ARTIFICIALLY AUGMENTED SAMPLES, SHRINKAGE, MSE REDUCTION AND PITMAN'S CLOSENESS CRITERION (PCC)

Based on papers: by YY (2006) and Gerard Biau and YY (2012)

Summary

• An inequality is provided that determines when shrinkage reduces the mean squared error (MSE) of an unbiased estimate.

• Artificially augmented samples are then used to obtain, among others, shrinkage estimates of the population's variance and covariance, which improve the unbiased estimates for all parameter values and for all probability models with marginals having finite second moments, and alternative jackknife estimates that complement the usual jackknife estimates in reducing the MSE.

• Results extended in GB &YY (2012). For a large class of distributions and large samples, it is shown that estimates of the variance σ^2 and of the standard deviation σ are more often Pitman closer to their target than the corresponding shrinkage estimates which improve the mean squared error. Our results indicate that Pitman closeness criterion, despite its controversial nature, should be regarded as a useful and complementary tool for the evaluation of estimates of σ^2 and of σ.

• You will get an idea about the "politics" in the Statistics field: decision theorists against Pitman's closeness criterion.

KEYWORDS: Augmented samples; Bias; Jackknife; Mean squared error; Multiple imputation; Pitman closeness; Shrinkage; U-statistics; Variance estimation; Standard deviation estimation.

1 Estimation Tools

Observe data $\mathbf{X}_n = (X_1, \ldots, X_n)$, with $X_1, \ldots, X_n$ independent, identically distributed (*i.i.d.*) random variables (r.vs) from a model/density $f(x, \theta)$ with the parameter $\theta(\in \Theta), \theta = \theta(f)$ unknown.

• $S_n = S(\mathbf{X_n})$: estimate of interest for θ.

• T_n : a generic estimate of θ.

• T_n is unbiased estimate of θ if $ET_n = \theta, \ \forall\ \theta \in \Theta$.

• Kernel: For a model f with parameter $\theta = \theta(f)$, a function $h : Eh(X_1, \ldots, X_m) = \theta\ \forall\ \theta \in \Theta$ is called a "kernel". W.l.o.g assume h is symmetric; otherwise it can be replaced by the symmetric kernel

$$\frac{1}{m!} \sum_{\mathcal{P}_m} h(x_{i_1}, \ldots, x_{i_m})$$

with $\mathcal{P}_m$ the $m!$ permutations $(i_1, \ldots, i_m)$ of $(1, \ldots, m)$; m is the order of the kernel.

• U-statistic: For any kernel h for $\theta = \theta(f)$ the corresponding U-statistic for estimating θ using sample $X_1, \ldots, X_n$, with size $n \geq m$ is obtained by averaging h "symmetrically" over all the observations,

$$U_n = U(X_1, \ldots, X_n) = \frac{1}{\binom{n}{m}} \sum_c h(X_{i_1}, \ldots, X_{i_m}) = E[h(X_1, \ldots, X_m) | X_{(1)}, \ldots, X_{(n)}], \quad (1)$$

where $\sum_c$ denotes summation over all $\binom{n}{m}$ combinations of m distinct elements $\{i_1, \ldots, i_m\}$ from $\{1, \ldots, n\}$, $X_{(1)}, \ldots, X_{(n)}$ is the order statistic.

Clearly, U_n is unbiased for θ since h is unbiased.

• If $S_n = S(X_1, \ldots, X_n)$ is unbiased for θ then it can be used as kernel.

Reference: Serfling, R. J. (1980 or more recent) *Approximation Theorems of Mathematical Statistics.* Wiley

Examples

a) Let $X_1, \ldots, X_n$ be *i.i.d. rvs* with $\theta = EX_1$, the mean. For the kernel $h(x) = x$ with $m = 1$, the corresponding U-statistic for θ is

$$U_n = \frac{1}{\binom{n}{1}} \sum_{i=1}^{n} h(X_i) = \bar{X}_n.$$

b) Let $X_1, \ldots, X_n$ be *i.i.d. rvs* with $\theta = Var(X_1)$, the variance. For the kernel of σ^2,

$$h(x_1, x_2) = \frac{x_1^2 + x_2^2 - 2x_1x_2}{2} = \frac{(x_1 - x_2)^2}{2}$$

the U-statistic for θ is

$$U_n = \frac{1}{\binom{n}{2}} \sum \sum_{1 \le i < j \le n} h(X_i, X_j) = \frac{1}{n-1} (\sum_{i=1}^{n} X_i^2 - n\bar{X}_n)^2 = s_n^2.$$

c) Let $X_1, \ldots, X_n$ be *i.i.d. rvs*, with $\theta = F(x_0) = P(X_1 \le x_0)$. For the kernel $h(x) = I(x \le x_0), I$ taking the value 1 if $x \le x_0$ and zero otherwise, the U-statistic for θ is

$$U_n = \frac{1}{\binom{n}{1}} \sum_{i=1}^{n} I(X_i \le x_0) = \hat{F}_n(x_0).$$

d) Recall from Neyman-Scott (Le Cam version) for X_1, Y_1, that $X_1 - Y_1$ has mean 0 and variance $2\sigma^2$ and therefore $E(X_1 - Y_1)^2 = 2\sigma^2$. Using this result for iid $X_1, \ldots, X_n$ with variance $\theta = \sigma^2$,

$$h_1(x_1, x_2) = \frac{(x_1 - x_2)^2}{2}$$

is a kernel from σ^2. Write the corresponding U-statistic for σ^2 based on $X_1, \ldots, X_n$ and h_1.

e) Independent, identically distributed vectors $(X_1, Y_1), \ldots, (X_n, Y_n)$ are observed with $\theta = Cov(X_1, Y_1)$, the covariance of X_1, Y_1.. Let

$$h_2((x_1, y_1), (x_2, y_2)) = \frac{(x_1 - x_2)(y_1 - y_2)}{2}.$$

Then, since (X_1, Y_1) is independent of (X_2, Y_2),

$$Eh_2((X_1, Y_1), (X_2, Y_2)) = .5[EX_1Y_1 - EX_1Y_2 - EX_2Y_1 + EX_2Y_2]$$

$$= .5[2EX_1Y_1 - 2EX_1EY_1] = Cov(X_1, Y_1).$$

Thus h_2 is a kernel for the Covariance of X_1, Y_1. Write the corresponding U-statistic for the Covariance of (X_1, Y_1) using h_2 and the sample.

• $\mathcal{R}(T_n, \theta)$: the cost (called "Risk") in estimating θ with T_n, calculated under f.

$\mathcal{R}(x, y)$ is a distance-measure with properties: *i)* $\mathcal{R}(x, x) = 0$,

ii) $\mathcal{R}(x, y) = \mathcal{R}(y, x)$, for every x, y in the domain of $\mathcal{R}$.

Example: $\mathcal{R}(T_n, \theta) = E(T_n - \theta)^2$ with the expected value taken under f.

$E(T_n - \theta)^2$ is the Mean Square(d) Error of T_n.

Question: If $E(S_n - \theta)^2 < E(T_n - \theta)^2$ do you think

$$P[|S_n - \theta| < |T_n - \theta|] > \frac{1}{2} \quad \text{or } < \frac{1}{2}?$$

To be seen ...

Definition 1.1 *Estimate S_n is inadmissible for $\theta \in \Theta$ with Risk function $\mathcal{R}$ if there an estimate $\tilde{S}_n$ such that*

$$\mathcal{R}(\tilde{S}_n, \theta) \leq \mathcal{R}(S_n, \theta) \; \forall \; \theta \in \Theta \tag{2}$$

and there is $\theta_0 \in \Theta$ for which (2) is strict.

To solve a statistical problem when $\mathcal{R}$-risk is the criterion of interest, attention should be restricted in admissible estimates.

Theorem 1.1 *Let $S_n = S(X_1, \ldots, X_n)$ be unbiased for $\theta = \theta(f)$. Then, the corresponding U-statistic, U_n, with kernel*

$$h = \frac{1}{n!} \sum_{\mathcal{P}} S(x_{i_1}, \ldots, x_{i_n})$$

is unbiased and

$$Var(U_n) \leq Var(S_n). \tag{3}$$

Proof: Since U_n is unbiased it is enough to prove

$$EU_n^2 \leq ES_n^2.$$

Since $U_n = E(S_n|X_{(1)}, \ldots, X_{(n)})$,

$$EU_n^2 = E[(E(S_n|X_{(1)}, \ldots, X_{(n)}))^2] \leq E[(E(S_n^2|X_{(1)}, \ldots, X_{(n)})] = ES_n^2.$$

with equality if and only if $P(U_n = S_n) = 1$.

Note: $X_{(1)}, \ldots, X_{(n)}$ is sufficient therefore $E(S_n|X_{(1)}, \ldots, X_{(n)})$ is independent of θ so an estimate.

• Therefore, if interested in the MSE of unbiased estimates of θ, we restrict attention to U-statistics for θ.

Definition 1.2 *Let S_n be an estimate of $\theta \in \Theta$. Then, $c \cdot S_n$ is a shrinkage estimate of $\theta, 0 < c < 1$.*

• Stein (1964) showed inadmissibility of the usual estimator $\frac{n-1}{n+1}s_n^2$ for the variance σ^2 of **a normal distribution** with unknown mean (Goodman, 1953, "A Simple Method for Improving some Estimators"), by providing a shrinkage estimate improving its MSE; sample $X_1, \ldots, X_n$ has mean $\bar{X}_n$,

$$s_n^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2.$$

• We will show inadmissibility of the unbiased estimate s_n^2 of σ^2 for all distributions having finite second moment. Similar results are proved for other unbiased estimates. The approach seems more interesting than the result.

Briefly:

1) An inequality is provided that determines when shrinkage reduces the mean squared error (MSE) of an unbiased estimate and a Lemma encourages the use of artificially augmented samples.

2) Artificially augmented samples (i.e. pseudo-values) are then used to obtain, among others, shrinkage estimates of the population's variance and covariance, which improve the unbiased estimates for all parameter values and for all probability models with marginals having finite second moments. A similar approach is used for U-statistics.

3) Alternative jackknife estimates are presented that complement the usual jackknife estimates in reducing the MSE.

Tools used: Augmented samples; Jackknife Estimates; Shrinkage Estimates; U-statistics.

Notation: We present first YY(2005). Lower case letters will be used to denote estimates, i.e. t_n is used instead of T_n.

2 Introduction (YY, 2006)

The estimation of a population's variance, σ^2, and covariance, $\sigma_{X,Y}$, is the stuff of statistical folkore. Often the unbiased estimate, s_n^2 of σ^2 is used, but for some probability models, shrinkage estimates, cs_n^2 reduce the mean squared error (MSE) of s_n^2 for every $\sigma, 0 < c < 1$, with n the sample size.

So far, there is no shrinkage estimate with a smaller MSE than s_n^2 that applies universally, for all values of σ, for all probability models and every sample size n (Stein 1964; Brown 1968; Arnold 1970; Lehmann 1983, p. 113). Such an estimate will be provided in these lectures. A similar situation holds for the unbiased estimate of $\sigma_{X,Y}$.

For an unbiased estimate, t_n, of a parameter θ with real values, **an increase in the sam-**

ple size n has usually the same effect as a successful shrinkage; both decrease the MSE. Questions arise as to **whether by artificially augmenting the sample, an estimate $\tilde{t}_n$ can be obtained with a smaller MSE than t_n for every θ value,** and as to **whether $\tilde{t}_n$ is a shrinkage estimate.**

In these lectures it is seen that for some parameters, this is indeed so, and that **even more is true**. $\tilde{t}_n^k$, **the average of the values of t_{n+k} on artificially augmented samples,** turns out to be a shrinkage estimate that has a smaller MSE than t_n not only for all θ values, **but also for all probability models,** $1 \leq k < n$.

In particular, in variance estimation, the average of $s^2_{n+1}(X_1, \ldots, X_n, X_i), i = 1, \ldots, n$, **turns out to be a shrinkage estimate** because for the U-statistic kernel $h(x_1, x_2)$, which determines σ^2 and s_n^2, $\mathbf{h}(\mathbf{x}, \mathbf{x}) = \mathbf{0}$. The obtained estimate, $\frac{(n+2)(n-1)}{n(n+1)} s_n^2$, has a smaller MSE than s_n^2 for all values of $\sigma, n \geq 2$ and for all probability models with finite second moments.

The same shrinkage coefficient, $\frac{(n+2)(n-1)}{n(n+1)}$ is obtained when averaging the values $t_{n+1,2}(X_1, ..., X_n, X_i), i = 1, \ldots, n$, of a U-statistic $t_{n,2}$ with symmetric kernel of order $m = 2$ vanishing at the diagonal, like, for example, those determined by $\sigma_{X,Y}$, Kendall's τ , and Gini's index g. The shrinkage estimate of $\sigma_{X,Y}$ also has a smaller MSE than the corresponding U-statistic for all covariance values, $n \geq 2$ and for all probability models with marginals having finite second moments. However, additional assumptions are needed for a similar result to hold when estimating either τ or g.

The results are presented for a U-statistic $t_{n,m}$ with a symmetric kernel of order $m \geq 2$ that vanishes when two arguments are repeated, and the shrinkage coefficients $c_{\delta_{n,k,m}}$ are obtained using $(n + k)$ artificially augmented samples, $1 \leq k < n$, where δ_n is a positive number that can be chosen to increase with n. For n large, our analysis suggests that shrinkage coefficients are to be obtained from $(n + k_n)$-augmented samples for the bias and the MSE improvement

to slowly decrease to 0 as n increases. Shrinkage coefficients are also obtained that are used to reduce the MSE of some other unbiased estimates.

An alternative $(n+1)$ jackknife estimate, $\tilde{t}_n^J$, is also provided, which together with the usual $(n-1)$ jackknife estimate, t_n^J, has the potential to reduce the MSE of a biased estimate, t_n, of θ. This is contrary to results on unaugmented jackkknife procedures (Shao and Tu 1995, sec. 2.5, p. 70, l. 3–5). For example, when the population's mean is unknown, the estimate $\tilde{t}_n^J = \frac{n-1}{n+1}s_n^2$ of σ^2 improves $t_n^J = s_n^2$ and $t_n = \frac{n-1}{n}s_n^2$ for various models. When t_n is smooth and n is large, conditions are provided that determine when $\tilde{t}_n^J, t_n^J$ and $\tilde{t}_n^1$ [the average of $t_{n+1}(X_1, \ldots, X_n, X_i), i = 1, ..., n$] have smaller MSE than t_n. It is expected that similar results will hold for $(n+k)$ jackknife estimates. In Section 2 a sufficient condition is provided for a shrinkage estimate to reduce the MSE of an unbiased estimate, t_n, of θ for all θ values and a family, $\mathcal{F}$, of probability models. In Section 3 the basis of the motivation to use $(n+k)$-artificially augmented samples and $\tilde{t}_n^k$ is presented. In Section 4, $\tilde{t}_n^k$ is used to obtain shrinkage estimates that improve the MSE of some U-statistics and other unbiased estimates. Finally, in Section 5, the alternative $(n+1)$ jackknife estimate $\tilde{t}_n^J$ is proposed and studied.

3 Shrinkage and MSE Reduction

Let $X_1, \ldots, X_n$ be a sample from an unknown cumulative distribution function F in a known class $\mathcal{F}$ of models, and let $t_n(X_1, ..., X_n)$ be an unbiased estimate of the unknown model parameter $\theta \in \Theta(\subseteq R)$ with finite second moment; θ may be, for example, the mean of F. The MSE, $E_F(c_n t_n - \theta)^2$, of the shrinkage estimate $c_n t_n,\ 0 < c_n < 1$, is minimized when

$$c_n(\theta, F) = \frac{\theta^2}{Et_n^2} = (1 + \frac{var(t_n)}{\theta^2})^{-1}. \tag{4}$$

Because $c_n(\theta, F)$ often depends on θ and F, this approach does not yield a universal shrinkage coefficient c_n that minimizes the MSE of t_n for every $\theta \in \Theta$, and for every $F \in \mathcal{F}$ when $\mathcal{F}$ consists of more than one model. An alternative goal is to determine shrinkage coefficients that reduce the MSE of t_n for every $\theta \in \Theta$ and for every $F \in \mathcal{F}$.

These coefficients are selected from the set $[\sup_{\Theta,\mathcal{F}} c_n(\theta, F), 1)$ that is nonempty if $inf_{\Theta,\mathcal{F}} \frac{var(t_n)}{\theta^2} > 0$ because

$$\sup_{\Theta,\mathcal{F}} c_n(\theta, F) = (1 + inf_{\Theta,\mathcal{F}} \frac{var(t_n)}{\theta^2})^{-1};$$

$\sup_{\Theta,\mathcal{F}}$ (resp. $\inf \sup_{\Theta,\mathcal{F}}$) denotes $\sup_{\theta\in\Theta,F\in\mathcal{F}}$ (resp. $\inf_{\theta\in\Theta,F\in\mathcal{F}}$).

We now characterize the shrinkage coefficients that reduce the MSE of t_n for a given θ and F.

Lemma 3.1

$$E(c_n t_n - \theta)^2 < E(t_n - \theta)^2 = var(t_n) \iff \frac{1 - c_n}{1 + c_n} < \frac{var(t_n)}{\theta^2}. \tag{5}$$

Proof. Use the relation

$$E(c_n t_n - \theta)^2 = c_n^2 var(t_n) + (1 - c_n)^2\theta^2. \qquad \square$$

From (5), it follows that when F is the true model, the unbiased estimate t_n can be improved with shrinkage for every $\theta \in \Theta$ if $\inf_{\theta\in\Theta} \frac{var(t_n)}{\theta^2}$ is bounded below by some known positive constant L_F that depends on F and n. This occurs when, for example, the Fisher information $I_{X_1}(\theta) = M/\theta^2, M > 0$, and the Cramer–Rao inequality holds for t_n at the model F. t_n can be improved with shrinkage for every $\theta \in \Theta$ and for every $F \in \mathcal{F}$ if $inf_{\Theta,\mathcal{F}} \frac{var(t_n)}{\theta^2}$ is bounded below by some known positive constant L that depends on n.

In (5) the lower bound $\frac{1-c_n}{1+c_n}$ is a decreasing function of c_n that should be suitably chosen; it should be large enough to cause moderate bias and for (5) to hold for every $\theta \in \Theta$ and every

$F \in \mathcal{F}$, with the corresponding MSE reduction to slowly decrease to 0 as n increases.

The estimate $\hat{c}_n = (1+\frac{\hat{V}_n}{t_n^2})^{-1}$ of $c_n(\theta, F)$ will not reduce the MSE of t_n for each n and each F, because $\hat{V}_n$ may not be a satisfactory estimate of $var(t_n)$ for all models F. This can be seen in examples for large samples, appearing at the end of the corresponding section in the section (YY, 2005).

4 Estimates based on artificially augmented samples

Pseudovalues of an estimate t_n of θ are used to, for example, estimate its variance or to obtain a new estimate with reduced bias or when data are missing. These pseudovalues are usually obtained by evaluating either t_n on B bootstrap samples (Efron 1979), t_{n-k} on $(n-k)$-reduced samples (Quenouille 1956), or t_n on samples obtained with multiple-imputation methods (Rubin 1987).

The class $\mathcal{A}_{n,k}$ of the (n+k) artificially augmented samples consists of the samples

$$\mathbf{X} = \mathbf{X}_{n+k} = (X_1, \ldots, X_n, X_{n+1} = X_{j_1}, \ldots, X_{n+k} = X_{j_k}), 1 \le j_1 < \ < j_k \le n,$$

and the pseudovalues $t_{n+k}(\mathbf{X}), \mathbf{X} \in \mathcal{A}_{n,k}$, are used to define the estimate

$$\tilde{t}_n^k = \binom{n}{k}^{-1} \sum_{\mathbf{X} \in \mathcal{A}_{n,k}} t_{n+k}(\mathbf{X}), \qquad 1 \le k \le n. \tag{6}$$

$\mathcal{A}_{n,k}, t_{n+k}(\mathbf{X}), \mathbf{X} \in \mathcal{A}_{n,k}$ and $\tilde{t}_n^k$ can all be thought of in terms of multiple imputation for a sample with size $(n+k)$ and k missing observations.

The proposition that follows encourages the use both of $(n+k)$-augmented samples and of the estimate $\tilde{t}_n^k$ in (6). The use of B bootstrap $(n+k)$-augmented samples is discouraged due to the additional randomization introduced by finite resampling (Yatracos 2002).

Definition 4.1 *For n independent, identically distributed random variables $X_1, \ldots, X_n$, their empirical cumulative distribution function*

$$\hat{F}_n(x) = \frac{1}{n}\sum_{i=1}^{n} I(X_i \leq x),\ x \in R,$$

$I(X_i \leq x) = 1$ if $X_i \leq x$ and 0 otherwise.

Proposition 4.1 . *Let $\hat{F}_n, \hat{F}_{n-k,i_1,\ldots,i_k}$ and $\hat{F}_{n+k,i_1,\ldots,i_k}$ denote, respectively, the empirical cumulative distributions of the original sample $\{X_1, \ldots, X_n\}, \{X_1, \ldots, X_n\} - \{X_{i_1}, \ldots, X_{i_k}\}$ and $\{X_1, \ldots, X_n, X_{i_1}, \ldots, X_{i_k}\}, 1 \leq k < n, 1 \leq i_l \neq i_m \leq n$. Then, for every x it holds that*

$$|\hat{F}_{n+k,i_1,\ldots,i_k}(x) - \hat{F}_n(x)| < |\hat{F}_{n-k,i_1,\ldots,i_k}(x) - \hat{F}_n(x)|. \tag{7}$$

Thus, for any x, $\hat{F}_{n+k,i_1,\ldots,i_k}(x)$ is closer than $\hat{F}_{n-k,i_1,\ldots,i_k}(x)$ to $\hat{F}_n(x)$ that contains all the information.

Proof. Let I denote the indicator function. Then, (7) follows from the relations

$$\hat{F}_{n+k,i_1,\ldots,i_k}(x) = \frac{n}{n+k}\hat{F}_n(x) + \frac{1}{n+k}\sum_{j=1}^{k} I(X_{i_j} \leq x) = \hat{F}_n(x) + \frac{1}{n+k}\sum_{j=1}^{k}[I(X_{i_j} \leq x) - \hat{F}_n(x)]$$

and

$$\hat{F}_{n-k,i_1,\ldots,i_k}(x) = \frac{n}{n+k}\hat{F}_n(x) - \frac{1}{n-k}\sum_{j=1}^{k} I(X_{i_j} \leq x) = \hat{F}_n(x) - \frac{1}{n-k}\sum_{j=1}^{k}[I(X_{i_j} \leq x) - \hat{F}_n(x)]$$

and (7) follows moving $\hat{F}_n(x)$ to the left side of the last two equations and taking absolute values. □

5 Shrinkage of U-Statistics

5.1 U-Statistics and Augmented Samples

For a symmetric kernel $h(x_1, x_2, ..., x_m)$ of degree m, such that

$$Eh(X_1, X_2, \ldots, X_m) = \theta,$$

the U-statistic of θ and the $(n+k)$-augmented sample estimate (6) are

$$t_{n,m} = \binom{n}{m}^{-1} \sum_{1 \le i_1 < \ldots < i_m \le n} h(X_{i_1}, \ldots, X_{i_m}) \tag{8}$$

and

$$\tilde{t}^k_{n,m} = \binom{n}{k}^{-1} \binom{n+k}{m}^{-1} \cdot \sum_{\mathbf{X} \in \mathcal{A}_{n,k}} \sum_{1 \le i_1 < \ldots < i_m \le n} h(X_{i_1}, \ldots, X_{i_m}) \tag{9}$$

5.2 Shrinkage Estimates

Let V_m denote the class of symmetric kernels of degree $m \geq 2$ that vanish when two of the arguments in the kernel are repeated. When $h \in V_m, \tilde{t}^k_{n,m}$ turns out to be a shrinkage estimate of $t_{n,m}$; the shrinkage coefficients are obtained in the next Proposition.

Proposition 5.1 *Let $t_{n,m}$ be as in (8)) with $h \in V_m, l = \min\{k, m\}$.*

a) *The (n +k)-augmented sample estimate is*

$$\tilde{t}^k_{n,m} = c_{n,k,m} t_{n,m} = \left[\binom{n+k}{m}^{-1} \sum_{j=0}^{l} \binom{k}{j} \binom{n-j}{m-j} \right] t_{n,m}. \tag{10}$$

b) The $(n+1)$-augmented sample estimate is

$$\tilde{t}^1_{n,m} = c_{n,1,m} t_{n,m} = \left[1 - \frac{m^2 - m}{n(n+1)} \right] \cdot t_{n,m}, \tag{11}$$

and the corresponding lower bound in (5) is

$$\frac{1 - c_{n,1,m}}{1 + c_{n,1,m}} = \frac{m^2 - m}{2n(n+1) - m^2 + m}, \tag{12}$$

and it holds that

$$c_{n,1,m} \leq c_{n,1,2}, \qquad m \geq 2. \tag{13}$$

c) When $m = 2$, the $(n+k)$-augmented sample estimate is

$$\tilde{t}^k_{n,2} = c_{n,k,2} t_{n,2} = \left[1 - \frac{2k}{(n+k)(n+k-1)} \right] t_{n,2}, \tag{14}$$

the corresponding lower bound in (5) is

$$\frac{1-c_{n,k,2}}{1+c_{n,k,2}} = \frac{k}{(n+k)(n+k-1)-k}, \tag{15}$$

and it holds for the coefficients that

$$0 < c_{n,k,2} \leq c_{n,k-1,2} \leq c_{n,1,2} \qquad 2 \leq k < n. \tag{16}$$

Proof: To Prove **a)**, let

$$\Delta_{n,k,m} = \sum_{j=0}^{l} \binom{k}{l} \binom{n-j}{m-j}.$$

In (8), $\binom{n}{k}\binom{n+k}{m}\tilde{t}^{k}_{n,m}$ has $\Delta_{n,k,m}\binom{n}{k}$ nonvanishing terms and equals

$$\binom{n}{k} \frac{\Delta_{n,k,m}}{\binom{n}{m}} \sum_{1 \leq i_1 < i_2 < \ldots < i_m \leq n} h(X_{i_1}, X_{i_2}, \ldots, X_{i_m}). \tag{17}$$

(9) follows from (8) and (17).

Proofs for **b)** and **c)** follow from the proof of part **a)**. □

Remark 1

a) To get a feeling for estimate (9), note for example that when $n = 3, m = 2, k = 2$ and $l = 2$, then $\tilde{t}^{2}_{3,2} = .8t_{3,2}$.

b) Equation (16) and Section 2 explain why $c_{n,1,2}$ is used when $m = 2$. From (14), it follows that, for n large, the bias and the MSE reduction decrease more slowly to 0 when using $c_{n,k_n,2}$ with k_n increasing (see also Sec. 5.4, Remark 5).

For $1 \leq j \leq m$, let

$$h_j(x_1, \ldots, x_j) = E[h(X_1, \ldots, X_m)|X_1 = x_1, \ldots, X_j = x_j], \qquad \zeta_j = var[h_j(X_1, \ldots, X_j)].$$

Then, it holds that (see, e.g., Serfling 1980, p. 183)

$$\binom{n}{m} var(t_{n,m}) = \sum_{j=1}^{m} \binom{m}{j} \binom{n-m}{m-j} \zeta_j, \tag{18}$$

and

$$0 \leq \zeta_1 \leq \zeta_2 \leq \ldots \leq \zeta_m = var[h(X_1, \ldots, X_m)]. \tag{19}$$

Proposition 5.2 *Let $t_{n,m}$ be as in (8) with $h \in V_m$.*

a) *The following statements are each sufficient for the shrinkage estimate $c_{\delta_{n,1,m}} t_{n,m}$ to have smaller MSE than $t_{n,m}$ for every $\theta \in \Theta$; δ_n is determined by (22) and can be chosen to increase with n.*

1. There is a known constant $L_F > 0$ such that

$$\binom{n}{m} \inf_{\theta \in \Theta} \frac{var(t_{n,m})}{\theta^2} > L_F \tag{20}$$

2. There is a known constant $L_F > 0$ and $j_0, 1 \leq j_0 \leq m$, such that

$$\inf_{\theta \in \Theta} \frac{\zeta_{j_0}}{\theta^2} > L_F. \tag{21}$$

b) *When $m = 2$, either (20) or (21) is sufficient for the shrinkage estimate $c_{\delta_{n,k,2}} t_{n,2}$ to have smaller MSE than $t_{n,2}$ for every $\theta \in \Theta$; δ_n is determined by (23) and can be chosen to increase with n.*

Proof: For part **a)** 1, use (5) and choose an increasing sequence δ_n such that for every n, it holds that

$$\binom{n}{m} \frac{1 - c_{\delta_n,1,m}}{1 + c_{\delta_n,1,m}} = \binom{n}{m} \frac{m^2 - m}{2\delta_n^2 + 2\delta_n - m^2 + m} < L_F. \tag{22}$$

The proof of part **a)** 2 follows from part **a)** 1, because from (18) and (19) it holds that

$$\binom{n}{m} var(t_{n,m}) \geq \zeta_m \geq \ldots \geq \zeta_1.$$

For part **b)** to prove sufficiency of (20), use (5) and choose an increasing sequence δ_n such that for every n, it holds that

$$\binom{n}{2} \frac{1 - c_{\delta_n,k,2}}{1 + c_{\delta_n,k,2}} = \binom{n}{2} \frac{k}{(\delta_n + k)(\delta_n + k - 1) - k} < L_F. \tag{23}$$

Sufficiency of (21) follows as for part **a)** 1. □

Remark 2.

When (21) holds for $j_0 \leq m - 1$, (19) implies that it also holds for $j_0 = m$.

Corollary 5.1 *If $L = \inf_{F\in\mathcal{F}} L_F$ is positive, then the estimate $c_{\delta_n,1,m}t_{n,m}$ (resp. $c_{\delta_{n,k,2}}t_{n,2}$) obtained using L instead of L_F in (22) [resp. (23)] has smaller MSE than $t_{n,m}$ (resp. $t_{n,2}$) for all θ's and for all models $F \in \mathcal{F}$.*

5.3 Applications

The kernels for the population variance and covariance, Kendall's τ and Gini's index are, respectively,

$$\frac{(x_1-x_2)^2}{2}, \qquad \frac{(x_1-x_2)(y_1-y_2)}{2}, \qquad sign((x_1-x_2)(y_1-y_2)), \qquad |x_1-x_2|^\gamma, \gamma>0;$$

$sign(x) = 1$ if $x > 0, sign(x) = -1$ if $x < 0$ and vanishes when $x = 0$. The ordering of the coefficients $c_{n,k,2}$ in Proposition 5.1 c) and (5) suggest using $(n+1)$-augmented sample estimates when n is small; $\delta_n = n^r$ is used herein.

The Population Variance σ_X^2 and the Population Covariance σ_{XY}.

Let $(X_1, Y_1), \ldots, (X_n, Y_n)$ be a two-dimensional sample with joint cumulative distribution function $F; \mu_X = EX_1, \mu_Y = EY_1, \sigma_X^2 = var(X_1), \sigma_Y^2 = var(Y_1), \sigma_{XY} = E(X_1-\mu_X)(Y_1-\mu_Y)$ and $\mu_{2,2} = E(X_1-\mu_X)^2(Y_1-\mu_Y)^2$.

Let $t_{n,2} = \frac{1}{n-1}\sum_{i=1}^n (X_i-\bar{X})(Y_i-\bar{Y})$ be the U-statistic estimating $\sigma_{XY}; \bar{X}, \bar{Y}$ are the averages of the X's and of the Y's. It holds that (Lee 1990, p. 14)

$$var(t_{n,2}) = \frac{\mu_{2.2}}{n} - \frac{(n-2)\sigma_{X,Y}^2 - \sigma_X^2\sigma_Y^2}{n(n-1)} = \frac{(n-1)(\mu_{2,2}-\sigma_{X,Y}^2) + \sigma_{X,Y}^2 + \sigma_X^2\sigma_Y^2}{n(n-1)} \tag{24}$$

and because

$$n(n-1)\frac{1-c_{n,1,2}}{1+c_{n,1,2}} = \frac{n^2-n}{n^2+n-1} < 1 < \frac{(n-1)(\mu_{2,2}-\sigma_{X,Y}^2) + \sigma_{X,Y}^2 + \sigma_X^2\sigma_Y^2}{\sigma_{X,Y}^2} \tag{25}$$

it follows from Corollary 5.1 that $\tilde{t}_{n,2}^1 = c_{n,1,2}t_{n,2}$ has a smaller MSE than $t_{n,2}$ for all values of $\sigma_{X,Y}$ and for any model F with $E_F X_1^2 < +\infty$ and $E_F Y_1^2 < +\infty, n \geq 2$.

When μ_X is unknown, σ_X^2 is usually estimated by $s_n^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$ even though $\frac{n-1}{n+1}s_n^2$ has a smaller MSE for normal populations but is not admissible (Stein, 1964). For nonnormal populations, s_n^2 may have smaller MSE than either $\frac{n-1}{n+1}s_n^2$ or $\hat{\sigma}^2 = \frac{n-1}{n}s_n^2$. Using (24) and (25), it follows that

$$\tilde{s}_n^2 = c_{n,1,2}s_n^2 = \frac{(n+2)(n-1)}{n(n+1)}s_n^2, \tag{26}$$

has smaller MSE that $t_{n,2} = s_n^2$ for all values of σ and for any model F with finite fourth moment and $n \geq 2$.

Kendall's τ

Let

$$t_{n,2} = \frac{2}{n(n-1)}\sum_{1\leq i<j\leq n} sign((X_i - X_j)(Y_i - Y_j))$$

be the U-statistic estimating

$$\tau = P[(X_2 - X_1)(Y_2 - Y_1) > 0] - P[(X_2 - X_1)(Y_2 - Y_1) < 0],$$

$\tau \in [-1+\epsilon, 1-\epsilon], 0 < \epsilon < 1$.

Then, there is positive integer $r = r(\epsilon)$ such that $\tilde{t}_{n,2}^1 = c_{n^r,1,2}t_n$ has a smaller MSE than t_n for $n \geq 2$. This follows from Proposition 5.2 **a)**, because it holds that (Lee 1990, p. 14)

$$var(t_{n,2}) = \frac{2}{n(n-1)} \cdot [2(n-2)var(E[sign(X_1 - X_2)(Y_1 - Y_2)|X_1 = x_1, Y_1 = y_1]) + 1 - \tau^2]$$

and that

$$\frac{n(n-1)}{2} \cdot \frac{var(t_{n,2})}{\tau^2} > \frac{n(n-1)}{2} \cdot var(t_{n,2}) > 1 - \tau^2 > 1 - (1-\epsilon)^2.$$

Gini's Index, g.

Let

$$t_{n,2} = \frac{2}{n(n-1)}\sum_{1\leq i<j\leq n} |X_i - X_j|^\gamma$$

be the U-statistic estimating $g = E|X_1 - X_2|^\gamma \in \mathcal{G}, \gamma > 0$. Assume that

$$\inf_{g \in \mathcal{G}} \frac{var|X_1 - X_2|^\gamma}{g^2} > L_F > 0.$$

Then, there is a positive integer $r = r(L_F)$ such that

$$\tilde{t}^1_{n,2} = c_{n^r,1,2} t_{n,2}$$

has a smaller MSE than t_n for $n \geq 2$. This follows from Proposition 5.2 **a)**, as in the previous example, because it holds that (Serfling 1980, p. 183)

$$var(t_{n,2}) = \frac{2}{n(n-1)}[2(n-2)varE(|X_1 - X_2|^\gamma | X_1 = x_1) + var|X_1 - X_2|^\gamma].$$

5.4 Additional Remarks

Remark 3

Equation (5), Proposition 5.2 and the results in the Applications Section 5.3 motivate the use of the shrinkage coefficient $c_{n^r,1,2}$ for any unbiased estimate t_n of θ for which

$$n^m \inf_{\theta \in \Theta} \frac{var(t_n)}{\theta^2} > L_F > 0,$$

with L_F known. r satisfies the inequality

$$\frac{1}{n^{2r-m} + n^{r-m} - n^{-m}} < L_F,$$

for $n \geq 2$ and $\tilde{t}_n = c_{n^r,1,2} t_n$ dominates t_n for all θ values, $n \geq 2$.

When L_F is not known, the shrinkage estimate asymptotically improves the unbiased estimate. For example, when estimating the mean μ of a distribution with $\bar{X}$ and the variance σ^2 is unknown, it follows from (5) and (15) with $k = 1$ that $\frac{(n-1)(n+2)}{n(n+1)}\bar{X}$ dominates $\bar{X}$ if $\frac{\mu^2}{\sigma^2} < \frac{n^2+n-1}{n}$, that is, if $\frac{\mu^2}{\sigma^2}$ is not "very large," which holds for n large.

Remark 4

When $c_{n^r,1,2}t_n$ is used instead of t_n, the amount of MSE reduction increases as the variance of t_n increases, and can be substantial irrespective of the sample size n. For example, in variance estimation for normal models, it holds that (Lehmann 1983, p. 113)

$$E\left[\tilde{c}\sum_{i=1}^{n}(X_i-\bar{x})^2-\sigma^2\right]^2=\sigma^4[(n^2-1)\tilde{c}^2-2(n-1)\tilde{c}+1] \tag{27}$$

and therefore the MSE reduction due to shrinkage is proportional to σ, which can take any positive value.

Remark 5

Rather than using $(n+k)$-augmented samples to obtain a shrinkage coefficient for $t_{n,m}$, a referee suggested finding $L>0$ such that $\inf_{\Theta,\mathcal{F}}\frac{var(t_{n,m})}{\theta^2}>L$ holds, then solve the equation $\frac{1-c}{1+c}=L$ to obtain the shrinkage estimate $ct_{n,m}$ that dominates $t_{n,m}$ for all $F\in\mathcal{F}$. However, the determination of a lower bound L is not straightforward, and the MSE reduction achieved with $ct_{n,m}$ may rapidly decrease to 0 as $n$increases. For example, in covariance estimation, it follows from (24) that for each model F, it holds that

$$\frac{var(t_{n,2})}{\sigma_{X,Y}^2}=\frac{1}{n(n-1)}+g(F,\sigma_{X,Y},n),$$

and it is not clear whether

$$\inf_{\Theta,\mathcal{F}}g(F,\sigma_{X,Y},n)>0$$

such that one can choose $L=\frac{1}{n(n-1)}$. In variance estimation, for the normal model with unknown mean, it holds that $\inf_{\Theta,\mathcal{F}}\frac{var(s_n^2)}{\sigma^4}>\frac{2}{n(n-1)}$ and one can choose $L_j=\frac{j}{n(n-1)}, j=1,2$. The solution of the equation $\frac{1-c}{1+c}=L_j$ is $c_{j,n}=\frac{n^2-n-j}{n^2-n+j}, j=1,2$, but observe that $c_{2,2}=0$.

In YY(2005), for sample sizes $n=5,10,15,20$ and for various c values, the corresponding value of $L=\frac{1-c}{1+c}$ and the MSE of the shrinkage estimate cs_n^2 are presented when $\sigma^2=1$. As n increases, the bias of cs_n^2 and the associated **MSE improvement** both decrease fast to 0, $c\in\{c_{n,k,2},c_{j,n};k=1,\ldots,4,j=1,2\}$. From (14), $1-c_{n,k,2}\sim\frac{2k}{n^2}$ and thus, for n large,

larger bias and MSE reduction can be achieved using coefficients $c_{n,k_n,2}$. One may choose, for example, $k_n = \gamma n, 0 < \gamma < 1$, to obtain

$$\frac{var(s_n^2) - E(c_{n,k_n,2}s_n^2 - \sigma^2)^2}{var(s_n^2)} \sim \frac{2}{n}\frac{\gamma}{(1+\gamma)^2}[2 - \frac{\gamma}{(1+\gamma)^2}].$$

6 Augmented Samples and the Jackknife

6.1 Jackknife Estimates and Pseudovalues

The $(n-1)$ jackknife estimate t_n^J (Quenouille, 1956) aims to reduce the bias of the estimate t_n of θ, and is the average of the pseudovalues $nt_n - (n-1)t_{n-1,i}, i = 1, ..., n$,

$$t_n^J = nt_n - \frac{1}{n}\sum_{i=1}^{n}(n-1)t_{n-1,i} = t_n + (n-1)(t_n - \frac{\sum_{i=1}^n t_{n-1,i}}{n}); \tag{28}$$

$t_{n-1,i} = t(X_1, \ldots, X_{i-1}, X_{i+1}, \ldots, X_n), i = 1, \ldots, n.$

Practice: Find the Jackknife estimate of the unbiased estimate of the variance, s_n^2, and of $\frac{n-1}{n}s_n^2$.

Equation (7) suggests using $(n+1)$-augmented samples to obtain the pseudo values

$$t_n + (n+1)(t_{n+1,i} - t_n) = (n+1)t_{n+1,i} - nt_n, \tag{29}$$

whose average

$$\tilde{t}_n^J = n^{-1}\sum_{i=1}^{n}[(n+1)t_{n+1,i} - nt_n] = t_n + (n+1)\left(\frac{\sum_{i=1}^n t_{n+1,i}}{n} - t_n\right), \tag{30}$$

is an alternative $(n+1)$ jackknife estimate; $t_{n+1,i} = t_{n+1}(X_1, \ldots, X_n, X_i),\ i = 1, \ldots, n.$

Note that in (28) and (30), the t_n corrections $(n-1)(t_n - \frac{\sum_{i=1}^n t_{n-1,j}}{n})$ and $(n+1)(\frac{\sum_{i=1}^n t_{n+1,j}}{n} - t_n)$ may have opposite signs and thus $\tilde{t}_n^J$ may increase the bias of t_n.

It should be mentioned that Hinkley (1978) and Beran (1984) used $(n+k)$-augmented samples, $k = 1, 2$, to study the properties of t_n^J but not for the purpose of deriving estimates.

Cabrera and Fernholz (1999) proposed a"target" estimate that, under model regularity conditions, has smaller bias and MSE than t_n.

6.2 MSE Reduction with t_n^J and $\tilde{t}_n^J$

From (28) [resp. (30)], it follows that t_n^J (resp. $\tilde{t}_n^J$) has a smaller MSE than t_n iff

$$E(t_n^J - t_n)^2 + 2(n-1)E(t_n - \theta)\left(t_n - \frac{\sum_{i=1}^n t_{n-1,i}}{n}\right) < 0 \tag{31}$$

$$\left[\text{resp. } E(\tilde{t}_n^J - t_n)^2 + 2(n+1)E(t_n - \theta)\left(\frac{\sum_{i=1}^n t_{n+1,i}}{n} - t_n\right) < 0\right]. \tag{32}$$

Because

$$E(t_n - \theta)\left(t_n - \frac{\sum_{i=1}^n t_{n-1,i}}{n}\right) \text{ and } E(t_n - \theta)\left(\frac{\sum_{i=1}^n t_{n+1,i}}{n} - t_n\right) \tag{33}$$

may have opposite signs, only one of (31) and (32) may hold. This is confirmed in the following example and for smooth functionals in Section 6.4.

Example. Let $X_1, \ldots, X_n$ be a sample from a **normal** distribution with unknown mean μ and variance σ^2, $\theta = \sigma^2$ and $\mathbf{t_n} = \hat{\sigma}^2$, the maximum likelihood estimate. Because t_n has smaller MSE than s_n^2 for every σ, (31) does not hold; s_n^2 is also the "target" estimate of σ^2 (Cabrera and Fernholz, 1999, sec. 4.1, p. 1093, l. 6 and 7). Among all estimates of σ^2 with form $c\sum_{i=1}^n (X_i - \bar{X})^2$, that with $c = (n+1)^{-1}$ minimizes the MSE (Goodman, 1953) and is the minimum risk-equivariant estimate (Lehmann 1983, p. 113). Because

$$\tilde{t}_n^J = \frac{1}{n+1}\sum_{i=1}^n (X_i - \bar{X})^2,$$

(32) holds.

Indeed, for any sample $X_1, \ldots, X_n, X_{n+1}$, it holds that

$$\bar{X}_{n+1} - \bar{X}_n = \frac{1}{n}(X_{n+1} - \bar{X}_n),$$

implying that

$$(n+1)t_{n+1} = nt_n + \frac{n}{n+1}(X_{n+1} - \bar{X}_n)^2, \tag{34}$$

For the $(n+1)$-augmented sample, $X_1, \ldots, X_n, X_{n+1} = X_i$, it follows from (34) that

$$(n+1)t_{n+1,i} = nt_n + \frac{n}{n+1}(X_i - \bar{X}_n)^2,$$

and therefore,

$$\tilde{t}_n^J = \frac{1}{n}\sum_{i=1}^{n}[(n+1)t_{n+1,i} - nt_n] = \frac{1}{n+1}\sum_{i=1}^{n}(X_i - \bar{X}_n)^2.$$

$\tilde{t}_n^J$ has a larger bias than t_n, because $Et_n = \sigma^2 - \frac{\sigma^2}{n}$ and $E\tilde{t}_n^J = \sigma^2 - \frac{2\sigma^2}{n+1}$

For the variance $\sigma^2 = \frac{m}{m-2}$ of the T_m distribution, estimates of the MSE based on 1,000 simulations indicate that $\tilde{t}_n^J$ has a smaller MSE than both t_n and t_n^J, $3 \le n \le 30, m = 3, 10, 20, 30,$ and that (31) does not hold. The graph of the results is presented in Figure 1 and Remark 7 in the last section (in YY 2005) provides the explanation.

Remark 6. The jackknife covariance estimate obtained with $(n+1)$-augmented samples is

$$\frac{1}{n+1}\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y}).$$

In the rest of YY2005, $t_n, t_n^J, \tilde{t}_n^J$ are examined asymptotically and conditions are given under which the augmented $(n+1)$-Jackknife $\tilde{t}_n^J$ improves the other estimates.

6.3 Von Mises Differentiable Statistical Functionals

Statistical Problem: $X_1, \ldots, X_n$ are $i.i.d,\ f(x|\theta)$, with *c.d.f.* $F(x, \theta)$. Let $\theta = T(F)$ be the parameter of interest and $\hat{\theta}_n$ its estimate. Assume $\hat{\theta}_n = T(\hat{F}_n)$, where $\hat{F}_n$ is the empirical *c.d.f.* of $X_1, \ldots, X_n$.

Example 6.1 Mean $\theta(F) = \mu(F) = T(F) = \int_R x dF(x)$, therefore

$$T(\hat{F}_n) = \int_R x d\hat{F}_n(x) = \frac{1}{n}\sum_{i=1}^{n} X_i = \bar{X}_n$$

since $d\hat{F}_n(x) = \frac{1}{n}$ when $x = X_i$ and is zero otherwise.

Example 6.2 Variance $\theta(F) = T(F) = \int_R (x - \mu(F))^2 dF(x)$, therefore

$$T(\hat{F}_n) = \int_R (x - \mu(\hat{F}_n))^2 d\hat{F}_n(x) = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X}_n)^2.$$

Example 6.3 p-th Quantile of $F = T(F) = F^{-1}(p)$ therefore

$$T(\hat{F}_n) = \hat{F}_n^{-1}(p) = p \text{ th sample quantile.}$$

Recall a Taylor Expansion:

$$f(x) - f(a) = (x-a)f'(a) + \frac{(x-a)^2}{2!}f^{(2)}(a) + \ldots + \frac{(x-a)^m}{m!}f^{(m)}(a) + R_m,$$

where R_m is the Remainder.

We can have a similar expansion for $T(\hat{F}_n)$:

$$T(\hat{F}_n) - T(F) = W_{mn} + R_{mn} = \sum_{j=1}^{m} \frac{d_j(F; \hat{F}_n - F)}{j!} + R_{mn},$$

where d_j indicates the j-th derivative to be determined and R_{mn} the stochastic Remainder that depends also on the sample size. Then we study $T(\hat{F}_n) - T(F)$ via W_{mn} and R_{mn}. For example, to show asymptotic normality of $T(\hat{F}_n) - T(F)$ we may use V_{1n} if it can be written as sum of $i.i.d.$ r.vs. and for the remainder it holds, as n increases to infinity,

$$n^{1/2} R_{1n} \xrightarrow{P} 0.$$

Informal Proposition (Von Mises) The type of asymptotic distribution of a differentiable statistical functional $T_n = T(\hat{F}_n)$ depends upon which is **the first nonvanishing term** in the Taylor development of the functional at the distribution F of the observations. If it is the linear term (the first), the limit distribution is normal under the usual assumptions for the Central Limit theorem. If the first non vanishing term is the one of order m, then the random variable $n^{m/2}[T(\hat{F}_n) - T(F)]$ converges in distribution to a random variable with finite variance.

The basic method of differentiation of a functional $T(F)$ is now described.

Definition 6.1 Let $\mathcal{F}$ be a space of distribution functions that includes the distribution functions $\{(1-\epsilon)F + \epsilon G; F \in \mathcal{F}, G \in \mathcal{F}; 0 \leq \epsilon \leq 1$. Consider a functional T defined on $(1-\epsilon)F + \epsilon G$ for all sufficiently small $\epsilon > 0$. If the limit

$$d_1 T(F; G - F) = \lim_{\epsilon \to 0^+} \frac{T[F + \epsilon(G - F)] - T(F)}{\epsilon}$$

exists, it is called the *Gateax differential* of T at F in the direction of G (or $(G - F)$).

Note that $d_1 T(F; G - F)$ is the usual derivative of ϵ from the right of zero, of the function $Q(\epsilon) = T[F + \epsilon(G - F)]$.

In general, the k-th order Gateaux differential of T at F in the direction G (or $G - F$) is

$$d_k T(F; G - F)) = \frac{d^k}{d\epsilon^k}\{T[F + \epsilon(G - F)] - T(F)\}|_{\epsilon=0+}.$$

Example 6.4 Let $T(F) = \int_R x dF(x), u \in R, G = G_{\delta_u}$ is the distribution of the Dirac function δ_u at u, observe that T is linear and

$$d_1 T(F; G_{\delta_u} - F) = \lim_{\epsilon \to 0^+} \frac{T[F + \epsilon(G_{\delta_u} - F)] - T(F)}{\epsilon} = \lim_{\epsilon \to 0^+} \frac{\epsilon T(G_{\delta_u} - F)}{\epsilon} = u - \mu.$$

$a_1(u) = u - \mu$ is the *Influence Function* for the mean, measuring the influence of u (or G_{δ_u}) used in a small perturbation of F.

An alternative approach to define a differential stronger than the Gateaux derivative follows the approach of differential in R^k with partial derivatives giving the directional derivatives. Let $\mathcal{D}$ be the linear space generated by the differences $G - H$ of members of the family of *c.d.f* $\mathcal{F}$, equipped with a norm $||\cdot||$. $\mathcal{D}$ can be represented as

$$\mathcal{D} = \{D = c(G - H); G, H \in \mathcal{F}, c \in R\}.$$

Definition 6.2 The functional T defined on $\mathcal{F}$ is said to have a differential at $F \in \mathcal{F}$ with respect to norm $||\cdot||$ if there exists functional $T(F; D)$ defined on $D \in \mathcal{D}$ **linear** in D :

$$T(G) - T(F) - T(F; G - F) = o(||G - F||)$$

as $||G - F|| \to 0$. $T(F; D)$ is the differential.

Remark 6.1 Linearity of $T(F; D)$ implies,

$$T(F; \sum_{i=1}^{k} a_i D_i) = \sum_{i=1}^{k} a_i T(F; D_i);$$

$a_i \in R, D_i \in \mathcal{D}, i = 1, \ldots, k$.

Remark 6.2 With the differential approach $T(\hat{F}_n) - T(F)$ is approximated by the random variable $T(F; \hat{F}_n - F)$ that will become a sum from previous Remark.

A proposition follows relating the differential $T(F; G - F)$ with the Gateaux derivative $d_1 T(F; G - F)$.

Proposition 6.1 If T has a differential $T(F; D)$ exists, then for any G the Gateaux derivative $d_1 T(F; G - F)$ exists and

$$d_1 T(F; G - F) = T(F; G - F).$$

Proof: Given G, observe that

$$(1 - \epsilon)F + \epsilon G - F = \epsilon(G - F).$$

Then, by linearity of T and since $||\epsilon(G - F)|| \to 0$ as ϵ decreases to zero,

$$T[(1 - \epsilon)F + \epsilon G] - T(F) = \epsilon T(F; G - F) + o(\epsilon||G - F||).$$

Therefore,

$$\lim_{\epsilon \to 0} \frac{T[(1 - \epsilon)F + \epsilon G] - T(F)}{\epsilon} = \lim_{\epsilon \to 0} \frac{\epsilon T(F; G - F) + \epsilon o(1)}{\epsilon} = T(F; G - F).$$

The differential $T(F; G - F)$ can be used to handle also the Remainder R_{1n}. More details in Serfling's book, Chapter on Von Mises differentiable statistical functions.

7 Pitman Closeness Criterion and Shrinkage

Given two estimates $\hat{\theta}_1$ and $\hat{\theta}_2$ of an unknown parameter θ, Pitman (1937) suggested that $\hat{\theta}_1$ should be regarded as a "closer" estimate of θ if

$$P\left(|\hat{\theta}_2 - \theta| > |\hat{\theta}_1 - \theta|\right) > 1/2.$$

This criterion, which is often called Pitman closeness, has an intuitive appeal and is in accordance with statistical tradition that preference should be expressed on a probability scale.

Much attention has been given to Pitman closeness criterion (PCC) properties in the 90's. It has been sharply criticized by some and vigorously defended by others on various counts. A good illustration of the debate is the paper by Robert, Hwang, and Strawderman (1993) and the subsequent discussion by Blyth; Casella and Wells; Ghosh, Keating, and Sen; Peddada; and Rao; in which different views in PCC's favor or against it are presented.

Leaving the controversy behind, the object of this communication (by Gerard Biau and YY) is to compare PCC with the familiar concept of mean squared error for variance estimation purposes. For a large class of distributions and large samples, it is shown herein that estimates of the variance σ^2 and of the standard deviation σ are more often "closer" to their target than the corresponding shrinkage estimates which improve the mean squared error. The same phenomenon is also observed for small and moderate sample sizes.

Our results indicate that PCC should be regarded as a useful and complementary tool for the evaluation of estimates of σ^2 and of σ, in agreement with Professor C. R. Rao's comment (in Robert *et al* 1993):

"I believe that the performance of an estimator should be examined under different criteria to understand the nature of the estimator and possibly to provide information to the decision-maker. I would include PCC in my list of criteria, except perhaps in the rare case where the

customer has a definite loss function".

To go straight to the point, suppose that $X_1, \ldots, X_n$ ($n \geq 2$) are independent, identically distributed (i.i.d.) real-valued random variables, with unknown mean and unknown finite positive variance σ^2. We consider here the estimation problem of the variance σ^2. Set

$$\bar{X}_n = \frac{1}{n} \sum_{i=1}^{n} X_i.$$

The sample variance estimate

$$S^2_{\mathrm{SV},n} = \frac{1}{n} \sum_{i=1}^{n} \left(X_i - \bar{X}_n\right)^2$$

and the unbiased estimate

$$S^2_{\mathrm{U},n} = \frac{1}{n-1} \sum_{i=1}^{n} \left(X_i - \bar{X}_n\right)^2$$

are both standard statistical procedures to estimate σ^2.

However, assuming squared error loss, more general estimates of the form

$$\delta_n \sum_{i=1}^{n} \left(X_i - \bar{X}_n\right)^2,$$

where $(\delta_n)_n$ is a positive sequence, are often preferred. For example, if $X_1, \ldots, X_n$ are sampled from a normal distribution, Goodman (1953) proved that we can improve upon $S^2_{\mathrm{SV},n}$ and $S^2_{\mathrm{U},n}$ uniformly by taking $\delta_n = 1/(n+1)$. This means, setting

$$S^2_{\mathrm{M},n} = \frac{1}{n+1} \sum_{i=1}^{n} \left(X_i - \bar{X}_n\right)^2,$$

that for all n and all values of the parameter,

$$E\left[S^2_{\mathrm{M},n} - \sigma^2\right]^2 < E\left[S^2_{\mathrm{SV},n} - \sigma^2\right]^2 \quad \text{and} \quad E\left[S^2_{\mathrm{M},n} - \sigma^2\right]^2 < E\left[S^2_{\mathrm{U},n} - \sigma^2\right]^2.$$

To see this, it suffices to note that, in the normal setting,

$$E\left[\delta_n \sum_{i=1}^{n} \left(X_i - \bar{X}_n\right)^2 - \sigma^2\right]^2 = \sigma^4 \left[\left(n^2-1\right)\delta_n^2 - 2(n-1)\delta_n + 1\right], \tag{35}$$

and that the right-hand side is uniformly minimized by $\delta_n^\star = 1/(n+1)$ (Lehmann and Casella, 1998, Chapter 2).

Since the values $\delta_n = 1/n$ and $\delta_n = 1/(n-1)$, corresponding to $S^2_{\mathrm{SV},n}$ and $S^2_{\mathrm{U},n}$, respectively, lie on the same side of $1/(n+1)$, it is often referred to $S^2_{\mathrm{M},n}$ as a shrinked version of $S^2_{\mathrm{SV},n}$ and $S^2_{\mathrm{U},n}$, respectively. Put differently, $S^2_{\mathrm{M},n} = c_n S^2_{\mathrm{SV},n}$ (respectively, $S^2_{\mathrm{M},n} = \tilde{c}_n S^2_{\mathrm{U},n}$) where, for each n, c_n (respectively, $\tilde{c}_n$) belongs to $(0,1)$.

Under different models and assumptions, inadmissibility results in variance and standard deviation estimation were proved using such estimates, among others, by Goodman (1953,1960), Stein (1964), Brown (1968), Arnold (1970) and Rukhin (1987) For a review of the topic, we refer the reader to Maatta and Casella (1999), who trace the history of the problem of estimating the variance based on a random sample from a normal distribution with unknown mean.

More recently, Yatracos (2005) provided shrinkage estimates of U-statistics based on artificially augmented samples and generalized, in particular, the variance shrinkage approach to non-normal populations by proving that, for all probability models with finite second moment, all values of σ^2 and all sample sizes $n \geq 2$, the estimate

$$S^2_{\mathrm{Y},n} = \frac{n+2}{n(n+1)} \sum_{i=1}^{n} \left(X_i - \bar{X}_n\right)^2$$

ameliorates the mean squared error of $S^2_{\mathrm{U},n}$. [Note that $S^2_{\mathrm{Y},n} = c_n S^2_{\mathrm{U},n}$ for some $c_n \in (0,1)$, so that $S^2_{\mathrm{Y},n}$ is in fact a shrinked version of $S^2_{\mathrm{U},n}$.]

Nevertheless, the variance shrinkage approach, which is intended to improve the mean squared error of estimates, should be carefully considered when performing point estimation. The rationale behind this observation is that the mean squared error is the average of the parameter estimation error over all samples whereas, in practice, we use an estimate's value based on one sample only and we care for the distance from its target.

To understand this remark, just consider the following example, due to Yatracos (2011).

Suppose again that $X_1, \ldots, X_n$ are independently normally distributed, with finite variance σ^2. Then, an easy calculation reveals that

$$P\left(\left|S^2_{\mathrm{M},n} - \sigma^2\right| > \left|S^2_{\mathrm{U},n} - \sigma^2\right|\right) = P\left(\chi^2_{n-1} < n - \frac{1}{n}\right), \tag{36}$$

where χ^2_{n-1} is a (central) chi-squared random variable with $n-1$ degrees of freedom (for a rigorous proof of this equality, see Lemma 10.1 in Section 10).

Figure 1 depicts the values of probability (36) for sample sizes ranging from 2 to 200. It is seen on this example that the probability slowly decreases towards the value 1/2, and that it may be significantly larger than 1/2 for small and even for moderate values of n.

Thus, Figure 1 indicates that, for a normal population, the standard unbiased estimate $S^2_{\mathrm{U},n}$ is Pitman closer to the target σ^2 than the shrinkage estimate $S^2_{\mathrm{M},n}$, despite the fact that, for all n,

$$E\left[S^2_{\mathrm{M},n} - \sigma^2\right]^2 < E\left[S^2_{\mathrm{U},n} - \sigma^2\right]^2.$$

Moreover, the advantage of $S^2_{\mathrm{U},n}$ with this respect becomes prominent for smaller values of n, and a similar phenomenon may be observed by comparing the probability performance of $S^2_{\mathrm{SV},n}$ vs $S^2_{\mathrm{M},n}$. In fact, our main Theorem 8.1 reveals (in the particular case of normal distribution) that

$$P\left(\left|S^2_{\mathrm{M},n} - \sigma^2\right| > \left|S^2_{\mathrm{U},n} - \sigma^2\right|\right) = \frac{1}{2} + \frac{5}{6\sqrt{\pi n}} + \mathrm{o}\left(\frac{1}{\sqrt{n}}\right)$$

and

$$P\left(\left|S^2_{\mathrm{M},n} - \sigma^2\right| > \left|S^2_{\mathrm{SV},n} - \sigma^2\right|\right) = \frac{1}{2} + \frac{13}{12\sqrt{\pi n}} + \mathrm{o}\left(\frac{1}{\sqrt{n}}\right),$$

that is, $S^2_{\mathrm{U},n}$ and $S^2_{\mathrm{SV},n}$ are both asymptotically Pitman closer to σ^2 than $S^2_{\mathrm{M},n}$. It is therefore clear, at least on these Gaussian examples, that we should be cautious when choosing to shrink the variance for point estimation purposes.

In the present paper, we generalize this discussion to a large class of distributions. Taking a more general point of view, we let $X_1, \ldots, X_n$ be a sample drawn according to some unknown

distribution with finite variance σ^2, and consider two candidates to estimate σ^2, namely

$$S_{1,n}^2 = \alpha_n \sum_{i=1}^{n} \left(X_i - \bar{X}_n\right)^2 \quad \text{and} \quad S_{2,n}^2 = \beta_n \sum_{i=1}^{n} \left(X_i - \bar{X}_n\right)^2.$$

Assuming mild moment conditions on the sample distribution, our main result (Theorem 8.1) offers an asymptotic development of the form

$$P\left(\left|S_{2,n}^2 - \sigma^2\right| \geq \left|S_{1,n}^2 - \sigma^2\right|\right) = \frac{1}{2} + \frac{\Delta}{\sqrt{n}} + \mathrm{o}\left(\frac{1}{\sqrt{n}}\right),$$

where the quantity Δ depends both on the moments of the distribution and the ratio of the sequences $(\alpha_n)_n$ and $(\beta_n)_n$. It is our belief that this probability should be reported in priority before deciding whether to use $S_{2,n}^2$ instead of $S_{1,n}^2$, depending on the sign and values of Δ. Standard distribution examples together with classical variance estimates are discussed, and similar results pertaining to the estimation of the standard deviation σ are also reported.

8 Main results

As for now, we let $X_1, \ldots, X_n$ ($n \geq 2$) be independent and identically distributed real-valued random variables, with unknown finite variance $\sigma^2 > 0$. Throughout the document, we let X be a generic random variable distributed as X_1 and make the following assumption on the distribution of X:

Assumption [A] Let $m = EX$. Then

(i) $EX^6 < \infty$ and $\tau > 0$, where

$$\tau^2 = E\left[\frac{X - m}{\sigma}\right]^4 - 1,$$

(ii) and

$$\limsup_{|u|+|v|\to\infty} \left|E \exp\left(iuX + ivX^2\right)\right| < 1.$$

The latter restriction, often called Cramér's condition, holds if the distribution of X is nonsingular or, equivalently, if that distribution has a nondegenerate absolutely continuous component—in particular, if X has a proper density function. A proof of this fact is given in Hall (1992, Chapter 2).

On the basis of the given sample $X_1, \ldots, X_n$, we wish to estimate σ^2. In this context, suppose that we are given two estimates $S_{1,n}^2$ and $S_{2,n}^2$, respectively defined by

$$S_{1,n}^2 = \alpha_n \sum_{i=1}^{n} \left(X_i - \bar{X}_n\right)^2 \quad \text{and} \quad S_{2,n}^2 = \beta_n \sum_{i=1}^{n} \left(X_i - \bar{X}_n\right)^2, \tag{37}$$

where $(\alpha_n)_n$ and $(\beta_n)_n$ are two positive sequences. Examples of such sequences have already been reported in the introduction section, and various additional illustrations will be discussed below. As a leading example, the reader should keep in mind the normal case, with $\alpha_n = 1/(n-1)$ (unbiased estimate) and $\beta_n = 1/(n+1)$ (minimum quadratic risk estimate). We first state our main result, whose proof relies on the technique of Edgeworth expansion (see, e.g., Hall, 1992, Chapter 2).

Theorem 8.1 *Assume that Assumption* $[\mathbf{A}]$ *is satisfied, and that the sequences* $(\alpha_n)_n$ *and* $(\beta_n)_n$ *in (37) satisfy the constraints*

$$(i)\ \beta_n < \alpha_n \quad \textit{and} \quad (ii)\ \frac{2}{\alpha_n + \beta_n} = n + a + \mathrm{o}(1) \quad \textit{as } n \to \infty,$$

where $a \in R$. *Then, for the estimates* $S_{1,n}^2$ *and* $S_{2,n}^2$ *in (37),*

$$P\left(\left|S_{2,n}^2 - \sigma^2\right| \geq \left|S_{1,n}^2 - \sigma^2\right|\right) = \frac{1}{2} + \frac{1}{\sqrt{2\pi n}}\left[\frac{a+1}{\tau} - \frac{1}{\tau^3}\left(\gamma^2 - \frac{\lambda}{6}\right)\right] + \mathrm{o}\left(\frac{1}{\sqrt{n}}\right)$$

as $n \to \infty$, *where*

$$\gamma = E\left[\frac{X-m}{\sigma}\right]^3 \quad \textit{and} \quad \lambda = E\left[\left(\frac{X-m}{\sigma}\right)^2 - 1\right]^3.$$

Some comments are in order to explain the meaning of the requirements of Theorem 8.1. Condition (i) may be interpreted by considering that $S_{2,n}^2$ is a shrinked version of $S_{1,n}^2$. For example,

in the normal population context, we typically have the ordering

$$\frac{1}{n+1} < \frac{1}{n} < \frac{1}{n-1},$$

which corresponds to the successive shrinked estimates $S^2_{\mathrm{M},n}$, $S^2_{\mathrm{SV},n}$ and $S^2_{\mathrm{U},n}$. To understand condition (ii), it is enough to note that an estimate of σ^2 of the form $\delta_n \sum_{i=1}^n (X_i - \bar{X}_n)^2$ is (weakly or strongly) consistent if, and only if, $\delta_n \sim 1/n$ as $n \to \infty$. Therefore, for consistent estimates $S^2_{1,n}$ and $S^2_{2,n}$, it holds $2/(\alpha_n + \beta_n) \sim n$, and condition (ii) just specifies this asymptotic development.

Finally, it is noteworthy to mention that all presented results may be adapted without too much effort to the known mean case, by replacing $\sum_{i=1}^n (X_i - \bar{X}_n)^2$ by $\sum_{i=1}^n (X_i - m)^2$ in the corresponding estimates. To see this, it suffices to observe that the proof of Theorem 8.1 starts with the following asymptotic normality result (see Proposition 10.1):

$$\sqrt{n}\, \frac{\frac{1}{n}\sum_{i=1}^n (Z_i - \bar{Z}_n)^2 - 1}{\tau} \xrightarrow{\mathcal{D}} \mathcal{N}(0,1) \quad \text{as } n \to \infty, \tag{38}$$

where

$$Z_i = \frac{X_i - m}{\sigma} \quad \text{and} \quad \tau^2 = E\left[\frac{X-m}{\sigma}\right]^4 - 1.$$

When the mean m is known, (38) has to be replaced by

$$\sqrt{n}\, \frac{\frac{1}{n}\sum_{i=1}^n Z_i^2 - 1}{\tau} \xrightarrow{\mathcal{D}} \mathcal{N}(0,1) \quad \text{as } n \to \infty,$$

and the subsequent developments are similar. We leave the interested reader the opportunity to adapt the results to this less interesting situation.

We are now in a position to discuss some application examples.

Example 1 Suppose that $X_1, \ldots, X_n$ are independently normally distributed with unknown positive variance σ^2. Elementary calculations show that, in this setting, $\tau^2 = 2$, $\gamma = 0$ and $\lambda = 8$.

The sample variance (maximum likelihood) $S^2_{\text{SV},n}$ has $\alpha_n = 1/n$, whereas the unbiased (jacknife) estimate $S^2_{\text{U},n}$ has $\alpha_n = 1/(n-1)$. The minimum risk estimate $S^2_{\text{M},n}$, which minimizes the mean squared error uniformly in n and σ^2, has $\beta_n = 1/(n+1)$ (Lehmann and Casella (1998, Chapter 2). Thus, $S^2_{\text{M},n}$ is a shrinked version of both $S^2_{\text{SV},n}$ and $S^2_{\text{U},n}$ (that is, $\beta_n < \alpha_n$), with

$$\frac{2}{\alpha_n + \beta_n} = \frac{2n^2 + 2n}{2n+1} = n + \frac{1}{2} + \text{o}(1) \quad \text{for } S^2_{\text{M},n} \text{ vs } S^2_{\text{SV},n}, \tag{39}$$

and

$$\frac{2}{\alpha_n + \beta_n} = \frac{n^2 - 1}{n} = n - \frac{1}{n} \quad \text{for } S^2_{\text{M},n} \text{ vs } S^2_{\text{U},n}. \tag{40}$$

Therefore, in this context, Theorem 8.1 asserts that

$$P\left(\left|S^2_{\text{M},n} - \sigma^2\right| > \left|S^2_{\text{SV},n} - \sigma^2\right|\right) = \frac{1}{2} + \frac{13}{12\sqrt{\pi n}} + \text{o}\left(\frac{1}{\sqrt{n}}\right)$$

and

$$P\left(\left|S^2_{\text{M},n} - \sigma^2\right| > \left|S^2_{\text{U},n} - \sigma^2\right|\right) = \frac{1}{2} + \frac{5}{6\sqrt{\pi n}} + \text{o}\left(\frac{1}{\sqrt{n}}\right).$$

Put differently, $S^2_{\text{SV},n}$ and $S^2_{\text{U},n}$ are both asymptotically Pitman closer to σ^2 than $S_{\text{M},n}$. It is also interesting to note that, according to (35), the maximum likelihood estimate has uniformly smaller risk than the unbiased estimate, i.e., for all n and all values of the parameter,

$$E\left[S^2_{\text{SV},n} - \sigma^2\right]^2 < E\left[S^2_{\text{U},n} - \sigma^2\right]^2.$$

Clearly, $S^2_{\text{SV},n}$ may be regarded as a shrinkage estimate of $S^2_{\text{U},n}$ and, with $\alpha_n = 1/(n-1)$ and $\beta_n = 1/n$, we obtain

$$\frac{2}{\alpha_n + \beta_n} = \frac{2n^2 - 2n}{2n-1} = n - \frac{1}{2} + \text{o}(1),$$

so that

$$P\left(\left|S^2_{\text{SV},n} - \sigma^2\right| > \left|S^2_{\text{U},n} - \sigma^2\right|\right) = \frac{1}{2} + \frac{7}{12\sqrt{\pi n}} + \text{o}\left(\frac{1}{\sqrt{n}}\right).$$

The take-home message here is that even if shrinkage improves the risk of squared error loss, it should nevertheless be carefully considered from a point estimation perspective. In particular,

the unbiased estimate $S^2_{\mathrm{U},n}$ is asymptotically Pitman closer to the target σ^2 than the shrinked (and mean squared optimal) estimate $S^2_{\mathrm{M},n}$. We have indeed

$$\lim_{n\to\infty} \sqrt{n}\left[P\left(\left|S^2_{\mathrm{M},n}-\sigma^2\right| > \left|S^2_{\mathrm{U},n}-\sigma^2\right|\right) - \frac{1}{2}\right] = \frac{5}{6\sqrt{\pi}},$$

despite the fact that, for all n,

$$E\left[S^2_{\mathrm{M},n}-\sigma^2\right]^2 < E\left[S^2_{\mathrm{U},n}-\sigma^2\right]^2.$$

This clearly indicates a potential weakness for any estimate obtained by minimizing a risk function, because extreme estimate's values that have small probability can drastically increase the risk function's value.

To continue the discussion, we may denote by ℓ a real number less than 1 and consider variance estimates of the general form

$$S^2_{\ell,n} = \frac{1}{n+\ell}\sum_{i=1}^{n}(X_i-\bar{X}_n)^2, \quad n > -\ell. \tag{41}$$

Clearly, $S^2_{\mathrm{M},n}$ is a shrinked version of $S^2_{\ell,n}$ and, in the normal setting, for all $n > -\ell$,

$$E\left[S^2_{\mathrm{M},n}-\sigma^2\right]^2 < E\left[S^2_{\ell,n}-\sigma^2\right]^2.$$

Next, applying Theorem 8.1 with

$$\alpha_n = \frac{1}{n+\ell} \quad \text{and} \quad \beta_n = \frac{1}{n+1},$$

we may write

$$\frac{2}{\alpha_n+\beta_n} = n + \frac{\ell+1}{2} + \mathrm{o}(1)$$

and, consequently,

$$P\left(\left|S^2_{\mathrm{M},n}-\sigma^2\right| > \left|S^2_{\ell,n}-\sigma^2\right|\right) = \frac{1}{2} + \frac{1}{4\sqrt{\pi n}}\left(\ell+\frac{13}{3}\right) + \mathrm{o}\left(\frac{1}{\sqrt{n}}\right).$$

The multiplier of the $1/\sqrt{n}$ term is positive for all $\ell > -13/3 \approx -4.33$. Thus, for $\ell \in (-13/3, 1)$, the estimate (41) is asymptotically Pitman closer to σ^2 than $S^2_{\mathrm{M},n}$, the minimum

quadratic risk estimate. Note that this result is in accordance with Pitman's observation that, in the Gaussian case, the best variance estimate with respect to PCC should have approximately $\alpha_n \approx 1/(n - 5/3)$ (Pitman, 1937, Paragraph 6).

Example 2 If $X_1, \ldots, X_n$ follow a Student's t-distribution with $\nu > 6$ degrees of freedom and unknown variance σ^2, then it is known (see, e.g., Yatracos, 2005, Remark 7) that $S^2_{\mathrm{M},n}$ improves both $S^2_{\mathrm{SV},n}$ and $S^2_{\mathrm{U},n}$ in terms of quadratic error. In this case, $m = 0$, $\gamma = 0$, whereas, for $0 < k < \nu$, even,

$$EX^k = \nu^{k/2} \prod_{j=1}^{k/2} \frac{2j-1}{\nu - 2j}.$$

Therefore,

$$\sigma^2 = \frac{\nu}{\nu-2}, \quad EX^4 = \frac{3\nu^2}{(\nu-2)(\nu-4)} \quad \text{and} \quad EX^6 = \frac{15\nu^3}{(\nu-2)(\nu-4)(\nu-6)}.$$

Consequently,

$$\tau^2 = \left(\frac{\nu-2}{\nu}\right)^2 \times \frac{3\nu^2}{(\nu-2)(\nu-4)} - 1 = \frac{2\nu-2}{\nu-4}$$

and

$$\begin{aligned}\lambda &= \left(\frac{\nu-2}{\nu}\right)^3 \times \frac{15\nu^3}{(\nu-2)(\nu-4)(\nu-6)} - 3\left(\frac{\nu-2}{\nu}\right)^2 \times \frac{3\nu^2}{(\nu-2)(\nu-4)} + 2 \\ &= \frac{8\nu(\nu-1)}{(\nu-4)(\nu-6)}.\end{aligned}$$

Hence, using identities (39)-(40), Theorem 8.1 takes the form

$$P\left(\left|S^2_{\mathrm{M},n} - \sigma^2\right| > \left|S^2_{\mathrm{SV},n} - \sigma^2\right|\right) = \frac{1}{2} + \frac{1}{6\sqrt{\pi n}} \left(\frac{\nu-4}{\nu-1}\right)^{1/2} \left(\frac{13\nu/2 - 27}{\nu-6}\right) + \mathrm{o}\left(\frac{1}{\sqrt{n}}\right)$$

and

$$P\left(\left|S^2_{\mathrm{M},n} - \sigma^2\right| > \left|S^2_{\mathrm{U},n} - \sigma^2\right|\right) = \frac{1}{2} + \frac{1}{6\sqrt{\pi n}} \left(\frac{\nu-4}{\nu-1}\right)^{1/2} \left(\frac{5\nu - 18}{\nu-6}\right) + \mathrm{o}\left(\frac{1}{\sqrt{n}}\right).$$

We see that all the constants in front of the $1/\sqrt{n}$ terms are positive for $\nu > 6$, despite the fact that

$$E\left[S^2_{\mathrm{M},n} - \sigma^2\right]^2 < E\left[S^2_{\mathrm{SV},n} - \sigma^2\right]^2 \quad \text{and} \quad E\left[S^2_{\mathrm{M},n} - \sigma^2\right]^2 < E\left[S^2_{\mathrm{U},n} - \sigma^2\right]^2.$$

Example 3 For non-normal populations, $S^2_{\mathrm{U},n}$ may have a smaller mean squared error than either $S^2_{\mathrm{SV},n}$ or $S^2_{\mathrm{M},n}$. In this general context, Yatracos (2005) proved that the estimate

$$S^2_{\mathrm{Y},n} = \frac{n+2}{n(n+1)} \sum_{i=1}^{n} \left(X_i - \bar{X}_n\right)^2$$

ameliorates the mean squared error of $S^2_{\mathrm{U},n}$ for all probability models with finite second moment, all values of σ^2 and all sample sizes $n \geq 2$. Here,

$$\alpha_n = \frac{1}{n-1} \quad \text{and} \quad \beta_n = \frac{n+2}{n(n+1)},$$

so that, for all $n \geq 2$, $\beta_n/\alpha_n < 1$ and

$$\frac{2}{\alpha_n + \beta_n} = \frac{n^3 - n}{n^2 + n - 1} = n - 1 + \mathrm{o}(1).$$

It follows, assuming that Assumption $[\mathbf{A}]$ is satisfied, that

$$P\left(\left|S^2_{\mathrm{Y},n} - \sigma^2\right| \geq \left|S^2_{\mathrm{U},n} - \sigma^2\right|\right) = \frac{1}{2} + \frac{1}{\sqrt{2\pi n}} \left[-\frac{1}{\tau^3}\left(\gamma^2 - \frac{\lambda}{6}\right)\right] + \mathrm{o}\left(\frac{1}{\sqrt{n}}\right).$$

For example, if X follows a normal distribution,

$$P\left(\left|S^2_{\mathrm{Y},n} - \sigma^2\right| > \left|S^2_{\mathrm{U},n} - \sigma^2\right|\right) = \frac{1}{2} + \frac{1}{3\sqrt{\pi n}} + \mathrm{o}\left(\frac{1}{\sqrt{n}}\right).$$

9 Standard deviation shrinkage

The previous section was concerned with the shrinkage estimation problem of the variance σ^2. Estimating the standard deviation σ is more involved, since, for example, it is not possible to find an estimate of σ which is unbiased for all population distributions (Lehmann and Casella, 1998, Chapter 2). Nevertheless, interesting results may still be reported when the sample observations $X_1, \ldots, X_n$ follow a normal distribution $\mathcal{N}(m, \sigma^2)$.

The most common estimates used to assess the standard deviation parameter σ typically have the form

$$\sqrt{S^2_{\mathrm{SV},n}} = \frac{1}{\sqrt{n}} \left[\sum_{i=1}^{n} (X_i - \bar{X}_n)^2 \right]^{1/2} \text{ or } \sqrt{S^2_{\mathrm{U},n}} = \frac{1}{\sqrt{n-1}} \left[\sum_{i=1}^{n} (X_i - \bar{X}_n)^2 \right]^{1/2}.$$

In all generality, both $\sqrt{S^2_{\mathrm{SV},n}}$ and $\sqrt{S^2_{\mathrm{U},n}}$ are biased estimates of σ. However, when the random variable X is normally distributed, a minor correction exists to eliminate the bias. To derive the correction, just note that, according to Cochran's theorem, $\sum_{i=1}^{n} (X_i - \bar{X}_n)^2/\sigma^2$ has a chi-squared distribution with $n-1$ degrees of freedom. Consequently, $[\sum_{i=1}^{n} (X_i - \bar{X}_n)^2]^{1/2}/\sigma$ has a chi distribution with $n-1$ degrees of freedom (Johnson, Kotz, and Balakrishnan, 1994, Chapter 18) whence

$$E \left[\sum_{i=1}^{n} (X_i - \bar{X}_n)^2 \right]^{1/2} = \frac{\sqrt{2}\,\Gamma\left(\frac{n}{2}\right)}{\Gamma\left(\frac{n-1}{2}\right)} \sigma,$$

where $\Gamma(.)$ is the gamma function. It follows that the quantity

$$\hat{\sigma}_{\mathrm{U},n} = \frac{\Gamma\left(\frac{n-1}{2}\right)}{\sqrt{2}\,\Gamma\left(\frac{n}{2}\right)} \left[\sum_{i=1}^{n} (X_i - \bar{X}_n)^2 \right]^{1/2}$$

is an unbiased estimate of σ. Besides, still assuming normality and letting $(\delta_n)_n$ be some generic positive normalization sequence, we may write

$$E \left[\delta_n \left[\sum_{i=1}^{n} \left(X_i - \bar{X}_n\right)^2 \right]^{1/2} - \sigma \right]^2 = \delta_n^2\, E \left[\sum_{i=1}^{n} (X_i - \bar{X}_n)^2 \right] - 2\sigma\delta_n\, E \left[\sum_{i=1}^{n} (X_i - \bar{X}_n)^2 \right]^{1/2} + \sigma^2,$$

hence

$$E \left[\delta_n \left[\sum_{i=1}^{n} \left(X_i - \bar{X}_n\right)^2 \right]^{1/2} - \sigma \right]^2 = \sigma^2 \left[(n-1)\delta_n^2 - 2\delta_n \frac{\sqrt{2}\,\Gamma\left(\frac{n}{2}\right)}{\Gamma\left(\frac{n-1}{2}\right)} + 1 \right].$$

Solving this quadratic equation in δ_n, we see that the right-hand side is uniformly minimized for the choice

$$\delta_n^{\star} = \frac{\sqrt{2}\,\Gamma\left(\frac{n}{2}\right)}{(n-1)\Gamma\left(\frac{n-1}{2}\right)} = \frac{\Gamma\left(\frac{n}{2}\right)}{\sqrt{2}\,\Gamma\left(\frac{n+1}{2}\right)}.$$

(see Goodman, 1953). Put differently, the estimate

$$\hat{\sigma}_{\mathrm{M},n} = \frac{\Gamma\left(\frac{n}{2}\right)}{\sqrt{2}\,\Gamma\left(\frac{n+1}{2}\right)} \left[\sum_{i=1}^{n} (X_i - \bar{X}_n)^2 \right]^{1/2}$$

improves uniformly upon $\sqrt{S^2_{\mathrm{SV},n}}$, $\sqrt{S^2_{\mathrm{U},n}}$ and $\hat{\sigma}_{\mathrm{U},n}$, which have, respectively,

$$\delta_n = \frac{1}{\sqrt{n}}, \quad \delta_n = \frac{1}{\sqrt{n-1}} \quad \text{and} \quad \delta_n = \frac{\Gamma\left(\frac{n-1}{2}\right)}{\sqrt{2}\,\Gamma\left(\frac{n}{2}\right)}.$$

Using the expansion

$$\frac{\Gamma\left(\frac{n+1}{2}\right)}{\Gamma\left(\frac{n}{2}\right)} = \sqrt{\frac{n}{2}}\left[1 - \frac{1}{4n} + \mathrm{o}\left(\frac{1}{n}\right)\right],$$

we may write

$$\frac{\Gamma\left(\frac{n}{2}\right)}{\sqrt{2}\,\Gamma\left(\frac{n+1}{2}\right)} = \frac{1}{\sqrt{n}}\left[1 + \frac{1}{4n} + \mathrm{o}\left(\frac{1}{n}\right)\right] \tag{42}$$

and

$$\frac{\Gamma\left(\frac{n-1}{2}\right)}{\sqrt{2}\,\Gamma\left(\frac{n}{2}\right)} = \frac{\sqrt{2}\,\Gamma\left(\frac{n+1}{2}\right)}{(n-1)\Gamma\left(\frac{n}{2}\right)} = \frac{1}{\sqrt{n-1}}\left[1 + \frac{1}{4n} + \mathrm{o}\left(\frac{1}{n}\right)\right]. \tag{43}$$

The relative positions of the estimates $\sqrt{S^2_{\mathrm{SV},n}}$, $\hat{\sigma}_{\mathrm{M},n}$, $\sqrt{S^2_{\mathrm{U},n}}$ and $\hat{\sigma}_{\mathrm{U},n}$ together with their coefficients are shown in Figure 2.

Theorem 9.1 below is the standard deviation counterpart of Theorem 8.1 for normal populations. Let

$$\hat{\sigma}^2_{1,n} = \alpha_n\left[\sum_{i=1}^{n}\left(X_i - \bar{X}_n\right)^2\right]^{1/2} \quad \text{and} \quad \hat{\sigma}^2_{2,n} = \beta_n\left[\sum_{i=1}^{n}\left(X_i - \bar{X}_n\right)^2\right]^{1/2} \tag{44}$$

be two candidates to the estimation of σ.

Theorem 9.1 *Assume that X has a normal distribution, and that the sequences $(\alpha_n)_n$ and $(\beta_n)_n$ in (44) satisfy the constraints*

$$(i)\ \beta_n < \alpha_n \quad \textit{and} \quad (ii)\ \left[\frac{2}{\alpha_n + \beta_n}\right]^2 = n + b + \mathrm{o}(1) \quad \textit{as } n \to \infty,$$

where $b \in R$. Then, for the estimates $\hat{\sigma}_{1,n}$ and $\hat{\sigma}_{2,n}$ in (44)

$$P\left(|\hat{\sigma}_{2,n} - \sigma| > |\hat{\sigma}_{1,n} - \sigma|\right) = \frac{1}{2} + \frac{1}{2\sqrt{\pi n}}\left(b + \frac{5}{3}\right) + \mathrm{o}\left(\frac{1}{\sqrt{n}}\right)$$

as $n \to \infty$.

As expressed by Figure 2, $\hat{\sigma}_{\mathrm{M},n}$ is a shrinked version of both $\sqrt{S^2_{\mathrm{U},n}}$ and $\hat{\sigma}_{\mathrm{U},n}$. Thus, continuing our discussion, we may first compare the performance, in terms of Pitman closeness, of $\hat{\sigma}_{\mathrm{U},n}$ vs $\hat{\sigma}_{\mathrm{M},n}$. These estimates have, respectively,

$$\alpha_n = \frac{\Gamma\left(\frac{n-1}{2}\right)}{\sqrt{2}\,\Gamma\left(\frac{n}{2}\right)} \quad \text{and} \quad \beta_n = \frac{\Gamma\left(\frac{n}{2}\right)}{\sqrt{2}\,\Gamma\left(\frac{n+1}{2}\right)}.$$

Using (42) and (43), we easily obtain

$$\left[\frac{2}{\alpha_n + \beta_n}\right]^2 = n - 1 + \mathrm{o}(1),$$

so that

$$P\left(|\hat{\sigma}_{\mathrm{M},n} - \sigma| > |\hat{\sigma}_{\mathrm{U},n} - \sigma|\right) = \frac{1}{2} + \frac{1}{3\sqrt{\pi n}} + \mathrm{o}\left(\frac{1}{\sqrt{n}}\right).$$

Similarly, with

$$\alpha_n = \frac{1}{\sqrt{n-1}} \quad \text{and} \quad \beta_n = \frac{\Gamma\left(\frac{n}{2}\right)}{\sqrt{2}\,\Gamma\left(\frac{n+1}{2}\right)},$$

we conclude

$$P\left(|\hat{\sigma}_{\mathrm{M},n} - \sigma| > \left|\sqrt{S^2_{\mathrm{U},n}} - \sigma\right|\right) = \frac{1}{2} + \frac{11}{24\sqrt{\pi n}} + \mathrm{o}\left(\frac{1}{\sqrt{n}}\right).$$

Remark 1 The methodology developed in the present paper can serve as a basis for analyzing other types of estimates. Suppose, for example, that $X_1, \ldots, X_n$ ($n \geq 2$) are independent identically distributed random variables with common density $f(x; \mu, \sigma) = \sigma^{-1} e^{-(x-\mu)/\sigma}$ ($x > \mu$), where $-\infty < \mu < \infty$ and $\sigma > 0$. On the basis of the given sample we wish to estimate the standard deviation σ. Denoting the order statistics associated with $X_1, \ldots, X_n$ by $X_{(1)}, \ldots, X_{(n)}$, one may write the maximum likelihood estimate of σ (which turns out to be minimum variance unbiased) in the form

$$T_{\mathrm{ML},n} = \frac{1}{n-1} \sum_{i=2}^{n} (X_{(i)} - X_{(1)}).$$

By sacrificing unbiasedness, we can consider as well the estimate

$$T_{\mathrm{M},n} = \frac{1}{n} \sum_{i=2}^{n} (X_{(i)} - X_{(1)})$$

which improves upon $T_{\text{ML},n}$ uniformly (Arnold, 1970) in terms of mean squared error. $T_{\text{M},n}$ is a shrinkage estimate of $T_{\text{ML},n}$ and, by an application of Lemma 10.1, we have

$$P\left(|T_{\text{M},n} - \sigma| > |T_{\text{ML},n} - \sigma|\right) = P\left(\Gamma_{n-1} < \frac{2n(n-1)}{2n-1}\right),$$

since $\sum_{i=2}^{n}(X_{(i)} - X_{(1)})/\sigma$ is distributed as a gamma random variable with $n-1$ degrees of freedom, denoted Γ_{n-1}. Recalling that $\Gamma_{n-1} \sim \sum_{i=1}^{n-1} Y_i$, where $Y_1, \ldots, Y_{n-1}$ are independent standard exponential random variables, we easily obtain, using the same Edgeworth-based methodology as was used to prove Theorem 8.1,

$$P\left(|T_{\text{M},n} - \sigma| > |T_{\text{ML},n} - \sigma|\right) = \frac{1}{2} + \frac{5}{6\sqrt{2\pi n}} + \mathrm{o}\left(\frac{1}{n}\right).$$

10 Proofs

10.1 Some preliminary results

Recall that $X_1, \ldots, X_n$ ($n \geq 2$) denote independent real-valued random variables, distributed as a generic random variable X with finite variance $\sigma^2 > 0$. Let $\Phi(x)$ be the cumulative distribution function of the standard normal distribution, that is, for all $x \in R$,

$$\Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-u^2/2} \mathrm{d}u.$$

We start by stating the following lemma, which is but a special case of Proposition 2.1 in Yatracos (2011)

Lemma 10.1 *Let T be a P-a.s. nonnegative real-valued random variable, and let $(\theta, c) \in R^+ \times (-1, 1)$ be two real numbers. Then*

$$P\left(|cT - \theta| \geq |T - \theta|\right) = P\left(T \leq \frac{2\theta}{1+c}\right).$$

Proof of Lemma 10.1 Just observe that

$$\begin{aligned} P\left(|cT-\theta| \geq |T-\theta|\right) &= P\left((cT-\theta)^2 \geq (T-\theta)^2\right) \\ &= P\left([(1+c)T-2\theta]\,(c-1)T \geq 0\right) \\ &= P\left((1+c)T-2\theta \leq 0\right) \\ &\quad \text{(since } T \text{ is } P\text{-a.s. nonnegative, } c<1 \text{ and } \theta \geq 0\text{)} \\ &= P\left(T \leq \frac{2\theta}{1+c}\right) \\ &\quad \text{(since } c > -1\text{)}. \end{aligned}$$

Proposition 10.1 *Assume that Assumption* $[\mathbf{A}]$ *is satisfied. Then, as* $n \to \infty$*,*

$$P\left(\sum_{i=1}^{n}(X_i-\bar{X}_n)^2 \leq (n+t)\sigma^2\right) = \Phi\left(\frac{t}{\tau\sqrt{n}}\right) + \frac{1}{\sqrt{2\pi n}}\,p_1\left(\frac{t}{\tau\sqrt{n}}\right)e^{-\frac{t^2}{2\tau^2 n}} + \mathrm{o}\left(\frac{1}{\sqrt{n}}\right),$$

uniformly in $t \in R$*, where*

$$p_1(x) = \frac{1}{\tau} + \frac{1}{\tau^3}\left(\gamma^2 - \frac{\lambda}{6}\right)(x^2-1),$$

with

$$\gamma = E\left[\frac{X-m}{\sigma}\right]^3 \quad \textit{and} \quad \lambda = E\left[\left(\frac{X-m}{\sigma}\right)^2 - 1\right]^3.$$

Proof of Proposition 10.1 Set

$$Z = \frac{X-m}{\sigma} \quad \text{and} \quad Z_i = \frac{X_i-m}{\sigma}, \quad i=1,\ldots,n,$$

and observe that, by the central limit theorem and Slutsky's lemma (van der Vaart, 1999, Chapter 2)

$$\sqrt{n}\,\frac{\frac{1}{n}\sum_{i=1}^{n}(Z_i-\bar{Z}_n)^2 - 1}{\tau} \xrightarrow{\mathcal{D}} \mathcal{N}(0,1) \quad \text{as } n \to \infty,$$

where

$$\bar{Z}_n = \frac{1}{n}\sum_{i=1}^{n} Z_i \quad \text{and} \quad \tau^2 = \mathrm{Var}Z^2 = E\left[\frac{X-m}{\sigma}\right]^4 - 1.$$

The result will be proved by making this limit more precise using an Edgeworth expansion (see, e.g., Hall, 1992, Chapter 2). To this aim, we first need some additional notation. Set $\mathbf{Z} = (Z, Z^2)$, $\mathbf{m} = E\mathbf{Z} = (0, 1)$ and, for $\mathbf{z} = (z^{(1)}, z^{(2)}) \in R^2$, let

$$A(\mathbf{z}) = \frac{z^{(2)} - (z^{(1)})^2 - 1}{\tau}.$$

Clearly, $A(\mathbf{m}) = 0$ and

$$\sqrt{n}\,\frac{\frac{1}{n}\sum_{i=1}^{n}(Z_i - \bar{Z}_n)^2 - 1}{\tau} = \sqrt{n}\,A(\bar{\mathbf{Z}}_\mathbf{n}).$$

For $j \geq 1$ and $i_j \in \{1, 2\}$, put

$$a_{i_1 \ldots i_j} = \frac{\partial^j A(\mathbf{z})}{\partial z^{(i_1)} \ldots \partial z^{(i_j)}}|_{\mathbf{z}=\mathbf{m}}.$$

For example,

$$a_2 = \frac{\partial^1 A(\mathbf{z})}{\partial z^{(2)}}|_{\mathbf{z}=\mathbf{m}} = \frac{1}{\tau}$$

and

$$a_{11} = \frac{\partial^2 A(\mathbf{z})}{\partial z^{(1)} \partial z^{(1)}}|_{\mathbf{z}=\mathbf{m}} = -\frac{2}{\tau}.$$

Let also

$$\mu_{i_1 \ldots i_j} = E\left[(\mathbf{Z} - \mathbf{m})^{(\mathbf{i_1})} \ldots (\mathbf{Z} - \mathbf{m})^{(\mathbf{i_j})}\right],$$

where $(\mathbf{Z} - \mathbf{m})^{(\mathbf{i})}$ denotes the i-th component of the vector $(\mathbf{Z} - \mathbf{m})$. Thus, with this notation, according to Hall (1992, Theorem 2.2), under the condition

$$\limsup_{|u|+|v| \to \infty} \left|E \exp\left(iuX + ivX^2\right)\right| < 1,$$

we may write, as $n \to \infty$,

$$P\left(\sqrt{n}\,A(\bar{\mathbf{Z}}_\mathbf{n}) \leq x\right) = \Phi(x) + \frac{1}{\sqrt{2\pi n}}\,p_1(x)e^{-x^2/2} + \mathrm{o}\left(\frac{1}{\sqrt{n}}\right),$$

uniformly in $x \in R$, where

$$p_1(x) = -A_1 - \frac{1}{6}A_2(x^2 - 1).$$

The coefficients A_1 and A_2 in the polynomial p_1 are respectively given by the formulae

$$A_1 = \frac{1}{2}\sum_{i=1}^{2}\sum_{j=1}^{2} a_{ij}\mu_{ij}$$

and

$$A_2 = \sum_{i=1}^{2}\sum_{j=1}^{2}\sum_{k=1}^{2} a_i a_j a_k \mu_{ijk} + 3\sum_{i=1}^{2}\sum_{j=1}^{2}\sum_{k=1}^{2}\sum_{\ell=1}^{2} a_i a_j a_{k\ell}\mu_{ik}\mu_{j\ell}.$$

Elementary calculations show that

$$a_2 = \tau^{-1}, \quad a_{11} = -2\tau^{-1}, \quad \text{and} \quad a_1 = a_{22} = a_{12} = a_{21} = 0.$$

Similarly,

$$\mu_{11} = 1, \quad \mu_{22} = \tau^2, \quad \mu_{12} = \mu_{21} = E[X-m]^3\sigma^{-3}, \quad \text{and}$$

$$\mu_{222} = E\left[(X-m)^2\sigma^{-2} - 1\right]^3.$$

Consequently,

$$A_1 = -\frac{1}{\tau} \quad \text{and} \quad A_2 = \frac{1}{\tau^3}\left(\lambda - 6\gamma^2\right),$$

with

$$\lambda = E\left[\left(\frac{X-m}{\sigma}\right)^2 - 1\right]^3 \quad \text{and} \quad \gamma = E\left[\frac{X-m}{\sigma}\right]^3.$$

Therefore

$$p_1(x) = \frac{1}{\tau} + \frac{1}{\tau^3}\left(\gamma^2 - \frac{\lambda}{6}\right)(x^2 - 1).$$

The conclusion follows by observing that, for all $t \in R$,

$$P\left(\sum_{i=1}^{n}(X_i - \bar{X}_n)^2 \le (n+t)\sigma^2\right) = P\left(\sqrt{n}A(\bar{\mathbf{Z}}_\mathbf{n}) \le \frac{t}{\tau\sqrt{n}}\right).$$

10.2 Proof of Theorem 8.1

Observe that $S_{2,n}^2 = c_n S_{1,n}^2$, where $c_n = \beta_n/\alpha_n \in (0,1)$ by assumption (i). Consequently, by Lemma 10.1,

$$\begin{aligned} P\left(\left|S_{2,n}^2 - \sigma^2\right| \geq \left|S_{1,n}^2 - \sigma^2\right|\right) &= P\left(S_{1,n}^2 \leq \frac{2\sigma^2}{1+\beta_n/\alpha_n}\right) \\ &= P\left(\sum_{i=1}^n (X_i - \bar{X}_n)^2 \leq \frac{2\sigma^2}{\alpha_n + \beta_n}\right) \\ &= P\left(\sum_{i=1}^n (X_i - \bar{X}_n)^2 \leq (n + a + \zeta_n)\sigma^2\right) \end{aligned}$$

(by assumption (ii)),

where $\zeta_n \to 0$ as $n \to \infty$. Let $\Phi(x)$ be the cumulative distribution function of the standard normal distribution, that is, for all $x \in R$,

$$\Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-u^2/2} \mathrm{d}u.$$

Thus, assuming $[\mathbf{A}]$ and using Proposition 10.1, we may write

$$P\left(\left|S_{2,n}^2 - \sigma^2\right| \geq \left|S_{1,n}^2 - \sigma^2\right|\right) = \Phi\left(\frac{a+\zeta_n}{\tau\sqrt{n}}\right) + \frac{1}{\sqrt{2\pi n}}\, p_1\left(\frac{a+\zeta_n}{\tau\sqrt{n}}\right) e^{-\frac{(a+\zeta_n)^2}{2\tau^2 n}} + \mathrm{o}\left(\frac{1}{\sqrt{n}}\right),$$

where

$$p_1(x) = \frac{1}{\tau} + \frac{1}{\tau^3}\left(\gamma^2 - \frac{\lambda}{6}\right)(x^2 - 1),$$

with

$$\tau^2 = E\left[\frac{X-m}{\sigma}\right]^4 - 1,$$

$$\gamma = E\left[\frac{X-m}{\sigma}\right]^3 \quad \text{and} \quad \lambda = E\left[\left(\frac{X-m}{\sigma}\right)^2 - 1\right]^3.$$

Using finally the Taylor series expansions, valid as $x \to 0$,

$$\Phi(x) = \frac{1}{2} + \frac{1}{\sqrt{2\pi}}\left(x + \mathrm{o}(x^2)\right) \quad \text{and} \quad e^x = 1 + \mathrm{o}(1),$$

we obtain

$$P\left(\left|S_{2,n}^2-\sigma^2\right| \geq \left|S_{1,n}^2-\sigma^2\right|\right)=\frac{1}{2}+\frac{1}{\sqrt{2\pi n}}\left[\frac{a+1}{\tau}-\frac{1}{\tau^3}\left(\gamma^2-\frac{\lambda}{6}\right)\right]+\mathrm{o}\left(\frac{1}{\sqrt{n}}\right),$$

as desired.

10.3 Proof of Theorem 9.1

By assumption (i), we may write $\hat{\sigma}_{2,n}=c_n\hat{\sigma}_{1,n}$, where $c_n=\beta_n/\alpha_n \in (0,1)$. Consequently, by Lemma 10.1,

$$\begin{aligned}
P\left(|\hat{\sigma}_{2,n}-\sigma| > |\hat{\sigma}_{1,n}-\sigma|\right) &= P\left(|\hat{\sigma}_{2,n}-\sigma| \geq |\hat{\sigma}_{1,n}-\sigma|\right) \\
&= P\left(\hat{\sigma}_{1,n} \leq \frac{2\sigma}{1+\beta_n/\alpha_n}\right) \\
&= P\left(\left[\sum_{i=1}^{n}(X_i-\bar{X}_n)^2\right]^{1/2} \leq \frac{2\sigma}{\alpha_n+\beta_n}\right) \\
&= P\left(\sum_{i=1}^{n}(X_i-\bar{X}_n)^2 \leq (n+b+\zeta_n)\sigma^2\right) \\
&\quad \text{(by assumption } (ii)\text{)},
\end{aligned}$$

where $\zeta_n \to 0$ as $n \to \infty$. The end of the proof is similar to the one of Theorem 8.1, recalling that, in the Gaussian setting, $\tau^2=2$, $\gamma=0$ and $\lambda=8$.