Estimation after Model Selection in a Gaussian Model

Edsel A. Peña

Department of Statistics

University of South Carolina

E-Mail: pena@stat.sc.edu

Joint with Prof. Vanja Dukić of the Univ. of Chicago.

Research support from the National Science Foundation.

1. Setting and Problem

- **Data Given:**

$$\boldsymbol{X} \equiv (X_1, X_2, \ldots, X_n) \text{ IID } F$$

where F is an unknown distribution function.

- **Model $\mathcal{M}$:**

$$F \in \mathcal{M} = \left\{ N(\mu, \sigma^2) : \mu \in \Re, \sigma^2 > 0 \right\}$$

$N(\mu, \sigma^2)$ = normal distribution with mean μ and variance σ^2.

- **Problems:**

  - Estimate σ^2;
  - Given $t \in \Re$, estimate

$$\tau(t) = F(t) = \Pr\{X_1 \leq t\} = \Phi\left(\frac{t-\mu}{\sigma}\right).$$

- Well-known (e.g., proved in Stat 201 ... cheers! if I teach it!) that the uniformly minimum variance unbiased (UMVU) estimator of σ^2 is

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

- UMVU estimator of $\tau(t)$, derived via Basu's Theorem and the Lehmann-Scheffe Theorem, is

$$\begin{aligned} \hat{\tau}_{UMVU}(t) &= \mathcal{T}\left(\frac{\sqrt{n-2}z_1(t)}{\sqrt{1-z_1(t)^2}}; n-2\right) I\{|z_1(t)| \leq 1\} \\ &\quad + I\{z_1(t) > 1\} \end{aligned}$$

with

$$z_1(t) = \frac{\sqrt{n}}{n-1}\left(\frac{t-\bar{X}}{S}\right),$$

and $\mathcal{T}(\cdot; k)$ being the Student's t-distribution function with k degrees-of-freedom.

- Maximum likelihood (ML) estimator of $\tau(t)$ is

$$\hat{\tau}_{MLE}(t) = \Phi\left(\frac{t-\bar{X}}{S}\right).$$

- Decision-theoretic framework: Loss functions utilized are

$$\begin{aligned} L_1(\hat{\sigma}^2, (\mu, \sigma^2)) &= \left(\frac{\hat{\sigma}^2 - \sigma^2}{\sigma^2}\right)^2; \\ L_2(\hat{\tau}(t), (\mu, \sigma^2)) &= (\hat{\tau}(t) - \Phi((t-\mu)/\sigma))^2; \\ L_3(\hat{\tau}, (\mu, \sigma^2)) &= \int L_2(\hat{\tau}(t), (\mu, \sigma^2))\Phi((dt-\mu)/\sigma). \end{aligned}$$

- Risk functions ("loss functions averaged-out over $\boldsymbol{X}$")

$$R_1(\hat{\sigma}^2, (\mu, \sigma^2)); \;\; R_2(\hat{\tau}(t), (\mu, \sigma^2)); \;\; R_3(\hat{\tau}, (\mu, \sigma^2))$$

  are the expected values of the respective loss functions with respect to $\boldsymbol{X}$ and when the true parameter values are (μ, σ^2). Since loss functions are quadratic, then

$$\text{Risk} = \text{Variance} + \text{Bias}^2.$$

- When using risk functions to evaluate estimators, and if we allow biased estimators, the sample variance S^2 is *not* the best. It is dominated by the ML and the minimum risk equivariant (MRE) estimators given by:

$$\hat{\sigma}^2_{MLE} = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2 \quad \text{and} \quad \hat{\sigma}^2_{MRE} = \frac{n}{n+1}\hat{\sigma}^2_{MLE}.$$

- Is the unbiasedness property (i.e., that the average equals the parameter) a *'sacred cow?'*

- All nontrivial Bayes estimators are biased ... so you know what will happen if biased estimators are *not* allowed! Also, can sometimes sacrifice some accuracy to gain precision!

Model $\mathcal{M}_0 = \mathcal{M}_{\mu_0}$

- Suppose it is known that $\mu = \mu_0$, so

$$F \in \mathcal{M}_0 = \left\{ N(\mu, \sigma^2) : \mu = \mu_0, \sigma^2 > 0 \right\}.$$

- Under model $\mathcal{M}_0$, the appropriate estimators are:

$$\hat{\sigma}^2_{UMVU}(\mu_0) = \frac{1}{n} \sum_{i=1}^{n} (X_i - \mu_0)^2;$$

$$\hat{\sigma}^2_{MRE}(\mu_0) = \frac{1}{n+2} \sum_{i=1}^{n} (X_i - \mu_0)^2;$$

$$\begin{aligned} \hat{\tau}_{UMVU}(t; \mu_0) &= \mathcal{T}\left(\frac{\sqrt{n-1} z_2(t)}{\sqrt{1 - z_2(t)^2}}; n-1 \right) I\{|z_2(t)| \le 1\} \\ &\quad + I\{z_2(t) > 1\}; \end{aligned}$$

$$z_2(t) = \frac{1}{\sqrt{n}} \left(\frac{t - \mu_0}{\hat{\sigma}_{UMVU}(\mu_0)} \right).$$

- Estimators developed under $\mathcal{M}$ are also candidate estimators under $\mathcal{M}_0$. Less efficient however since they do *not* exploit the added structure of $\mathcal{M}_0$. For instance, under $\mathcal{M}_0$,

$$\text{Eff}\left(\hat{\sigma}^2_{UMVU}(\mu_0) : \hat{\sigma}^2_{UMVU} \right) = 1 + \frac{1}{n-1}.$$

Model $\mathcal{M}_p$: An Intermediate Model

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

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

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

- **Problem:** Under this intermediate model, how should we estimate σ^2 and $\tau(t) = F(t)$? What are the consequences of using the estimators developed under $\mathcal{M}$, which are also candidate estimators under $\mathcal{M}_p$?

- Can we exploit the structure of $\mathcal{M}_p$ to obtain better estimators of σ^2 and $\tau(t)$? What happens when $p \to \infty$?

- Model $\mathcal{M}_p$ can be viewed as having the p sub-models

$$\mathcal{M}_{\mu_1}, \mathcal{M}_{\mu_2}, \ldots, \mathcal{M}_{\mu_p},$$

with σ^2 a common parameter among these sub-models.

2. Motivation and Importance

- Estimation of the variance, the precision parameter, and the distribution function are important from a practical point of view, as well as theoretically.

- Model $\mathcal{M}_p$ corresponds to practical settings where there are a finite number of possible populations, and a sample $\boldsymbol{X}$ is taken from one of them. Setting of the Neyman-Pearson Lemma and of multiple decision problems.

- C. Stein (1964) developed the approach of hypothesizing that model $\mathcal{M}_0$ may hold, and by doing *pre-test* to accept or reject this hypothesis and deciding on which estimator to use based on this test, was able to show that $\hat{\sigma}^2_{MRE}$ is also an *inadmissible* estimator of σ^2!

- Our primary motivating situations leading to this problem are:

- **Issue of 'inference after model selection:'** 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)*?
- **Survival analysis and reliability settings:** Only known that the family of failure times is either Weibull or gamma (also, Cox PH or accelerated failure model). How to estimate the survivor distribution?
- **Multiple regression:** A subset of predictors is chosen, then other inferences, such as prediction, is performed.
- **Smooth goodness-of-fit tests:** An embedding approach is used and it is desired to determine the size of the embedding class *adaptively.*
- **Others:** In nonparametric regression or function estimation, bandwidths in kernel smoothing are determined *adaptively.*

3. Intuitive Strategies

Strategy I: Simply utilize estimators developed under $\mathcal{M}$, or a fully nonparametric model.

Strategy II (Classical):

> **Step 1 (Model Selection):** Choose the most plausible sub-model using the data $\boldsymbol{X} = (X_1, X_2, \ldots, X_n)$.
>
> **Step 2 (Inference):** Use the best estimators or test procedures in the chosen sub-model, but with these estimators or tests still using the same data $\boldsymbol{X}$. Resulting procedures become *adaptive*.

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. Resulting procedures are again *adaptive*.

Question: Which strategy leads to better procedures, and how could we justify formally each of these intuitive strategies?

4. Classical Estimators Under $\mathcal{M}_p$

- Likelihood Function:

$$L(\mu, \sigma^2) = \prod_{i=1}^{p} L_i(\mu_i, \sigma^2)^{I\{\mu=\mu_i\}};$$

For $i = 1, 2, \ldots, p$,

$$L_i(\mu_i, \sigma^2) = \left(\frac{1}{\sqrt{2\pi}}\right)^n \left(\frac{1}{\sigma^2}\right)^{n/2} \exp\left\{-\frac{n\hat{\sigma}_i^2}{2\sigma^2}\right\};$$

$$\hat{\sigma}_i^2 = \frac{1}{n}\sum_{j=1}^{n}(X_j - \mu_i)^2.$$

- **Model Selector:** $\hat{M} = \hat{M}(X_1, X_2, \ldots, X_n)$

$$\hat{M} = \arg\max_{1\le i\le p} L_i(\mu_i, \hat{\sigma}_i^2) = \arg\min_{1\le i\le p} \hat{\sigma}_i^2 = \arg\min_{1\le i\le p} |\bar{X} - \mu_i|.$$

  - $\hat{M}$ chooses the sub-model leading to the smallest estimate of σ^2, or the sub-model whose mean is closest to the sample mean.
  - Selector also has the interpretation of being the 'highest posterior probability (hpp)' model selector associated with a noninformative prior on (μ, σ^2).
  - Other selectors could also be utilized! E.g., Akaike's AIC, Schwarz Bayesian information criterion (BIC).

- MLE of (μ, σ^2) under $\mathcal{M}_p$:

$$(\hat{\mu}_{p,MLE}, \hat{\sigma}^2_{p,MLE}) = (\hat{\mu}_{\hat{M}}, \hat{\sigma}^2_{\hat{M}}) = \sum_{i=1}^{p} I\{\hat{M} = i\}(\mu_i, \hat{\sigma}^2_i),$$

so the MLE of σ^2 is

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

- A two-stage adaptive estimator of the 'classical' form.

- **Remark:** If $\hat{M} = i$, the adaptive estimator $\hat{\sigma}^2_{\hat{M}}$ does not have the same properties as ith sub-model estimator $\hat{\sigma}^2_i$.

- An alternative estimator of the same flavor as above is to use the sub-model's MRE's given by

$$\hat{\sigma}^2_{MRE,i} = \frac{n}{n+2}\hat{\sigma}^2_i, \quad i = 1, 2, \ldots, p,$$

to obtain

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

- **Remark:** Label 'MRE' is a misnomer since this estimator need not be the minimum risk equivariant estimator under model $\mathcal{M}_p$.

- Adaptive estimator (of 'classical form') of $\tau(t) = F(t)$ under model $\mathcal{M}_p$ is:

$$\hat{\tau}_{p,MLE}(t) = \Phi\left(\frac{t-\mu_{\hat{M}}}{\hat{\sigma}_{\hat{M}}}\right) = \sum_{i=1}^{p} I\{\hat{M} = i\}\Phi\left(\frac{t-\mu_i}{\hat{\sigma}_i}\right)$$

- Another adaptive estimator obtained from the UMVUs of sub-models is as follows: Let

$$z_{3i}(t) = \frac{1}{\sqrt{n}}\left(\frac{t-\mu_i}{\hat{\sigma}_i}\right), \quad i = 1, 2, \ldots, p,$$

and for $i = 1, 2, \ldots, p$,

$$\begin{aligned} \hat{\tau}_{UMVU,i}(t) &= \mathcal{T}\left(\frac{\sqrt{n-1}z_{3i}(t)}{\sqrt{1-z_{3i}(t)^2}}; n-1\right) I\{|z_{3i}(t)| \leq 1\} \\ &\quad + I\{z_{3i}(t) > 1\}. \end{aligned}$$

An estimator of $\tau(t)$ is

$$\hat{\tau}_{p,UMVU}(t) = \hat{\tau}_{UMVU,\hat{M}}(t) = \sum_{i=1}^{p} I\{\hat{M} = i\}\hat{\tau}_{UMVU,i}(t).$$

- **Remark:** Notice that the properties of these adaptive estimators are not easily obtainable because of the interplay between the model selector $\hat{M}$ and the sub-model estimator, both of which are using the same sample data. This interplay makes these situations tough to handle.

5. Bayes Estimators Under $\mathcal{M}_p$

- **Joint Prior Distribution for** (μ, σ^2)**:**

$$\pi(\mu, \sigma^2 | \tilde{\boldsymbol{\theta}}, \beta, \kappa) = \left\{ \prod_{i=1}^{p} \tilde{\theta}_i^{I\{\mu=\mu_i\}} \right\} \frac{\beta^{\kappa-1}}{\Gamma(\kappa-1)} \left(\frac{1}{\sigma^2}\right)^{\kappa} \exp\left(-\frac{\beta}{\sigma^2}\right)$$

- Independent priors between μ and σ^2, and for σ^2, an inverted gamma prior.

- **Joint Posterior Distribution:** By Bayes rule, with $M_i = I\{\mu = \mu_i\}$,

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

$$C \;=\; \frac{1}{\Gamma(n/2+\kappa-1)} \left\{ \sum_{i=1}^{p} \frac{\tilde{\theta}_i}{(n\hat{\sigma}_i^2/2+\beta)^{n/2+\kappa-1}} \right\}^{-1}.$$

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

$$\theta_i(\kappa, \beta, n, \boldsymbol{x}) = \frac{\tilde{\theta}_i (n\hat{\sigma}_i^2/2+\beta)^{-(n/2+\kappa-1)}}{\Sigma_{j=1}^{p} \tilde{\theta}_j \left(n\hat{\sigma}_j^2/2+\beta\right)^{-(n/2+\kappa-1)}}$$

- As $n \to \infty$, and when viewed as a function of $\boldsymbol{X}$, the posterior probability of the correct sub-model converges almost surely to 1.

- **Posterior Density of σ^2:**

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

- **Bayes Estimator of σ^2:**

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

- Estimator is a weighted combination, in contrast to the 'classical forms' of earlier estimators. Note the weights are *data-dependent* or *adaptive*!

- **Non-Informative Prior:**
  - Uniform prior for sub-models: $\tilde{\theta}_i = 1/p, i = 1, 2, \ldots, p$.
  - Jeffrey's prior on each sub-model: $\beta \to 0; \kappa \to 1$.

- The model selector $\hat{M}$ can be interpreted as the highest posterior probability selector corresponding to this non-informative prior distribution.

- **Sub-Models Posterior Probabilities:**

$$\theta_i^*(n, \boldsymbol{x}) = \frac{(\hat{\sigma}_i^2)^{(-n/2)}}{\sum_{j=1}^p (\hat{\sigma}_j^2)^{(-n/2)}}$$

- **Limiting Bayes Estimator of σ^2:**

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

Remark: Estimator actually examined in the sequel did **not** have the multiplier $\left(\frac{n}{n-2}\right)$!

- **Sub-Models Limiting Bayes Estimators:**

$$\tilde{\sigma}_{LB,i}^2 = \left(\frac{n}{n-2}\right) \hat{\sigma}_i^2, \quad i = 1, 2, \ldots, p,$$

- Another adaptive estimator of σ^2 can be formed from these limiting Bayes estimators via

$$\hat{\sigma}_{p,ALB}^2 = \tilde{\sigma}_{LB,\hat{M}}^2 = \left(\frac{n}{n-2}\right) \sum_{i=1}^p I\{\hat{M} = i\} \hat{\sigma}_i^2.$$

Referred to as the *adaptive limiting Bayes estimator.*

- **Remark:** The estimators $\hat{\sigma}_{p,MRE}^2$, $\hat{\sigma}_{p,MLE}^2$, and $\hat{\sigma}_{p,ALB}^2$ belong to the same class of estimators. Consequently, it suffices to derive results for $\hat{\sigma}_{p,MLE}^2$ since results for the other two estimators becomes immediately obtainable.

- **Bayes Estimator of $\tau(t)$:** Posterior mean of $\tau(t)$. Can be shown to be

$$\hat{\tau}_{p,Bayes}(t;\kappa,\beta,\boldsymbol{\theta}) = \sum_{i=1}^{p} \theta_i(\kappa,\beta,n,\boldsymbol{x})\times \mathcal{T}\left(\frac{\sqrt{\kappa-1+\frac{n}{2}}(t-\mu_i)}{\sqrt{\frac{n}{2}\hat{\sigma}_i^2+\beta}};\ 2\left(\kappa-1+\frac{n}{2}\right)\right).$$

- For the non-informative prior ($\tilde{\theta}_i = 1/p$, $\beta \to 0$, $\kappa \to 1$), the limiting Bayes estimator of $\tau(t)$ is

$$\hat{\tau}_{p,LB}(t) = \sum_{i=1}^{p} \left\{\frac{\left(\hat{\sigma}_i^2\right)^{-n/2}}{\sum_{j=1}^{p}\left(\hat{\sigma}_j^2\right)^{-n/2}}\right\} \mathcal{T}\left(\frac{t-\mu_i}{\hat{\sigma}_i};\ n\right).$$

- Which is an adaptively weighted combination of the limiting Bayes estimators in the sub-models.

- Bayes framework therefore leads to estimators that are adaptively weighted combinations of the sub-model estimators.

- Classical framework (MLE, for example) produces two-step estimators: Step 1 is the process of choosing the model; and Step 2 is the process of using the estimator in the chosen model.

Recap: Estimators of σ^2 under $\mathcal{M}_p$

- **Developed under $\mathcal{M}$:**
  - $\hat{\sigma}^2_{UMVU} = S^2 = \frac{1}{n-1}\Sigma_{i=1}^n (X_i - \bar{X})^2$
  - $\hat{\sigma}^2_{MLE} = \frac{1}{n}\Sigma_{i=1}^n (X_i - \bar{X})^2$ and $\hat{\sigma}^2_{MRE} = \frac{n}{n+1}\hat{\sigma}^2_{MLE}$

- **Developed under $\mathcal{M}_p$:**
  - $\hat{\sigma}^2_i = \frac{1}{n}\Sigma_{j=1}^n (X_j - \mu_i)^2, \quad i = 1, 2, \ldots, p.$
  - $\hat{M} = \arg\min_{1 \le i \le p} \hat{\sigma}^2_i = \arg\min_{1 \le i \le p} |\bar{X} - \mu_i|$
  - $\hat{\sigma}^2_{p,MLE} = \hat{\sigma}^2_{\hat{M}} = \Sigma_{i=1}^p I\{\hat{M} = i\}\hat{\sigma}^2_i$
  - $\hat{\sigma}^2_{MRE,i} = \frac{n}{n+2}\hat{\sigma}^2_i, \quad i = 1, 2, \ldots, p.$
  - $\hat{\sigma}^2_{p,MRE} = \hat{\sigma}^2_{MRE,\hat{M}} = \Sigma_{i=1}^p I\{\hat{M} = i\}\hat{\sigma}^2_{MRE,i}$
  - $\hat{\sigma}^2_{p,ALB} = \tilde{\sigma}^2_{LB,\hat{M}} = \left(\frac{n}{n-2}\right)\Sigma_{i=1}^p I\{\hat{M} = i\}\hat{\sigma}^2_i$
  - $\hat{\sigma}^2_{p,LB} = \Sigma_{j=1}^p \left\{\frac{(\hat{\sigma}^2_i)^{-(n/2)}}{\Sigma_{j=1}^p (\hat{\sigma}^2_j)^{-(n/2)}}\right\}\hat{\sigma}^2_i$

- **Question:** Which among these σ^2 estimators is best in terms of their risk function?

6. Comparison of σ^2 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 $\hat{\sigma}^2_{UMVU}$:

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

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

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

- **Notation:** $Z \sim N(0,1)$, $\boldsymbol{Z} = (Z_1, Z_2, \ldots, Z_n)' \sim N_n(\mathbf{0}, \boldsymbol{I})$, and $\boldsymbol{\mu} = (\mu_1, \mu_2, \ldots, \mu_p)'$. With μ_{i_0} the *true mean* with $i_0 \in \{1, 2, \ldots, p\}$, let

$$\boldsymbol{\Delta} = \frac{\boldsymbol{\mu} - \mu_{i_0}\mathbf{1}}{\sigma}$$

where $\mathbf{1} = (1, 1, \ldots, 1)'$.

- **Proposition:** Under $\mathcal{M}_p$ with μ_{i_0} the true mean,

$$\frac{\hat{\sigma}^2_i}{\sigma^2} \stackrel{d}{=} \frac{1}{n}\left(W + V_i^2\right), i = 1, 2, \ldots, p;$$

$W \sim \chi^2_{n-1}$; $\boldsymbol{V} = (V_1, \ldots, V_p)' \sim N_p(-\sqrt{n}\boldsymbol{\Delta}, \boldsymbol{J} \equiv \mathbf{1}\mathbf{1}')$; with W and $\boldsymbol{V}$ stochastically independent.

- Representation of $\boldsymbol{V}$:

$$\boldsymbol{V} = Z\mathbf{1} - \sqrt{n}\boldsymbol{\Delta}$$

- **Implication:** Distributional characteristics of $\hat{\sigma}_i^2/\sigma^2$ depends on $(\boldsymbol{\mu}, \sigma^2)$ *only* through $\boldsymbol{\Delta}$ (and n)! This simplifies comparisons, both theoretical and simulated.

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

- **Theorem:** Under $\mathcal{M}_p$ with μ_{i_0} the true mean,

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

$W \sim \chi^2_{n-1}$, $Z \sim N(0,1)$, and with $W \perp\!\!\!\perp Z$; and

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

[Convention: $\Delta_{(0)} = -\infty$ and $\Delta_{(p+1)} = +\infty$.]

- Representation leads to exact mean and variance:

- **Mean:**

$$
\begin{aligned}
\text{EpMLE}(\boldsymbol{\Delta}) &\equiv \mathbf{E}\left\{\frac{\hat{\sigma}^2_{p,MLE}}{\sigma^2}\right\} \\
&= 1 - \frac{2}{\sqrt{n}} \sum_{i=1}^{p} \Delta_{(i)}[\phi(L(\Delta_{(i)}, \boldsymbol{\Delta})) - \phi(U(\Delta_{(i)}, \boldsymbol{\Delta}))] + \\
&\quad \sum_{i=1}^{p} \Delta^2_{(i)}[\Phi(U(\Delta_{(i)}, \boldsymbol{\Delta})) - \Phi(L(\Delta_{(i)}, \boldsymbol{\Delta}))];
\end{aligned}
$$

- **Variance:**

$$
\begin{aligned}
\text{VpMLE}(\boldsymbol{\Delta}) &\equiv \mathbf{Var}\left\{\frac{\hat{\sigma}^2_{p,MLE}}{\sigma^2}\right\} \\
&= \frac{1}{n}\left\{2\left(1-\frac{1}{n}\right) + \frac{1}{n}\left[\sum_{i=1}^{p} \zeta_{(i)}(4) - \left(\sum_{i=1}^{p} \zeta_{(i)}(2)\right)^2\right]\right\};
\end{aligned}
$$

with

$$
\xi(k;\Omega_{(i)}) \equiv \mathbf{E}\left\{Z^k I(\Omega_{(i)})\right\} = \int_{L(\Delta_{(i)},\boldsymbol{\Delta})}^{U(\Delta_{(i)},\boldsymbol{\Delta})} z^k \phi(z)dz,
$$

and, for $m \in \mathcal{Z}_+$,

$$
\begin{aligned}
\zeta_{(i)}(m) &\equiv \mathbf{E}\left\{I(\Omega_{(i)})(Z - \sqrt{n}\Delta_{(i)})^m\right\} \\
&= \sum_{k=0}^{m} (-1)^{(m-k)} \begin{pmatrix} m \\ k \end{pmatrix} \left(\sqrt{n}\Delta_{(i)}\right)^{(m-k)} \xi(k;\Omega_{(i)}).
\end{aligned}
$$

- These lead to *exact* expressions of the risk functions of $\hat{\sigma}^2_{p,MLE}$, and of $\hat{\sigma}^2_{p,MRE}$ and $\hat{\sigma}^2_{p,ALB}$.

- When $p = 2$, expressions simplify. The mean becomes

$$\begin{aligned}\text{EpMLE}(\Delta) &= 1 - \left(\frac{2}{\sqrt{n}}|\Delta|\right)\left\{\phi\left(\frac{\sqrt{n}}{2}|\Delta|\right) - \right. \\ &\left. \left(\frac{\sqrt{n}}{2}|\Delta|\right)\left[1 - \Phi\left(\frac{\sqrt{n}}{2}|\Delta|\right)\right]\right\}.\end{aligned}$$

- Follows from this that $\hat{\sigma}^2_{p,MLE}$ is *negatively biased* for σ^2.

- **Question:** What happens when the number of sub-models increases indefinitely?

- **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

  (i) $\text{Eff}\left(\hat{\sigma}^2_{p,MLE} : \hat{\sigma}^2_{UMVU}\right) \to \frac{2n^2}{(n-1)(2n-1)} > 1$;

  (ii) $\text{Eff}\left(\hat{\sigma}^2_{p,MRE} : \hat{\sigma}^2_{UMVU}\right) \to \frac{2(n+2)^2}{(n-1)(2n+7)} > 1$;

  (iii) $\text{Eff}\left(\hat{\sigma}^2_{p,MRE} : \hat{\sigma}^2_{p,MLE}\right) \to \frac{(2n-1)(n+2)^2}{(2n+7)n^2} > 1$; and

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

  Also, in the limit, $\hat{\sigma}^2_{p,ALB}$ is dominated by $\hat{\sigma}^2_{UMVU}$.

- From (iv), the advantage of exploiting $\mathcal{M}_p$ could be *lost forever* when p increases!

Representation: Limiting Bayes Estimator

- **Theorem:** Under $\mathcal{M}_p$ with μ_{i_0} the true mean,

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

where

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

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

$$\theta_i(\boldsymbol{T}) = \frac{(1+T_i^2)^{-(n/2)}}{\sum_{j=1}^{p}(1+T_j^2)^{-(n/2)}}, \quad i = 1, 2, \ldots, p.$$

- However, even with this nice-looking representation, it is 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 $\hat{\sigma}^2_{p,LB}$.

Table 1: 2nd-order approximation and simulation results for the mean and variance functions of $\hat{\sigma}^2_{p,LB}/\sigma^2$, and the risk function of $\hat{\sigma}^2_{p,LB}$ for different combinations of p, $\mathbf{\Delta}$, and n. For each combination, 10000 simulation replications were performed.

Combinations of p and $\mathbf{\Delta}$	n	Mean Appr.	Mean Sim.	Variance Appr.	Variance Sim.	Risk Appr.	Risk Sim.
$\mathbf{\Delta}$=(-0.25,0,0.25) p=3	3	0.802	0.961	0.450	0.650	0.489	0.651
	10	0.956	0.989	0.182	0.204	0.184	0.204
	30	0.988	0.994	0.066	0.067	0.066	0.067
$\mathbf{\Delta}$=(-0.5,0,0.5) p=3	3	0.844	0.937	0.401	0.652	0.425	0.656
	10	0.994	0.985	0.204	0.206	0.204	0.206
	30	1.024	1.004	0.074	0.067	0.074	0.067
$\mathbf{\Delta}$=(0,0.25, 0.50) p=3	3	0.951	1.036	0.596	0.745	0.598	0.746
	10	1.005	1.006	0.201	0.205	0.201	0.205
	30	1.004	1.004	0.068	0.068	0.068	0.068
$\mathbf{\Delta}$=(0,0.5,1) p=3	3	1.071	1.106	0.726	0.886	0.731	0.898
	10	1.025	1.021	0.220	0.220	0.220	0.221
	30	1.012	1.004	0.070	0.068	0.071	0.068
$\mathbf{\Delta}$=(-0.25:0.0625:0.25) p=9	3	0.857	0.977	0.495	0.648	0.515	0.649
	10	0.970	0.984	0.185	0.203	0.186	0.204
	30	0.986	0.993	0.065	0.066	0.065	0.066
$\mathbf{\Delta}$=(-0.25:0.03125:0.25) p=17	3	0.867	0.974	0.504	0.670	0.522	0.671
	10	0.972	0.988	0.186	0.202	0.187	0.202
	30	0.987	0.994	0.065	0.067	0.065	0.067
$\mathbf{\Delta}$=(0:0.0625:0.5) p=9	3	0.992	1.031	0.642	0.742	0.642	0.743
	10	1.021	1.033	0.206	0.221	0.206	0.222
	30	1.013	1.013	0.069	0.071	0.069	0.071
$\mathbf{\Delta}$=(0:0.03125:0.5) p=17	3	1.000	1.042	0.652	0.755	0.652	0.756
	10	1.024	1.030	0.207	0.215	0.207	0.216
	30	1.015	1.018	0.069	0.071	0.069	0.071

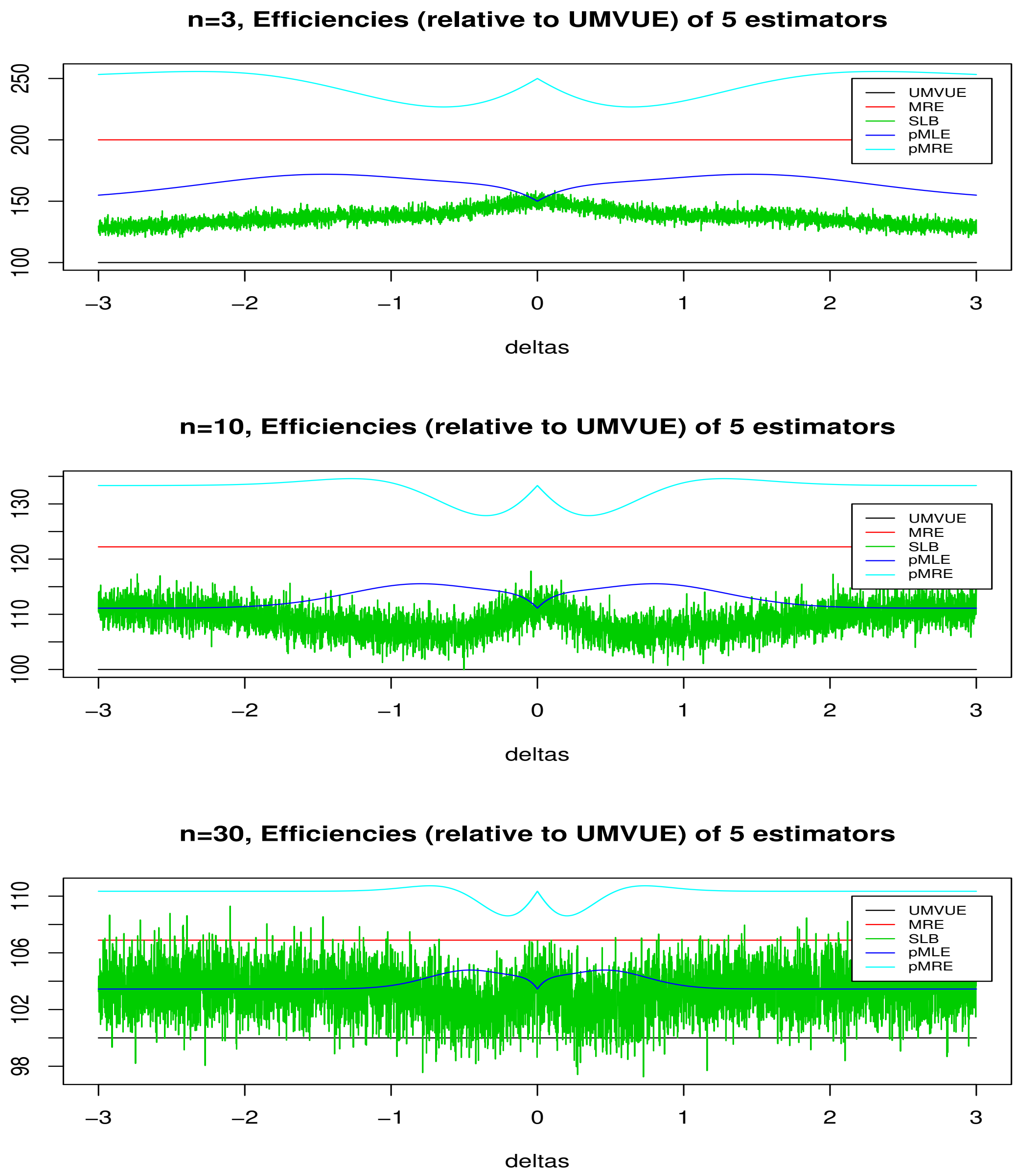


Figure 1: Relative efficiencies of σ^2 estimators for $p = 2$ and $\mathbf{\Delta} = (0, \Delta)$ for $n \in \{3, 10, 30\}$. For the (p,LB) the (connected) scatterplot represents the simulated estimates of relative efficiency based on the 10000 replications for each Δ.

Table 2: Relative efficiencies of the variance estimators for different combinations of p, $\mathbf{\Delta}$, and n. For the (p,LB) estimator, 95% empirical confidence intervals are shown in parenthesis.

Combinations of p and $\mathbf{\Delta}$	n	Efficiency % UMVU	MRE	LB	pMLE	pMRE	ALB
$\mathbf{\Delta}$=(-0.25,0,0.25) p=3	3	100	200	156 (148, 165)	173	222	14
	10	100	122	110 (110, 117)	117	123	71
	30	100	106	105 (102, 107)	105	107	90
$\mathbf{\Delta}$=(-0.5,0,0.5) p=3	3	100	200	162 (153, 168)	183	209	17
	10	100	122	107 (107, 114)	119	124	72
	30	100	106	97 (99, 105)	106	110	88
$\mathbf{\Delta}$=(0,0.25, 0.50) p=3	3	100	200	137 (133, 146)	164	226	12
	10	100	122	103 (102, 109)	114	126	65
	30	100	106	98 (99, 105)	104	108	87
$\mathbf{\Delta}$=(0,0.5,1) p=3	3	100	200	109 (111, 125)	165	221	13
	10	100	122	98 (97, 105)	114	128	65
	30	100	106	102 (100, 106)	104	110	86
$\mathbf{\Delta}$=(-0.25:2^{-4}:0.25) p=9	3	100	200	154 (148, 163)	173	221	14
	10	100	122	109 (110, 118)	117	122	71
	30	100	106	102 (102, 108)	105	105	91
$\mathbf{\Delta}$=(-0.25:2^{-5}:0.25) p=17	3	100	200	150 (148, 161)	173	221	14
	10	100	122	113 (109, 118)	116	122	71
	30	100	106	104 (102, 108)	105	105	91
$\mathbf{\Delta}$=(0:2^{-4}:0.5) p=9	3	100	200	136 (132, 146)	163	225	12
	10	100	122	100 (100, 108)	114	125	65
	30	100	106	98 (97, 103)	104	107	87
$\mathbf{\Delta}$=(0:2^{-5}:0.5) p=17	3	100	200	134 (132, 146)	163	225	12
	10	100	122	102 (100, 107)	114	125	65
	30	100	106	99 (96, 103)	104	107	87

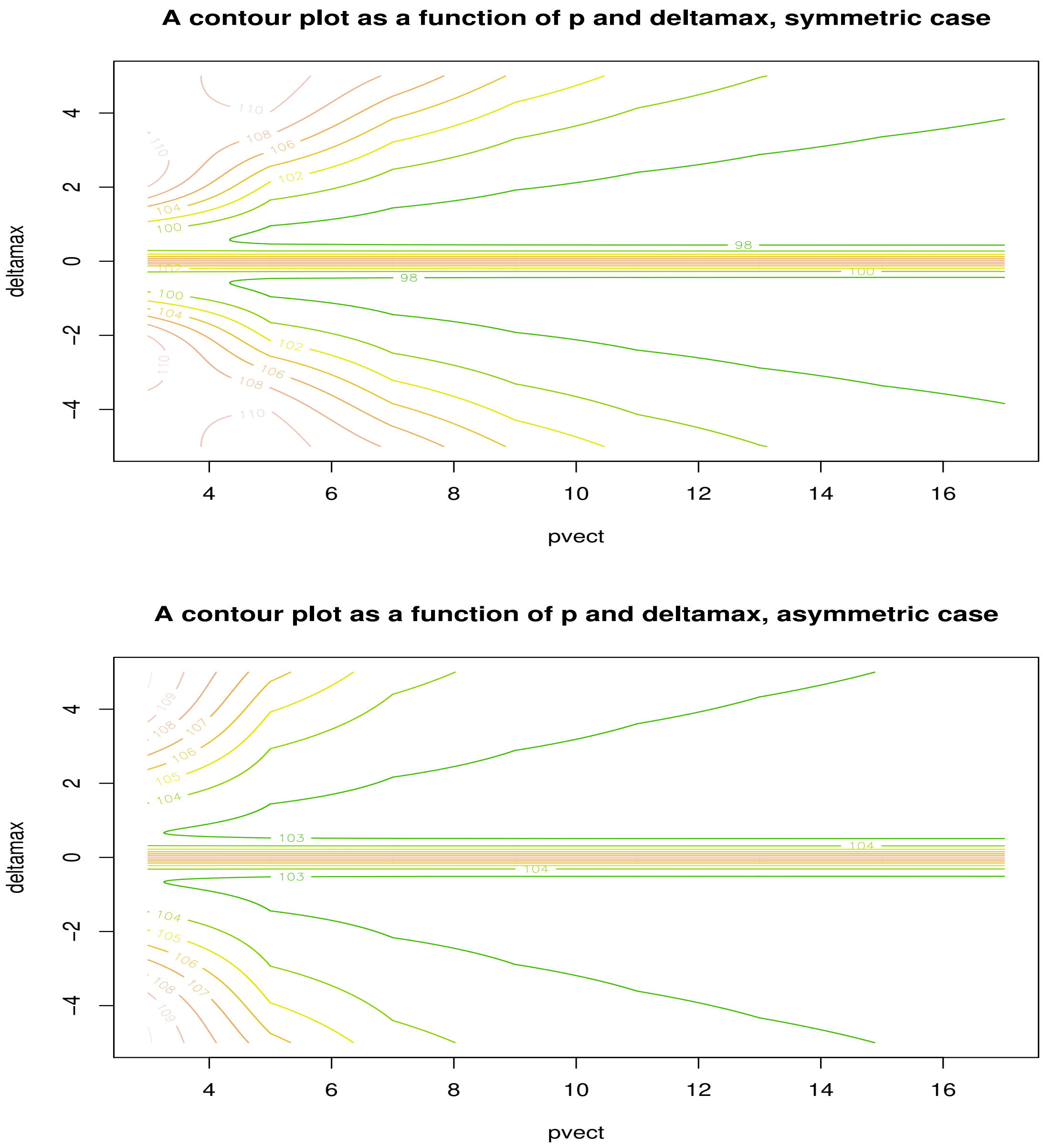


Figure 2: Relative efficiencies of $\hat{\sigma}^2_{p,MRE}$ (pMRE) wrt $\hat{\sigma}^2_{MRE}$ (MRE) in a symmetric and asymmetric Δ cases, as a function of $\Delta_{\max}$ and p for sample size of $n = 10$. Symmetric case of form $\boldsymbol{\Delta} = [-\Delta_{max} : \Delta_{max}/(p-1) : \Delta_{max}]$; while asymmetric case of form $\boldsymbol{\Delta} = [0 : \Delta_{max}/(2(p-1)) : \Delta_{max}]$.

7. Recap: $\tau(t) = F(t)$ Estimators

- **$\mathcal{M}$-UMVU:**

$$z_1(t) = \frac{\sqrt{n}}{n-1}\left(\frac{t-\bar{X}}{S}\right)$$

$$\begin{aligned}\hat{\tau}_{UMVU}(t) \;&=\; \mathcal{T}\left(\frac{\sqrt{n-2}z_1(t)}{\sqrt{1-z_1(t)^2}}; n-2\right) I\{|z_1(t)| \leq 1\} \\ &\quad + I\{z_1(t) > 1\}\end{aligned}$$

- **$\mathcal{M}_p$-UMVU:**

$$z_{3i}(t) = \frac{1}{\sqrt{n}}\left(\frac{t-\mu_i}{\hat{\sigma}_i}\right), \quad i = 1, 2, \ldots, p$$

$$\begin{aligned}\hat{\tau}_{UMVU,i}(t) \;&=\; \mathcal{T}\left(\frac{\sqrt{n-1}z_{3i}(t)}{\sqrt{1-z_{3i}(t)^2}}; n-1\right) I\{|z_{3i}(t)| \leq 1\} \\ &\quad + I\{z_{3i}(t) > 1\}\end{aligned}$$

$$\hat{\tau}_{p,UMVU}(t) = \hat{\tau}_{UMVU,\hat{M}}(t) = \sum_{i=1}^{p} I\{\hat{M} = i\}\hat{\tau}_{UMVU,i}(t)$$

- **$\mathcal{M}_p$-Limiting Bayes:**

$$\hat{\tau}_{p,LB}(t) \;=\; \sum_{i=1}^{p}\left\{\frac{\left(\hat{\sigma}_i^2\right)^{-n/2}}{\sum_{j=1}^{p}\left(\hat{\sigma}_j^2\right)^{-n/2}}\right\} \;\mathcal{T}\left(\frac{t-\mu_i}{\hat{\sigma}_i};\, n\right)$$

- **$\mathcal{M}_p$-Adaptive Limiting Bayes:**

$$\hat{\tau}_{p,ALB}(t) = \sum_{i=1}^{p} I\{\hat{M} = i\}\mathcal{T}\left(\frac{t-\mu_i}{\hat{\sigma}_i}; n\right).$$

- Also obtained distributional representations for these estimators which show that their distributions depend on $(t, \boldsymbol{\mu}, \sigma^2)$ *only through* $(\xi(t), \boldsymbol{\Delta})$, where

$$\xi(t) = \frac{t - \mu_{i_0}}{\sigma} = \text{standardized } t\text{-value}.$$

- But, even with the representations, *no exact* expressions of their risk functions were obtained.

- Comparisons for the $\tau(t)$-estimators were therefore performed through simulations.

Results of Comparisons

- For the τ-estimators, efficiencies are relative to $\hat{\tau}_{UMVU}$.

- To compare globally the τ-estimators, we approximated the risk functions arising from the global loss function

$$L_3(\hat{\tau}, (\mu, \sigma^2)) = \int \left[\hat{\tau}(t) - \Phi\left(\frac{t-\mu}{\sigma}\right)\right]^2 \Phi\left(\frac{dt-\mu}{\sigma}\right).$$

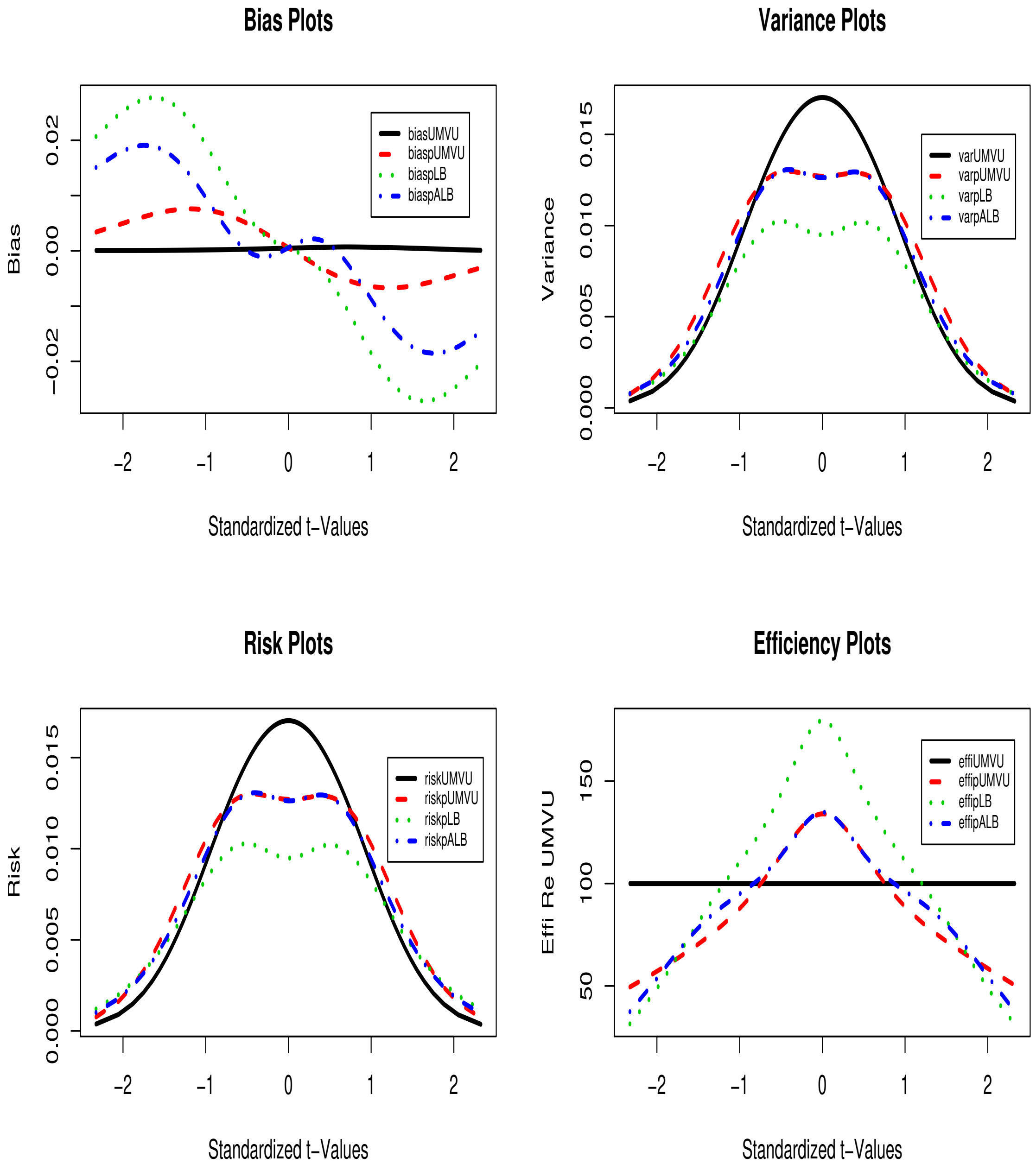
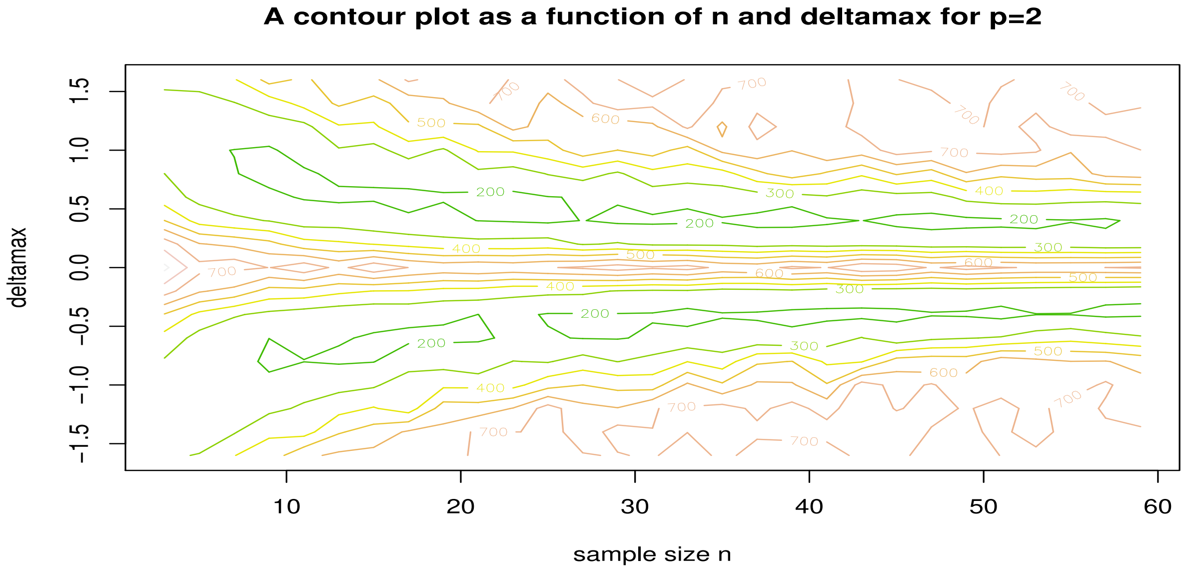


Figure 3: Pointwise biases, variances, risks, and relative efficiencies of the four distribution estimators $\hat{\tau}$, for $\mathbf{\Delta} = (-1, 0, 1)$ and sample size $n = 10$. For each standardized time point $\xi(t)$, 10000 simulation replications were performed.

Table 3: Relative global efficiencies (rel. to the UMVU estimator $\hat{\tau}^2_{UMVU}$) of the three distribution estimators for different combinations of p, $\mathbf{\Delta}$, and n. 10000 simulation replications were performed for each combination.

Combinations of p and $\mathbf{\Delta}$	n	pUMVU Eff %	LB Eff %	ALB Eff %	Combinations of p and $\mathbf{\Delta}$	pUMVU Eff %	LB Eff %	ALB Eff %
$\mathbf{\Delta}$=(-0.25,0,0.25) p=3	3 10 30	316 190 107	704 419 194	395 197 107	$\mathbf{\Delta}$=(-1, 0, 1) p=3	105 108 352	181 138 389	117 111 359
$\mathbf{\Delta}$=(-0.5,0,0.5) p=3	3 10 30	169 94 83	367 158 110	194 96 84	$\mathbf{\Delta}$=(-1: 2^{-1} :1) p=5	109 87 82	199 115 108	123 89 82
$\mathbf{\Delta}$=(0,0.25, 0.50) p=3	3 10 30	274 190 158	428 220 163	334 197 159	$\mathbf{\Delta}$=(-1: 2^{-2} :1) p=9	110 95 88	214 120 103	125 98 89
$\mathbf{\Delta}$=(0,0.5,1) p=3	3 10 30	180 152 145	239 166 181	209 157 146	$\mathbf{\Delta}$=(-1: 2^{-3} :1) p=17	111 98 97	227 123 103	125 101 97
$\mathbf{\Delta}$=(-0.25:2^{-4}:0.25) p=9	3 10 30	321 205 128	779 530 288	403 213 129	$\mathbf{\Delta}$=(-1: 2^{-4} :1) p=33	110 99 99	230 125 103	124 102 99
$\mathbf{\Delta}$=(-0.25:2^{-5}:0.25) p=17	3 10 30	319 204 130	774 560 313	400 212 131	$\mathbf{\Delta}$=(-1: 2^{-5} :1) p=65	111 99 99	239 126 103	125 102 100
$\mathbf{\Delta}$=(0:2^{-4}:0.5) p=9	3 10 30	274 196 174	431 204 131	334 204 175	$\mathbf{\Delta}$=(-1: 2^{-6} :1) p=129	110 99 99	237 127 103	124 102 100
$\mathbf{\Delta}$=(0:2^{-5}:0.5) p=17	3 10 30	276 203 176	432 209 125	336 211 178	$\mathbf{\Delta}$=(-1: 2^{-7} :1) p=257	111 99 99	238 127 103	125 102 100
$\mathbf{\Delta}$=(0, 1) p=2	3 10 30	176 186 525	259 227 553	204 193 539	$\mathbf{\Delta}$=(-1: 2^{-8} :1) p=513	111 99 99	239 127 103	125 102 100



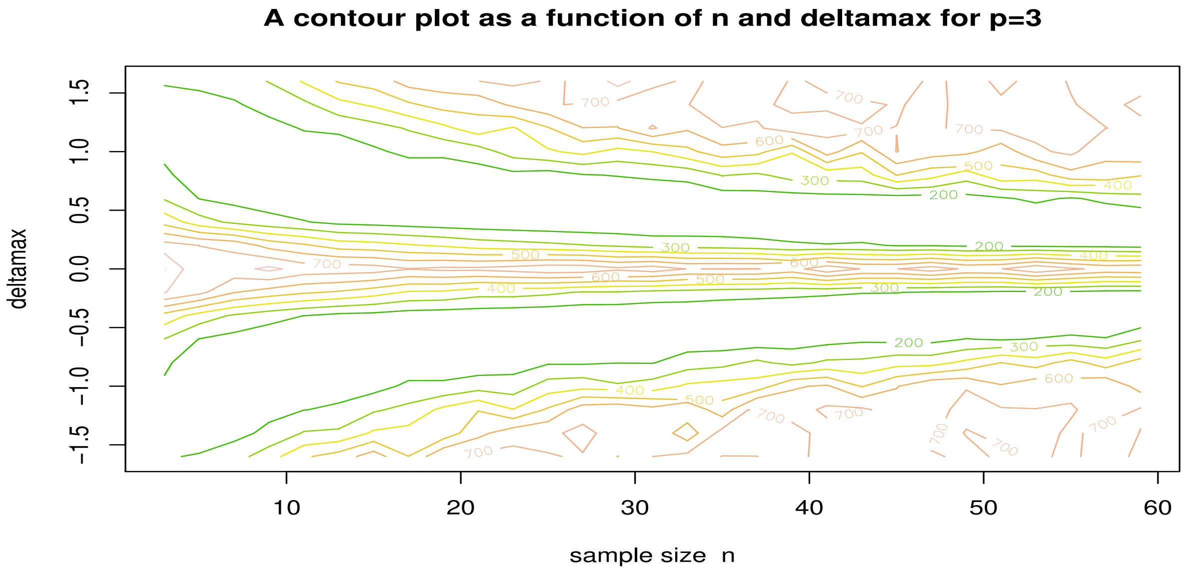


Figure 4: Relative global efficiencies of pLB with respect to UMVU in a asymmetric ($p = 2$) and symmetric ($p = 3$) Δ cases, as a function of $\Delta_{\max}$ and sample size n. Scenario 1: Corresponds to an asymmetric case of form $\mathbf{\Delta} = (0, \Delta_{max})$. Scenario 2: Corresponds to a symmetric case of form $\mathbf{\Delta} = (-\Delta_{max}, 0, \Delta_{max})$. For each combination of $(n, \mathbf{\Delta})$, 10000 simulation replications were performed.

8. Concluding Remarks

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

- **Approach I:** Utilize procedures for a wider model subsuming the sub-models. Could lead to loss of efficiency.

- **Approach II:** Utilize a two-step approach: First step is to select the sub-model using the data; second step is to use a procedure (e.g., estimator) for the chosen sub-model, again using the same data.

  Should recognize that properties of the two-step procedure will be different from the sub-model properties of the procedures.

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

  Resulting procedure is an adaptively weighted combination of the (Bayes) procedures in the sub-models.

- Approaches (II) and (III) appear 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.

- Hard to conclude which of approaches (II) or (III) is preferable. In the Gaussian model considered, approach (II) performed better in estimating the variance σ^2, but approach (III) performed better in estimating the distribution function.

  This calls for further studies in more complicated settings, such as those that motivated our study.

- To conclude,

  Observe Caution!

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