



Taylor & Francis
Taylor & Francis Group



The Exact Amount of t-ness That the Normal Model Can Tolerate

Author(s): Nils Lid Hjort

Source: *Journal of the American Statistical Association*, Jun., 1994, Vol. 89, No. 426 (Jun., 1994), pp. 665-675

Published by: Taylor & Francis, Ltd. on behalf of the American Statistical Association

Stable URL: <https://www.jstor.org/stable/2290869>

JSTOR is a not-for-profit service that helps scholars, researchers, and students discover, use, and build upon a wide range of content in a trusted digital archive. We use information technology and tools to increase productivity and facilitate new forms of scholarship. For more information about JSTOR, please contact support@jstor.org.

Your use of the JSTOR archive indicates your acceptance of the Terms & Conditions of Use, available at <https://about.jstor.org/terms>



JSTOR

Taylor & Francis, Ltd. and American Statistical Association are collaborating with JSTOR to digitize, preserve and extend access to *Journal of the American Statistical Association*

The Exact Amount of t -ness That the Normal Model Can Tolerate

Nils Lid Hjort*

Suppose that the normal model is used for data $Y_1, \dots, Y_n$, but that the true distribution is a t distribution with location and scale parameters ξ and σ and m degrees of freedom. The normal model corresponds to $m = \infty$. Using a local asymptotic framework where m is allowed to increase with n , two classes of estimands are identified. One small class, which in particular contains the functions of ξ alone, is affected by t -ness only to the second order, and maximum likelihood estimation in the two- or three-parameter models becomes equivalent. For all other estimands, it is shown that if $m \geq 1.458\sqrt{n}$, then maximum likelihood estimation using the incorrect normal model is still more precise than using the correct three-parameter model. This is further shown to be true in regression models with t -distributed residuals. We also propose and analyze compromise estimators that in various ways interpolate between the normal and the nonnormal models. A separate section extends the t -ness results to general normal scale mixtures, in which case the tolerance radius around the normal error distribution takes the form of an upper bound $.3429/\sqrt{n}$ for the variance of the scale mixture distribution. Proving our results requires somewhat nonstandard "corner asymptotics," because behavior of estimators must be studied when the crucial parameter $\gamma = 1/m$ is close to 0, which is not an inner point of the parameter space, and the maximum likelihood estimator of m is equal to ∞ with positive probability.

KEY WORDS: Choice of model; Corner asymptotics; Deliberate bias; Guarding against heavier tails; Normal scale mixtures; Parametric robustness; Robust regression; Tolerance radius.

1. INTRODUCTION AND SUMMARY

The most popular model for iid data $Y_1, \dots, Y_n$ is to postulate normality; that is, assuming $f(y) = \phi((y - \xi)/\sigma)/\sigma$ for suitable parameters ξ and σ . In many situations the normal density is too light-tailed to constitute a serious description, however. A remedy then is to use

$$f(y, \xi, \sigma, m) = g_m\left(\frac{y - \xi}{\sigma}\right) \frac{1}{\sigma},$$

where $g_m(t)$ is the t density with m degrees of freedom. The narrower normal model corresponds to $m = \infty$, and it is naturally felt that for large m the discrepancy between normality and t -ness should not matter. One would also expect inference methods based on the formally incorrect normality assumption to work better than inference methods based on the wider three-parameter model, for large values of m , because the estimation variability increases with the inclusion of m .

This article studies the problem of choosing between "narrow model" and "wide model" estimators. Using the narrow method means introducing a certain bias due to incorrect modeling, whereas using the wide method means allowing additional parameter estimation noise. Thus choosing between the estimators amounts to a statistical balancing act with perhaps deliberate bias against sampling variability.

Suppose, for example, that the parameter to be estimated is the upper quartile $q = F^{-1}(.75)$. Concentrating on maximum likelihood (ML) estimators, the two methods amount to

$$\hat{q}_{\text{narr}} = \hat{\xi}_{\text{narr}} + .675\hat{\sigma}_{\text{narr}} \quad \text{and} \\ \hat{q}_{\text{wide}} = \hat{\xi}_{\text{wide}} + G^{-1}(.75, \hat{m}_{\text{wide}})\hat{\sigma}_{\text{wide}}.$$

Here $G(\cdot, m) = G_m(\cdot)$ is the t -distribution function with m degrees of freedom, with inverse $G^{-1}(\cdot, m) = G_m^{-1}(\cdot)$, whereas $\hat{\xi}_{\text{narr}}$ and $\hat{\sigma}_{\text{narr}}$ are the ML estimators under the narrow two-parameter model and $\hat{\xi}_{\text{wide}}$, $\hat{\sigma}_{\text{wide}}$, $\hat{m}_{\text{wide}}$ are the ML estimators under the wide three-parameter model. (The narrow model estimators are, of course, the ordinary empirical mean and empirical standard deviation statistics, whereas ML estimates under the wide model must be computed by numerical maximization techniques.) How large must m be for $\hat{q}_{\text{narr}}$ to be more precise than $\hat{q}_{\text{wide}}$? Suppose, for a second example, that the parameter to be estimated is sd, the standard deviation for Y_i 's. We should compare

$$\hat{\text{sd}}_{\text{narr}} = \left[\frac{1}{n} \sum_{i=1}^n (Y_i - \bar{Y}_n)^2 \right]^{1/2} \quad \text{and} \\ \hat{\text{sd}}_{\text{wide}} = \sqrt{\frac{\hat{m}_{\text{wide}}}{\hat{m}_{\text{wide}} - 2}} \hat{\sigma}_{\text{wide}}.$$

For what range of m values is the narrow estimator more precise than the much more laborious wide estimator, and for which m values will it be advantageous to use the latter? Finally, a third example is that of estimating the probability $p(y) = \Pr\{Y_i \leq y\}$, in which case the two estimators to be compared are

$$\hat{p}_{\text{narr}}(y) = \Phi((y - \hat{\xi}_{\text{narr}})/\hat{\sigma}_{\text{narr}}) \quad \text{and} \\ \hat{p}_{\text{wide}}(y) = G((y - \hat{\xi}_{\text{wide}})/\hat{\sigma}_{\text{wide}}, \hat{m}_{\text{wide}}).$$

These problems have a surprisingly sharp and simple solution. A natural large-sample framework is established in Section 2, where results are reached for the large-sample behavior of ML estimators for (ξ, σ) in the narrow model and of ML estimators for (ξ, σ, γ) in the wide model, where $\gamma = 1/m$. These are used in Section 3 to solve the problem. It is convenient to answer the question in terms of two classes of estimands. For a certain small class of estimands, identified

* Nils Lid Hjort is Professor of Statistics, Department of Mathematics and Statistics, University of Oslo, N-0316 Oslo, Norway. The author thanks the reviewers and in particular the associate editor for constructive suggestions that led to a strengthened paper. A part of the work reported here was carried out while the author was on sabbatical leave at the University of Oxford, with partial support from the Royal Norwegian Research Council.

precisely in Section 3, and which in particular contains all functions of the mean alone, t -ness turns out to be a secondary issue only, and "narrow" and "wide" estimation becomes large-sample equivalent. For the much larger class of *all other estimands*, it turns out that if only $m \geq 1.458\sqrt{n}$, then the narrow method is more precise than the wide method, in terms of large-sample limits of mean squared errors. If t -ness is present with $m < 1.458\sqrt{n}$, then the wide method is more precise. Thus $1.458\sqrt{n}$ is effectively the "tolerance distance" from the normal model with respect to t -ness.

A fuller story is told in Section 4, where a general class of compromise estimators is considered. These interpolate between the narrow normal and the wider t -model. We single out a few of these that are designed to work well both under normal and nonnormal conditions. The regression case, where residuals can have a t distribution, is treated in Section 5. Results from the previous sections can be extended, and the tolerance distance becomes again precisely $1.458\sqrt{n}$. Our compromise estimators in this case can be viewed as a basis for performing robust regression analysis, guarding against heavier-than-normal tails.

Although this article's primary focus is on the t extension of the normal to model error term distributions, quite general results valid when the error distribution is of the normal scale mixture type $\mathcal{N}\{0, 1\}/S$, where S has a distribution suitably centered around 1, are described in Section 6. The tolerated amount of disturbance around the normal model takes the very simple form $\text{var } S \leq .3429/\sqrt{n}$. The t case corresponds to $S = (\chi_m^2/m)^{1/2}$, for which the variance is approximately equal to $1/(2m)$. Thus the general tolerance bound agrees with the specific t -based one. Finally, some complementing remarks are offered in Section 7, including consequences for confidence intervals and discussion of situations with additional departures from the basic normal regression model.

The problems about balancing modeling bias and estimation variability for incorrectly specified parametric models are obviously of a general nature and can be studied for other important models as well as on a general basis. Such a study was indeed reported on by Hjort (1993), who provided further background, a general theory, and explicit results for a generous list of commonly used statistical models. But the present t -ness departure case is nonregular and cumbersome, and cannot be handled as a special case of the general regular theory. What makes this problem special is that the model must be studied when the crucial parameter $\gamma = 1/m$ is close to 0, which is not an inner point in the parameter space. In particular, the ML estimator of γ is 0 (i.e., the ML estimator of m is infinity) with positive probability, the familiar ML asymptotics break down, and special methods are called for.

Classical robustness theory has been concerned in part with construction of inference methods that work well in certain *nonparametric* neighborhoods of a basic model; for example, the normal regression model. This article is a contribution to the theory of *parametric* robustness, where one studies departures from a basic model in specific directions. There are perhaps two major themes in this general area: (1)

construction of new inference methods that work well both under the idealized center model and in the wider parametric model, and (2) studies of how well or not well standard procedures tolerate various amounts of disturbance in these directions. Both themes are worked on here.

Certain robustness aspects are easier to study more fully in the parametric context. The modeling bias versus estimation variance balance structure of model choices is more directly apparent, and one may more easily study general estimands. Our article provides an explicit illustration of the fact that deliberate bias can be useful, as was more generally described by Bickel (1984) and by Hjort (1993). Lange, Little, and Taylor (1989) studied multivariate t distributions as an alternative to the normal ones, partly with the degrees of freedom equal to a fixed number (they reported that " $m = 4$ works well") and partly treated as an unknown parameter. The latter case corresponds to what we call "wide estimation" here. Finally, we point out that the idea of compromise estimators is not new. Versions of such were considered by Bickel (1983, 1984), and the so-called configural polysampling approach to robustness uses ideas similar in spirit to those of Section 4 here to compromise between the normal and the "slash" model (see Morgenthaler and Tukey 1991 for several papers on the subject). We are able to consider completely general estimands with relative ease, however, whereas the Morgenthaler and Tukey methods do not seem to be developed beyond the basic location and scale parameters.

2. LARGE-SAMPLE FRAMEWORK FOR THE PROBLEM

The wide model has parameters ξ , σ , and m . Let us reparameterize to $\gamma = 1/m$, so that the density becomes

$$f(y, \xi, \sigma, \gamma) = \frac{c(\gamma)}{\sigma} \left[1 + \gamma \left(\frac{y - \xi}{\sigma} \right)^2 \right]^{-[1/2 + 1/(2\gamma)]},$$

$$c(\gamma) = \frac{\sqrt{\gamma}}{\sqrt{\pi}} \frac{\Gamma(2^{-1} + (2\gamma)^{-1})}{\Gamma((2\gamma)^{-1})}. \quad (2.1)$$

We are interested in this model for γ in the vicinity of 0. Using precise Taylor expansions and approximations to the $\log \Gamma(\cdot)$ function, one finds that

$$\log f(y, \xi, \sigma, \gamma) = \log f(y, \xi, \sigma, 0) + \gamma \left(\frac{1}{4} z^4 - \frac{1}{2} z^2 - \frac{1}{4} \right) - \frac{1}{2} \gamma^2 \left(\frac{1}{3} z^6 - \frac{1}{2} z^4 \right) + O(\gamma^3), \quad (2.2)$$

in which $z = (y - \xi)/\sigma$. Having $\gamma = 0$ corresponds to $m = \infty$ and gives back the ordinary normal model.

Let $\mu = \mu(f) = \mu(\xi, \sigma, \gamma)$ be some parameter estimand of interest. We assume that μ is smooth with continuous derivatives throughout the inner parameter space $(\xi, \sigma, \gamma) \in (-\infty, \infty) \times (0, \infty) \times (0, \infty)$ and that the right derivative exists at $\gamma = 0$, $\lim_{\gamma \rightarrow 0+} [\mu(\xi, \sigma, \gamma) - \mu(\xi, \sigma, 0)]/\gamma$. We concentrate on ML procedures and wish to study the performance of the two estimators

$$\hat{\mu}_{\text{narr}} = \mu(\hat{\xi}_{\text{narr}}, \hat{\sigma}_{\text{narr}}, 0) \quad \text{and} \quad \hat{\mu}_{\text{wide}} = \mu(\hat{\xi}, \hat{\sigma}, \hat{\gamma}), \quad (2.3)$$

where for simplicity of notation the subscript “wide” is dropped for the ML estimators in the three-parameter model.

These could be compared in an asymptotic framework in which Y_i 's come from some fixed $f(y, \xi_0, \sigma_0, \gamma)$ and $\gamma > 0$. In this case $\sqrt{n}(\hat{\mu} - \mu)$ has a limit distribution. The situation is different for the narrow model procedure. Here $\sqrt{n}(\hat{\mu}_{\text{narr}} - \mu)$ can be represented as a sum of two terms. The first is $\sqrt{n}[\mu(\hat{\xi}_{\text{narr}}, \hat{\sigma}_{\text{narr}}, 0) - \mu(\xi_0, \sigma_0, 0)]$, which has a limit distribution with generally smaller variability than that of the wide model procedure; and the second is $-\sqrt{n}[\mu(\xi_0, \sigma_0, \gamma) - \mu(\xi_0, \sigma_0, 0)]$, which tends to plus or minus infinity, reflecting a bias that for very large n will dominate completely. This merely goes to show that with very large sample sizes, one is penalized for any bias and should use the wide model. This result is not very informative, however, and suggests that a large-sample framework that uses a local neighborhood of $\gamma = 0$ that shrinks when the sample size grows is much more adequate. Consider, therefore, model P_n , the n th model, under which

$$Y_1, \dots, Y_n \text{ are iid from } f_n(y) = f(y, \xi_0, \sigma_0, \delta/\sqrt{n}). \quad (2.4)$$

The primary justification for working with neighborhoods of exactly this size is simply that it will give informative quantitative results. Here (ξ_0, σ_0) is a fixed but arbitrary parameter point. The true parameter to be estimated is $\mu_{\text{true}} = \mu(\xi_0, \sigma_0, \delta/\sqrt{n})$.

To assess the behavior of the estimators of μ , we need to know what happens to narrow and wide estimators of the two- and three-model parameters. Consider the score functions for the wide model, evaluated at the null point $(\xi_0, \sigma_0, 0)$. Letting γ carefully tend to 0 in expressions for the three partial log-derivatives of f leads to

$$\begin{aligned} \begin{pmatrix} U(y) \\ V(y) \\ W(y) \end{pmatrix} &= \begin{pmatrix} \partial \log f(y, \xi_0, \sigma_0, 0)/\partial \xi \\ \partial \log f(y, \xi_0, \sigma_0, 0)/\partial \sigma \\ \partial \log f(y, \xi_0, \sigma_0, 0)/\partial \gamma \end{pmatrix} \\ &= \begin{pmatrix} z/\sigma_0 \\ (z^2 - 1)/\sigma_0 \\ \frac{1}{4}z^4 - \frac{1}{2}z^2 - \frac{1}{4} \end{pmatrix}, \end{aligned} \quad (2.5)$$

where $z = (y - \xi_0)/\sigma_0$; refer to (2.2). We shall also need the accompanying 3×3 size information matrix, the covariance matrix of these three, when Y has the $f(y, \xi_0, \sigma_0, 0)$ distribution (i.e., it is simply $\mathcal{N}\{\xi_0, \sigma_0^2\}$). One finds that

$$\begin{aligned} J_{\text{wide}} &= \text{var}_0 \begin{pmatrix} Z/\sigma_0 \\ (Z^2 - 1)/\sigma_0 \\ \frac{1}{4}Z^4 - \frac{1}{2}Z^2 - \frac{1}{4} \end{pmatrix} \\ &= \begin{pmatrix} 1/\sigma_0^2 & 0 & 0 \\ 0 & 2/\sigma_0^2 & 2/\sigma_0 \\ 0 & 2/\sigma_0 & 7/2 \end{pmatrix}. \end{aligned}$$

Note that the upper left 2×2 block $J_{\text{narr}} = \text{diag}(1/\sigma_0^2, 2/\sigma_0^2)$ is the information matrix of the narrow model, evaluated at (ξ_0, σ_0) . For future reference, we note that

$$\begin{aligned} J_{\text{wide}}^{-1} &= \begin{pmatrix} \sigma_0^2 & 0 & 0 \\ 0 & \frac{7}{6}\sigma_0^2 & -\frac{2}{3}\sigma_0 \\ 0 & -\frac{2}{3}\sigma_0 & \frac{2}{3} \end{pmatrix}, \\ J_{\text{narr}}^{-1} &= \begin{pmatrix} \sigma_0^2 & 0 \\ 0 & \sigma_0^2/2 \end{pmatrix}. \end{aligned} \quad (2.6)$$

Lemma. Let $\bar{U}_n$ denote the average of $U(Y_i)$'s, and similarly for $\bar{V}_n$ and $\bar{W}_n$. Under the sequence of models P_n of (2.4),

$$\begin{pmatrix} \sqrt{n}\bar{U}_n \\ \sqrt{n}\bar{V}_n \\ \sqrt{n}\bar{W}_n \end{pmatrix} \rightarrow_d \begin{pmatrix} 0 + K \\ (2/\sigma_0)\delta + L \\ (7/2)\delta + M \end{pmatrix}$$

as $n \rightarrow \infty$, where $(K, L, M)' \sim \mathcal{N}_3\{0, J_{\text{wide}}\}$.

Proof. This follows essentially from the triangular version of the Lindeberg theorem. A key observation is that

$$f_n(y) \doteq f(y, \xi_0, \sigma_0, 0)[1 + W(y)\delta/\sqrt{n}].$$

This implies that $(U(Y_i), V(Y_i), W(Y_i))'$ has expected value $(0, (2/\sigma_0)\delta/\sqrt{n}, \frac{7}{2}\delta/\sqrt{n})$ plus $O(n^{-1})$ terms, and that its variance matrix is $J_{\text{wide}} + O(\delta/\sqrt{n})$. (See also Hjort 1993, sec. 2.)

Proposition 1. Under model P_n of (2.4), the behavior of the narrow-model-based estimators is given by

$$\begin{pmatrix} \sqrt{n}(\hat{\xi}_{\text{narr}} - \xi_0) \\ \sqrt{n}(\hat{\sigma}_{\text{narr}} - \sigma_0) \end{pmatrix} \doteq_d J_{\text{narr}}^{-1} \begin{pmatrix} \sqrt{n}\bar{U}_n \\ \sqrt{n}\bar{V}_n \end{pmatrix} \rightarrow_d \begin{pmatrix} 0 + \sigma_0^2 K \\ \sigma_0 \delta + \frac{1}{2}\sigma_0^2 L \end{pmatrix},$$

in which $A_n \doteq_d B_n$ means that $A_n - B_n$ tends to 0 in probability.

Proof. This is essentially the familiar Taylor expansion argument, carried out in the present local neighborhood framework. Note the bias term $(0, \sigma_0 \delta)'$. The details are contained in more general arguments given by Hjort (1993, sec. 2). More direct methods of proof could also have been used, because $\hat{\xi}_{\text{narr}}$ and $\hat{\sigma}_{\text{narr}}$ are relatively easy to work with.

The wide method case is much more complicated because of the corner problem. Introduce

$$\begin{pmatrix} A \\ B \\ C \end{pmatrix} = J_{\text{wide}}^{-1} \begin{pmatrix} K \\ L \\ M \end{pmatrix} \sim \mathcal{N}_3\{0, J_{\text{wide}}^{-1}\}. \quad (2.7)$$

Note that $K = (1/\sigma_0^2)A$, $L = (2/\sigma_0^2)B + (2/\sigma_0)C$. The limit in Proposition 1, written in terms of $(A, B, C)'$, becomes

$$\begin{aligned} \begin{pmatrix} \sqrt{n}(\hat{\xi}_{\text{narr}} - \xi_0) \\ \sqrt{n}(\hat{\sigma}_{\text{narr}} - \sigma_0) \end{pmatrix} &\rightarrow_d \begin{pmatrix} 0 + A \\ B + \sigma_0(C + \delta) \end{pmatrix} \\ &\sim \mathcal{N}_2\left\{\begin{pmatrix} 0 \\ \sigma_0 \delta \end{pmatrix}, \begin{pmatrix} \sigma_0^2 & 0 \\ 0 & \frac{1}{2}\sigma_0^2 \end{pmatrix}\right\}. \end{aligned} \quad (2.8)$$

Because I want my readers to join me for the main story, I defer the proof of the following proposition to the Appendix.

Proposition 2. Under model P_n of (2.4), the behavior of the wide-model-based estimators is given by

$$\begin{pmatrix} \sqrt{n}(\hat{\xi} - \xi_0) \\ \sqrt{n}(\hat{\sigma} - \sigma_0) \\ \sqrt{n}(\hat{\gamma} - \delta/\sqrt{n}) \end{pmatrix} \rightarrow_d \begin{pmatrix} A \\ B \\ C \end{pmatrix} \quad \text{if } C \geq -\delta, \\ \rightarrow_d \begin{pmatrix} A \\ B + \sigma_0(C + \delta) \\ -\delta \end{pmatrix} \quad \text{if } C \leq -\delta.$$

3. CALCULATING THE TOLERANCE DISTANCE

Our program is to use the delta method of linearization in conjunction with Propositions 1 and 2 to reach limit distribution results for the narrow and wide estimators, and then to compute and compare mean squared errors.

First, consider the narrow method. Using Proposition 1, we find

$$\begin{aligned} & \sqrt{n}[\mu(\hat{\xi}_{\text{narr}}, \hat{\sigma}_{\text{narr}}, 0) - \mu(\xi_0, \sigma_0, \delta/\sqrt{n})] \\ &= \sqrt{n}[\mu(\hat{\xi}_{\text{narr}}, \hat{\sigma}_{\text{narr}}, 0) - \mu(\xi_0, \sigma_0, 0)] \\ &\quad - \sqrt{n}[\mu(\xi_0, \sigma_0, \delta/\sqrt{n}) - \mu(\xi_0, \sigma_0, 0)] \\ &\doteq_d \frac{\partial \mu}{\partial \xi} \sqrt{n}(\hat{\xi}_{\text{narr}} - \xi_0) + \frac{\partial \mu}{\partial \sigma} \sqrt{n}(\hat{\sigma}_{\text{narr}} - \sigma_0) \\ &\quad - \sqrt{n} \frac{\partial \mu}{\partial \gamma} \delta/\sqrt{n} \\ &\rightarrow_d \Lambda_{\text{narr}} = \frac{\partial \mu}{\partial \xi} A + \frac{\partial \mu}{\partial \sigma} [B + \sigma_0(C + \delta)] - \frac{\partial \mu}{\partial \gamma} \delta, \end{aligned}$$

where the partial derivatives are computed at the null model $(\xi_0, \sigma_0, 0)$. The limit variable is normal with

$$E\Lambda_{\text{narr}} = b\delta = \left(\sigma_0 \frac{\partial \mu}{\partial \sigma} - \frac{\partial \mu}{\partial \gamma} \right) \delta$$

and

$$\text{var } \Lambda_{\text{narr}} = \tau_0^2 = \left[\left(\frac{\partial \mu}{\partial \xi} \right)^2 + \frac{1}{2} \left(\frac{\partial \mu}{\partial \sigma} \right)^2 \right] \sigma_0^2. \quad (3.1)$$

In particular, the narrow method has risk $E\Lambda_{\text{narr}}^2 = b^2\delta^2 + \tau_0^2$; see Section 7.2 for some consequences of this. This “narrow result” is really contained in general results of Hjort (1993).

Next, consider the wide method. Using Proposition 2, we find that

$$\begin{aligned} & \sqrt{n}[\mu(\hat{\xi}, \hat{\sigma}, \hat{\gamma}) - \mu(\xi_0, \sigma_0, \sigma/\sqrt{n})] \\ &\doteq_d \frac{\partial \mu}{\partial \xi} \sqrt{n}(\hat{\xi} - \xi_0) + \frac{\partial \mu}{\partial \sigma} \sqrt{n}(\hat{\sigma} - \sigma_0) \\ &\quad + \left[\left(\frac{\partial \mu}{\partial \gamma} \right) + O(1/\sqrt{n}) \right] \sqrt{n}(\hat{\gamma} - \delta/\sqrt{n}) \\ &\rightarrow_d \Lambda_{\text{wide}} = \frac{\partial \mu}{\partial \xi} A + \frac{\partial \mu}{\partial \sigma} B + \frac{\partial \mu}{\partial \gamma} C \quad \text{if } C \geq -\delta, \\ &= \frac{\partial \mu}{\partial \xi} A + \frac{\partial \mu}{\partial \sigma} [B + \sigma_0(C + \delta)] - \frac{\partial \mu}{\partial \gamma} \delta \\ &\quad \text{if } C \leq -\delta. \end{aligned}$$

This is not a normal distribution. We calculate its mean squared error by conditioning on the value of C . Using (2.6) and (2.7) and ordinary techniques, we find that

$$\begin{pmatrix} A \\ B \end{pmatrix} | \{C = c\} \sim \mathcal{N}_2 \left\{ \begin{pmatrix} 0 \\ -\sigma_0 c \end{pmatrix}, \begin{pmatrix} \sigma_0^2 & 0 \\ 0 & \frac{1}{2} \sigma_0^2 \end{pmatrix} \right\}. \quad (3.2)$$

Further calculations show that

$$\begin{aligned} \Lambda_{\text{wide}} | \{C = c\} &\sim \mathcal{N}\{-bc, \tau_0^2\} \quad \text{if } c \geq -\delta, \\ &\sim \mathcal{N}\{b\delta, \tau_0^2\} \quad \text{if } c \leq -\delta. \end{aligned}$$

Accordingly, $E[\Lambda_{\text{wide}}^2 | C = c]$ is $b^2c^2 + \tau_0^2$ if $c \geq -\delta$ and $b^2\delta^2 + \tau_0^2$ if $c \leq -\delta$, and

$$E\Lambda_{\text{wide}}^2 = b^2E[C^2I\{C \geq -\delta\} + \delta^2I\{C \leq -\delta\}] + \tau_0^2.$$

We are now in a position to find out when the narrow and risky estimator is better than the wide and safe one. From (2.7), we can write $C = \kappa N$, where $\kappa^2 = \frac{2}{3}$ and N is normal $(0, 1)$. From (3.1), assuming that $b \neq 0$, it is clear that the narrow method is better than the wide one if and only if

$$\delta^2 \leq E[\kappa^2 N^2 I\{N \geq -\delta/\kappa\} + \delta^2 I\{N \leq -\delta/\kappa\}],$$

or

$$\begin{aligned} a^2 &\leq E[N^2 I\{N \geq -a\} + a^2 I\{N \leq -a\}] \\ &= \Phi(a) - a\phi(a) + a^2(1 - \Phi(a)), \end{aligned} \quad (3.3)$$

using $a = \delta/\kappa$. But this is equivalent to $0 \leq a \leq .8399$, as is borne out by numerical computations. This means that $0 \leq \delta \leq .8399\sqrt{2/3} = .6858$, and we have reached the following conclusion.

Result. (a) The case where $b = \sigma_0(\partial\mu/\partial\sigma) - (\partial\mu/\partial\gamma) = 0$ is rather trivial; this typically corresponds to a parameter estimand μ functionally independent of σ and γ at $\gamma = 0$. In this case $\hat{\mu}_{\text{wide}}$ and $\hat{\mu}_{\text{narr}}$ are asymptotically equivalent, regardless of δ . (b) In the more interesting case $b \neq 0$, the narrow-model-based estimator is better than or as good as the wider-model-based estimator (with respect to limiting mean squared error) if and only if $\delta \leq .6858$, or $\gamma \leq .6858/\sqrt{n}$, or degrees of freedom $m \geq \sqrt{n}/.6858 = 1.4582\sqrt{n}$.

Lange et al. (1989, sec. 3 and app. B) provided information matrix calculations for the t model (also in the multivariate case), and case (a) of our Result can be deduced from these. We note that the mean parameter $\mu = \xi$ (and more generally $\mu = x_i'\beta$ in the regression case to be considered in Sec. 5) has $b = 0$, so these important estimands are not at all affected by moderate t -ness.

4. A FULLER STORY: COMPROMISE ESTIMATORS

The two estimators (2.3) that have been considered so far have both somewhat extreme attitudes. The first is a firm believer in normality; the second is a firm disbeliever in normality. This section looks at some compromising methods designed to work well under both “close to normal” and

“distinctively nonnormal” conditions. (See also Hjort 1993, sec. 5.)

We have shown in Proposition 2 that $\sqrt{n}(\hat{\gamma} - \delta/\sqrt{n})$ tends to $\max\{C, -\delta\}$ in distribution, where $C \sim \mathcal{N}\{0, \kappa^2\}$ and $\kappa^2 = \frac{2}{3}$. Now shift attention to $T_n = \sqrt{n}\hat{\gamma}/\kappa$, the natural statistic for testing $\gamma = 0$ (normality) against $\gamma > 0$ (t -ness).

Theorem. Consider the general estimator

$$\mu^* = [1 - w(T_n)]\hat{\mu}_{\text{narr}} + w(T_n)\hat{\mu}_{\text{wide}}, \quad (4.1)$$

where $w(T_n)$ is some appropriate weight function, assumed only to be continuous at 0 (where the limit $T \vee 0$ of T_n has positive probability) and almost everywhere on $(0, \infty)$. Then $\sqrt{n}(\mu^* - \mu_{\text{true}})$ tends in distribution to a Λ with mean squared error

$$E\Lambda^2 = \frac{2}{3}b^2R(a) + \tau_0^2, \quad \text{where}$$

$$R(a) = E_a[(w(T)T - a)^2I\{T > 0\} + a^2I\{T \leq 0\}], \quad (4.2)$$

and where $a = \delta/\kappa$ is the normalized measure of t -ness.

Proof. We have $T_n \rightarrow_d a + \kappa^{-1}\max\{C, -\delta\} = T \vee 0$, where $T = a + C/\kappa \sim \mathcal{N}\{a, 1\}$, and

$$\Lambda_{\text{narr}} = b\delta + \frac{\partial\mu}{\partial\xi}A + \frac{\partial\mu}{\partial\sigma}[B + \sigma_0\kappa(T - a)]$$

and

$$\begin{aligned} \Lambda_{\text{wide}} &= \frac{\partial\mu}{\partial\xi}A + \frac{\partial\mu}{\partial\sigma}B + \frac{\partial\mu}{\partial\gamma}\kappa(T - a) \quad \text{if } T > 0, \\ &= \frac{\partial\mu}{\partial\xi}A + \frac{\partial\mu}{\partial\sigma}(B + \sigma_0\kappa T) - \frac{\partial\mu}{\partial\gamma}a\kappa \quad \text{if } T \leq 0. \end{aligned} \quad (4.3)$$

These will next be seen to be special cases of a much more general phenomenon. By the continuous mapping theorem,

$$\begin{aligned} \sqrt{n}(\mu^* - \mu_{\text{true}}) &\rightarrow \Lambda \\ &= [1 - w(T \vee 0)]\Lambda_{\text{narr}} + w(T \vee 0)\Lambda_{\text{wide}}. \end{aligned}$$

Some calculations lead to

$$\begin{aligned} \Lambda|\{T = t\} &\sim \mathcal{N}\{b\kappa(a - w(t)t), \tau_0^2\} \quad \text{if } t > 0, \\ &\sim \mathcal{N}\{b\kappa a, \tau_0^2\} \quad \text{if } t \leq 0. \end{aligned}$$

The claim of the theorem follows from this.

Observe that $R(a)$ is the risk function, under squared error loss, for the estimator $\hat{a}(T \vee 0) = w(T \vee 0)(T \vee 0)$ for a nonnegative parameter a , based on observing the single variable $T \vee 0$, where $T \sim \mathcal{N}\{a, 1\}$. There is accordingly a one-to-one correspondence between estimators μ^* of type (4.1) for a general $\mu(\xi, \sigma, \gamma)$ and estimators $a^*(t) = w(t)t$ for a in the structurally very simple one-observation problem. The behavior of any given μ^* can be studied quite simply in terms of its associated $R(a)$ function, and any reasonable a -estimator method can be transported to a reasonable μ estimator, via $w(t) = a^*(t)/t$.

What are interesting values of a ? We have $a = \delta/\kappa$ and $m = 1/\gamma = \sqrt{n}/\delta = \sqrt{1.5n}/a$, and T_n used in (4.1) detects

nonnormality ($m < \infty$) with probability $\Phi(a - 1.645)$ (using level 5%). This means that T_n detects a values beyond 4 with probability at least .99. We may think of a values beyond 4, or $m \leq .306\sqrt{n}$, as being t departures from normality that should be clearly visible from data. This tentatively suggests that estimators of the type (4.1) should be used with $w(t)$ close to 1 for $t \geq 4$ and with small risk behavior for $R(a)$ when $a \leq 4$.

A briefly annotated list of interesting choices for $w(T_n)$ in (4.1) follows next.

- The narrow method uses $w(t) = 0$ and corresponds to using $\hat{a}_{\text{narr}}(t) = 0$ to estimate a . Its risk is $R_{\text{narr}}(a) = a^2$, which is good for a small (m large) but disastrous for a large (m small).
- The wide method has $w(t) = 1$ and corresponds to $\hat{a}_{\text{wide}}(t) = t \vee 0$ to estimate a . Its risk is

$$\begin{aligned} R_{\text{wide}}(a) &= E_a[(T - a)^2I\{T \geq 0\} + a^2I\{T \leq 0\}] \\ &= \Phi(a) - a\phi(a) + a^2(1 - \Phi(a)); \end{aligned}$$

compare (3.3). It starts at .50 at 0 and climbs towards 1. This estimator is minimax. Its risk is above .99 for $a \geq 2.67$. Again, if $a \leq .8399$, then the narrow is best, and if $a > .8399$, then the wide is best.

- Try out $w(t) = w$, a constant. We may compute the resulting $R(a)$, and minimize with respect to the choice of w . The best choice, expressed in terms of the parameter point a , is

$$w_0(a) = \frac{a^2\Phi(a) + a\phi(a)}{(a^2 + 1)\Phi(a) + a\phi(a)}.$$

A simple idea is then to insert T_n for a ; that is, using $\hat{a}_{\text{ratio}} = w_0(t)t$ to estimate a and using (4.1) with $w_0(T_n)$ to estimate μ . The risk $R_{\text{ratio}}(a)$ starts at .249 and is better than $R_{\text{wide}}(a)$ for $a \leq 1.32$ and is never much worse. Its maximum is 1.223, at $a = 2.90$, after which it decreases toward 1. The narrow is better than the present one only for $a \leq .68$, and quickly becomes much worse after that.

- Some natural Bayesian/empirical Bayesian ideas are as follows. Assume that a is distributed like $|\mathcal{N}\{0, \tau^2\}|$; that is, with prior distribution $\pi(a) = (2/\tau)\phi(a/\tau)$ on $[0, \infty)$. The Bayes solution associated with the loss function implicit in (4.2) can be seen to be the familiar $E\{a|T = t\}$ if $t > 0$ and an arbitrary value can be used when $T \vee 0 = 0$; that is, when information on T is $T \leq 0$. In the present case the Bayes solution becomes

$$\begin{aligned} \hat{a}_\tau(t) &= E\{a|t\} = \nu t + \sqrt{\nu}\phi(\sqrt{\nu}t)/\Phi(\sqrt{\nu}t), \\ &\quad \text{where } \nu = \tau^2/(\tau^2 + 1). \end{aligned}$$

Because $E_aT = a$ and $Ea^2 = \tau^2$, a simple empirical estimate for ν is $T^2/(T^2 + 1)$. This leads to

$$\begin{aligned} \hat{a}_{\text{eb}}(t) &= \frac{t^2}{t^2 + 1}t \\ &\quad + \frac{t}{\sqrt{t^2 + 1}}\phi\left(\frac{t}{\sqrt{t^2 + 1}}\right)/\Phi\left(\frac{t}{\sqrt{t^2 + 1}}\right). \end{aligned}$$

As for performance; the risk $R_{eb}(a)$ starts at .337 and is better than $R_{wide}(a)$ for $a \leq 2.09$ and is never much worse. It is not quite as good as the narrow method when $a \leq .67$, but quickly becomes much better after that. It reaches its maximum value of only 1.147 at $a = 3.75$, and decreases toward 1 thereafter.

- (e) The limit of the aforementioned Bayes rules, when $\tau \rightarrow \infty$, is $\hat{a}_{vag}(t) = t + \phi(t)/\Phi(t)$. This is also the Bayes solution under a vague flat prior on the halflife. It is minimax like the $\hat{a}_{wide}$ but has a differently shaped risk function; see Figure 1.
- (f) Finally, we could mention pretest and related estimators. The if-else or pretest estimator uses $w(t) = 0$ if $t \leq d$ and $w(t) = 1$ if $t > d$ in (4.1) and corresponds to $\hat{a}_{pre}(t) = 0$ if $t \leq d$ and $\hat{a}_{pre}(t) = t$ if $t > d$, for a suitable cutoff value d . The theory of Section 3 suggests that $d = .8399$ is a good choice, for example. It has risk

$$R_{pre}(a) = \Phi(a - d) + a^2[1 - \Phi(a - d)] - (a - d)\phi(d - a).$$

These are inadmissible estimators in the decision theoretic sense; each can be uniformly improved on (except in the borderline case $d = 0$, which is the wide method, and case $d = \infty$, which is the narrow method). A related but smoother version is the limited translation variety $\hat{a}_{lim}(t) = 0$ if $t \leq d$ and $\hat{a}_{lim}(t) = t - d$ if $t > d$. This corresponds to using $w(t) = 0$ if $t \leq d$ and $w(t) = 1 - d/t$ if $t > d$. The risk function becomes

$$R_{lim}(a) = (1 + d^2)\Phi(a - d) + a^2[1 - \Phi(a - d)] - (a + d)\phi(a - d),$$

with maximum value $1 + d^2$ occurring at infinity. The latter estimators are presumably reasonable approximations to minimax solutions subject to doing well at 0, following the work of Bickel (1983) in a more regular situation.

It is worth mentioning that in the general but regular case treated by Hjort (1993), where $a = \delta/\kappa$ can vary freely on

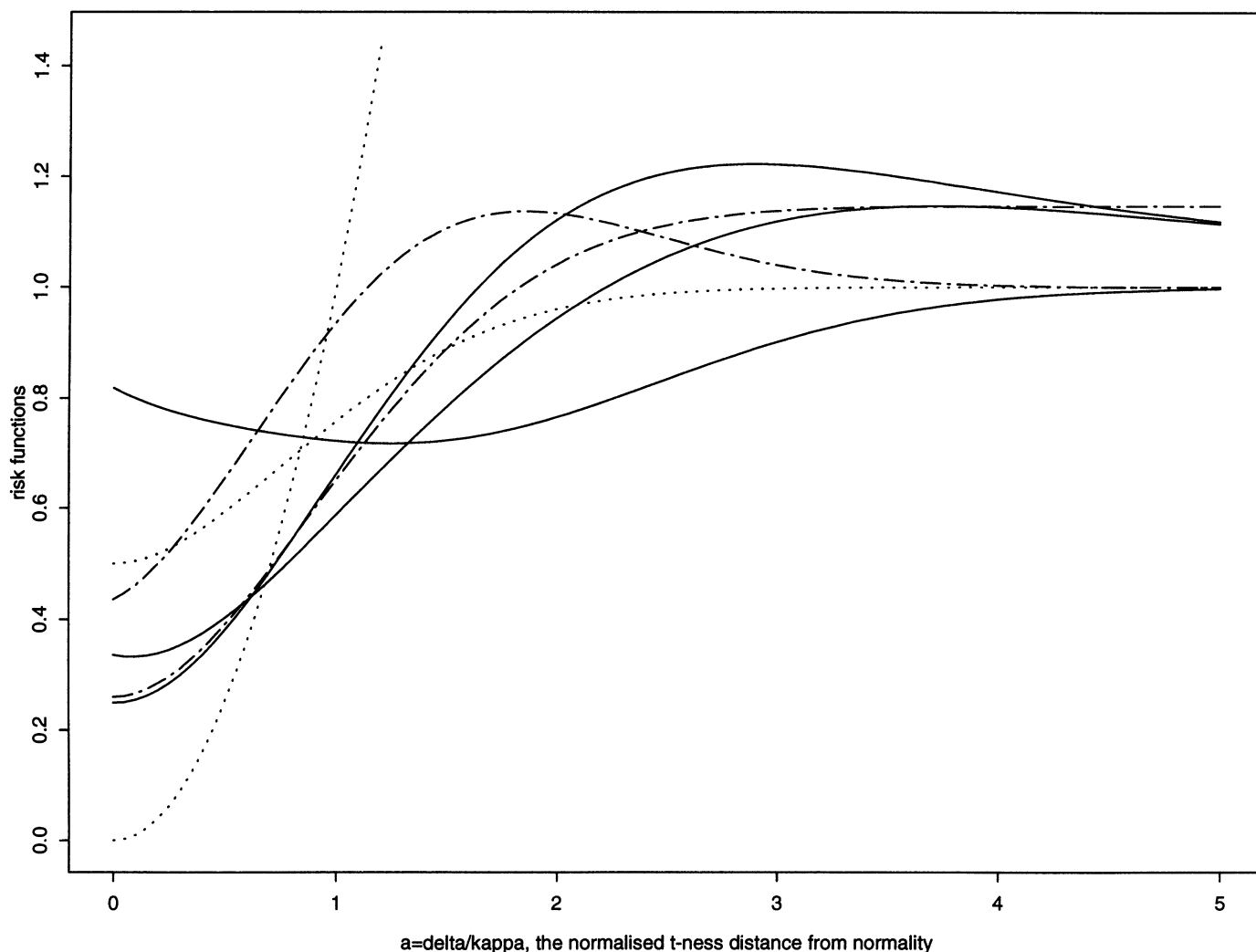


Figure 1. Risk Functions $R(a)$ for Seven Procedures, Corresponding to Seven Choices of Weight Function $w(\cdot)$ in (4.1). Risks for the wide and the narrow methods start at .500 and .000 and are shown by dotted lines. The three risks shown by solid lines are the ratio method (starting at .249), the empirical Bayes method (starting at .337), and the Bayes solution under uniform prior on the halflife (starting at .819). Finally the pretest method with $d = .8399$ (starting at .436) and the limited translation procedure with d chosen to make maximum risk $1 + d^2$ equal to that of the empirical Bayes method are shown by dashed lines.

the line, then aforementioned methods (e) and (f) become equivalent, as do methods (c) and (d).

Several risk functions $R(a)$ are plotted in Figure 1. These were computed by numerical integration where necessary. Overall, both the ratio method (c) and the empirical Bayes method (d) seem to be quite satisfactory solutions: they are almost as good as the wide method under distinctively non-normal conditions and are considerably better than the wide method under close-to-normal conditions. The limited translation procedure of (f) is also quite good and is relevant if one's aim is to minimize the maximum risk subject to bounding the risk at the narrow model.

In the end, one should also compute b and τ_0 of (3.1), or (5.2) in the regression case to be discussed, to see the specific consequences in terms of risk function (4.2) for the estimated under study; see the examples in Section 7.1.

5. EXTENSION TO REGRESSION MODELS

Suppose that

$$Y_i = x_{i,1}\beta_1 + \cdots + x_{i,p}\beta_p + \sigma Z_i = x'_i\beta + \sigma Z_i, \quad (5.1)$$

where $\beta_1, \dots, \beta_p$ are regression parameters and the Z_i 's are iid from a t_m distribution. How risky are the ordinary methods, which all have $m = \infty$ as underlying assumption? How statistically noisy are the more ambitious $p + 2$ -parameter methods that use ML estimates for $\beta_1, \dots, \beta_p, \sigma, m$? For example, should we use

$$\hat{\mu}_{\text{narr}} = \sqrt{2/\pi} \hat{\sigma}_{\text{narr}} \quad \text{or} \\ \hat{\mu}_{\text{wide}} = \sqrt{1/\pi} \left[\Gamma\left(\frac{1}{2}\right) \hat{m}_{\text{wide}} - \frac{1}{2} \right] / \Gamma\left(\frac{1}{2}\right) \hat{m}_{\text{wide}} \hat{\sigma}_{\text{wide}}$$

to estimate $\mu = E|Y(x) - x'\beta|$, the expected distance from regression curve to data point? What is an effective yet safe way of estimating regression deciles?

Let us briefly indicate how results from earlier sections extend to this situation. Let β_0 and σ_0 be arbitrary but fixed, and let $\gamma = 1/m$ tend to 0 like $\gamma = \gamma_n = \delta/\sqrt{n}$. The score function becomes

$$\begin{pmatrix} U(y_i) \\ V(y_i) \\ W(y_i) \end{pmatrix} = \begin{pmatrix} \partial \log f(y_i, \beta_0, \sigma_0, 0) / \partial \beta \\ \partial \log f(y_i, \beta_0, \sigma_0, 0) / \partial \sigma \\ \partial \log f(y_i, \beta_0, \sigma_0, 0) / \partial \gamma \end{pmatrix} \\ = \begin{pmatrix} z_i x_i / \sigma_0 \\ (z_i^2 - 1) / \sigma_0 \\ \frac{1}{4} z_i^4 - \frac{1}{2} z_i^2 - \frac{1}{4} \end{pmatrix},$$

in which $z_i = (y_i - x'_i \beta_0) / \sigma_0$. The $(p + 2) \times (p + 2)$ information matrix becomes

$$J_{\text{wide}} = \lim_{n \rightarrow \infty} \frac{1}{n} \sum_{i=1}^n \text{var}_0 \begin{pmatrix} Z_i x_i / \sigma_0 \\ (Z_i^2 - 1) / \sigma_0 \\ \frac{1}{4} Z_i^4 - \frac{1}{2} Z_i^2 - \frac{1}{4} \end{pmatrix} \\ = \begin{pmatrix} \mathbf{D} / \sigma_0^2 & 0 & 0 \\ 0 & 2 / \sigma_0^2 & 2 / \sigma_0 \\ 0 & 2 / \sigma_0 & 7 / 2 \end{pmatrix},$$

in which it is assumed that the $p \times p$ design matrix $\mathbf{D}$, the limit in probability of $\mathbf{D}_n = n^{-1} \sum_{i=1}^n x_i x'_i$, exists. We note that

$$J_{\text{wide}}^{-1} = \begin{pmatrix} \sigma_0^2 \mathbf{D}^{-1} & 0 & 0 \\ 0 & \frac{7}{6} \sigma_0^2 & -\frac{2}{3} \sigma_0 \\ 0 & -\frac{2}{3} \sigma_0 & \frac{2}{3} \end{pmatrix}, \\ J_{\text{narr}}^{-1} = \begin{pmatrix} \sigma_0^2 \mathbf{D}^{-1} & 0 \\ 0 & \sigma_0^2 / 2 \end{pmatrix}.$$

The parallel to Section 2's Lemma is that

$$\begin{pmatrix} \sqrt{n} \bar{U}_n \\ \sqrt{n} \bar{V}_n \\ \sqrt{n} \bar{W}_n \end{pmatrix} \rightarrow_d \begin{pmatrix} 0 + K \\ (2/\sigma_0)\delta + L \\ (7/2)\delta + M \end{pmatrix},$$

where $(K, L, M)' \sim \mathcal{N}_{p+2}\{0, J_{\text{wide}}\}$. This is true by the triangular Lindeberg theorem under the usual condition $(1/n) \max_{i \leq n} (x_{i,j} - \bar{x}_i)^2 \rightarrow 0$ for each i . For the familiar normality-based (least squares-type) estimators, one finds that

$$\begin{pmatrix} \sqrt{n}(\hat{\beta}_{\text{narr}} - \beta_0) \\ \sqrt{n}(\hat{\sigma}_{\text{narr}} - \sigma_0) \end{pmatrix} \doteq_d J_{\text{narr}}^{-1} \begin{pmatrix} \sqrt{n} \bar{U}_n \\ \sqrt{n} \bar{V}_n \end{pmatrix} \rightarrow_d \begin{pmatrix} 0 + \sigma_0^2 K \\ \sigma_0 \delta + \frac{1}{2} \sigma_0^2 L \end{pmatrix} \\ = \begin{pmatrix} 0 + \mathbf{D}A \\ B + \sigma_0(C + \delta) \end{pmatrix} \\ \sim \mathcal{N}_{p+1} \left\{ \begin{pmatrix} 0 \\ \sigma_0 \delta \end{pmatrix}, \begin{pmatrix} \sigma_0^2 \mathbf{D}^{-1} & 0 \\ 0 & \frac{1}{2} \sigma_0^2 \end{pmatrix} \right\},$$

writing $(A, B, C)'$ for $J_{\text{wide}}^{-1}(K, L, M)'$, which is $\mathcal{N}_{p+2}\{0, J_{\text{wide}}^{-1}\}$. Next, regarding the ML estimators $\hat{\beta}, \hat{\sigma}, \hat{\gamma}$ in the wider $p + 2$ -parameter model, Proposition 2 with proof can be lifted mutatis mutandis and becomes

$$\begin{pmatrix} \sqrt{n}(\hat{\beta} - \beta_0) \\ \sqrt{n}(\hat{\sigma} - \sigma_0) \\ \sqrt{n}\hat{\gamma} / \kappa \end{pmatrix} \rightarrow_d \begin{pmatrix} A \\ B \\ T \end{pmatrix} \quad \text{if } T \geq 0, \\ \rightarrow_d \begin{pmatrix} A \\ B + \sigma_0 \kappa T \\ 0 \end{pmatrix} \quad \text{if } T \leq 0.$$

The rest of the story is very similar to that of Sections 3 and 4. The limit variables Λ_{narr} and Λ_{wide} are as in (4.3), only with $(\partial \mu / \partial \beta)' \mathbf{D} A$ replacing $(\partial \mu / \partial \xi)' A$, and A is now $\mathcal{N}_p\{0, \sigma_0^2 \mathbf{D}^{-1}\}$ and not merely $\mathcal{N}\{0, \sigma_0^2\}$. Conditioning on T , we find in the end that the main result (4.2) is true, with

$$b = \sigma_0 \frac{\partial \mu}{\partial \sigma} - \frac{\partial \mu}{\partial \gamma}, \\ \tau_0^2 = \left[\left(\frac{\partial \mu}{\partial \beta} \right)' \mathbf{D}^{-1} \left(\frac{\partial \mu}{\partial \beta} \right) + \frac{1}{2} \left(\frac{\partial \mu}{\partial \sigma} \right)^2 \right] \sigma_0^2. \quad (5.2)$$

Section 3's main result about $m \geq 1.458 \sqrt{n}$ for all estimands

with $b \neq 0$ is also true verbatim. And for the problem of performing linear regression analysis when the residuals could have fatter tails than the normal, a natural proposal is to use

$$\mu^* = [1 - w(\sqrt{1.5n}\hat{\gamma})]\mu(\hat{\beta}_{\text{narr}}, \hat{\sigma}_{\text{narr}}, 0) + w(\sqrt{1.5n}\hat{\gamma})\mu(\hat{\beta}, \hat{\sigma}, \hat{\gamma}), \quad (5.3)$$

where $w(\cdot)$ is as in (c) or (d) of Section 4.

6. GENERAL NORMAL SCALE MIXTURES

The primary focus of this article has been the t extension of the normal to model fatter tails in the error distribution. This choice is partly dictated by tradition and familiarity; other models might more successfully fit many data sets with the heavier tails syndrome, but everybody knows the t . In addition, Lange et al. (1989) and others have reported good robustness results using the t . In any case it is comforting that results quite similar to those of previous sections can be obtained for the large class of general normal scale mixtures, and that these results agree qualitatively with those for t -ness, as we now demonstrate.

Suppose that $Y_i = \xi + \sigma Z_i$ where the Z_i 's are iid and distributed as $\mathcal{N}\{0, 1\}/S$, say, where S is independent of the nominator. Let us merely assume that S has a density $h_\gamma(s)$ on $(0, \infty)$ and that this distribution becomes increasingly concentrated around 1 as $\gamma \rightarrow 0$. For technical convenience, we stipulate that $E_\gamma S = 1 + k_1\gamma + O(\gamma^2)$, $E_\gamma(S-1)^2 = k_2\gamma + O(\gamma^2)$, $E_\gamma(S-1)^3 = O(\gamma^2)$, and $E_\gamma(S-1)^4 = O(\gamma^2)$; these are implied by moment convergence of $(S-1)/(k_2\gamma)^{1/2}$ to the standard normal, for example. Results will be seen to depend only on the local mean and the local variance of S ; that is, on k_1 and k_2 alone. The limiting case gives a unit point mass at $S = 1$ that corresponds to a perfectly normal error distribution. Again, the program is to consider a sequence of contiguous alternatives to the normal where $\gamma = \delta/\sqrt{n}$. The t case hitherto considered is the special case where $S = (\chi_m^2/m)^{1/2}$ and $\gamma = 1/m$ and for which $k_1 = -\frac{1}{4}$ and $k_2 = \frac{1}{2}$.

The density of Z_i is $f_\gamma(z) = \int_0^\infty \phi(zs)sh_\gamma(s) ds$. Expand for s close to 1 to find

$$\phi(zs)s = \phi(z) \left[1 - (z^2 - 1)(s - 1) + \frac{1}{2}(z^3 - 3z)z(s - 1)^2 + O((s - 1)^3) \right].$$

With some further details, this leads to $f_\gamma(z) = \phi(z)[1 + \gamma R(z) + \gamma^2 R_2(z) + O(\gamma^3)]$, where

$$R(z) = \frac{1}{2}k_2(z^4 - 6z^2 + 3) + k_3(z^2 - 1),$$

$$\text{writing } k_3 = \frac{3}{2}k_2 - k_1, \quad (6.1)$$

and $R_2(z)$ is a certain polynomial of degree 8 and with mean 0 under the normal. This gives

$$\log f_\gamma(z) = \log \phi(z) + \gamma R(z) - \frac{1}{2}\gamma^2 S(z) + O(\gamma^3), \quad (6.2)$$

paralleling (and generalizing) (2.2) and (A.1) in the Appendix. Here $S(z) = R(z)^2 - R_2(z)$, and in the final analysis (see the t case proof in the Appendix) it only matters that $S(Z)$ has mean equal to the variance of $R(Z)$ in the null case.

One can now go through the chains of reasoning from previous sections and obtain analog results for the present general model. There is a new J_{wide} matrix, and the analog of Proposition 2 features $B + k_3\sigma_0(C + \delta)$, where $(A, B, C)' \sim \mathcal{N}_3\{0, J_{\text{wide}}^{-1}\}$ (and $k_3 = 1$ for the t case). We omit details but report that the narrow and the wide limiting risks are

$$E\Lambda_{\text{narr}}^2 = (b^*\delta)^2 + \tau_0^2$$

and

$$E\Lambda_{\text{wide}}^2 = (b^*)^2 E[C^2 I\{C \geq -\delta\} + \delta^2 I\{C \leq -\delta\}] + \tau_0^2,$$

featuring a new $b^* = k_3\sigma_0(\partial\mu/\partial\sigma) - (\partial\mu/\partial\gamma)$, while τ_0^2 still is as in (3.1). The variance κ^2 of C is now $1/(6k_2^2)$. As in Sections 3 and 4, it is convenient to use $a = \delta/\kappa$ as a measure of normalized distance from normality. The arguments of Section 3 lead to the following.

General Result. (a) If the $\mu = \mu(\xi, \sigma, \gamma)$ estimand is such that $b^* = 0$, then $\hat{\mu}_{\text{wide}}$ and $\hat{\mu}_{\text{narr}}$ are asymptotically equivalent, regardless of δ . (b) If the μ estimand is such that $b^* \neq 0$, then narrow-model-based estimation is better than or as good as wide-model-based estimation (with respect to limiting mean squared error) if and only if $a \leq .8399$, or $\gamma \leq .8399\kappa/\sqrt{n}$, or $k_2\gamma = \text{var } S \leq .3429/\sqrt{n}$.

Exactly the same tolerance radius is found in the regression context $Y_i = x_i'\beta + \sigma Z_i$ when Z_i is distributed as a standard normal divided by S , after paralleling the work of Section 5. The answer is perhaps surprisingly simple, involving only the variance of the scale mixture, but comes about because our large-sample framework is concerned only with local alternatives to the normal. The general result also indicates that conclusions of previous sections based on the t -ness model, including the compromise recipes of Section 5, remain qualitatively correct in a broader robustness context.

There are also parametric families with both heavier and lighter than normal tails. In the technical report version of this article, calculations are given for a quasi-extension of the t allowing negative degrees of freedom. A better example is the exponential power family discussed by Box and Tiao (1983, sec. 3.2). Calculating the tolerance distance and devising compromise estimators is much easier than for the t case, applying general methods of Hjort (1993), due to the fact that the normal in these cases is an inner member of the family. Finally, we point out that there are other natural extensions of the basic normal model that also involve problems with corners of parameter spaces and where this article's methods should be useful. One example is the normal contamination model.

7. ADDITIONAL REMARKS

7.1 Some Estimands

To illustrate both the general formulas and the relative importance of bias and estimation noise, let us go through a short list of important estimands.

1. Let $\mu = x'\beta$, the regression curve at a specific point. Then $b = 0$ and all compromise estimators become asymptotically equivalent, with $\tau_0^2 = \frac{1}{2} x' \mathbf{D}^{-1} x \sigma_0^2$ as the limiting normalized risk. Thus familiar least squares estimates are sufficiently precise even in the presence of t -ness, and the same is true in other cases where the estimand depends only on $\beta_1, \dots, \beta_p$.

2. Let $\mu = E|Y(x) - x'\beta|$, our starting example of Section 5. Then $\mu = \sigma E|Z|$, where Z is t_m distributed, and some determined analysis gives $b = \frac{1}{2} \sigma_0 \phi(0)$, $\tau_0^2 = 2\phi(0)^2 \sigma_0^2$. This gives

$$\text{risk} = \frac{\sigma_0^2}{\pi} \left[\frac{1}{12} R(a) + 1 \right]$$

for the limit distribution version of n times mean squared error for μ^* ; see (4.2).

3. Let μ be the p th quantile of the distribution for $Y(x)$ at x . It is, for example, often useful and illuminating to draw the nine regression deciles (corresponding to $p = j/10$) in the same diagram, as functions of x . Then $\mu = x'\beta + \sigma G^{-1}(p, m)$ in the notation of Section 1. One can work out a suitable expression for $\partial\mu/\partial\gamma$ and then find b and τ_0 of (5.2). The end result is

$$\text{risk} = \left\{ \frac{2}{3} [z_p + A(z_p)/\phi(z_p)]^2 R(a) + x' \mathbf{D}^{-1} x + \frac{1}{2} z_p^2 \right\} \sigma_0^2,$$

in which $z_p = \Phi^{-1}(p)$ and $A(t) = \int_{-\infty}^t \phi(z) W(z) dz$, and $W(z)$ is as in (2.5). Formula (5.3), with a weight function deemed appropriate in view of the discussion of Section 4, provides a regression quantile estimator that is robust against fatter-than-normal tails of the error distribution. More generally, robust methods for regression quantiles have been provided by Koenker and Bassett (1978), Koenker and Portnoy (1987), and Hjort and Pollard (1993, sec. 3D).

4. The case of a probability $\mu = \Pr\{Y(x) \leq y\} = G((y - x'\beta)/\sigma, m)$ is similar to but simpler than case 3. The same expression for risk emerges, with $z(y) = (y - x'\beta_0)/\sigma_0$ replacing z_p .

7.2 False Confidence

We proved in Section 3 that $\sqrt{n}(\hat{\mu}_{\text{narr}} - \mu_{\text{true}})$ tends to $\mathcal{N}\{b\delta, \tau_0^2\}$ under the (2.4) sequence of models. Traditional normality-based inference uses in essence that the limit is $\mathcal{N}\{0, \tau_0^2\}$. Accordingly, $b^2\delta^2$ is the invisible extra burden associated with using the normality-based estimator when in fact the wider model (2.4) is true. Consequences of this include that traditional normality-based confidence intervals and testing procedures behave incorrectly; the intervals have adequate length but are incorrectly placed, and the tests have too-high significance levels. If $CI_{\text{narr}} = \hat{\mu} \pm 1.645\hat{\tau}_0/\sqrt{n}$, for example, then the coverage probability converges to $\Pr\{|\mathcal{N}\{b\delta/\tau_0, 1\}| \leq 1.645\}$, which is strictly less than 90% unless $b = 0$ or $\delta = 0$ (see also Hjort 1993, sec. 4H).

7.3 How Far Away is $1.458\sqrt{n}$?

One can test normality ($\gamma = 0$) against t -ness ($\gamma > 0$) using $T_n = \sqrt{1.5n}\hat{\gamma}$ (see Sect. 4). The limit distribution under normality is $\max\{N, 0\}$, where N is standard normal. The

$T_n > 1.645$ test has (asymptotic) level 5% and power $\Phi(a - 1.645)$. One way of quantifying the distance from normality to the first intolerable t -distribution is in terms of $\Phi(0.8399 - 1.645) = .210$, the probability of detecting this amount of t -ness. The corresponding detection probability figure is .329 for the case of a 10% level test. These rather low power figures, even for the strong test that looks specifically for t deviation, suggest that $1.458\sqrt{n}$ is not very far away from the normal, and that this moderate amount might not be easily visible. Statisticians often look at quantile-quantile plots to check approximate normality, and simulations demonstrate that these plots often look fairly linear with this amount of t -ness. As a little experiment, consider the natural quantile test statistic

$$D_n = \max_i \sqrt{n+2} |Y_{(i)} - \hat{\xi}_{\text{narr}} - \hat{\sigma}_{\text{narr}} \Phi^{-1}(i/(n+1))|$$

for normality against nonnormal tails, where the maximum is over $.025 \leq i/(n+1) \leq .975$, say (i.e., it is scaled in such a way that it has a limit distribution in terms of a zero-mean Gaussian process); this is the formal version of looking for nonstraightness of the quantile plot. The power of the 5% version of this test for $n = 100$ at the suggested tolerance limit with 14.58 degrees of freedom is about 13%. The corresponding power figure is 23% for a 10% level test.

The bottom line is that standard normal-based procedures for estimands other than the simple mean parameter (more precisely, those where b of (3.1) or (5.2) is not 0) are in danger. The risk calculations of Section 4 and the confidence undershooting pointed out in Section 7.2 are relevant for quite moderate doses of fat tails. This agrees also with the general scale mixture calculations of Section 6. Remedies include full estimation in the wide model, which is what Lange et al. (1989) did, or compromise, as with (5.3).

There are other measures of distance from normality to t -ness that could fruitfully be used here (see Hjort 1993, secs. 4B and 4C for other proposals and interpretations in a general parametric robustness context). The L_1 distance $\int |f_\gamma - f_0| dy$ is approximately $.434/\sqrt{n}$ at the tolerance threshold, for example. Kass (1989, p. 211) mentioned as a piece of statistical trivia that the t density that in the sense of a certain curvature measure lies halfway between the Cauchy and the normal has $m = 3$ degrees of freedom, and some calculations give that on the L_1 -distance scale, the corresponding answer has $m = 2.3$ degrees of freedom. The point is again that $m \geq 1.458\sqrt{n}$ means being reasonably close to normality.

7.4 Additional Deviances from the Normal

In many situations one would have to worry about additional ways in which the normal regression model could be wrong; for example, in the direction of variance heterogeneity or nonlinearity of the mean. This calls for an extension of our framework to moderate misspecification in several directions, where the first is t -ness. Our results can be suitably generalized and become more nuanced. There will be a sacrosanct region around the normal, inside of which the precision of *each normality-based estimator* is at least as good as that of the corresponding wide-model estimator. *Each*

given estimand will have its own, larger tolerance region, however, where narrow estimation is more precise than wide estimation (see Hjort 1993, sec. 5I).

7.5 Other Loss Functions

Our risk functions and tolerance limits have been computed starting from the squared error criterion, where squared bias and variance are exchangeable currencies. Similar results can be reached for other loss functions but become more complicated. If absolute loss $\sqrt{n}|\mu^* - \mu_{\text{true}}|$ is used, then there is again a one-one correspondence between estimators (4.1) of μ and estimators $\hat{a}(t) = w(t)t$ of the nonnegative a based on $T \vee 0$, where $T \sim \mathcal{N}\{a, 1\}$, but the absolute loss for μ transforms into the different loss function $\int_0^\infty L(\rho(\hat{a}(t) - a))\phi(t - a) dt$ for a , where $L(z) = E|z + \mathcal{N}\{0, 1\}| = z + 2\phi(z) - 2[1 - \Phi(z)]$ and $\rho = |b|/\tau_0$. The tolerance limit now depends on the parameter estimated via the value of ρ (see Hjort 1993, sec. 5H).

APPENDIX: PROOF OF PROPOSITION 2

By (2.2), the log-likelihood function can be written as

$$\begin{aligned} L_n(\xi, \sigma, \gamma) &= \sum_{i=1}^n \log f(y_i, \xi, \sigma, \gamma) \\ &= -n \log(2\pi)^{1/2} - n \log \sigma - \frac{1}{2} \sum_{i=1}^n (y_i - \xi)^2 / \sigma^2 \\ &\quad + \gamma \sum_{i=1}^n R((y_i - \xi)/\sigma) \\ &\quad - \frac{1}{2} \gamma^2 \sum_{i=1}^n S((y_i - \xi)/\sigma) + O_p(n\gamma^3), \end{aligned} \quad (\text{A.1})$$

in which

$$R(z) = \frac{1}{4} z^4 - \frac{1}{2} z^2 - \frac{1}{4} \quad \text{and} \quad S(z) = \frac{1}{3} z^6 - \frac{1}{2} z^4.$$

The limit in probability of $(1/n)L_n(\xi, \sigma, \delta/\sqrt{n})$, under sequence (2.4), is seen to be $-\log(2\pi)^{1/2} - \log \sigma - \frac{1}{2}[(\xi - \xi_0)^2 + \sigma_0^2]/\sigma^2$, uniformly over compact sets. It follows that the sequence of ML estimators $\hat{\xi}, \hat{\sigma}$ must converge in probability to the parameter values that maximize this limit; i.e., to the underlying ξ_0, σ_0 . By working with the γ -related part of (A.1), one shows similarly that $\hat{\gamma} = \hat{\delta}/\sqrt{n}$ must converge to 0 in probability.

In what follows, let $I_n(\xi, \sigma, \gamma)$ be the 3×3 matrix with elements $\sum_{i=1}^n (\partial^2/\partial \xi^2) \log f(y_i, \xi, \sigma, \gamma)$ and so on. If $\tilde{\xi}_n, \tilde{\sigma}_n$, and $\tilde{\gamma}_n$ tend to ξ_0, σ_0 , and 0 in probability, still under the (2.4) sequence of models, then $-(1/n)I_n(\tilde{\xi}_n, \tilde{\sigma}_n, \tilde{\gamma}_n) \rightarrow_p J_{\text{wide}}$. This holds because direct inspection shows

$$\begin{aligned} &-\frac{1}{n} I_n(\tilde{\xi}_n, \tilde{\sigma}_n, \tilde{\gamma}_n) \\ &= -\frac{1}{n} I_n(\xi_0, \sigma_0, 0) + O_p(|\tilde{\xi}_n - \xi_0| + |\tilde{\sigma}_n - \sigma_0| + \tilde{\gamma}_n), \end{aligned}$$

and the first term here can be shown to converge to J_{wide} , under (2.4), using ordinary methods.

There are two possibilities regarding the maximizers of (A.1). Data $(y_1, \dots, y_n)$ are such that maximum occurs either for some $\hat{\gamma} > 0$ or occurs for $\hat{\gamma} = 0$. In the first case the ML values are solutions to $\partial L_n/\partial \xi = 0$, $\partial L_n/\partial \sigma = 0$, $\partial L_n/\partial \gamma = 0$, and the familiar Taylor argument yields

$$\begin{pmatrix} \sqrt{n}(\hat{\xi} - \xi_0) \\ \sqrt{n}(\hat{\sigma} - \sigma_0) \\ \sqrt{n}(\hat{\gamma} - 0) \end{pmatrix} = [-I_n(\tilde{\xi}, \tilde{\sigma}, \tilde{\gamma})/n]^{-1} \begin{pmatrix} \sqrt{n}\bar{U}_n \\ \sqrt{n}\bar{V}_n \\ \sqrt{n}\bar{W}_n \end{pmatrix}$$

for suitable $(\tilde{\xi}, \tilde{\sigma}, \tilde{\gamma})$ somewhere between $(\xi_0, \sigma_0, 0)$ and $(\hat{\xi}, \hat{\sigma}, \hat{\gamma})$; see the definition in (2.5). In the second case $L_n(\xi, \sigma, \gamma)$ decreases in $\gamma \geq 0$, and the ML estimators are $(\hat{\xi}_{\text{narr}}, \hat{\sigma}_{\text{narr}}, 0)$. Let Ω_n be the set of $(y_1, \dots, y_n)$ for which the first case happens. Then $\sqrt{n}(\hat{\xi} - \xi_0, \hat{\sigma} - \sigma_0, \hat{\gamma} - 0)'$ becomes

$$\begin{aligned} &J_{\text{wide}}^{-1} \begin{pmatrix} \sqrt{n}\bar{U}_n \\ \sqrt{n}\bar{V}_n \\ \sqrt{n}\bar{W}_n \end{pmatrix} + O_p(n^{-1/2}) \\ \text{or} \quad &\begin{pmatrix} J_{\text{narr}}^{-1} & \sqrt{n}\bar{U}_n \\ & \sqrt{n}\bar{V}_n \\ & 0 \end{pmatrix} + O_p(n^{-1/2}) \end{aligned}$$

as Ω_n is in command or not. It turns out that Ω_n happens or not according to whether

$$\Delta_n = -\frac{2}{3} \sigma_0 \sqrt{n} \bar{V}_n + \frac{2}{3} \sqrt{n} \bar{W}_n > 0 \quad \text{or} \quad \leq 0$$

to a first-order approximation. The Δ_n here is the third component of $J_{\text{wide}}^{-1} \sqrt{n}(\bar{U}_n, \bar{V}_n, \bar{W}_n)'$, and the precise statement is that $I(\Omega_n) - I\{\Delta_n > 0\}$ goes to 0 in probability under the (2.4) regime of models. Using this result, the Lemma, and (2.7)–(2.8) in tandem yields the statement of Proposition 2, by the continuous mapping theorem on $\sqrt{n}(\bar{U}_n, \bar{V}_n, \bar{W}_n)'$.

To prove that Ω_n and $\{\Delta_n > 0\}$ are asymptotically equivalent events, consider once more the second half of (A.1), which can be expressed as

$$\delta \frac{1}{\sqrt{n}} \sum_{i=1}^n R((y_i - \xi)/\sigma) - \frac{1}{2} \delta^2 \frac{1}{n} \sum_{i=1}^n S((y_i - \xi)/\delta)$$

plus a remainder of size $O_p(\delta^3/\sqrt{n})$. This is a parabola in $\delta \geq 0$, with maximum occurring to the right of 0 or at 0 depending on the sign of the R average. (The S average will be positive with probability tending to 1 in the parameter region of interest.) Accordingly, apart from an $O_p(n^{-1/2})$ term,

$$\sqrt{n}\hat{\gamma} = \hat{\delta} = \frac{1}{\sqrt{n}} \sum_{i=1}^n R((y_i - \hat{\xi})/\hat{\sigma}) / \frac{1}{n} \sum_{i=1}^n S((y_i - \hat{\xi})/\hat{\sigma})$$

provided that nominator is positive, and $\hat{\delta} = 0$ otherwise. But

$$\begin{aligned} \frac{1}{n} \sum_{i=1}^n R\left(\frac{y_i - \hat{\xi}}{\hat{\sigma}}\right) &= \frac{1}{n} \sum_{i=1}^n R\left(z_i - \frac{\hat{\xi} - \xi_0}{\sigma_0} - \frac{y_i - \xi_0}{\sigma_0^2} (\hat{\sigma} - \sigma_0)\right) \\ &\quad + O_p(n^{-1}) \\ &= \frac{1}{n} \sum_{i=1}^n R(z_i) - \frac{1}{n} \sum_{i=1}^n R'(z_i) \frac{\hat{\xi} - \xi_0}{\sigma_0} \\ &\quad - \frac{1}{n} \sum_{i=1}^n R'(z_i) z_i \frac{\hat{\sigma} - \sigma_0}{\sigma_0} + O_p(n^{-1}), \end{aligned}$$

where $z_i = (y_i - \xi_0)/\sigma_0$, and similarly for the S -function term. Judicious calculations based on this show that $(1/n) \sum_{i=1}^n S((y_i - \hat{\xi})/\hat{\sigma})$ goes to $7/2$ and that

$$\begin{aligned} \frac{1}{\sqrt{n}} \sum_{i=1}^n R((y_i - \hat{\xi})/\hat{\sigma}) &= \sqrt{n} \bar{W}_n - (2/\sigma_0) \sqrt{n} (\hat{\sigma} - \sigma_0) + O_p(n^{-1/2}) \\ &= \frac{7}{3} \sqrt{n} \bar{W}_n - \frac{7}{3} \sigma_0 \sqrt{n} \bar{V}_n + O_p(n^{-1/2}) \end{aligned}$$

on the set Ω_n . This finally means that $\sqrt{n}\hat{\gamma} = \Delta_n + O_p(n^{-1/2})$ in the $\Delta_n + O_p(n^{-1/2}) > 0$ case and is 0 in the $\Delta_n + O_p(n^{-1/2}) \leq 0$ case. This proves what was needed.

[Received April 1991. Revised July 1993.]

REFERENCES

- Bickel, P. J. (1983), "Minimax Estimation of the Mean of a Normal Distribution Subject to Doing Well at a Point," in *Recent Advances in Statistics, A Festschrift for Herman Chernoff*, eds. Rizvi and Siegmund, New York: Academic Press, pp. 511–528.
- (1984), "Parametric Robustness: Small Biases Can be Worthwhile," *The Annals of Statistics*, 12, 864–879.
- Box, G. E. P., and Tiao, G. C. (1973), *Bayesian Inference in Statistical Analysis*, Menlo Park, CA: Addison-Wesley.
- Hjort, N. L. (1993), "Estimation in Moderately Misspecified Models," submitted to *Journal of the American Statistical Association*.
- Hjort, N. L., and Pollard, D. (1993), "Asymptotics for Minimisers of Convex Processes," submitted to *Annals of Statistics*.
- Kass, R. E. (1989), "The Geometry of Asymptotic Inference" (with discussion), *Statistical Science*, 3, 188–234.
- Koenker, R. W., and Bassett, G. W. (1978), "Regression Quantiles," *Econometrica*, 46, 33–50.
- Koenker, R. W., and Portnoy, S. (1987), " L Estimation for Linear Models," *Journal of the American Statistical Association*, 82, 851–857.
- Lange, K. L., Little, R. J. A., and Taylor, J. M. G. (1989), "Robust Statistical Modeling Using the t Distribution," *Journal of the American Statistical Association*, 84, 881–896.
- Morgenthaler, S., and Tukey, J. W. (eds.) (1991), *Configural Polysampling: A Route to Practical Robustness*, New York: John Wiley.