I got more data, my model is more refined, but my estimator is getting worse! Am I just dumb?

Xiao-Li Meng

Department of Statistics, Harvard University

- Meng and Xie (2012) "How can I find more information for my estimation?", for *Econometric Reviews*.

1. Do more (correct) data imply better estimator?
   - Not so even for LSE
   - Self Efficiency
   - Self-efficient Estimating Equation
2. Does making more (correct) assumptions help?
   - Not so even for bivariate normal
   - Preserving the Second Bartlett Identity
   - Using a Guiding Working Model (GWM)
3. Interaction between data pattern and model assumptions
   - AR(1) model
4. Is Jeffreys prior really non-informative?

Surprise: Even least-squares estimator ...

A Heteroscedastic Regression Model

$$Y_i = \beta X_i + \epsilon_i, \quad \epsilon_i \sim N(0, \sigma^2 X_i^\eta), \quad i = 1, \ldots, n$$

Surprise: Even least-squares estimator ...

A Heteroscedastic Regression Model

$$Y_i = \beta X_i + \epsilon_i, \quad \epsilon_i \sim N(0, \sigma^2 X_i^\eta), \quad i = 1, \ldots, n$$

Least-squares estimator:

$$\hat{\beta}^{LSE} = \frac{\sum_{i=1}^n X_i Y_i}{\sum_{i=1}^n X_i^2}$$

Surprise: Even least-squares estimator ...

A Heteroscedastic Regression Model

$$Y_i = \beta X_i + \epsilon_i, \quad \epsilon_i \sim N(0, \sigma^2 X_i^\eta), \quad i = 1, \ldots, n$$

Least-squares estimator:

$$\hat{\beta}^{LSE} = \frac{\sum_{i=1}^n X_i Y_i}{\sum_{i=1}^n X_i^2}$$

"Sandwich" estimator of var:

$$\hat{V}^{LSE} = \frac{\sum_{i=1}^n X_i^2 (Y_i - X_i \hat{\beta}^{LSE})^2}{\left[\sum_{i=1}^n X_i^2\right]^2}$$

Surprise: Even least-squares estimator ...

A Heteroscedastic Regression Model

$$Y_i = \beta X_i + \epsilon_i, \quad \epsilon_i \sim N(0, \sigma^2 X_i^\eta), \quad i = 1, \ldots, n$$

Least-squares estimator:

$$\hat{\beta}^{LSE} = \frac{\sum_{i=1}^n X_i Y_i}{\sum_{i=1}^n X_i^2}$$

"Sandwich" estimator of var:

$$\hat{V}^{LSE} = \frac{\sum_{i=1}^n X_i^2 (Y_i - X_i \hat{\beta}^{LSE})^2}{\left[\sum_{i=1}^n X_i^2\right]^2}$$

But $\hat{\beta}^{LSE}$ is not *self-efficient* (Meng, 1994) when $\eta \neq 0$:

$$\mathrm{V}(\hat{\beta}^{LSE} | X, \theta) = \sigma^2 \frac{\sum_{i=1}^n X_i^{2+\eta}}{\left[\sum_{i=1}^n X_i^2\right]^2}$$

Surprise: Even least-squares estimator ...

A Heteroscedastic Regression Model

$$Y_i = \beta X_i + \epsilon_i, \quad \epsilon_i \sim N(0, \sigma^2 X_i^\eta), \quad i = 1, \ldots, n$$

Least-squares estimator:

$$\hat{\beta}^{LSE} = \frac{\sum_{i=1}^n X_i Y_i}{\sum_{i=1}^n X_i^2}$$

"Sandwich" estimator of var:

$$\hat{V}^{LSE} = \frac{\sum_{i=1}^n X_i^2 (Y_i - X_i \hat{\beta}^{LSE})^2}{\left[\sum_{i=1}^n X_i^2\right]^2}$$

But $\hat{\beta}^{LSE}$ is not *self-efficient* (Meng, 1994) when $\eta \neq 0$:

$$\mathrm{V}(\hat{\beta}^{LSE} | X, \theta) = \sigma^2 \frac{\sum_{i=1}^n X_i^{2+\eta}}{\left[\sum_{i=1}^n X_i^2\right]^2}$$

Compare, when $\eta = 0$:

$$\mathrm{V}(\hat{\beta}^{LSE} | X, \theta) = \sigma^2 \frac{1}{\sum_{i=1}^n X_i^2}$$

What went wrong?

Take $\eta = 2$ and $X_i = (101 - i)^{-1}$, $i = 1, \ldots, 100$.

$$V_{64}^{LSE} = 0.0214 < V_{100}^{LSE} = 0.4049.$$

What went wrong?

Take $\eta = 2$ and $X_i = (101 - i)^{-1}$, $i = 1, \ldots, 100$.

$$V_{64}^{LSE} = 0.0214 < V_{100}^{LSE} = 0.4049.$$

But for MLE, the properly weighted LSE,

$$V_{64}^{MLE} = 0.0156 > V_{100}^{MLE} = 0.01.$$

What went wrong?

Take $\eta = 2$ and $X_i = (101 - i)^{-1}$, $i = 1, \ldots, 100$.

$$V_{64}^{LSE} = 0.0214 < V_{100}^{LSE} = 0.4049.$$

But for MLE, the properly weighted LSE,

$$V_{64}^{MLE} = 0.0156 > V_{100}^{MLE} = 0.01.$$

- With $n = 64$, LSE is 73% efficient compared with MLE.
- With $n = 100$, LSE is only 2.5% efficient.

What went wrong?

Take $\eta = 2$ and $X_i = (101 - i)^{-1}$, $i = 1, \ldots, 100$.

$$V_{64}^{LSE} = 0.0214 < V_{100}^{LSE} = 0.4049.$$

But for MLE, the properly weighted LSE,

$$V_{64}^{MLE} = 0.0156 > V_{100}^{MLE} = 0.01.$$

- With $n = 64$, LSE is 73% efficient compared with MLE.
- With $n = 100$, LSE is only 2.5% efficient.
- Those observations with large variabilities received more weight than they deserve.

Please throw away some data ...

Weighting the Heteroscedastic Regression Model: $W_i = X_i^{-\eta/2}$

$$W_i Y_i = \beta(W_i X_i) + \tilde{\epsilon}_i, \quad \tilde{\epsilon}_i \sim N(0, \sigma^2), \quad i = 1, \ldots, n$$

Please throw away some data ...

Weighting the Heteroscedastic Regression Model: $W_i = X_i^{-\eta/2}$

$$W_i Y_i = \beta(W_i X_i) + \tilde{\epsilon}_i, \quad \tilde{\epsilon}_i \sim N(0, \sigma^2), \quad i = 1, \ldots, n$$

$$\hat{\beta}_{MLE} = \frac{\sum_{i=1}^n X_i^{1-\eta} Y_i}{\sum_{i=1}^n X_i^{2-\eta}}$$

Please throw away some data ...

Weighting the Heteroscedastic Regression Model: $W_i = X_i^{-\eta/2}$

$$W_i Y_i = \beta(W_i X_i) + \tilde{\epsilon}_i, \quad \tilde{\epsilon}_i \sim N(0, \sigma^2), \quad i = 1, \ldots, n$$

$$\hat{\beta}_{MLE} = \frac{\sum_{i=1}^n X_i^{1-\eta} Y_i}{\sum_{i=1}^n X_i^{2-\eta}}$$

$$\mathrm{V}(\hat{\beta}_{MLE}|X, \theta) = \sigma^2 \frac{1}{\sum_{i=1}^n X_i^{2-\eta}}$$

Statistics

Please throw away some data ...

Weighting the Heteroscedastic Regression Model: $W_i = X_i^{-\eta/2}$

$$W_i Y_i = \beta(W_i X_i) + \tilde{\epsilon}_i, \quad \tilde{\epsilon}_i \sim N(0, \sigma^2), \quad i = 1, \ldots, n$$

$$\hat{\beta}_{MLE} = \frac{\sum_{i=1}^n X_i^{1-\eta} Y_i}{\sum_{i=1}^n X_i^{2-\eta}}$$

$$\mathrm{V}(\hat{\beta}_{MLE}|X, \theta) = \sigma^2 \frac{1}{\sum_{i=1}^n X_i^{2-\eta}}$$

So it is justifiable to throw away some data points if you don't know how to use them most effectively because

Please throw away some data ...

Weighting the Heteroscedastic Regression Model: $W_i = X_i^{-\eta/2}$

$$W_i Y_i = \beta(W_i X_i) + \tilde{\epsilon}_i, \quad \tilde{\epsilon}_i \sim N(0, \sigma^2), \quad i = 1, \ldots, n$$

$$\hat{\beta}_{MLE} = \frac{\sum_{i=1}^n X_i^{1-\eta} Y_i}{\sum_{i=1}^n X_i^{2-\eta}}$$

$$\mathrm{V}(\hat{\beta}_{MLE} | X, \theta) = \sigma^2 \frac{1}{\sum_{i=1}^n X_i^{2-\eta}}$$

So it is justifiable to throw away some data points if you don't know how to use them most effectively because

When the optimal W_i's have large variation, **setting small W_i's to zero better approximates the optimal weighting scheme than "blindly" using equal weights.**

Don't sue me ...

WHEN IT SEEMS DESIRABLE TO IGNORE DATA

Herman Chernoff
Massachusetts Institute of Technology

ABSTRACT

An experiment designed to detect the relative motion of two astronomical objects raised the problem of testing, against shift alternatives, the hypotehsis H_0 that two energy distributions are equivalent. The relevant data consist of independent Poisson counts X_{ij} with means $\lambda_j p_{ij} T_{ij}$ where λ_j is the intensity of radiation from the jth object, p_{ij} is the probability that a random photon from the jth object has energy in a small interval centered about e_i, and T_{ij} is the time duration allocated to the count X_{ij}. The hypothesis H_0 implies that $p_{i1} = p_{i2}$ for $i = 1, 2, \ldots, m$.

A natural test uses the statistic $\Sigma e_i(\hat{p}_{i2} - \hat{p}_{i1})$ where the $\hat{p}_{ij}$ are estimates of p_{ij}. For intervals where the p_{ij} were anticipated to be small, the experimenter chose small T_{ij} values

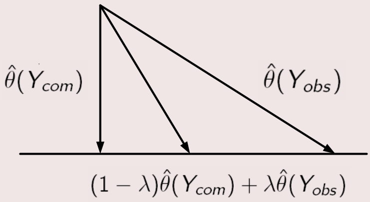
Need a bit more: Self-efficiency

Definition of Self-efficiency (Meng, 1994):

Let W_c be a data set, and W_o be a subset of W_c created by a selection mechanism. A statistical estimation procedure $\hat{\theta}(\cdot)$ for θ is said to be **self-efficient** (with respect to the selection mechanism) if $\theta(W_c)$ has the smallest MSE in the linear class $\{(1-\lambda)\hat{\theta}(W_c) + \lambda\hat{\theta}(W_o), \ \ \lambda \in R\}$.

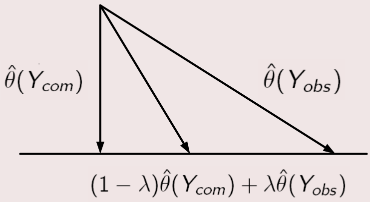
A Geometrical Characterization of Self-efficiency

Orthogonality



A Geometrical Characterization of Self-efficiency

Orthogonality

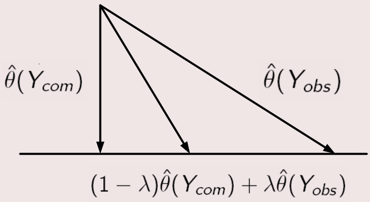


Pythagoras Identity

$$E(\hat{\theta}(Y_{obs}) - \theta)^2 = E(\hat{\theta}(Y_{com}) - \theta)^2 + E(\hat{\theta}(Y_{obs}) - \hat{\theta}(Y_{com}))^2$$

A Geometrical Characterization of Self-efficiency

Orthogonality



Pythagoras Identity

$$E(\hat{\theta}(Y_{obs}) - \theta)^2 = E(\hat{\theta}(Y_{com}) - \theta)^2 + E(\hat{\theta}(Y_{obs}) - \hat{\theta}(Y_{com}))^2$$

It is closely related to Rao-Blackwellization

$$\mathrm{E}[\hat{\theta}(Y_{obs})|\hat{\theta}(Y_{com})] = \hat{\theta}(Y_{com}).$$

Self-efficient Estimating Equations

Estimating Equation

Let the estimators $\hat{\theta}(Y_{com})$ and $\hat{\theta}(Y_{obs})$ be derived from

$$S_{com}(Y_{com};\theta) = 0 \quad \text{and} \quad S_{obs}(Y_{obs};\theta) = 0,$$

which satisfy certain regularity conditions.

Self-efficient Estimating Equations

Estimating Equation

Let the estimators $\hat{\theta}(Y_{com})$ and $\hat{\theta}(Y_{obs})$ be derived from

$$S_{com}(Y_{com};\theta) = 0 \quad \text{and} \quad S_{obs}(Y_{obs};\theta) = 0,$$

which satisfy certain regularity conditions.

The Characterization

The estimating procedure $\hat{\theta}(\cdot)$ is self-efficient if and only if

$$\left[E\left(-\frac{\partial S_{obs}}{\partial \theta}\right)\right]^{-1} E\left(S_{obs} S_{com}^{\top}\right) = \left[E\left(-\frac{\partial S_{com}}{\partial \theta}\right)\right]^{-1} E\left(S_{com} S_{com}^{\top}\right)$$

Self-efficient Estimating Equations

Estimating Equation

Let the estimators $\hat{\theta}(Y_{com})$ and $\hat{\theta}(Y_{obs})$ be derived from

$$S_{com}(Y_{com};\theta) = 0 \quad \text{and} \quad S_{obs}(Y_{obs};\theta) = 0,$$

which satisfy certain regularity conditions.

The Characterization

The estimating procedure $\hat{\theta}(\cdot)$ is self-efficient if and only if

$$\left[E\left(-\frac{\partial S_{obs}}{\partial \theta}\right)\right]^{-1} E\left(S_{obs}S_{com}^{\top}\right) = \left[E\left(-\frac{\partial S_{com}}{\partial \theta}\right)\right]^{-1} E\left(S_{com}S_{com}^{\top}\right)$$

- Can be viewed as a generalization of the second Bartlett identity.

Examples of Self-efficient Procedures

Holds for Arbitrary Pattern of the Observed Data (asymptotically)

- Maximum Likelihood Estimators
- Bayesian Estimators

Examples of Self-efficient Procedures

Holds for Arbitrary Pattern of the Observed Data (asymptotically)

- Maximum Likelihood Estimators
- Bayesian Estimators

Holds for "Regular Pattern" of the Observed Data

If Y_{com} is an i.i.d. sequence and Y_{obs} is a random subset of it (i.e., MCAR), i.e.,

$$\begin{aligned} Y_{com} &= (Y_1, \cdots, Y_n) \\ Y_{obs} &= (Y_{i_1}, \cdots, Y_{i_m}) . \end{aligned}$$

Then any estimating equation with the form

$$S(Y_{com}; \theta) = \sum_{i=1}^{n} U(Y_i; \theta)$$

is self-efficient.

Checking Self-efficiency of LSE under ARCH(1)

ARCH(1) Regression Model (Eagle, 1982)

$$Y_t = X_t\beta + \epsilon_t, \qquad t = 1, \ldots, N$$

$$\mathrm{V}[\epsilon_t|\epsilon_j, j < t] = \alpha_0 + \alpha_1\epsilon_{t-1}^2$$

with $\alpha_0 > 0$, $\alpha_1 \geq 0$, and $\epsilon_0 \equiv Y_0$ an unknown fixed parameter.

Checking Self-efficiency of LSE under ARCH(1)

ARCH(1) Regression Model (Eagle, 1982)

$$Y_t = X_t\beta + \epsilon_t, \qquad t = 1, \ldots, N$$

$$\mathrm{V}[\epsilon_t | \epsilon_j, j < t] = \alpha_0 + \alpha_1 \epsilon_{t-1}^2$$

with $\alpha_0 > 0$, $\alpha_1 \geq 0$, and $\epsilon_0 \equiv Y_0$ an unknown fixed parameter.

Given $Y_{obs} = \{Y_{t_1}, \ldots, Y_{t_n}\}$ and assume $X_t \equiv 1$.

- Let $V_{t_1,t_n} = \mathrm{V}(\hat{\beta}^{LSE})$ based on Y_{obs}, and $n/N \to r > 0$.

Checking Self-efficiency of LSE under ARCH(1)

ARCH(1) Regression Model (Eagle, 1982)

$$Y_t = X_t\beta + \epsilon_t, \qquad t = 1, \ldots, N$$

$$\mathrm{V}[\epsilon_t|\epsilon_j, j < t] = \alpha_0 + \alpha_1\epsilon_{t-1}^2$$

with $\alpha_0 > 0$, $\alpha_1 \geq 0$, and $\epsilon_0 \equiv Y_0$ an unknown fixed parameter.

Given $Y_{obs} = \{Y_{t_1}, \ldots, Y_{t_n}\}$ and assume $X_t \equiv 1$.

- Let $V_{t_1,t_n} = \mathrm{V}(\hat{\beta}^{LSE})$ based on Y_{obs}, and $n/N \to r > 0$.
- Then LSE is (asymptotically) self-efficient with respect to randomly selecting a consecutive segment if and only if

$$\lim_{n\to\infty} \frac{V_{1,N}}{V_{t_1,t_n}} = r.$$

- This holds if and only if $\alpha_1 < 1$.

So what about making more assumptions?

So what about making more assumptions?

Estimating correlation ρ based on bivariate normal data $\{(x_i, y_i), i = 1, \ldots, n\}$

So what about making more assumptions?

Estimating correlation ρ based on bivariate normal data $\{(x_i, y_i), i = 1, \ldots, n\}$

$$\hat{\rho}_n^{MLE} = \frac{s_{xy} - n\overline{xy}}{s_x s_y} \equiv h_n(\hat{\phi}_n, s_{xy}),$$

So what about making more assumptions?

Estimating correlation ρ based on bivariate normal data $\{(x_i, y_i), i = 1, \ldots, n\}$

$$\hat{\rho}_n^{MLE} = \frac{s_{xy} - n\overline{xy}}{s_x s_y} \equiv h_n(\hat{\phi}_n, s_{xy}),$$

where $\hat{\phi}_n = \{\overline{x}, \overline{y}, s_x, s_y\}$ is MLE for $\phi = \{\mu_x, \mu_y, \sigma_x, \sigma_y\}$.

So what about making more assumptions?

Estimating correlation ρ based on bivariate normal data $\{(x_i, y_i), i = 1, \ldots, n\}$

$$\hat{\rho}_n^{MLE} = \frac{s_{xy} - n\overline{xy}}{s_x s_y} \equiv h_n(\hat{\phi}_n, s_{xy}),$$

where $\hat{\phi}_n = \{\overline{x}, \overline{y}, s_x, s_y\}$ is MLE for $\phi = \{\mu_x, \mu_y, \sigma_x, \sigma_y\}$.

- What if we know both margins are $N(0, 1)$, and hence $\phi = \phi_0 = \{0, 0, 1, 1\}$?

So what about making more assumptions?

Estimating correlation ρ based on bivariate normal data $\{(x_i, y_i), i = 1, \ldots, n\}$

$$\hat{\rho}_n^{MLE} = \frac{s_{xy} - n\overline{xy}}{s_x s_y} \equiv h_n(\hat{\phi}_n, s_{xy}),$$

where $\hat{\phi}_n = \{\overline{x}, \overline{y}, s_x, s_y\}$ is MLE for $\phi = \{\mu_x, \mu_y, \sigma_x, \sigma_y\}$.

- What if we know both margins are $N(0,1)$, and hence $\phi = \phi_0 = \{0, 0, 1, 1\}$?
- Should it be obvious then to replace $\hat{\phi}_n$ by its true value?

$$\hat{\rho}_n^{PLUG} = \frac{s_{xy} - 0}{1} = \frac{1}{n}\sum_{i=1}^{n} x_i y_i.$$

So what about making more assumptions?

Estimating correlation ρ based on bivariate normal data $\{(x_i, y_i), i = 1, \ldots, n\}$

$$\hat{\rho}_n^{MLE} = \frac{s_{xy} - n\overline{xy}}{s_x s_y} \equiv h_n(\hat{\phi}_n, s_{xy}),$$

where $\hat{\phi}_n = \{\overline{x}, \overline{y}, s_x, s_y\}$ is MLE for $\phi = \{\mu_x, \mu_y, \sigma_x, \sigma_y\}$.

- What if we know both margins are $N(0,1)$, and hence $\phi = \phi_0 = \{0, 0, 1, 1\}$?
- Should it be obvious then to replace $\hat{\phi}_n$ by its true value?

$$\hat{\rho}_n^{PLUG} = \frac{s_{xy} - 0}{1} = \frac{1}{n}\sum_{i=1}^{n} x_i y_i.$$

- This is an unbiased but *terrible* estimator!

How bad is "partial plug-in"?

How bad is “partial plug-in”?

The MLE for ρ *conditioning* on $\phi = \phi_0$ is a root of

How bad is "partial plug-in"?

The MLE for ρ *conditioning* on $\phi = \phi_0$ is a root of

$$\rho^3 - s_{xy}\rho^2 - (s_x^2 + s_y^2 - 1)\rho + s_{xy} = 0.$$

How bad is "partial plug-in"?

The MLE for ρ *conditioning* on $\phi = \phi_0$ is a root of

$$\rho^3 - s_{xy}\rho^2 - (s_x^2 + s_y^2 - 1)\rho + s_{xy} = 0.$$

$$\mathrm{V}(\hat{\rho}_n^{CMLE}) = \frac{1}{n}\frac{(1-\rho^2)^2}{1+\rho^2}$$

How bad is "partial plug-in"?

The MLE for ρ *conditioning* on $\phi = \phi_0$ is a root of

$$\rho^3 - s_{xy}\rho^2 - (s_x^2 + s_y^2 - 1)\rho + s_{xy} = 0.$$

$$\mathrm{V}(\hat{\rho}_n^{CMLE}) = \frac{1}{n}\frac{(1-\rho^2)^2}{1+\rho^2}$$

Compare with

$$\mathrm{V}(\hat{\rho}_n^{MLE}) = \frac{1}{n}(1-\rho^2)^2$$

How bad is "partial plug-in"?

The MLE for ρ *conditioning* on $\phi = \phi_0$ is a root of

$$\rho^3 - s_{xy}\rho^2 - (s_x^2 + s_y^2 - 1)\rho + s_{xy} = 0.$$

$$\mathrm{V}(\hat{\rho}_n^{CMLE}) = \frac{1}{n}\frac{(1-\rho^2)^2}{1+\rho^2}$$

Compare with

$$\mathrm{V}(\hat{\rho}_n^{MLE}) = \frac{1}{n}(1-\rho^2)^2$$

$$\mathrm{V}(\hat{\rho}_n^{PLUG}) = \frac{1}{n}(1+\rho^2)$$

How bad is "partial plug-in"?

The MLE for ρ *conditioning* on $\phi = \phi_0$ is a root of

$$\rho^3 - s_{xy}\rho^2 - (s_x^2 + s_y^2 - 1)\rho + s_{xy} = 0.$$

$$\mathrm{V}(\hat{\rho}_n^{CMLE}) = \frac{1}{n}\frac{(1-\rho^2)^2}{1+\rho^2}$$

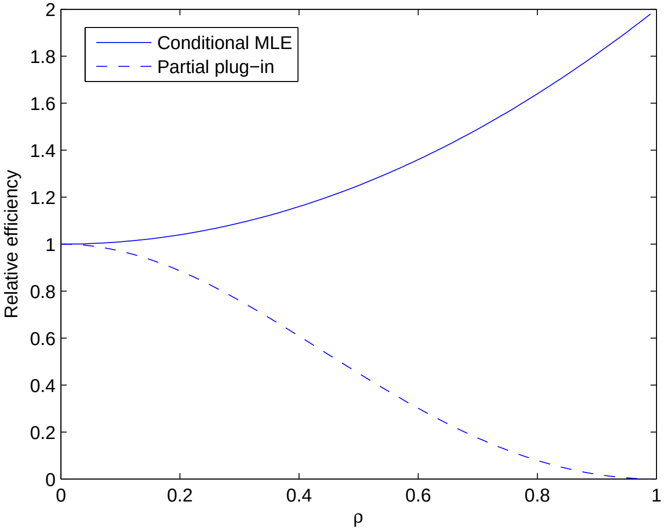
Compare with

$$\mathrm{V}(\hat{\rho}_n^{MLE}) = \frac{1}{n}(1-\rho^2)^2$$

$$\mathrm{V}(\hat{\rho}_n^{PLUG}) = \frac{1}{n}(1+\rho^2)$$

- NOTE: The validity of these estimators and their variances does not depend on the normality.

It made a difference, but in the wrong direction!



Puzzle of the Day:

- **How could the information about the marginal variances help estimate the correlation, which is invariant to scale (and location) transformation?**

How does MLE get it correctly?

Fisher information for $\theta = (\theta_1, \theta_2)$

$$I(\theta) = \begin{pmatrix} i_{11} & i_{12} \\ i_{21} & i_{22} \end{pmatrix} \quad \& \quad I^{-1}(\theta) = \begin{pmatrix} i^{(11)} & i^{(12)} \\ i^{(21)} & i^{(22)} \end{pmatrix}$$

How does MLE get it correctly?

Fisher information for $\theta = (\theta_1, \theta_2)$

$$I(\theta) = \begin{pmatrix} i_{11} & i_{12} \\ i_{21} & i_{22} \end{pmatrix} \quad \& \quad I^{-1}(\theta) = \begin{pmatrix} i^{(11)} & i^{(12)} \\ i^{(21)} & i^{(22)} \end{pmatrix}$$

Fisher information for θ_1:

$$I(\theta_1) = [i^{(11)}]^{-1} = i_{11} - \frac{i_{12}^2}{i_{22}}.$$

How does MLE get it correctly?

Fisher information for $\theta = (\theta_1, \theta_2)$

$$I(\theta) = \begin{pmatrix} i_{11} & i_{12} \\ i_{21} & i_{22} \end{pmatrix} \quad \& \quad I^{-1}(\theta) = \begin{pmatrix} i^{(11)} & i^{(12)} \\ i^{(21)} & i^{(22)} \end{pmatrix}$$

Fisher information for θ_1:

$$I(\theta_1) = [i^{(11)}]^{-1} = i_{11} - \frac{i_{12}^2}{i_{22}}.$$

Fisher information for θ_1 *given* θ_2:

$$I(\theta_1|\theta_2) = i_{11}.$$

How does MLE get it correctly?

Fisher information for $\theta = (\theta_1, \theta_2)$

$$I(\theta) = \begin{pmatrix} i_{11} & i_{12} \\ i_{21} & i_{22} \end{pmatrix} \quad \& \quad I^{-1}(\theta) = \begin{pmatrix} i^{(11)} & i^{(12)} \\ i^{(21)} & i^{(22)} \end{pmatrix}$$

Fisher information for θ_1:

$$I(\theta_1) = [i^{(11)}]^{-1} = i_{11} - \frac{i_{12}^2}{i_{22}}.$$

Fisher information for θ_1 *given* θ_2:

$$I(\theta_1|\theta_2) = i_{11}.$$

The *gain* in information for θ_1 due to the knowledge of θ_2:

$$\mathcal{G}(\theta_1|\theta_2) \equiv I(\theta_1|\theta_2) - I(\theta_1) = \frac{i_{12}^2}{i_{22}} = i_{11} r_{1,2}^2 \geq 0.$$

Making Bartlett adjustment before plug-in

Adjust EE $S(Y;\theta)$ to $S_A(Y;\theta) = A(\theta)S(Y;\theta)$ to ensure Bartlett Identity (BI)

$$V(S_A(Y;\theta)) = \mathrm{E}\left[-\frac{\partial S_A(Y;\theta)}{\partial\theta}\right].$$

This implies $A(\theta) = J^\top(\theta)V^{-1}(S(Y;\theta))$ where

$$J^\top(\theta) = \mathrm{E}\left[-\left(\frac{\partial S(Y;\theta)}{\partial\theta}\right)^\top\right].$$

Making Bartlett adjustment before plug-in

Adjust EE $S(Y;\theta)$ to $S_A(Y;\theta) = A(\theta)S(Y;\theta)$ to ensure Bartlett Identity (BI)

$$V(S_A(Y;\theta)) = \mathrm{E}\left[-\frac{\partial S_A(Y;\theta)}{\partial\theta}\right].$$

This implies $A(\theta) = J^\top(\theta)V^{-1}(S(Y;\theta))$ where

$$J^\top(\theta) = \mathrm{E}\left[-\left(\frac{\partial S(Y;\theta)}{\partial\theta}\right)^\top\right].$$

- To plug-in $\theta_2 = \theta_2^{(0)}$, take the rows of $S_A(Y;\theta)$ corresponding to $\left(\frac{\partial S(Y;\theta)}{\partial\theta_1}\right)^\top$ in $J^\top(\theta)$ as the estimating equations for θ_1, and then plug-in $\theta_2 = \theta_2^{(0)}$.

A conditional normal working model

Let $Y = \{Y_{t_1}, \dots, Y_{t_n}\}$, $\mathcal{F}_{j-1} = \sigma\{Y_{t_i}, i < j\}$, and assume

$$Y_{t_j} | \mathcal{F}_{j-1}; \theta \sim N(\mu_j(\theta), \tau_j^2(\theta)), \quad j = 1, \dots, n.$$

A conditional normal working model

Let $Y = \{Y_{t_1}, \ldots, Y_{t_n}\}$, $\mathcal{F}_{j-1} = \sigma\{Y_{t_i}, i < j\}$, and assume

$$Y_{t_j} | \mathcal{F}_{j-1}; \theta \sim N(\mu_j(\theta), \tau_j^2(\theta)), \quad j = 1, \ldots, n.$$

Let $d_j(\theta) = \frac{Y_{t_j} - \mu_j(\theta)}{\tau_j(\theta)}$, then the MLE of θ is a root of

$$S_n(\theta) \equiv S_n^{(\mu)}(\theta) + S_n^{(\tau)}(\theta) = 0$$

A conditional normal working model

Let $Y = \{Y_{t_1}, \ldots, Y_{t_n}\}$, $\mathcal{F}_{j-1} = \sigma\{Y_{t_i}, i < j\}$, and assume

$$Y_{t_j} | \mathcal{F}_{j-1}; \theta \sim N(\mu_j(\theta), \tau_j^2(\theta)), \quad j = 1, \ldots, n.$$

Let $d_j(\theta) = \frac{Y_{t_j} - \mu_j(\theta)}{\tau_j(\theta)}$, then the MLE of θ is a root of

$$S_n(\theta) \equiv S_n^{(\mu)}(\theta) + S_n^{(\tau)}(\theta) = 0$$

$$S_n^{(\mu)}(\theta) = \sum_{j=1}^{n} d_j(\theta) \left[\frac{\mu_j'(\theta)}{\tau_j(\theta)}\right]; \quad S_n^{(\tau)}(\theta) = \sum_{j=1}^{n} \left(d_j^2(\theta) - 1\right) \left[\frac{\tau_j'(\theta)}{\tau_j(\theta)}\right].$$

A conditional normal working model

Let $Y = \{Y_{t_1}, \ldots, Y_{t_n}\}$, $\mathcal{F}_{j-1} = \sigma\{Y_{t_i}, i < j\}$, and assume

$$Y_{t_j}|\mathcal{F}_{j-1}; \theta \sim N(\mu_j(\theta), \tau_j^2(\theta)), \quad j = 1, \ldots, n.$$

Let $d_j(\theta) = \frac{Y_{t_j} - \mu_j(\theta)}{\tau_j(\theta)}$, then the MLE of θ is a root of

$$S_n(\theta) \equiv S_n^{(\mu)}(\theta) + S_n^{(\tau)}(\theta) = 0$$

$$S_n^{(\mu)}(\theta) = \sum_{j=1}^{n} d_j(\theta) \left[\frac{\mu_j'(\theta)}{\tau_j(\theta)}\right]; \ S_n^{(\tau)}(\theta) = \sum_{j=1}^{n} \left(d_j^2(\theta) - 1\right) \left[\frac{\tau_j'(\theta)}{\tau_j(\theta)}\right].$$

- $\mathrm{E}[S_n^{(\mu)}(\theta)] = 0$ only requires $\mu_j(\theta)$ correctly specified.

A conditional normal working model

Let $Y = \{Y_{t_1}, \ldots, Y_{t_n}\}$, $\mathcal{F}_{j-1} = \sigma\{Y_{t_i}, i < j\}$, and assume

$$Y_{t_j} | \mathcal{F}_{j-1}; \theta \sim N(\mu_j(\theta), \tau_j^2(\theta)), \quad j = 1, \ldots, n.$$

Let $d_j(\theta) = \frac{Y_{t_j} - \mu_j(\theta)}{\tau_j(\theta)}$, then the MLE of θ is a root of

$$S_n(\theta) \equiv S_n^{(\mu)}(\theta) + S_n^{(\tau)}(\theta) = 0$$

$$S_n^{(\mu)}(\theta) = \sum_{j=1}^{n} d_j(\theta) \left[\frac{\mu_j'(\theta)}{\tau_j(\theta)}\right]; \quad S_n^{(\tau)}(\theta) = \sum_{j=1}^{n} \left(d_j^2(\theta) - 1\right) \left[\frac{\tau_j'(\theta)}{\tau_j(\theta)}\right].$$

- $\mathrm{E}[S_n^{(\mu)}(\theta)] = 0$ only requires $\mu_j(\theta)$ correctly specified.
- $\mathrm{E}[S_n^{(\tau)}(\theta)] = 0$ also requires $\tau_j(\theta)$ correctly specified.

A conditional normal working model

Let $Y = \{Y_{t_1}, \ldots, Y_{t_n}\}$, $\mathcal{F}_{j-1} = \sigma\{Y_{t_i}, i < j\}$, and assume

$$Y_{t_j} | \mathcal{F}_{j-1}; \theta \sim N(\mu_j(\theta), \tau_j^2(\theta)), \quad j = 1, \ldots, n.$$

Let $d_j(\theta) = \frac{Y_{t_j} - \mu_j(\theta)}{\tau_j(\theta)}$, then the MLE of θ is a root of

$$S_n(\theta) \equiv S_n^{(\mu)}(\theta) + S_n^{(\tau)}(\theta) = 0$$

$$S_n^{(\mu)}(\theta) = \sum_{j=1}^{n} d_j(\theta) \left[\frac{\mu_j'(\theta)}{\tau_j(\theta)}\right]; \quad S_n^{(\tau)}(\theta) = \sum_{j=1}^{n} (d_j^2(\theta) - 1) \left[\frac{\tau_j'(\theta)}{\tau_j(\theta)}\right].$$

- $\mathrm{E}[S_n^{(\mu)}(\theta)] = 0$ only requires $\mu_j(\theta)$ correctly specified.
- $\mathrm{E}[S_n^{(\tau)}(\theta)] = 0$ also requires $\tau_j(\theta)$ correctly specified.
- $\mathrm{Cov}(S_n^{(\mu)}(\theta), S_n^{(\tau)}(\theta)) = 0$ as long as $\mathrm{E}[d_j^3 | \mathcal{F}_{j-1}] = 0$, under which $S_n(\theta)$ also satisfies BI.

Two Kinds of Information Additivity

Let $\xi_j(\theta) = \mu_j'(\theta)/\tau_j(\theta)$, $\eta_j(\theta) = \tau_j'(\theta)/\tau_j(\theta)$,

$$\mathcal{I}_n(\theta) = \mathcal{I}_n^{(\mu)}(\theta) + \mathcal{I}_n^{(\tau)}(\theta). \qquad (A_1)$$

$$\mathcal{I}_n^{(\mu)}(\theta) = \sum_{j=1}^{n} \mathrm{E}[\xi_j(\theta)\xi_j^\top(\theta)]; \qquad (A_2)$$

$$\mathcal{I}_n^{(\tau)}(\theta) = 2\sum_{j=1}^{n} \mathrm{E}[\eta_j(\theta)\eta_j^\top(\theta)] \qquad (A_2)$$

Two Kinds of Information Additivity

Let $\xi_j(\theta) = \mu_j'(\theta)/\tau_j(\theta)$, $\eta_j(\theta) = \tau_j'(\theta)/\tau_j(\theta)$,

$$\mathcal{I}_n(\theta) = \mathcal{I}_n^{(\mu)}(\theta) + \mathcal{I}_n^{(\tau)}(\theta). \qquad (A_1)$$

$$\mathcal{I}_n^{(\mu)}(\theta) = \sum_{j=1}^{n} \mathrm{E}[\xi_j(\theta)\xi_j^\top(\theta)]; \qquad (A_2)$$

$$\mathcal{I}_n^{(\tau)}(\theta) = 2\sum_{j=1}^{n} \mathrm{E}[\eta_j(\theta)\eta_j^\top(\theta)] \qquad (A_2)$$

- Model-Reduction Additivity: (A_1)

Two Kinds of Information Additivity

Let $\xi_j(\theta) = \mu_j'(\theta)/\tau_j(\theta)$, $\eta_j(\theta) = \tau_j'(\theta)/\tau_j(\theta)$,

$$\mathcal{I}_n(\theta) = \mathcal{I}_n^{(\mu)}(\theta) + \mathcal{I}_n^{(\tau)}(\theta). \qquad (A_1)$$

$$\mathcal{I}_n^{(\mu)}(\theta) = \sum_{j=1}^{n} \mathrm{E}[\xi_j(\theta)\xi_j^\top(\theta)]; \qquad (A_2)$$

$$\mathcal{I}_n^{(\tau)}(\theta) = 2\sum_{j=1}^{n} \mathrm{E}[\eta_j(\theta)\eta_j^\top(\theta)] \qquad (A_2)$$

- Model-Reduction Additivity: (A_1)
- Data-Augmentation Additivity: (A_2)

AR(1) model with spacing

AR(1) model with spacing

AR(1) model for the complete data;

$$Y_t = \rho Y_{t-1} + \epsilon_t, \quad \epsilon_t \overset{\text{iid}}{\sim} N(0, \sigma^2), \quad t = 1, \ldots, N,$$

AR(1) model with spacing

AR(1) model for the complete data;

$$Y_t = \rho Y_{t-1} + \epsilon_t, \quad \epsilon_t \overset{\text{iid}}{\sim} N(0, \sigma^2), \quad t = 1, \ldots, N,$$

$$\Longrightarrow Y_t | Y_s \sim N(\rho^{t-s} Y_s,\ k_{t-s}(\rho)\sigma^2),\ \text{with } k_l(\rho) = \sum_{j=0}^{l-1} \rho^{2j}.$$

AR(1) model with spacing

AR(1) model for the complete data;

$$Y_t = \rho Y_{t-1} + \epsilon_t, \quad \epsilon_t \overset{\text{iid}}{\sim} N(0, \sigma^2), \quad t = 1, \ldots, N,$$

$$\Longrightarrow Y_t | Y_s \sim N(\rho^{t-s} Y_s,\ k_{t-s}(\rho)\sigma^2), \text{ with } k_l(\rho) = \sum_{j=0}^{l-1} \rho^{2j}.$$

For observed data $\{Y_{t_j}, j = 1, \ldots, n\}$; $\theta = (\rho, \sigma^2, Y_0)$

$$\mathcal{I}_n^{(\mu)}(\theta) = \frac{1}{\sigma^2} \begin{pmatrix} Y_0^2 A_{1,n}(\rho) + \sigma^2 B_{2,n}(\rho) & 0 & t_1\gamma_1(\rho) Y_0 \\ 0 & 0 & 0 \\ t_1\gamma_1(\rho) Y_0 & 0 & \rho\gamma_1(\rho) \end{pmatrix}$$

$$\mathcal{I}_n^{(\tau)}(\theta) = \frac{1}{2\sigma^4} \begin{pmatrix} \sigma^4 \sum_{j=1}^n \delta_j^2(\rho) & \sigma^2 \sum_{j=1}^n \delta_j(\rho) & 0 \\ \sigma^2 \sum_{j=1}^n \delta_j(\rho) & n & 0 \\ 0 & 0 & 0 \end{pmatrix}$$

Suppose $t_{j+1} - t_j = s$. Then

- $$\mathcal{R}_n^{(s=1)}(\rho|\sigma^2) \equiv \frac{\mathcal{G}_n(\rho|\sigma^2)}{\mathcal{I}_n(\rho)} < \frac{1}{n-1},$$

Interaction between Data Pattern and Model Assumptions

Suppose $t_{j+1} - t_j = s$. Then

- $$\mathcal{R}_n^{(s=1)}(\rho|\sigma^2) \equiv \frac{\mathcal{G}_n(\rho|\sigma^2)}{\mathcal{I}_n(\rho)} < \frac{1}{n-1},$$

- $$\mathcal{R}_\infty^{(s=2)}(\rho|\sigma^2) = \frac{1-\rho^2}{2(1+\rho^2)} \leq \frac{1}{2}$$

Interaction between Data Pattern and Model Assumptions

Suppose $t_{j+1} - t_j = s$. Then

- $$\mathcal{R}_n^{(s=1)}(\rho|\sigma^2) \equiv \frac{\mathcal{G}_n(\rho|\sigma^2)}{\mathcal{I}_n(\rho)} < \frac{1}{n-1},$$

- $$\mathcal{R}_\infty^{(s=2)}(\rho|\sigma^2) = \frac{1-\rho^2}{2(1+\rho^2)} \leq \frac{1}{2}$$

- $$\mathcal{R}_\infty^{(s=3)}(\rho|\sigma^2) = \frac{2(1-\rho^2)(1+2\rho^2)^2}{9\rho^2(\rho^4+\rho^2+1)}; \textit{unbounded!}$$

Interaction between Data Pattern and Model Assumptions

Suppose $t_{j+1} - t_j = s$. Then

- $$\mathcal{R}_n^{(s=1)}(\rho|\sigma^2) \equiv \frac{\mathcal{G}_n(\rho|\sigma^2)}{\mathcal{I}_n(\rho)} < \frac{1}{n-1},$$

- $$\mathcal{R}_\infty^{(s=2)}(\rho|\sigma^2) = \frac{1-\rho^2}{2(1+\rho^2)} \leq \frac{1}{2}$$

- $$\mathcal{R}_\infty^{(s=3)}(\rho|\sigma^2) = \frac{2(1-\rho^2)(1+2\rho^2)^2}{9\rho^2(\rho^4+\rho^2+1)}; \textit{unbounded!}$$

Because

$$\mathrm{V}(Y_{t+s}|Y_t) = (1+\rho^2+\cdots+\rho^{2(s-1)})\sigma^2$$

Optimal spacing for estimating ρ with given n

For large n

$$\lim_{n\to\infty} \frac{\mathcal{I}_n^{(s)}(\rho)}{n} = \frac{s^2\rho^{2(s-1)}}{1-\rho^{2s}} \equiv H(s,\rho).$$

Optimal spacing for estimating ρ with given n

For large n

$$\lim_{n\to\infty} \frac{\mathcal{I}_n^{(s)}(\rho)}{n} = \frac{s^2\rho^{2(s-1)}}{1-\rho^{2s}} \equiv H(s,\rho).$$

For given ρ, $H(s,\rho)$ is maximized at

$$s_{max}(\rho) = \frac{1.59362...}{-\log\rho^2}.$$

Optimal spacing for estimating ρ with given n

For large n

$$\lim_{n\to\infty} \frac{\mathcal{I}_n^{(s)}(\rho)}{n} = \frac{s^2\rho^{2(s-1)}}{1-\rho^{2s}} \equiv H(s,\rho).$$

For given ρ, $H(s,\rho)$ is maximized at

$$s_{max}(\rho) = \frac{1.59362...}{-\log \rho^2}.$$

For optimizing integer s, we have

- $s = 1$ is optimal as long as $|\rho| \le 3^{-1/2} = 0.577...$;

Optimal spacing for estimating ρ with given n

For large n

$$\lim_{n\to\infty} \frac{\mathcal{I}_n^{(s)}(\rho)}{n} = \frac{s^2\rho^{2(s-1)}}{1-\rho^{2s}} \equiv H(s,\rho).$$

For given ρ, $H(s,\rho)$ is maximized at

$$s_{max}(\rho) = \frac{1.59362...}{-\log\rho^2}.$$

For optimizing integer s, we have

- $s = 1$ is optimal as long as $|\rho| \leq 3^{-1/2} = 0.577...$;
- $s = 2$ is optimal when

$$0.577 < |\rho| \leq \sqrt{\frac{\sqrt{105}-5}{10}} = 0.724...$$

Optimal Spacing and Relative Gain

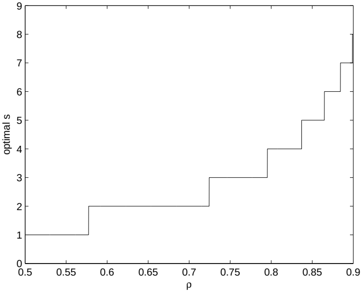


Figure: Optimal spacing.

Optimal Spacing and Relative Gain

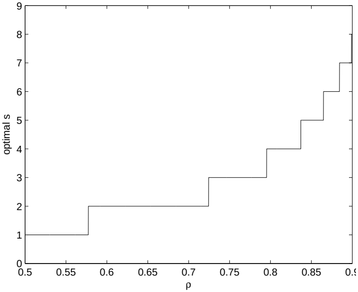


Figure: Optimal spacing.

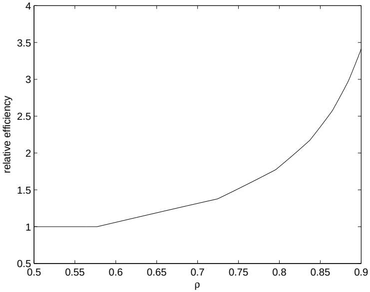


Figure: Relative efficiency.

Log Relative Efficiency: Fixed n vs Fixed N

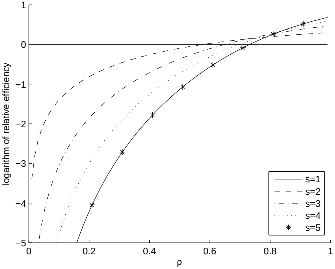


Figure: Fixed n; $N = ns$

Log Relative Efficiency: Fixed n vs Fixed N

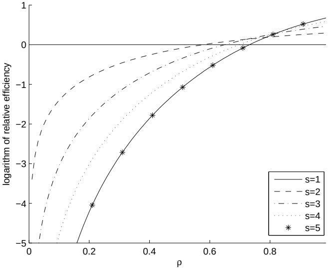


Figure: Fixed n; $N = ns$

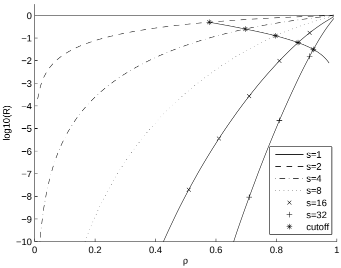


Figure: Fixed N, $n = N/s$

Bayesians are not automatically immune ...

Bayesians are not automatically immune ...

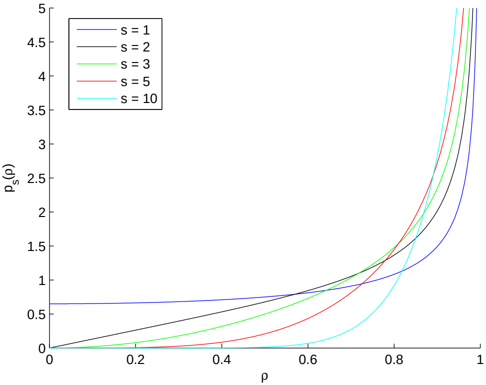
For large n, $\mathcal{I}_n \approx ns^2\rho^{s-1}(1-\rho^{2s})^{-1/2}$, where s is the spacing.

- Jeffreys prior is the same as $\rho^{2s} \sim \mathrm{Beta}(1/2, 1/2)$.

Bayesians are not automatically immune ...

For large n, $\mathcal{I}_n \approx ns^2\rho^{s-1}(1-\rho^{2s})^{-1/2}$, where s is the spacing.

- Jeffreys prior is the same as $\rho^{2s} \sim \mathrm{Beta}(1/2, 1/2)$.



Chinglish of the Day ...

Chinglish of the Day ...

- If it seems like a good Lintuition, don't jump on it!