A review of recent robust inference: an approach based on maxbias

Victor J. Yohai
University of Buenos Aires, Department of Matematics
Ciudad Universitaria
Buenos Aires, Argentina
vyohai@sinectis.com.ar

Ruben H. Zamar
University of British Columbia, Department of Statistics
333-6356 Agricultural Road
Vancouver, Canada
ruben@stat.ubc.ca

1. Introduction

Sampling variability has constituted the main focus of the statistical theory and practice in the past century. The attention devoted by statisticians to this problem is justified in part by its obvious importance (specially for small samples) and also by the fact that sampling variability can be easily modeled and measured. In general, estimates face two sources of uncertainty: *sampling variability* (or variance) and *bias*. We notice that assessing the variance of robust estimates is not a trivial matter. Most asymptotic results apply under restrictive conditions which are not satisfied in situations when robust estimates are actually needed. More generally applicable variance estimates can be obtained using bootstrap and robust bootstrap recently introduced by Salian and Zamar (2000).

The bias caused by poor/uneven data quality, data contamination, gross errors, missing values, etc. has received much less attention. Unlike variability, bias is difficult to model and measure. We think that the current balance of research emphasis on these topics do not reflect their actual relative importance. For moderate and large data sets, the uncertainty due to *data quality* (bias) clearly dominates that due to *data quantity* (sampling variability). Notice that standard errors are usually of order $O(1/\sqrt{n})$ while biases are of order $O(1)$.

The robustness theory developed in the second half of past century can be used to address, at least in part, the issues of bias and data quality.

2. Assesing the Bias of Location M-Estimates

We will restrict attention to location M-estimates, $\widehat{\mu}_n$, that satisfies the equation

$$\sum \psi\left(\left(y_i - \widehat{\mu}_n\right)/\hat{\sigma}_n\right) = 0, \tag{1}$$

where ψ is an appropriate (monotone) score function and $\hat{\sigma}_n$ is a robust scale estimate that converges to $\sigma(F)$. Let $\eta(t, F, \psi) = -\int_{-\infty}^{\infty} \psi\left(\left(y - t\right)/\hat{\sigma}(F)\right) dF(y)$. Under very mild regularity conditions (see Huber, 1981) $\widehat{\mu}_n$ converges to the value $t = T(F, \psi)$ satisfying $\eta(t, F, \psi) = 0$.

We will assume that $\{y_1, \ldots, y_n\}$ are independent and identically distributed random variables with common distribution F in the family

$$\mathcal{V}(F_0, \epsilon) = \{F : F = (1-\epsilon)F_0 + \epsilon H, \quad H \ \text{ arbitrary}\}, \tag{2}$$

where $0 < \epsilon < 0.5$, $F_0(y) = \Phi((y-\mu_0)/\sigma_0)$, and $\Phi = N(0,1)$. The robustness model (2), called *Tukey's contamination neighborhood,* is a simple and flexible device to model datasets of uneven quality. According to this model $(1-\epsilon)100\%$ of the data follows a normal location-scale model and $\epsilon 100\%$ of the data comes from an arbitrary, unspecified source. We can choose ϵ and H to represent situations of asymmetry/heavy tailness of the error distribution, isolated outliers and cluster of outliers. We can interpret $(1-\epsilon)100\%$ as the minimum possible percentage of good quality measurements in our dataset. In summary, $(1-\epsilon)$ is a *parameter* measuring the *data quality.* It has also been recently observed that these neighborhoods can be used to model missing data.

In the case of robust estimates, the bias caused by outliers and other departures from symmetry can be assessed using the concept of maximum asymptotic bias. Suppose we have a large sample from a distribution F containing at most a fraction $\epsilon 100\%$ of contamination. Suppose we wish to bound the absolute difference

$$D(F, \psi) = |T(F, \psi) - T(F_0, \psi)|$$

between the M-location $T(F, \psi)$ of the contaminated distribution and the M-location $T(F_0, \psi)$ of the core (uncontaminated) distribution. Martin and Zamar (1993) showed that

$$\left|\left[T(F, \psi) - T(F_0, \psi)\right]/\sigma_0\right| \leq B(\epsilon)$$

where $B(\epsilon)$ is the solution in t to $\eta(t,\ \bar{F}, \psi) = 0$, with $\bar{F} = (1-\epsilon)F + \epsilon\delta_\infty$. Therefore, $D(F, \psi)$ is bounded by $\sigma_0 B(\epsilon)$.

Berrendero and Zamar, 2001 notices that, in practice σ_0 is unknown and must be estimated by a robust scale $\hat{\sigma}_n$ and that, unfortunately the quantity $\hat{\sigma}(F)B(\epsilon)$ is not an upper bound for $D(F, \psi)$ because $\hat{\sigma}(F)$ may underestimate σ_0. For instance, if $F = 0.90N(0,1) + 0.10\delta_{0.15}$, $T(F, \psi) = M(F)$, the median of F, and $\hat{\sigma}(F)$ is the MAD of F (median absolute deviations about the median), then

$$|M(F) - M(F_0)| = 0.1397 > MAD(F)B(0.10) = 0.8818 \times 0.1397 = 0.1232.$$

A quantity, $K(\epsilon)$ such that $\hat{\sigma}(F)K(\epsilon)$ is a bound for $D(F)$ will be called *bias bound*. The bias bound is a new theoretical concept which highlights the practical potential of maxbias curves.

Martin and Zamar (1993) also give formulas for the explosion and implosion maxbiases of the scale functional $\hat{\sigma}^+(\epsilon) = \sup_{F \in V_\epsilon(F_0)} \hat{\sigma}(F)/\sigma_0$ and $\hat{\sigma}^-(\epsilon) = \inf_{F \in V_\epsilon(F_0)} \hat{\sigma}(F)/\sigma_0$. Then ,

$$D(F, \psi) \leq \sigma_0 B(\epsilon) = \hat{\sigma}(F) B(\epsilon) \, (\sigma_0/\hat{\sigma}(F)) \leq \hat{\sigma}(F) B(\epsilon)/\hat{\sigma}^-(\epsilon)$$

and so

$$K(\epsilon) = B(\epsilon)/\hat{\sigma}^-(\epsilon)$$

is a bias bound for $T(F, \psi)$.

3. Globally Robust Confidence Intervals

What are desirable features for a "robust confidence interval" to have? In our opinion, a robust confidence interval should be *stable* and *informative*. The robust confidence interval should be *stable* in the sense of keeping a high coverage level (at or above the nominal) not only at the central model but also over the contamination neighborhood. The robust confidence interval should also be *informative* in the sense of keeping a reasonable average length over the entire neighborhood. These two features are more precisely stated in the next definition. Adrover, Salibian and Zamar (2000) give the following definition:

Definition 1: A confidence interval (L_n, U_n) for θ is called *globally robust* of level $(1 - \alpha)$ if it satisfies the following conditions:

1. (*Stable interval*) The minimum asymptotic coverage over the ϵ-contamination neighborhood is $(1 - \alpha)$:

$$\lim_{n \to \infty} \inf_{F \epsilon \mathcal{F}_\epsilon(F_\theta)} P_F \left(L_n < \theta < U_n \right) \geq (1 - \alpha).$$

2. (*Informative interval*) The maximum asymptotic length of the interval is bounded over the ϵ-contamination neighborhood:

$$\lim_{n \to \infty} \sup_{F \epsilon \mathcal{F}_\epsilon(F_\theta)} [U_n - L_n] < \infty.$$

Adrover et al. (2000) show that classical Student-t confidence intervals

$$\overline{X}_n \pm t_{(n-1)}(1 - \alpha/2) S_n/\sqrt{n},$$

fail Parts 1 and 2 of Definition 1. We might try to remedy this problem by replacing $\overline{X}_n$ by arobust asymptotically normal location estimate T_n and S_n by an appropriate estimate of the standard error of T_n . Such interval would satisfy Part 2 but fails to satisfy Part 1 of Definition

1. Adrover et al. (2000) show that globally robust $(1-\alpha)100\%$ confidence intervals can be defined using a bias bound $\bar{\theta}_n$ for T_n as follows:

$$T_n \pm \bar{q}_n.$$

Here $\bar{q}_n$ is implicitely defined by the equation

$$\Phi\left(\frac{\bar{q}_n - \bar{\theta}_n}{S_n}\right) + \Phi\left(\frac{\bar{q}_n + \bar{\theta}_n}{S_n}\right) - 1 = 1 - \alpha.$$

REFERENCES

Adrover, J.G., Salibian, M. and Zamar, R.H. (2000). Robust inference for the simple linear regression model. Unpublished manuscript.

Berrendero, J.R. and Zamar, R.H. (2001). Maximum bias curves for robust regression with non–elliptical regressors. To appaear in the *Ann. Statist.*.

Huber, P.J. (1981). *Robust Statistics. Wiley.*

Martin, R.D. and Zamar, R. (1993) Bias robust estimation of scale. *Ann. Statist.* **21**, 991–1017.

RÉSUMÉ

L'approche robuste pour l'analyse des données utilise des models qui ne déterminent pas complètement la distribution des données, mais qui prétendent que la distribution réelle est dans un voisinage d'un certain model paramétrique. En conséquence, l'inférence robuste devrait être valide pour n'importe quelle distribution dans ce voisinage. L'inférence robuste a deux sources majeures d'incertitude: (i) la variabilité échantillonnale et (ii) le biais causé par les valeurs aberrantes et autres contaminations. L'estimation de la variabilité échantillonnale par la théorie asymptotique usuelle nécessite généralement l'hypothèse de symétrie de la distribution du terme d'erreur ou l'hypothèse de la connaissance exacte d'un paramètre d'échelle. Pour la plupart des problèmes où des méthodes robustes sont requises, aucune de ces hypothèses n'est satisfaite. Une autre approche pour l'estimation de la variabilité échantillonnale est de faire un rééchantillonnage bootstrap de l'estimateur robuste. Cependant, cette méthode a deux faiblesses en matière d'analyse robuste: elle est calculatoirement coûteuse (et parfois impracticable) et les quantiles de son estimateur ne sont pas robustes. On présente une procédure bootstrap qui surmonte ces problèmes.

Les procédures d'inférence robustes ignorent d'ordinaire l'incertitude du biais. Il s'ensuit que la probabilité réelle de couverture d'intervalles de confiance s'avère plus petite que le seuil nominal. Similairement, les erreurs de type-I réelles des tests sont plus élevées que les nominales. Cet article montre comment les courbes de biais maximum peuvent être utilisées pour tenir compte de l'incertitude du biais et obtenir des tests valides et des intervals de confiance nominaux dans tout le voisinage. La procédure est mise en pratique par application à des problèmes de location et de régression.