right)^\tau, \quad \theta \in \Theta$$

If $m = k$, a solution to $G_n(\theta) = 0$ is a GEE estimator.
If $m > k$, a solution to $G_n(\theta) = 0$ may not exist.

If a solution to $G_n(\theta) = 0$ does not exist because $m > k$, should we delete $m - k$ equations? If so, which ones should be removed?

Example

Consider the following estimation problem.
Let $\widehat{\phi}_j$ be a consistent estimator of ϕ_j, $j = 1, ..., m$.
Suppose that we have an addtional condition that

$$\phi_j = \alpha + \beta t_j, \quad j = 1, ..., m,$$

where α and β are unknown parameters and t_i's are known distinct constants.
If we obtain estimators $\widehat{\alpha}$ and $\widehat{\beta}$, then we can estimate ϕ_j by $\widehat{\alpha} + \widehat{\beta} t_j$, which may be better than $\widehat{\phi}_j$, $j = 1, ..., m$.
How do we estimate α and β?
If we choose two j_1 and j_2, then consistent estimators of α and β are

$$\widehat{\beta} = \frac{\widehat{\phi}_{j_1} - \widehat{\phi}_{j_2}}{t_{j_1} - t_{j_2}}, \quad \widehat{\alpha} = \widehat{\phi}_{j_1} + \widehat{\beta} t_{j_1}$$

But which j_1 and j_2 should we take?

Intuitively, we can use the least squares method: we treat $\widehat{\phi}_j$, $j = 1,...,m$, as "data" and fit a regression with t_j's as "covariate values". This is equivalent to minimizing

$$\frac{1}{m}\sum_{j=1}^{m}[\widehat{\phi}_j - (\alpha + \beta t_j)]^2 = G_n^{\tau}(\theta)G_n(\theta)$$

with

$$G_n(\theta) = \begin{pmatrix} \widehat{\phi}_1 \\ \vdots \\ \widehat{\phi}_m \end{pmatrix} - \begin{pmatrix} \alpha + \beta t_1 \\ \vdots \\ \alpha + \beta t_m \end{pmatrix}$$

Idea: if we cannot find α and β such that $\alpha + \beta t_j = \widehat{\phi}_j$ for all j, then we try to find α and β such that the least squares $G(\theta)^{\tau}G(\theta)$ is as small as possible.

In this example, the least squares estimators have explicit forms:

$$\widehat{\beta} = \frac{\sum_j (t_j - \bar{t})\widehat{\phi}_j}{\sum_j (t_j - \bar{t})^2}, \quad \widehat{\alpha} = \frac{1}{m}\sum_j \widehat{\phi}_j - \widehat{\beta}\bar{t}, \quad \bar{t} = \frac{1}{m}\sum_j t_j$$

The generalized method of moments (GMM)

If $G_n(\widehat{\theta}) = 0$, then $G_n^\tau(\widehat{\theta})G(\widehat{\theta}) = 0$ and is minimized; hence $G_n(\widehat{\theta}) = 0$ and $G_n^\tau(\widehat{\theta})G(\widehat{\theta}) = \min_\theta G_n^\tau(\theta)G(\theta)$ are equivalent.

If $G_n(\theta) = 0$ has no solution, we can still minimize $G_n^\tau(\theta)G_n(\theta)$, using a data driven procedure, not trying to determine which equations should be included.

GMM algorithm

A GMM estimate of θ can be obtained using the following two-step algorithm (the second step is to gain efficiency).

1. Obtain $\widehat{\theta}^{(1)}$ by minimizing $G_n^\tau(\theta)G_n(\theta)/2$ over $\theta \in \Theta$.
2. Let $\widehat{W}$ be the inverse matrix of the $m \times m$ matrix whose (j, j') element is equal to

$$\frac{1}{n}\sum_{i=1}^{n} \psi_j(x_i, \widehat{\theta}^{(1)})\psi_{j'}(x_i, \widehat{\theta}^{(1)})$$

The GMM estimate $\widehat{\theta}$ is obtained by minimizing

$$G_n^\tau(\theta)\widehat{W}G_n(\theta)/2 \quad \text{over } \theta \in \Theta$$

Asymptotic properties of GMM estimators

Using a similar argument to the one for GEE, we can show that there exists a sequence $\widehat{\theta}_n$ of GMM solutions that is consistent for θ.

Let $Q_n(\theta) = G_n^\tau(\theta) W G_n(\theta)/2$ and assume first that W is a fixed matrix. Then

$$-Q_n'(\theta) \approx Q_n''(\theta)(\widehat{\theta}_n - \theta)$$

where

$$Q_n'(\theta) = \partial Q_n(\theta)/\partial\theta = G_n'^\tau(\theta) W G_n(\theta),$$

$G_n'(\theta) = \partial G_n(\theta)/\partial\theta$ and

$$Q_n''(\theta) = \partial Q_n'(\theta)/\partial\theta = G_n''(\theta) W G_n(\theta) + G_n'^\tau(\theta) W G_n'(\theta)$$

$G_n''(\theta) = \partial^2 G_n(\theta)/\partial\theta\partial\theta^\tau$.
By the LLN and the fact that $G_n(\theta) \to_p 0$,

$$G_n'(\theta) \to_p B \quad \text{and} \quad Q_n''(\theta) \to_p B^\tau W B$$

By the CLT,

$$\sqrt{n} G_n(\theta) \to_d N(0, \Sigma) \quad \Sigma = \mathrm{Var}(g(X_1, \theta)).$$

Consequently,

$$\sqrt{n}(\widehat{\theta}_n - \theta) \to_d N(0, (B^\tau WB)^{-1} B^\tau W \Sigma WB (B^\tau WB)^{-1})$$

Note that

$$(B^\tau WB)^{-1} B^\tau W \Sigma WB (B^\tau WB)^{-1} \geq (B^\tau \Sigma^{-1} B)^{-1}$$

and the equality holds if and only if $W = \Sigma^{-1}$.

This implies that we should use $W = \Sigma^{-1}$.

But Σ is unknown.

Since $\widehat{\theta}_n$ is consistent with any W, we can first obtain an estimator $\widehat{\theta}_n^{(1)}$ with $W = I$ and then estimate Σ by

$$\widehat{\Sigma} = \frac{1}{n} \sum_{i=1}^n g(X_i, \widehat{\theta}_n^{(1)}) g(X_i, \widehat{\theta}_n^{(1)})^\tau$$

Then $\widehat{\Sigma}$ is a consistent estimator of Σ and we can use $W = \widehat{\Sigma}^{-1}$ in the 2nd step of GMM.

The resulting GMM estimator $\widehat{\theta}_n$ satisfies

$$\sqrt{n}(\widehat{\theta}_n - \theta) \to_d N(0, (B^\tau \Sigma^{-1} B)^{-1})$$

and is asymptotically the most efficient estimator among all GMM estimators with different choices of W.

Example

Let $X_1, ..., X_n$ be i.i.d. with $\theta = E(X_1)$, $\theta^2 = \text{Var}(X_1)$, and $E(X_1^4) < \infty$.
Consider moment estimators of θ.
If we use the first order moment, then the moment estimator of θ is the sample mean $\bar{X}$.
If we use the second order moment, then the moment estimator of θ is the solution of $2\theta^2 = M_2 = n^{-1}\sum_{i=1}^n X_i^2$.
Which estimator is more efficient (asymptotically)?
Note that the two equations

$$\bar{X} - \theta = 0, \qquad M_2 - 2\theta^2 = 0$$

cannot be solved simultaneously.
If we apply GMM, then we solve

$$\min_\theta (\bar{X} - \theta, M_2 - 2\theta^2) W \begin{pmatrix} \bar{X} - \theta \\ M_2 - 2\theta^2 \end{pmatrix} = 0$$

According to the GMM theory, this estimator is at least asymptotically as efficient as and is likely asymptotically more efficient than either $\bar{X}$ or $(M_2/2)^{1/2}$.