R O B U S T N E S S - FALL 2015

ASSIGNMENT 1

Problem 1: Suppose that $y_1, y_2, ..., y_n$ are i.i.d. $N\left(\mu, \mu^2\right)$, for some unknown $\mu > 0$. Consider the following estimates for μ :

- The sample mean $\overline{y}$
- The sample standard deviation, s
- The ML-estimator, $\widehat{\mu}$
- The sample median, $\widehat{m}$

(A) Derive the ML-estimator for μ.

(B) Is the model, $N\left(\mu, \mu^2\right)$, translation equivariant? Scale equivariant?

(C) Are the estimators above translation equivariant? Scale equivariant?

(D) Propose a measure $d(T, \mu)$ to assess the performance of the given estimators, which remains invariant under changes in the data that preserve the model. **Hint:** make use of the equivariance properties of the considered estimators.

(E) Design and conduct a simulation study to compare the four estimates when the model holds. Show that by using the performance measure $d(T, \mu)$ you can restrict attention to a single canonical value of μ, for example $\mu = 1$. Report your results using tables, figures, etc. Briefly comment on your findings.

E) Suppose that $\alpha 100\%$ of the data are outliers generated from a normal distribution with a large mean (e.g. 5μ) and a small variance (e.g. 0.1μ). Take $\alpha = 0.01, 0.05$ and 0.10. Design and conduct a simulation study to compare the performance of the four estimates in this case. Report your results using tables, figures, etc. Briefly comment on your findings.

Problem 2:

(a) Compute the asymptotic Gaussian efficiency (against the sample mean) for Huber optimal location M-estimate that uses MAD as a fixed scale.

(b) Verify that $c = 1.345$ yields 95% asymptotic efficiency at the Normal distribution model. What value of c yields 99% asymptotic efficiency? What value of c yields 90% asymptotic efficiency?

(c) Suppose that the data $x_1, x_2, ..., x_n$ come from a student t distribution with k degrees of freedom, $F_k(x)$. Show that in this case

$$MAD(x_1, x_2, ..., x_n) \rightarrow \frac{F_k^{-1}(3/4)}{\Phi^{-1}(3/4)}.$$

(d) Compute the asymptotic Student t distribution efficiency (against the sample mean) for Huber optimal location M-estimate that uses MAD as a fixed scale. Consider the same normalized student t distributions as in the course slides. **NOTE:** you must take into account the limiting values for the MAD at the different Student t distributions.

(e) Calculate the asymptotic Student t distribution efficiency (against the sample mean) for median. Consider the same normalized student t distributions as in the course slides. Present the results from (d) and (e) in a table and briefly comment on these results.

Problem 3: One-step Newton-Raphson location M-estimates.

Suppose we wish to solve the location M estimating equation

$$\frac{1}{n}\sum_{i=1}^{n}\psi\left(\frac{y_i - t}{\hat{\sigma}_n}\right) = 0$$

where $\hat{\sigma}_n$ is square-root-of-n consistent, that is, $\sqrt{n}(\hat{\sigma}_n - \sigma)$ is bounded in probability (e.g. the MAD). Suppose that $\widehat{\mu}_0$ is an initial estimate for μ, which is also square-root consistent (e.g. the median).

(**A**) Show that the one-step Newton-Raphson location M-estimate is equal to

$$\widehat{\mu}_1 = \widehat{\mu}_0 + \frac{\frac{1}{n}\sum_{i=1}^{n}\psi\left(\frac{y_i - \widehat{\mu}_0}{\hat{\sigma}_n}\right)}{\frac{1}{n\hat{\sigma}_n}\sum_{i=1}^{n}\psi'\left(\frac{y_i - \widehat{\mu}_0}{\hat{\sigma}_n}\right)}.$$

and show that $\widehat{\mu}_1$ is consistent and fully efficient (has the same asymptotic variance as the fully iterated location M-estimate $\widehat{\mu}$).

(**B**) Design and conduct a simulation study to investigate the small finite sample properties of $\widehat{\mu}_1$ (compared with $\widehat{\mu}$), regarding efficiency.

(**C**) Design and conduct a simulation study to investigate the small finite sample properties of $\widehat{\mu}_1$ (compared with $\widehat{\mu}$), regarding robustness.

Hint: If $\widehat{\mu}_0$ is very robust but rather inefficient (e.g. the sample median) then $\widehat{\mu}_1$ should be as efficient as $\widehat{\mu}$ for clean data and more robust than $\widehat{\mu}$ for outlier-contaminated data. Do the results of your simulation study support this conjecture?