Estimation after Model Selection

Vanja M. Dukić
Department of Health Studies
University of Chicago
E-Mail: vanja@uchicago.edu

Edsel A. Peña*
Department of Statistics
University of South Carolina
E-Mail: pena@stat.sc.edu

ENAR 2003 Talk
March 31, 2003
Tampa Bay, FL

Research support from NSF

Motivating Situations

- Suppose you have a random sample $\boldsymbol{X} = (X_1, X_2, \ldots, X_n)$ (possibly censored) from an unknown distribution F which belongs to either the Weibull class or the gamma class. What is the best way to estimate $F(t)$ or some other parameter of interest?

- Suppose it is known that the unknown df F belongs to either of p models $\mathcal{M}_1, \mathcal{M}_2, \ldots, \mathcal{M}_p$, which are possibly nested. What is the best way of estimating a parameter common to each of these models?

Intuitive Strategies

Strategy I: Utilize estimators developed under larger model $\mathcal{M}$, or implement a fully nonparametric approach.

Strategy II (Classical): [**Step 1** (Model Selection):] Choose most plausible model using the data, possibly via information measures. [**Step 2** (Inference):] Use estimators in the chosen sub-model, but with these estimators still using the same data $\boldsymbol{X}$.

Strategy III (Bayesian): Determine adaptively (i.e., using $\boldsymbol{X}$) the plausibility of each of the sub-models, and form a weighted combination of the sub-model estimators or tests. Referred also as **model averaging**.

Relevance and Issues

What are the consequences of first selecting a sub-model and then performing inference such as estimation or testing hypothesis, with these two steps utilizing the *same* sample data (i.e., *double-dipping)*?

Is it always better to do model-averaging, that is, a Bayesian framework, or equivalently, under what circumstances is model averaging preferable over a classical two-step approach?

When the number of possible models increases, would it be better to simply utilize a wider, possibly nonparametric, model?

A Concrete Gaussian Model

- **Data:**

  $$\boldsymbol{X} \equiv (X_1, X_2, \ldots, X_n) \text{ IID } F \in \mathcal{M} = \left\{ N(\mu, \sigma^2) : \mu \in \Re, \sigma^2 > 0 \right\}$$

- Uniformly minimum variance unbiased (UMVU) estimator of σ^2 is the sample variance

  $$\widehat{\sigma}^2_{UMVU} = S^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X})^2.$$

- Decision-theoretic framework with loss function

  $$L_1(\widehat{\sigma}^2, (\mu, \sigma^2)) \;=\; \left(\frac{\widehat{\sigma}^2 - \sigma^2}{\sigma^2} \right)^2.$$

- Risk function: For the quadratic loss L_1,

$$\text{Risk}(\hat{\sigma}^2) = \text{Variance}\left(\frac{\hat{\sigma}^2}{\sigma^2}\right) + \left[\ \text{Bias}\left(\frac{\hat{\sigma}^2}{\sigma^2}\right)\right]^2$$

- S^2 is *not* the best. Dominated by ML and the minimum risk equivariant (MRE) estimators:

$$\hat{\sigma}^2_{MLE} = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2$$

$$\hat{\sigma}^2_{MRE} = \left(\frac{n}{n+1}\right)\hat{\sigma}^2_{MLE}$$

Model $\mathcal{M}_p$: Our 'Test' Model

- Suppose we do not know the exact value of μ, but we do know it is one of p possible values. This leads to model $\mathcal{M}_p$:

$$\mathcal{M}_p = \left\{ N(\mu, \sigma^2) : \mu \in \{\mu_1, \ldots, \mu_p\}, \sigma^2 > 0 \right\}$$

  where $\mu_1, \mu_2, \ldots, \mu_p$ are known constants.

- Under $\mathcal{M}_p$, how should we estimate σ^2? What are the consequences of using the estimators developed under $\mathcal{M}$?

- Can we exploit structure of $\mathcal{M}_p$ to obtain better estimators of σ^2?

$\boxed{\text{Classical Estimators Under } \mathcal{M}_p}$

- **Sub-Model MLEs and MREs:**

$$\hat{\sigma}_i^2 = \frac{1}{n}\sum_{j=1}^{n}(X_j - \mu_i)^2; \quad \hat{\sigma}^2_{MRE,i} = \frac{1}{n+2}\sum_{j=1}^{n}(X_j - \mu_i)^2$$

- **Model Selector:** $\widehat{M} = \widehat{M}(\boldsymbol{X})$

$$\widehat{M} = \arg\min_{1 \le i \le p} \hat{\sigma}_i^2 = \arg\min_{1 \le i \le p} |\bar{X} - \mu_i|.$$

- $\widehat{M}$ chooses the sub-model leading to the smallest estimate of σ^2, or whose mean is closest to the sample mean.

- **MLE** of σ^2 under $\mathcal{M}_p$ (a two-step *adaptive* estimator):

$$\hat{\sigma}^2_{p,MLE} = \hat{\sigma}^2_{\widehat{M}} = \sum_{i=1}^{p} I\{\widehat{M} = i\}\hat{\sigma}^2_i.$$

- **An alternative Estimator:** Use the sub-model's MRE to obtain

$$\hat{\sigma}^2_{p,MRE} = \hat{\sigma}^2_{MRE,\widehat{M}} = \sum_{i=1}^{p} I\{\widehat{M} = i\}\hat{\sigma}^2_{MRE,i}.$$

- Properties of *adaptive* estimators not easily obtainable due to interplay between the model selector $\widehat{M}$ and the sub-model estimator.

Bayes Estimators Under $\mathcal{M}_p$

- **Joint Prior for (μ, σ^2):**
  - Independent priors
  - Prior for μ: Multinomial$(1, \tilde{\theta})$
  - Prior for σ^2: Inverted Gamma(κ, β)

- **Posterior Probabilities of Sub-Models:**

$$\theta_i(\boldsymbol{x}) = \frac{\tilde{\theta}_i \Big(n\widehat{\sigma}_i^2/2 + \beta \Big)^{-(n/2+\kappa-1)}}{\sum_{j=1}^{p} \tilde{\theta}_j \Big(n\widehat{\sigma}_j^2/2 + \beta \Big)^{-(n/2+\kappa-1)}}$$

- **Posterior Density of σ^2:**

$$\pi(\sigma^2 \mid \boldsymbol{x}) = C \sum_{i=1}^{p} \tilde{\theta}_i \left(\frac{1}{\sigma^2}\right)^{-(\kappa+n/2)} \exp\left[-\frac{1}{\sigma^2}\left(n\hat{\sigma}_i^2/2 + \beta\right)\right].$$

- **Bayes (Weighted) Estimator of σ^2:**

$$\hat{\sigma}^2_{p,Bayes}(\boldsymbol{X}) = \sum_{i=1}^{p} \theta_i(\boldsymbol{X}) \times \left\{\left(\frac{n}{n+2(\kappa-2)}\right)\hat{\sigma}_i^2 + \left(\frac{2(\kappa-2)}{n+2(\kappa-2)}\right)\left(\frac{\beta}{\kappa-2}\right)\right\}.$$

- **Non-Informative Priors:** Uniform prior for sub-models: $\tilde{\theta}_i = 1/p, i = 1, 2, \ldots, p$; $\beta \to 0$.

- One particular limiting Bayes estimator is:

$$\hat{\sigma}^2_{p,LB1} = \sum_{i=1}^{p} \left[\frac{(\hat{\sigma}^2_i)^{-n/2}}{\sum_{j=1}^{p} (\hat{\sigma}^2_j)^{-n/2}} \right] \hat{\sigma}^2_i$$

  an *adaptively* weighted estimator formed from the sub-model estimators.

- But, based on the simulation studies, a better one is that formed from the sub-model MREs:

$$\hat{\sigma}^2_{p,PLB1} = \left(\frac{n}{n+2} \right) \hat{\sigma}^2_{p,LB1}$$

$\boxed{\textbf{Comparing the Estimators}}$

- $R\left(\hat{\sigma}^2_{UMVU}, (\mu, \sigma^2)\right) = \frac{2}{n-1}$.

- $R\left(\hat{\sigma}^2_{MRE}, (\mu, \sigma^2)\right) = \frac{2}{n+1}$.

- Efficiency measure relative to $\hat{\sigma}^2_{UMVU}$:

$$\text{Eff}(\hat{\sigma}^2 : \hat{\sigma}^2_{UMVU}) = \frac{R(\hat{\sigma}^2_{UMVU}, (\mu, \sigma^2))}{R(\hat{\sigma}^2, (\mu, \sigma^2))}.$$

- $\text{Eff}(\hat{\sigma}^2_{MRE} : \hat{\sigma}^2_{UMVU}) = \frac{n+1}{n-1} = 1 + \frac{2}{n-1}$.

$\boxed{\text{Properties of } \mathcal{M}_p\text{-Based Estimators}}$

- **Notation:** Let $Z \sim N(0,1)$ and with μ_{i_0} the *true mean*, define
$$\boldsymbol{\Delta} = \frac{\boldsymbol{\mu} - \mu_{i_0}\mathbf{1}}{\sigma}.$$

- **Proposition:** Under $\mathcal{M}_p$,
$$\frac{\widehat{\sigma}_i^2}{\sigma^2} \stackrel{d}{=} \frac{1}{n}\left(W + V_i^2\right), i = 1, 2, \ldots, p;$$
with W and $\boldsymbol{V}$ independent, and
$$W \sim \chi^2_{n-1};$$
$$\boldsymbol{V} = Z\mathbf{1} - \sqrt{n}\boldsymbol{\Delta} \sim N_p(-\sqrt{n}\boldsymbol{\Delta}, \boldsymbol{J} \equiv \mathbf{1}\mathbf{1}').$$

- **Notation:** Given $\mathbf{\Delta}$, let $\Delta_{(1)} < \Delta_{(2)} < \ldots < \Delta_{(p)}$ be the ordered values. $\mathbf{\Delta}$ always has a zero component.

- **Theorem:** Under $\mathcal{M}_p$,

$$\frac{\widehat{\sigma}^2_{p,MLE}}{\sigma^2} \stackrel{d}{=} \frac{1}{n}\{W + \sum_{i=1}^{p} I\{L(\Delta_{(i)}, \mathbf{\Delta}) < Z < U(\Delta_{(i)}, \mathbf{\Delta})\}(Z - \sqrt{n}\Delta_{(i)})^2\Bigg\};$$

with

$$L(\Delta_{(i)}, \mathbf{\Delta}) = \frac{\sqrt{n}}{2}\left[\Delta_{(i)} + \Delta_{(i-1)}\right];$$
$$U(\Delta_{(i)}, \mathbf{\Delta}) = \frac{\sqrt{n}}{2}\left[\Delta_{(i)} + \Delta_{(i+1)}\right].$$

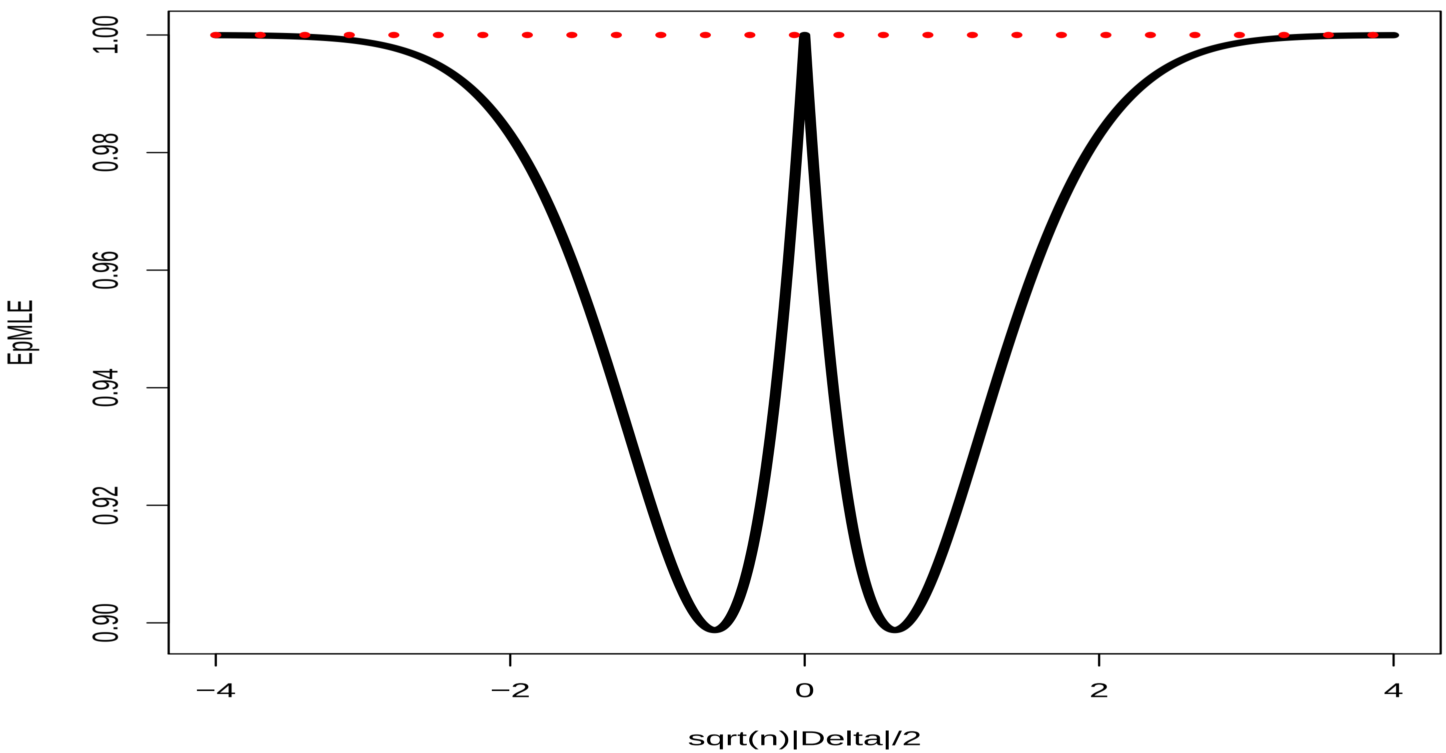
- **Mean:**

$$\text{EpMLE}(\boldsymbol{\Delta}) \equiv \mathbf{E}\left\{\frac{\hat{\sigma}^2_{p,MLE}}{\sigma^2}\right\}$$

$$= \; 1 - \frac{2}{\sqrt{n}}\sum_{i=1}^{p}\Delta_{(i)}[\phi(L(\Delta_{(i)},\boldsymbol{\Delta})) - \phi(U(\Delta_{(i)},\boldsymbol{\Delta}))] + \sum_{i=1}^{p}\Delta^2_{(i)}[\Phi(U(\Delta_{(i)},\boldsymbol{\Delta})) - \Phi(L(\Delta_{(i)},\boldsymbol{\Delta}))];$$

- Case of $p = 2$.

$$\text{EpMLE}(\Delta) = 1 - \left(\frac{2}{\sqrt{n}}|\Delta|\right) \times \left\{\phi\left(\frac{\sqrt{n}}{2}|\Delta|\right) - \left(\frac{\sqrt{n}}{2}|\Delta|\right)\left[1 - \Phi\left(\frac{\sqrt{n}}{2}|\Delta|\right)\right]\right\}$$



- $\widehat{\sigma}^2_{p,MLE}$ is negatively biased for σ^2 (*even though each sub-model estimator is unbiased*). Effect of double-dipping.

- **Variance:**

$$\text{VpMLE}(\boldsymbol{\Delta}) \equiv \textbf{Var}\left\{\frac{\widehat{\sigma}^2_{p,MLE}}{\sigma^2}\right\}$$

$$= \frac{1}{n}\left\{2\left(1-\frac{1}{n}\right)+\frac{1}{n}\left[\sum_{i=1}^{p}\zeta_{(i)}(4)-\left(\sum_{i=1}^{p}\zeta_{(i)}(2)\right)^2\right]\right\};$$

$$\zeta_{(i)}(m) \equiv \mathbf{E}\left\{I\{L(\Delta_{(i)},\boldsymbol{\Delta}) < Z \leq U(\Delta_{(i)},\boldsymbol{\Delta})\}(Z-\sqrt{n}\Delta_{(i)})^m\right\}.$$

- These formulas enable computations of the theoretical risk functions of the classical $\mathcal{M}_p$-based estimators.

An Iterative Estimator

- Consider the Class: $\mathcal{C} = \{\tilde{\sigma}^2(c) \equiv c\hat{\sigma}^2_{p,MLE} : \ c \geq 0\}$

- The risk function of $\tilde{\sigma}^2(c)$, which is a quadratic function in c, could be minimized wrt c. The minimizing value is

$$c^*(\mathbf{\Delta}) = EpMLE(\mathbf{\Delta})/\{VpMLE(\mathbf{\Delta}) + [EpMLE(\mathbf{\Delta})]^2\}$$

- Given a c^*, $\mathbf{\Delta} = (\mu - \mu_{i_0}\mathbf{1}_p)/\sigma$ could be estimated via

$$\widehat{\mathbf{\Delta}} = \frac{(\mu - \mu_{\widehat{M}}\mathbf{1}_p)}{\tilde{\sigma}(c^*)}$$

- This in turn could be used to obtain a new estimate of $c^*(\mathbf{\Delta})$

$\boxed{\textbf{Algorithm for } \tilde{\sigma}^2_{p,ITER}}$

- **Step 0 (Initialization):** Set a value for tol (say, tol = 10^{-8}) and set $c_{old} = 1$.

- **Step 1:** Define $\tilde{\sigma}^2 = (c_{old})\hat{\sigma}^2_{p,MLE}$.

- **Step 2:** Compute $\hat{\boldsymbol{\Delta}} = (\mu - \mu_{\widehat{M}} \mathbf{1}_p)/\tilde{\sigma}$.

- **Step 3:** Compute $c_{new} = \frac{EpMLE(\hat{\boldsymbol{\Delta}})}{VpMLE(\hat{\boldsymbol{\Delta}})+[EpMLE(\hat{\boldsymbol{\Delta}})]^2}$.

- **Step 4:** If $|c_{old} - c_{new}| < tol$ set $\tilde{\sigma}^2_{p,ITER} = \tilde{\sigma}^2$ then stop; else $c_{old} = c_{new}$ then back to Step 1.

Impact of Number of Sub-Models

- **Theorem:** With $n > 1$ fixed, if as $p \to \infty$, $\max_{2 \le i \le p} |\Delta_{(i)} - \Delta_{(i-1)}| \to 0$, $\Delta_{(1)} \to -\infty$, and $\Delta_{(p)} \to \infty$, then

$$\text{Eff}\left(\hat{\sigma}^2_{p,MRE} : \hat{\sigma}^2_{MRE}\right) \to \frac{2(n+2)^2}{(n+1)(2n+7)} < 1.$$

- Therefore, the advantage of exploiting the structure of $\mathcal{M}_p$ could be *lost forever* when p increases!

$\boxed{\text{Representation: Weighted Estimators}}$

- 'Umbrella' Estimator: For $\alpha > 0$, define

$$\widehat{\sigma}^2_{p,LB}(\alpha) = \sum_{i=1}^{p} \left\{ \frac{(\widehat{\sigma}^2_i)^{-\alpha}}{\sum_{j=1}^{p} (\widehat{\sigma}^2_j)^{-\alpha}} \right\} \widehat{\sigma}^2_i .$$

- **Theorem:** Under $\mathcal{M}_p$,

$$\frac{\widehat{\sigma}^2_{p,LB}(\alpha)}{\sigma^2} \stackrel{d}{=} \frac{W}{n} \{1 + H(\boldsymbol{T}; \alpha)\};$$

$$\boldsymbol{T} = (T_1, T_2, \ldots, T_p)' = \boldsymbol{V} / \sqrt{W};$$

$$H(\boldsymbol{T};\alpha) = \sum_{i=1}^{p} \theta_i(\boldsymbol{T};\alpha) T_i^2;$$

$$\theta_i(\boldsymbol{T};\alpha) = \frac{(1+T_i^2)^{-\alpha}}{\sum_{j=1}^{p}(1+T_j^2)^{-\alpha}}.$$

- Even with this representation, still difficult to obtain exact expressions for the mean and variance.

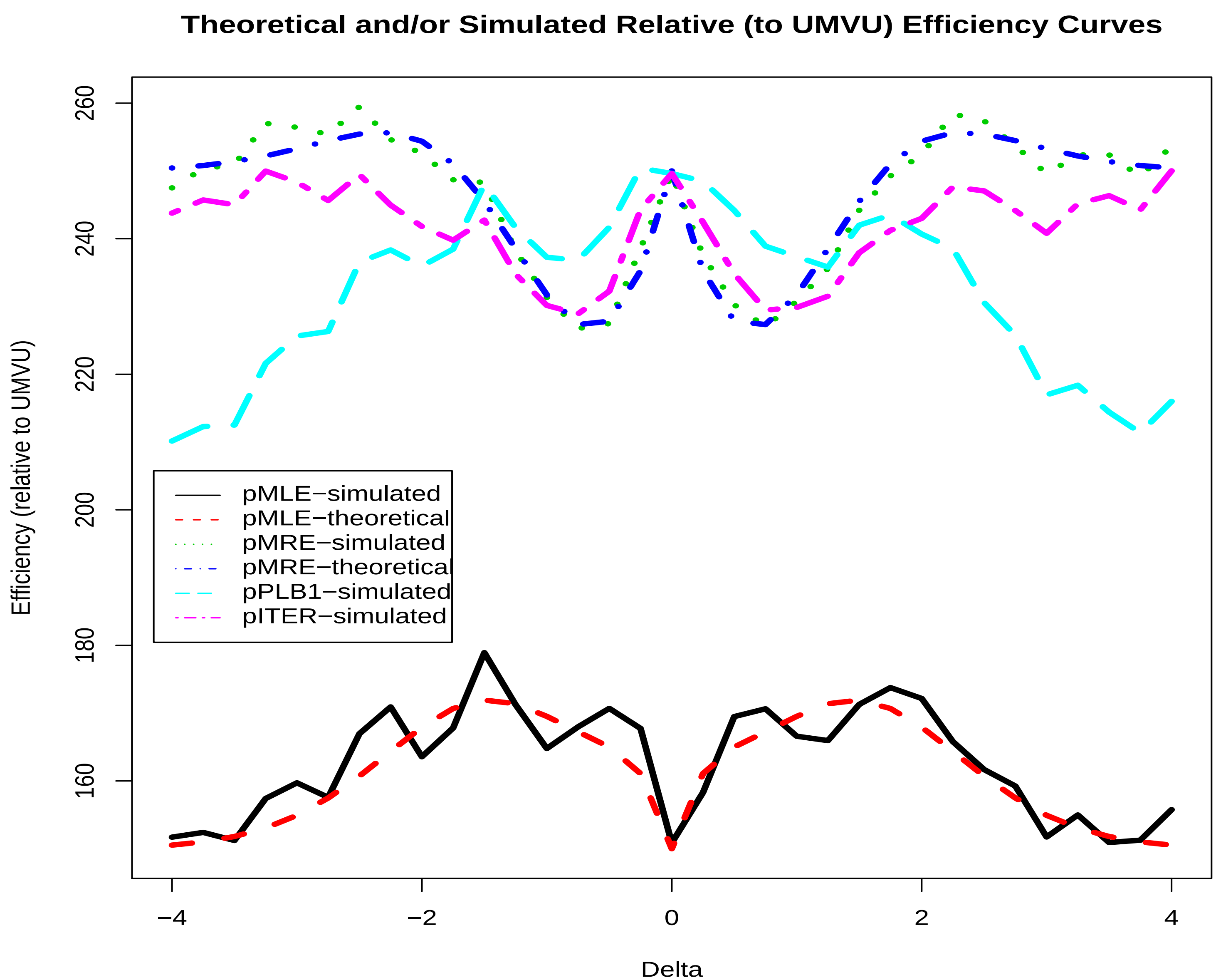
- Developed 2nd-order approximations, but were not so satisfactory when $n \leq 15$.

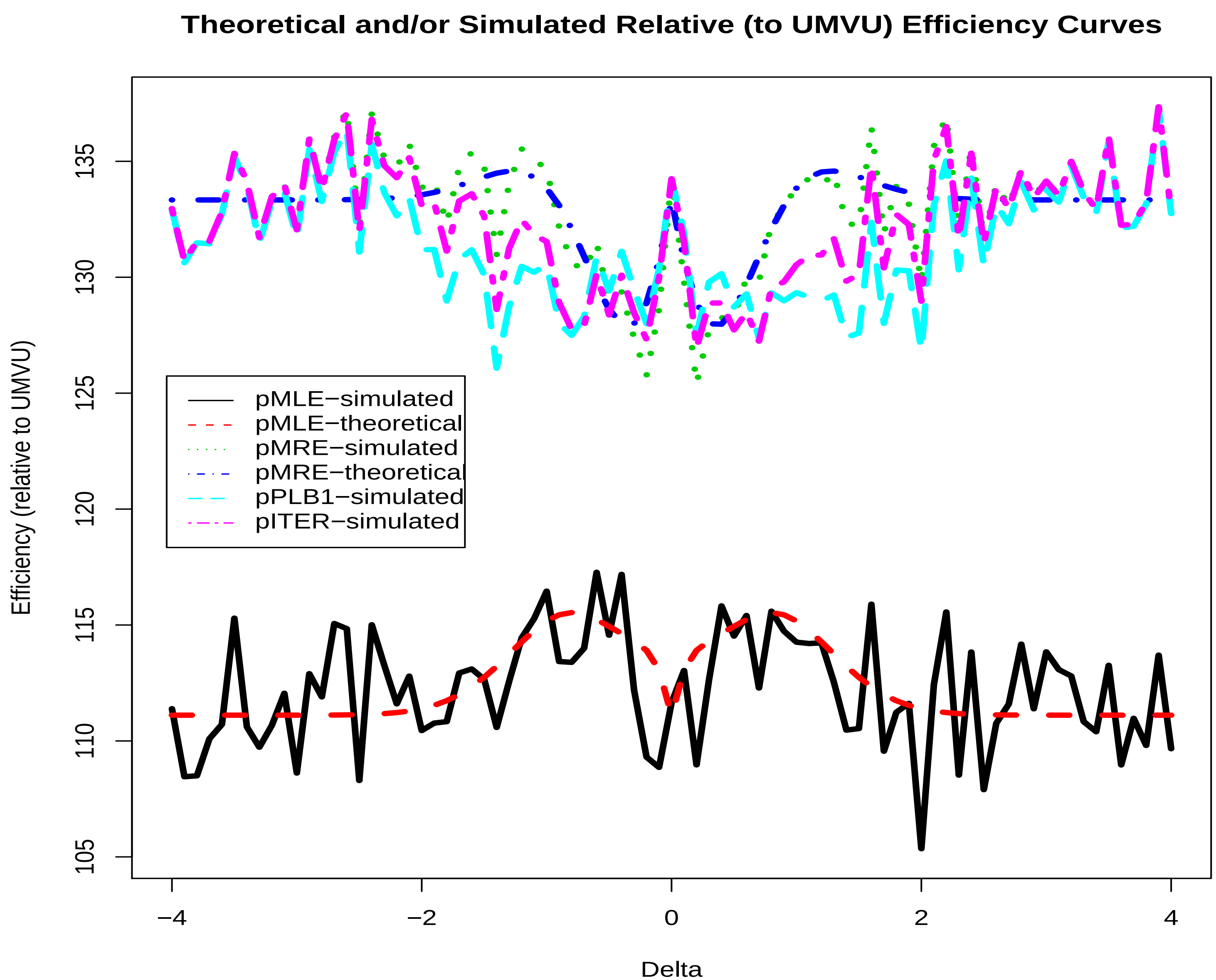
- In the comparisons, we resorted to simulations to approximate the risk function of the weighted estimators.

$\boxed{\textbf{Some Simulation Results}}$

Figures 1 and 2

Simulated and Theoretical Risk Curves
for $n = 3$ and $n = 10$
(Based on 10000 replications per Δ)


Theoretical and/or Simulated Relative (to UMVU) Efficiency Curves
Efficiency (relative to UMVU)
260
240
220
200
180
160
−4
−2
0
2
4
Delta
pMLE−simulated
pMLE−theoretical
pMRE−simulated
pMRE−theoretical
pPLB1−simulated
pITER−simulated



Theoretical and/or Simulated Relative (to UMVU) Efficiency Curves
Efficiency (relative to UMVU)
105
110
115
120
125
130
135
pMLE−simulated
pMLE−theoretical
pMRE−simulated
pMRE−theoretical
pPLB1−simulated
pITER−simulated
−4
−2
0
2
4
Delta

Table: Relative efficiency (wrt UMVU) for symmetric Δ and increasing p with limits $[-1,1]$ and $n = 3, 10, 30$ using 1000 replications. Except for the first set, denoted by (*), where the mean vector is $\{0, 1\}$, the other mean vectors are of form $[-1 : 2^{-k} : 1]$ whose $p = 2^{(k+1)} + 1$. A last letter of 's' on the label means 'theoretical', whereas an 's' means 'simulated.'

n	k	p	pMLEs	pMLEt	pMREs	pMREt	pPLB1s	pITERs
3	*	2	171	170	238	232	247	238
10	*	2	118	115	139	134	133	135
30	*	2	101	104	109	111	108	109
3	0	3	208	195	219	216	260	224
10	0	3	116	120	136	134	127	129
30	0	3	111	104	115	111	114	114
3	1	5	185	185	203	199	248	212
10	1	5	114	119	119	124	120	118
30	1	5	111	106	115	110	112	113
3	2	9	188	182	198	195	243	209
10	2	9	117	118	120	120	127	123
30	2	9	102	106	104	107	103	103
3	3	17	183	181	190	194	235	200
10	3	17	111	117	118	119	123	119
30	3	17	113	105	115	106	115	115
3	4	33	184	181	193	194	239	204
10	4	33	117	117	116	119	125	121
30	4	33	102	105	105	105	105	105
3	5	65	159	181	194	194	226	199
10	5	65	124	117	120	119	132	127
30	5	65	106	105	105	105	107	107

Concluding Remarks

- In models with sub-models, and interest is to infer about a common parameter, possible approaches are:

- **Approach I:** Use a wider model subsuming the sub-models, possibly a fully nonparametric model. Possibly inefficient, though might be easier to ascertain properties.

- **Approach II:** A two-step approach: Select sub-model using data; then use procedure for chosen sub-model, again using same data.

- **Approach III:** Utilize a Bayesian framework. Assign a prior to the sub-models, and (conditional) priors on the parameters within the sub-models. Leads to model-averaging.

- Approaches (II) and (III) are preferable over approach (I); but when the number of sub-models is large, approach (I) may provide better estimators and a simpler determination of the properties.

- If the sub-models are quite different and the model selector can choose the correct model easily, or the sub-models are not too different that an erroneous choice of the model by the selector will not matter much, approach (II) appears preferable. In the in-between situation, approach (III) seems preferable.

- For the specific Gaussian model considered, the iterative estimator actually performed in a robust fashion.

- To conclude,

Observe Caution!

when doing inference after model selection especially when *double-dipping* on the data!