

Resampling Permutations in Regression without Second Moments*

RAOUL LEPAGE

Michigan State University

AND

KRZYSZTOF PODGÓRSKI

Indiana Purdue University

For given observations we consider the resampling distribution obtained by permuting residuals versus the *distribution of errors conditional on their order statistics*. We observe that *with high probability* both are approximately equal on *contrast vectors* after suitable normalization. No moment/distribution assumptions on the underlying errors other than exchangeability are required. These results remain valid if we consider reduced regression models derived from the original one by removing data for which the corresponding residuals take extreme values. We obtain general conditions under which confidence regions produced by such methods are accurate. In the absence of finite second moments randomness of the limiting bootstrap distributions has been considered as a failure of the traditional bootstrap and no practical meaning has been given to the phenomenon. For our method this type of randomness is still present in the limit but it has clear probabilistic interpretation as a conditional distribution which can be used, e.g., to obtain conditional confidence sets. We study this phenomenon in detail for the case of independent errors with distribution from the domain of attraction of stable laws. © 1996 Academic Press, Inc.

1. INTRODUCTION

This paper considers a multiple regression model

$$Y = \mathbf{X}\beta + \varepsilon, \quad (1)$$

where $\mathbf{X}$ is an $n \times d$ nonrandom design matrix, β is a $d \times 1$ vector of parameters, and ε is an $n \times 1$ random vector of errors which has an

Received January 24, 1994; revised October 10, 1995.

1991 AMS subject classification: 62G09.

Key words and phrases: resampling, linear regressions, permutations, bootstrap, contrasts, stable.

* Research partially supported by ONR Grant N00014-91-J-1087.

exchangeable distribution, i.e., a distribution which is invariant on permutations of coordinates. Let us denote the vector of the least squares residuals by $\varepsilon^\perp = Y - \hat{Y}$, where $\hat{Y}$ is the projection of Y to the column space of $\mathbf{X}$.

For reasons discussed later we also consider reduced regression models which come from (1) by deleting rows of the design matrix. For a vector $e \in \mathbb{R}^n$ let $R_i = R_i(e)$ be the rank of e_i among $(e_i)_{i=1}^n$ counted from the bottom and defined in such a way that $(R_i)_{i=1}^n$ is a permutation of all $\{1, \dots, n\}$ regardless of ties in e . The last condition can be preserved, for example, if among tied coordinates of e a larger rank will be assigned to a coordinate with a larger index. For any subset $M \subseteq \{1, \dots, n\}$ we define a reduced regression model

$$Y_M = \mathbf{X}_M \beta + \varepsilon_M, \quad (2)$$

where Y_M , $\mathbf{X}_M$, and ε_M consist of those rows of Y , $\mathbf{X}$, and ε indexed by i whose R_i is in M . In this paper we will consider cases when the deletion is based on ranks $(R_i)_{i=1}^n$ of either errors or computable residuals. For example by taking $M = \{2, \dots, n-1\}$ we exclude the two observations with the smallest and the largest errors or with the smallest and the largest residuals. Notice that with a fixed set M for different values of the vector of errors ε we can have different reduced models (2) and thus the design matrix $\mathbf{X}_M$ becomes random.

About the reduced design matrices $\mathbf{X}_M$ we assume that they preserve rank d . This assumption is not necessary but simplifies statements of results. Otherwise we would have to divide the underlying probability space on disjoint sets where $\mathbf{X}_M$ has constant rank and state our results conditioning on such a partition. On the other hand, the assumption is natural since, according to the results, our method works when d/m is small.

A vector v is called a *contrast* if its entries add to zero and, equivalently, if the Euclidean scalar product $v \cdot \mathbf{1} = 0$, where $\mathbf{1}$ denotes the vector of 1's. The more important assumption on the design $\mathbf{X}$ requires that the vector $\mathbf{1}$ belongs to column space of $\mathbf{X}$ and thus of any $\mathbf{X}_M$ for a subset M of $\{1, \dots, n\}$. Without losing generality we can assume that $\mathbf{1}$ is the first column of $\mathbf{X}$. In such a case contrast vectors come to (2) in a natural way. Namely, for $i > 1$ the least squares estimators of β_i are of the form $\hat{\beta}_i = v_i \cdot Y_M$ for contrasts v_i depending also on $R_{\bar{M}} = \{R_i \notin M\}$. Indeed, v_i is the i th row of the matrix $\mathbf{M} = (\mathbf{X}_M^T \mathbf{X}_M)^{-1} \mathbf{X}_M^T$ for which we have $\mathbf{M} \mathbf{X}_M = \mathbf{I}$ and this last equality implies that any row v_i for $i > 1$ is orthogonal to the first column which is equal to $\mathbf{1}$.

Ordinarily one is interested in approximating the joint distribution of $v \cdot \varepsilon$ for several v since it is equal to the estimation error, i.e., $v_i \cdot \varepsilon = (v_i \cdot Y) - (v_i \cdot \mathbf{X} \beta) = \hat{\beta}_i - \beta_i$, $i = 1, \dots, d$. For i.i.d. errors with finite

variance, the central limit theorem is popularly used for this purpose. Bootstrap methods apply to this case as well (see Efron, 1979; Freedman, 1981; Wu, 1986), but how are we to know if the underlying assumptions have practical application to the data at hand? See LePage (1990).

Instead of approximating or estimating the unconditional distribution of $v \cdot \varepsilon$, we propose, without moment assumptions, to estimate the sampling distribution of $v \cdot \varepsilon$ *conditional* on the sigma field $\mathcal{F}$ generated by the *order statistics* of the coordinates of ε . For this purpose we will use randomly permuted residuals. To be more precise let us introduce the following notation. We use $v \cdot \varepsilon | \mathcal{F}$ to denote the above conditional distribution. Let π denote a uniformly distributed random permutation applied to the coordinates of n -space. Suppose that the distribution of π , conditional on ε , is also uniform over all $n!$ permutations, so that π is independent of ε . We will observe that, provided d/n is small, the distribution $v \cdot \pi \varepsilon^\perp | Y$ provides with *high probability* a close approximation of $v \cdot \pi \varepsilon | \mathcal{F}$ given that the both distributions are scaled by dividing by the conditional constant $\sqrt{E^{\mathcal{F}}(v \cdot \pi \varepsilon)^2}$. Here and henceforth $E^{\mathcal{F}}$ denotes expectation conditional on $\mathcal{F}$. Notice that $v \cdot \pi \varepsilon | \mathcal{F}$ is equal to $v \cdot \varepsilon | \mathcal{F}$ since ε has exchangeable coordinates. As will be seen in the most important cases the scaling $\sqrt{E^{\mathcal{F}}(v \cdot \pi \varepsilon)^2}$ will not influence closeness of the conditional distributions. So we state the following proposal.

PROPOSAL 1. *For any finite set of contrast vectors $\{v_k, k \leq l\}$, estimate the joint $\mathcal{F}$ -conditional sampling distributions of contrasts $v_k \cdot \varepsilon$ by the distributions of $v_k \cdot \pi \varepsilon^\perp$ conditional on Y .*

Illustration. To explicate the proposed method let us consider a simple three-point regression $y = 1 + x + \varepsilon$ with $(x_1, x_2, x_3) = (-0.75, -0.25, 1)$ and $(\varepsilon_1, \varepsilon_2, \varepsilon_3) = (-0.75, 1.5, 1.75)$. Figure 1 below shows the least square fit to the data and five other fits resulting from permuting the three residuals.

The six-point distribution $v \cdot \pi \varepsilon^\perp | Y$, where v is the contrast corresponding to slope resides on $(-0.83, -0.51, -0.45, 0, 0.83, 0.95)$ while $v \cdot \varepsilon | \mathcal{F}$, which could be obtained by permuting actual errors, occupies $(-1.5, -1.42, 9.23, 0.5, 1, 1.19)$. Of course this simple example illustrates only the implementation of the method. For examples in support of Proposal 1 and other theoretical results we refer to Section 5.

Proposal 1 provides strong support for the approach taken by Freedman and Lane (1983) who develop *descriptive* tests of linear hypotheses. To quote from them, “our reference sets are derived by permuting residuals, and our significance level is a descriptive statistic rather than a probability.” It will be seen below that $v \cdot \pi \varepsilon^\perp | Y$ is an estimate of a *conditional* distribution insensitive to the moment assumptions. Thus it is

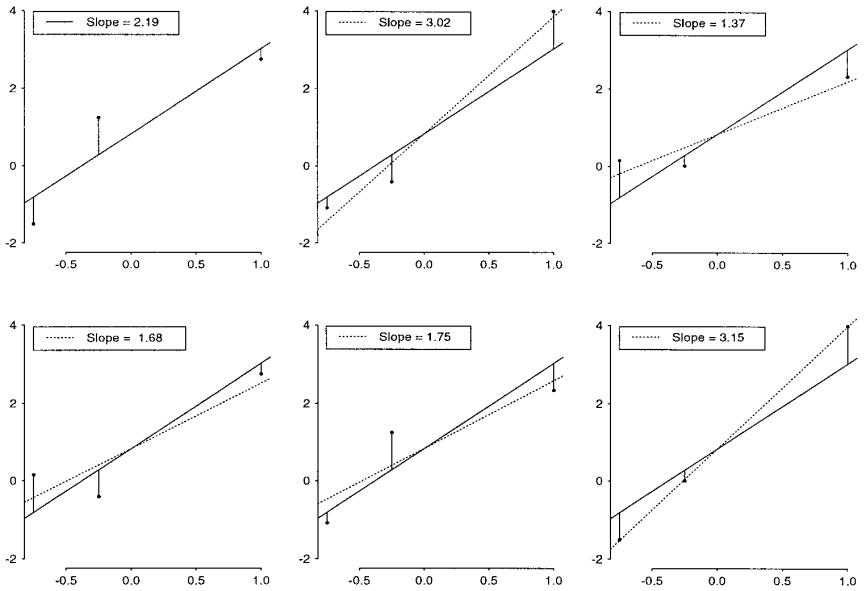


FIG. 1. Permuting residuals in three-point regression.

possible to achieve robust estimation of the *conditional* sampling distributions of least squares estimators, which are conditionally unbiased in the case of contrasts.

As we will see Proposal 1 can be applied to a quite general class of exchangeable distributions. However, sometimes the conditional distribution $v \cdot \varepsilon | \mathcal{F}$ itself will possess certain unpleasant properties, as in the case of long-tailed distributions when some coordinates of ε can take relatively very large values. It can happen for example that confidence regions will not be in the form of single intervals but sums of widely separated intervals. By our result the distribution $v \cdot \pi \varepsilon^\perp | Y$ will inherit this. In such a case to “improve” the distribution of interest it would be reasonable to exclude observations with outsized disturbance by errors and then apply our method of resampling by permutations to this reduced model. Our results extend also to this situation.

Let us describe this in a more rigorous way. We consider a reduced model as defined by (2) and denote its residuals $\varepsilon_M^\perp = Y_M - \hat{Y}_M$, where by $\hat{Y}_M$ we denote now the projection of Y_M into the column space of the reduced design $\mathbf{X}_M$. In the problem of estimation of the coefficients β we use contrast vectors that are rows of $(\mathbf{X}_M^\top \mathbf{X}_M)^{-1} \mathbf{X}_M^\top$ and the matrix $\mathbf{X}_M$ is now dependent on values of $R_{\bar{M}}$. For that reason we allow a contrast vector $v = v(R_{\bar{M}}) \in \mathbb{R}^m$ to be dependent on ε through ranks $R_{\bar{M}}$ of either errors

or, more suitably, some computable residuals (see *Rank Problem* in Section 2). Further, let $\mathcal{F}_M$ denote the sigma field generated by $R_{\bar{M}}$, together with the order statistics of ε_M , while $\mathcal{G}_M$ is the sigma field generated by $R_{\bar{M}}$ and Y_M . Notice that ranks $R_{\bar{M}}$ provide information concerning data excluded from the model, Y_M are observations remaining after this exclusion, and $\mathcal{G}_M$ represents information which should be given to us to employ a reduced model based on extreme values of errors. We will use the same letter π for a random permutation in $\mathbb{R}^m$ disregarding dependence on a dimension. With this notation we will find that if the ratio d/m is small then, as before, the distribution $v \cdot \pi \varepsilon_M^\perp | \mathcal{G}_M$ approximates the distribution $v \cdot \varepsilon_M | \mathcal{F}_M$ when appropriately scaled by $\sqrt{E^{\mathcal{G}_M}(v \cdot \pi \varepsilon_M)^2}$. Once again scaling usually will not disturb the asymptotic closeness of these distributions. Also we will see that under suitable assumptions the residuals can contain information about ranks of errors. Thus we can state the generalization of our proposal.

PROPOSAL 2. *For any finite set of contrast vectors $\{v_k; k \leq l\}$, which may depend upon ranks $R_{\bar{M}}$ of errors or, more suitably, residuals, estimate the joint $\mathcal{F}_M$ -conditional sampling distribution of contrasts $v_k \cdot \varepsilon_M$ by the joint $\mathcal{G}_M$ -conditional distribution of $v_k \cdot \pi \varepsilon_M^\perp$.*

In general, the problem of handling deviations from classical assumptions on linear regression can be approached from different points of view. Robust estimators diminish the role of long tails, while identifying and removing data with outsized errors may enable least squares analysis (see also Rousseeuw and Leroy, 1987, for some related discussion). Proposal 2 seems to be more in the spirit of this second approach. One of its most interesting features is that although we require d/n to be small the results by themselves are not asymptotic in the sense that by conditional analysis of the reduced models we can get tighter confidence intervals when using smaller number of observations. This is exploited in the statistical diagnostic procedure described in Section 5, Example 2. In certain reasonable examples our resulting confidence regions are, in fact, tighter than those based on standard implementations of M -estimation (see Section 5, Example 3).

With regard to the traditional bootstrap in the context of regression with long tailed errors the work of Arcones and Giné (1989) proves that taking smaller resample size of the order $o(n) n / \log \log n$ allows recovery of the asymptotic unconditional distribution of estimation errors for the sample mean case (see also Hall and LePage, 1995). This approach, if it could be applied to least squares, would enable one to develop confidence intervals based on the unconditional asymptotic sampling distribution of the least squares estimators. However (see Fig. 2, Section 3), this will in general lead

to larger confidence interval widths than the conditional approach taken in Proposition 1. This is because conditioning on the order statistics removes one source of variation. We believe that this inefficiency which should be present in more general situations is the major weakness of the reduced resample bootstrap for the long-tailed case and provides an additional argument in favor of recovery of conditional distributions. The work of Kinatader (1990) also deals with the traditional bootstrap and conditionally (but not as a conditional bootstrap!) interprets the limiting bootstrap distribution through the invariance principle for stable laws. Our work was partially inspired by these results but our method of permutation resampling, by recovering conditional distribution of estimation errors, has a natural statistical interpretation which does not exist for the traditional bootstrap. Moreover, the technical requirements for our proofs are greater since permutations do not asymptotically reduce to Poisson coefficients as with bootstrap.

Finally let us emphasize that our work does not aim at the presentation of the ready-to-use statistical methods although one is outlined in Section 5. We believe rather that the conditional approach developed here and introduced in LePage (1990) deliver a new and useful tool of statistical inference which is at least competitive with other methods. This tool, however, should be further thoroughly studied to find the range of its potential applicability.

2. MAIN RESULTS

We assume the notation of the introduction. As before, ε is assumed to have exchangeable coordinates and to be independent of a random uniformly distributed permutation π . Our first goal is to show that for any contrast vector $v = v(R_{\bar{M}})$ dependent on ranks $R_{\bar{M}} = R_{\bar{M}}(\varepsilon)$ the conditional distribution $v \cdot \pi \varepsilon_M^\perp | \mathcal{G}_M$ approximates the conditional distribution $v \cdot \varepsilon_M | \mathcal{F}_M$ when appropriately scaled by $\sqrt{E^{\mathcal{G}_M}(v \cdot \pi \varepsilon_M)^2}$. Here as before $\mathcal{G}_M = \sigma(Y_M, R_{\bar{M}})$ and $\mathcal{F}_M$ is generated by $R_{\bar{M}}$ and the order statistics of ε_M . In Section 4 it will be proven that both $v \cdot \pi \varepsilon_M^\perp | \mathcal{G}_M$ and $v \cdot \varepsilon_M | \mathcal{F}_M$ are centered at zero while the latter is equal to $v \cdot \pi \varepsilon_M | \mathcal{G}_M$. Our first result can be stated as follows.

PROPOSITION 1. *Let $M \subseteq \{1, \dots, n\}$ have cardinality m . If $\mathbf{1}$ is in the column space of $\mathbf{X}_M$ which has rank d and π is a random permutation of coordinates in $\mathbb{R}^m$ then for each contrast vector v dependent on $R_{\bar{M}}(\varepsilon)$ we have*

$$E_{\mathcal{F}} \frac{E^{\mathcal{G}_M}(v \cdot \pi \varepsilon_M - v \cdot \pi \varepsilon_M^\perp)^2}{E^{\mathcal{G}_M}(v \cdot \pi \varepsilon_M)^2} = \frac{d-1}{m-1}.$$

More generally, for contrast vectors v_k , $k = 1, \dots, r$, dependent on $R_{\bar{M}}(\varepsilon)$ we have

$$E^{\mathcal{F}} \frac{\sum_{k=1}^r E^{\mathcal{G}_M}(v_k \cdot \pi \varepsilon_M - v_k \cdot \pi \varepsilon_M^\perp)^2}{\sum_{k=1}^r E^{\mathcal{G}_M}(v_k \cdot \pi \varepsilon_M)^2} = \frac{d-1}{m-1}.$$

For the proofs of this and other results of this work we refer to Section 4.

We now discuss two problems in relation to the above result. The *Rank Problem* which is to justify substituting the ranks of the observed residuals for the ranks of unobserved errors $R_{\bar{M}}(\varepsilon)$ and the *Scaling Problem* which is to justify ignoring the unobserved scaling $\sqrt{E^{\mathcal{F}_M}(v \cdot \varepsilon_M)^2}$ in the construction of confidence intervals.

Rank Problem. Numerical simulations of the linear regression models with long tailed distributions for $d=2$ (see also Section 5) suggest that the extreme values of errors can be in many cases identified by some preliminary examination of data. Our goal here is to provide a formal procedure for such identifications which can be applied with mathematical rigor.

We consider here an unspecified subset $M \subseteq \{1, \dots, n\}$ but the most natural is to consider an M which excludes indices corresponding to extreme errors, for example the largest and the smallest ones. We assume that we know a $(1-\alpha)$ confidence region $\Gamma \subseteq \mathbb{R}^{d-1}$ for $(\beta_2, \dots, \beta_d)$. Such a confidence region can be based on conditional or unconditional distributions (see Section 5 for the description of one possible method of finding Γ). The following result is independent of how Γ has been chosen. Consider the $n \times (d-1)$ matrix $\mathbf{X} \ominus \mathbf{1} = [X_2, \dots, X_d] - [\bar{X}_2 \mathbf{1}, \dots, \bar{X}_d \mathbf{1}]$ and for $\gamma \in \mathbb{R}^{d-1}$ a vector of residuals $\varepsilon^\perp(\gamma) = (Y - \bar{Y} \cdot \mathbf{1}) - \mathbf{X} \ominus \mathbf{1} \gamma$ (now not necessarily against the least squares fit).

PROPOSITION 2. *Let M be a fixed subset of $\{1, \dots, n\}$. If Γ is an $(1-\alpha)$ confidence region for $(\beta_2, \dots, \beta_d)$ and for any $\gamma, \gamma' \in \Gamma$ we have*

$$R_{\bar{M}}(\varepsilon^\perp(\gamma)) = R_{\bar{M}}(\varepsilon^\perp(\gamma')), \quad (3)$$

then with probability $1-\alpha$ for any $\gamma \in \Gamma$ we have

$$R_{\bar{M}}(\varepsilon^\perp(\gamma)) = R_{\bar{M}}(\varepsilon).$$

In the above result as well as in Proposition 1 a set M is given a priori unconditionally on data. From a practical point of view it would be more natural to decide for M after some examination of the data when it would

become more clear which extreme ranks should be used to define a suitable reduced model. For example having a confidence region Γ for $(\beta_2, \dots, \beta_n)$ we could look for an M with some extreme ranks for which (3) holds. In such a case M depends on Γ and thus is obtained conditionally on some function ε . We can exploit Proposition 1 to provide arguments for this type of a conditional procedure. The result is stated below.

COROLLARY 1. *Let $M = M(\varepsilon) \subseteq \{1, \dots, n\}$ be a set function of errors such that the number of elements in any of its values is at least $m_0 > 1$. Then*

$$E^{\mathcal{F}} \left(\frac{E^{\mathcal{G}_M}(v \cdot \pi \varepsilon_M - v \cdot \pi \varepsilon_M^\perp)^2}{E^{\mathcal{G}_M}(v \cdot \pi \varepsilon_M)^2} \right)^{1/2} \leq \sqrt{\frac{d-1}{m_0-1}}.$$

Scaling Problem. First, let us explain the possible difficulties when scaling is ignored. For simplicity of notation consider the case $M = \{1, \dots, n\}$, generalization being obvious. For a contrast vector v , let

$$\gamma(\varepsilon) = E^e(v \cdot \pi \varepsilon - v \cdot \pi \varepsilon^\perp)^2 / E^e(v \cdot \pi \varepsilon)^2,$$

$$C_1 = \sqrt{E^{\mathcal{F}}(v \cdot \varepsilon)^2},$$

where in the latter we do not explicitly show dependence on the order statistics of ε . Denoting the distribution functions of $v \cdot \varepsilon | \mathcal{F}$, $v \cdot \pi \varepsilon^\perp | Y$ by F_1 , F_2 , respectively, we have the following relations.

PROPOSITION 3. *For every positive δ, r*

$$P(\gamma(\varepsilon)/\delta > r) \leq \frac{d-1}{n-1} (r\delta)^{-1}, \quad (4)$$

and for all $x \in \mathbb{R}$

$$F_1(x) \in [F_2(x - C_1 \delta) - \gamma(\varepsilon) \delta^{-2}, F_2(x + C_1 \delta) + \gamma(\varepsilon) \delta^{-2}]. \quad (5)$$

By (4) the random variable $\gamma(\varepsilon)$ is small with high probability. Thus the accuracy of our proposed approximation of F_1 by F_2 depends on $C_1 \delta$. For example, to use a quantile of F_2 for obtaining a confidence interval we want $C_1 \delta$ to be small relative to the quantile and thus to the length of the confidence interval. In the next paragraph we will explore this problem in more detail.

For each $p \in (0, 1)$ let K_p^1 , K_p^2 denote p -quantiles of F_1 , F_2 , respectively. By (4) for each δ_0 there exist a subset Ω_0 of underlying probability space

with $P(\Omega_0) > 1 - \delta_0$ and a constant $M > 0$ dependent neither on n nor on d such that on Ω_0 we have

$$\gamma(\varepsilon) \leq Md/n.$$

By (5) on Ω_0 for each $\delta > 0$ we have

$$F_1(x) \in \left[F_2\left(x - \sqrt{\frac{Md}{\delta}} \frac{C_1}{\sqrt{n}}\right) - \delta, F_2\left(x + \sqrt{\frac{Md}{\delta}} \frac{C_1}{\sqrt{n}}\right) + \delta \right]$$

for all $x \in \mathbb{R}$. This implies that for any p , $\delta > 0$ such that $p \pm \delta \in (0, 1)$ we have

$$K_{p-\delta}^2 - \sqrt{\frac{Md}{\delta}} \frac{C_1}{\sqrt{n}} \leq K_p^1 \leq K_{p+\delta}^2 + \sqrt{\frac{Md}{\delta}} \frac{C_1}{\sqrt{n}}.$$

Thus it would be desirable to have $C_1/\sqrt{n}$, at least asymptotically, small relative to K_p^1 . We can say equivalently that we are interested in the cases when quantiles of $\sqrt{n} W_n$ are convergent to infinity, where

$$W_n = \frac{(v \cdot \pi \varepsilon) | \varepsilon}{\sqrt{E^e(v \cdot \pi \varepsilon)^2}}. \quad (6)$$

It is worth to note here that the same asymptotic behavior of W_n is needed for the scaling problem based on Corollary 1.

In some cases our approximations will be justified regardless of the form of contrast vectors. This will happen if $s_{\varepsilon_n}^2/n$ is convergent in probability to zero, where $s_{\varepsilon_n}^2$ is the sample variance of ε_n . Indeed, by simple application of Proposition 1 we can obtain Theorem 1. In its formulation we do not indicate explicitly the dependence of F_1, F_2 on n and also it should be remembered that F_1, F_2 being conditional distributions are dependent on a random value of ε_n .

THEOREM 1. *If $s_{\varepsilon_n}^2/n$ converges in probability to zero then for every $\delta > 0$ and for sufficiently large n on a set of probability not less than $1 - \delta$ we have*

$$F_1(x) \in [F_2(x - \delta) - \delta, F_2(x + \delta) + \delta]$$

for all $x \in \mathbb{R}$.

Notice that the assumption required in this result is satisfied for errors in the domain of attractions of stable laws when $\alpha \in (1, 2]$.

More generally, the behavior of (6) can be very nicely described for errors sampled independently from a distribution belonging to the domain of attraction of an arbitrary stable law by means of a conditional invariance principle to which we now turn.

3. CONDITIONAL INVARIANCE PRINCIPLE FOR ERRORS ATTRACTED TO STABLE LAWS

Here we present only the motivation for a conditional invariance principle and describe the form of the conditional limit law. General results on this subject and their proofs can be found in LePage, Podgórski, and Ryznar (1994). The idea is to employ a simple construction via very specially chosen errors belonging to the domain of attraction of a stable law and then follow these special errors as $n \rightarrow \infty$.

Let $\delta = (\delta_i)_{i \in \mathbb{N}}$ be a sequence of independent random signs, i.e., $P(\delta_i = \pm 1) = \frac{1}{2}$, let $U = (U_i)_{i \in \mathbb{N}}$ be a sequence of independent random variables with uniform distribution on $[0, 1]$, and let $\Gamma = (\Gamma_i)_{i \in \mathbb{N}}$ be a sequence of arrival times of a Poisson process with intensity one. We will assume that δ , U , Γ are mutually independent. It is well known that if $G(x) = x^{-\alpha} \mathbf{1}_{[1, \infty)}(x)$, then the distribution of random vector

$$\varepsilon_n = (\varepsilon_{n,i})_{i=1}^n = \pi(\delta_i \Gamma_i^{-1/\alpha})_{i=1}^n / \Gamma_{n+1}^{-1/\alpha}, \quad (7)$$

where π is a random permutation independent of (δ, Γ) , is equal to the distribution of i.i.d. random variables from the domain of attraction of a symmetric stable law with index of stability α . In fact for $x > 1$ we have $P(|\varepsilon_{n,i}| > x) = x^{-\alpha}$ (see also LePage, 1980).

On a sequence $v_n = (v_{i,n})_{i=1}^n$ of contrast vectors we impose the form

$$v_{n,i} = v\left(\frac{i}{n}\right) - \frac{1}{n} \sum_{k=1}^n v\left(\frac{k}{n}\right), \quad (8)$$

where $v: [0, 1] \rightarrow \mathbb{R}$ is a continuous function such that $\int v = 0$.

To express the limit distribution of $W_n = [(v_n \cdot \pi \varepsilon_n) | \varepsilon_n] / \sqrt{E^{e_n}(v_n \cdot \pi \varepsilon_n)^2}$ we will need a notion and some basic properties of the stochastic integral of a function v with respect to a Levy motion. For details see for example Schilder (1970) or Kuelbs (1973). By a Levy motion we mean here a càdlàg version of the stochastic process $(A_t)_{t \in [0, 1]}$ continuous in probability with independent and homogeneous increments such that for any $t \in [0, 1]$ a distribution of A_t is symmetric stable with characteristic function $\phi_t(u) = \exp(-t |u|^\alpha)$. A random variable $\int v dA$ has a stable distribution with the characteristic function $\phi_v(t) = \exp(-|t|^\alpha \int |v|^\alpha)$. Let σ -field $\mathcal{J}$ be a σ -field generated by values of jumps of trajectories of this process. LePage (1980) has given the following series representation of a càdlàg Lévy motion

$$A_t = K_\alpha^{-1} \sum_{i=1}^{\infty} \delta_i \Gamma_i^{-1/\alpha} \mathbf{1}_{[U_i, 1]}(t) = K_\alpha^{-1} \sum_{i=1}^{\infty} \delta_i \Gamma_i^{-1/\alpha} \mathbf{1}_{[0, t]}(U_i), \quad (9)$$

where $K_\alpha = \int_0^\infty \alpha(1 - \cos r) r^{-1-\alpha} dr$. For such a version of a Lévy motion $\mathcal{J} = \sigma(\delta, \Gamma)$ and

$$Z = K_\alpha^{-2} \sum_{i=1}^{\infty} \Gamma_i^{-2/\alpha} \quad (10)$$

is the sum of the squared values of jumps of this process. This series is convergent with probability one (see LePage, 1980).

With the definitions given by (7)–(10) we have the following limit behavior of W_n .

THEOREM 2. *For each function with bounded variation $v: [0, 1] \rightarrow \mathbb{R}$ with probability one the conditional distributions*

$$W_n = \frac{v_n \cdot \pi \varepsilon_n}{\sqrt{E^{\varepsilon_n}(v_n \cdot \pi \varepsilon_n)^2}} \Big|_{\varepsilon_n}$$

converge weakly to the distribution

$$\frac{\int v dA}{\sqrt{Z \int v^2}} \Big|_{\mathcal{J}}.$$

Remark 3.1. Proposition 1 implies that the joint distributions $(v_n \cdot \pi \varepsilon_n^\perp, v_n \cdot \pi \varepsilon_n) | \varepsilon_n$ scaled by $\sqrt{E^{\varepsilon_n}(v_n \cdot \pi \varepsilon_n)^2}$ converge to $(\int v dA, \int v dA) | \mathcal{J}$ scaled by $\sqrt{Z \int v^2}$.

Remark 3.2. It can be proven that the results are true in a more general situation. Namely instead of the special form of errors (7) we can consider arbitrary independent errors from the domain of attraction of an arbitrary, not necessary symmetric, stable law. Since proofs of such generalizations as well as more complete discussion of the invariance principle can be found in LePage, Podgórski, and Ryznar (1994) we omit the proof of Theorem 2.

One of statistical consequences of the above result is illustrated on Fig. 2 on which a histogram of widths of 90% conditional confidence intervals of $\int v dA | \mathcal{J}$ is confronted with the width of 90% unconditional interval for $\int v dA$.

We observe that conditional confidence intervals are usually much narrower than one based on the unconditional sample distribution. This is an important advantage of any method aiming at recovery of conditional distributions against methods (such as reduced resample size bootstrap) intending to recover sampling unconditional distributions.

As a corollary to Theorem 2 we obtain the desired property for the quantiles of the distributions of $\sqrt{n} W_n$.

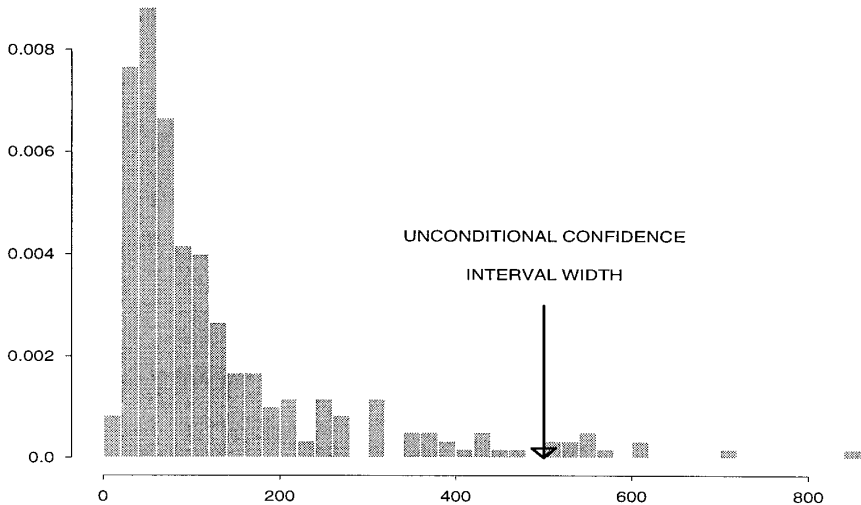


FIG. 2. Histogram of widths of confidence intervals based on conditional distribution versus the width of unconditional confidence interval.

COROLLARY 2. *For a nonzero function with bounded variation $v: [0, 1] \rightarrow \mathbb{R}$ the quantiles of $\sqrt{n} W_n$, different from the median, converge to plus or minus infinity with probability one.*

The special form of contrast vectors assumed above is important for obtaining a result of this type. We will see in the next example that in some extreme cases, i.e., for a very unusual choice of contrast vectors, the quantiles converge to zero. In such a case quantile of F_1 are not reliable approximations of corresponding ones for F_2 .

EXAMPLE 1. Assume that $\mathbf{v}_n = (v_1, \dots, v_k, 0, \dots, 0) \in \mathbb{R}^n$, where $k \leq n$ does not depend on n and $(v_1, \dots, v_k)$ is a fixed contrast vector. Further let $\boldsymbol{\varepsilon}_n = (\varepsilon_1, \dots, \varepsilon_n)$ where ε_i , $i \in \mathbb{N}$ are i.i.d. random variables such that $s_{\varepsilon_n}/\sqrt{n}$ diverges in probability to infinity. In such a case $\mathbf{v}_n \cdot \pi \boldsymbol{\varepsilon}_n | \boldsymbol{\varepsilon}_n$ converges in distribution to the distribution of $\mathbf{v}_k \cdot \boldsymbol{\varepsilon}_k$. But the assumption that $s_{\varepsilon_k}/\sqrt{n}$ diverges in probability to infinity implies that $\sqrt{n}/\sqrt{E^*(\mathbf{v}_n \cdot \pi \boldsymbol{\varepsilon}_n)^2}$ will be convergent to zero and so also the quantiles of $\sqrt{n} W_n$. See Section 4 for more rigorous arguments. As a particular example of a random variable satisfying the required condition one can consider errors with the distribution given by the coordinates of the vector in (7) for $\alpha \in (0, 1)$.

Theorem 1 and Theorem 2 validate approximating the unknown distribution of errors of contrasts by permuting residuals under special assumptions on distributions of errors. Failure of this approximation in Example 1 is due to the asymptotically pathological choice of contrast.

4. PROOFS

Special cases of the following lemma, when v is not random, are apparently well known: Chernoff (1973, Section 4.1); Freedman and Lane (1983, p. 296, Lemma 1). Although formulated here in probabilistic language it seems to illustrate rather a geometrical property of the group of permutations. We present the simple proof of this result.

LEMMA 1. *Let v be a random contrast vector. Assume that v , a random $n \times 1$ -vector ξ and a σ -field $\mathcal{G}$ all are independent of a random permutation π . Then*

$$E^{\mathcal{G}}(v \cdot \pi \xi)^2 = E^{\mathcal{G}}(\|v\|^2 s_{\xi}^2).$$

Proof. We have

$$\begin{aligned} E^{\mathcal{G}}(v \cdot \xi)^2 &= E^{\mathcal{G}} \sum_{\sigma \in \Sigma_n} E^{\sigma(v, \mathcal{G})}(v \cdot \sigma \xi)^2 / n! \\ &= E^{\mathcal{G}} \sum_{\sigma \in \Sigma_n} \left(\sum_{a=1}^n v_a^2 \xi_{\sigma(a)}^2 + \sum_{a \neq b} v_a v_b \xi_{\sigma(a)} \xi_{\sigma(b)} \right) / n! \\ &= E^{\mathcal{G}} \left(\left(\sum_{a=1}^n v_a^2 \right) \bar{\xi}^2 + \frac{1}{n(n-1)} \left(\sum_{a \neq b} v_a v_b \right) \left(\sum_{a \neq b} \xi_a \xi_b \right) \right) \\ &= E^{\mathcal{G}} \left\{ \|v\|^2 \bar{\xi}^2 + \frac{1}{n-1} (\|v\|^2 - (v \cdot \mathbf{1})^2) (\bar{\xi}^2 - n \bar{\xi}^2) \right\}, \end{aligned}$$

where Σ_n denotes the set of all permutations of coordinates of vectors in $\mathbb{R}^n$. Since $(v \cdot \mathbf{1}) = 0$ we obtain

$$E^{\mathcal{G}}(v \cdot \pi \xi)^2 = E^{\mathcal{G}} \left(\|v\|^2 \frac{n}{n-1} (\bar{\xi}^2 - \bar{\xi})^2 \right) = E^{\mathcal{G}}(\|v\|^2 s_{\xi}^2). \quad \blacksquare$$

Now we will prove that $v \cdot \pi \varepsilon_M^\perp | \mathcal{G}_M$ and $v \cdot \varepsilon_M | \mathcal{F}_M$ are centered at zero. First notice that for σ in Σ_m we have $(\varepsilon_M, R_{\bar{M}}) \stackrel{d}{=} (\sigma \varepsilon_M, R_{\bar{M}})$. To see that let consider an event $\{\varepsilon_M \in A, R_{\bar{M}} = r_{\bar{M}}\}$. Define a permutation $\tilde{\sigma} \in \Sigma_n$ such that it permutes ε_i for $r_i \in M$ the same way as σ and leaves ε_i unchanged for $r_i \notin M$. Then $R_{\bar{M}}(\tilde{\sigma} \varepsilon) = R_{\bar{M}}(\varepsilon)$ on a set $\{R_{\bar{M}}(\varepsilon) = r_{\bar{M}}\}$; thus by exchangeability of ε we have

$$\begin{aligned} P(\varepsilon_M(\varepsilon) \in A, R_{\bar{M}}(\varepsilon) = r_{\bar{M}}) &= P(\varepsilon_M(\tilde{\sigma} \varepsilon) \in A, R_{\bar{M}}(\tilde{\sigma} \varepsilon) = r_{\bar{M}}) \\ &= P(\sigma \varepsilon_M(\varepsilon) \in A, R_{\bar{M}}(\varepsilon) = r_{\bar{M}}). \end{aligned}$$

This implies that $v \cdot \varepsilon_M | \mathcal{F}_M \stackrel{d}{=} v \cdot \pi \varepsilon_M | \mathcal{G}_M$. Indeed, for each bounded and measurable function h and a random permutation π independent of ε we have

$$\begin{aligned} E^{\mathcal{F}_M} h(v \cdot \varepsilon_M) &= \frac{1}{m!} E^{\mathcal{F}_M} \sum_{\sigma \in \Sigma_m} h(v(R_{\bar{M}}) \cdot \sigma \varepsilon_m) = E^{\mathcal{F}_M} h(v(R_{\bar{M}}) \cdot \pi \varepsilon_M) \\ &= E^{\mathcal{G}_M} h(v \cdot \pi \varepsilon_M), \end{aligned}$$

where in the last equality we use that $\mathcal{F}_M$ is a sub- σ -field of $\mathcal{G}_M$ and the fact that the last conditional expectation is $\mathcal{F}_M$ -measurable. By independence of π and ε we have

$$\begin{aligned} E^{\mathcal{G}_M}(v \cdot \pi \varepsilon_m^\perp) &= E^{\mathcal{G}_M}(\pi v(R_{\bar{M}}) \cdot \varepsilon_M^\perp) = \frac{1}{m!} E^{\mathcal{G}_M} \sum_{\sigma \in \Sigma_m} (\sigma v(R_{\bar{M}}) \cdot \varepsilon_M^\perp) \\ &= E^{\mathcal{G}_M}(E^{R_{\bar{M}}}(\pi v(R_{\bar{M}})) \cdot \varepsilon_M^\perp) = 0, \end{aligned}$$

where in the last equality we use the fact that $E(\pi u) = \bar{u} \mathbf{1} = n^{-1}(u \cdot \mathbf{1}) \mathbf{1}$ for any $n \times 1$ vector u and thus $E(\pi u) = 0$ if u is a contrast vector. Similarly, we have

$$E^{\mathcal{F}_M}(v \cdot \pi \varepsilon_m) = E^{\mathcal{F}_M}(\pi v(R_{\bar{M}}) \cdot \varepsilon_M) = E^{\mathcal{F}_M}(E^{R_{\bar{M}}}(\pi v(R_{\bar{M}})) \cdot \varepsilon_M) = 0.$$

The previous lemma and some elementary properties of conditional expectation are the only tools in the proof of our main result.

Proof of Proposition 1. By Lemma 1 we have that

$$\begin{aligned} E^{\mathcal{G}_M}[(v \cdot \pi \varepsilon_M) - (v \cdot \pi \varepsilon_M^\perp)]^2 &= E^{\mathcal{G}_M}(v \cdot \pi(\hat{\varepsilon}_M))^2 \\ &= E^{\mathcal{G}_M}(\|v\|^2 \cdot s_{\varepsilon_M}^2) \\ &= s_{\varepsilon_M}^2 \|v\|^2, \end{aligned} \tag{11}$$

where the last equality follows from measurability of $s_{\varepsilon_M}^2$ and $\|v\|^2$ with respect to $\mathcal{G}_M$. Here and henceforth $\hat{\varepsilon}_M$ denotes the projection of ε_M to the column space of $\mathbf{X}_M$. By similar arguments we have also

$$E^{\mathcal{G}_M}(v \cdot \pi \varepsilon_M)^2 = s_{\varepsilon_M}^2 \|v\|^2. \tag{12}$$

Using the fact that $(\varepsilon_M, R_{\bar{M}}) \stackrel{d}{=} (\sigma \varepsilon_M, R_{\bar{M}})$ and since both $s_{\varepsilon_M}^2$ and $E^{\varepsilon_M} s_{\pi \varepsilon_M}^2$ are $\mathcal{F}_M$ measurable, we have

$$E^{\mathcal{F}} s_{\varepsilon_M}^2 = E^{\mathcal{F}} E^{\mathcal{F}_M} s_{\varepsilon_M}^2 = E^{\mathcal{F}} E^{\mathcal{F}_M} s_{\pi \varepsilon_M}^2 = E^{\varepsilon_M} s_{\pi \varepsilon_M}^2$$

and, thus,

$$E^{\mathcal{F}} \frac{E^{\mathcal{G}_M}[(v \cdot \pi \varepsilon_M) - (v \cdot \pi \varepsilon_M^\perp)]^2}{E^{\mathcal{G}_M}(v \cdot \pi \varepsilon_M)^2} = \frac{E^{\varepsilon_M} s_{\pi \varepsilon_M}^2}{s_{\varepsilon_M}^2}. \quad (13)$$

The proof can be completed using an equality $E^{\mathcal{F}} s_{\varepsilon}^2 = ((d-1)/(n-1)) s_{\varepsilon}^2$ obtained by Box and Watson (1962, Section 4.1) in their study of F -distribution approximations for $s_{\varepsilon}^2/s_{\varepsilon}^2$ and also mentioned by Freedman and Lane (1983); see Wu (1986, p. 267). In our case it may be obtained from Lemma 1 by expanding in an orthonormal basis $\{u_\alpha\}$ of $\mathbf{X}_M$ with $u_1 = \mathbf{1}/m$ as

$$\begin{aligned} E^{\varepsilon_M} s_{\pi \varepsilon_M}^2 &= \frac{1}{m-1} E^{\varepsilon_M} (\|\pi \hat{\varepsilon}_M\|^2 - (\pi \hat{\varepsilon}_M \cdot \mathbf{1})^2/m) \\ &= \frac{1}{m-1} E^{\varepsilon_M} \left(\sum_{a=1}^d (u_a \cdot \pi \varepsilon_M)^2 - (\varepsilon_M \cdot \mathbf{1})^2/m \right) \\ &= \frac{1}{m-1} \sum_{a=2}^d E^{\varepsilon_M} (u_a \cdot \pi \varepsilon_M)^2 \\ &= \frac{1}{m-1} \sum_{a=2}^d E^{\varepsilon_M} (\|u_a\|^2) s_{\varepsilon_M}^2 = \frac{d-1}{m-1} s_{\varepsilon_M}^2, \end{aligned} \quad (14)$$

where Lemma 1 was used in the fourth equality.

To obtain the extension to $(v_i \cdot \varepsilon | \mathcal{F})_{i=1}^r$ for many contrast vectors it is enough to notice that by (11) and (12) we have

$$\frac{\sum_{k=1}^r E^{\mathcal{G}_M}[(v_k \cdot \pi \varepsilon_M) - (v_k \cdot \pi \varepsilon_M^\perp)]^2}{\sum_{k=1}^r E^{\mathcal{G}_M}(v_k \cdot \pi \varepsilon_M)^2} = \frac{s_{\varepsilon_M}^2 \sum_{k=1}^r E^{\mathcal{G}_M} \|v_k\|^2}{s_{\varepsilon_M}^2 \sum_{k=1}^r E^{\mathcal{G}_M} \|v_k\|^2} = \frac{s_{\varepsilon_M}^2}{s_{\varepsilon_M}^2}.$$

Now conclusion follows immediately from (14). $\blacksquare$

Likewise we can obtain the result on the relative second moments of the differences $\pi \varepsilon_M - \pi \varepsilon_M^\perp$. For simplicity of notation let us formulate this for the case $M = \emptyset$ and let $\mathbf{P}$ denotes the projection to the column space of $\mathbf{X} \ominus \mathbf{1}$. We have

$$E^{\mathcal{F}} \frac{E^{\varepsilon} \|\mathbf{P}(\pi \varepsilon - \pi \varepsilon^\perp)\|^2}{E^{\varepsilon} \|\mathbf{P}(\pi \varepsilon)\|^2} = \frac{d-1}{n-1}.$$

Indeed, for any vector u we have by the definition of s_u^2 that $s_u^2 = s_{\mathbf{P}u}^2$ and $\|u\|^2 = (u \cdot \mathbf{1})^2/n + (n-1)s_u^2$. Thus,

$$E^{\mathcal{F}} \|\mathbf{P}(\pi \varepsilon - \pi \varepsilon^\perp)\|^2 = (n-1) E^{\mathcal{F}} s_{\pi \varepsilon}^2 = (n-1) E^{\mathcal{F}} s_{\varepsilon}^2$$

and

$$E^{\mathcal{F}} \|\pi \mathbf{P} \varepsilon\|^2 = (n-1) E^{\mathcal{F}} s_{\varepsilon}^2.$$

Proof of Proposition 2. First note that ε and $(\varepsilon - \bar{\varepsilon}\mathbf{1})$ have the same ranks. Then with probability at least $1 - \alpha$ the true parameter $\gamma_0 = (\beta_2, \dots, \beta_d)$ is in Γ . Since $\varepsilon^\perp(\gamma_0) = Y - \bar{Y}\mathbf{1} - \mathbf{X} \ominus \mathbf{1}\gamma_0 = \varepsilon - \bar{\varepsilon}\mathbf{1}$ thus $R_{\bar{M}}(\varepsilon) = R_{\bar{M}}(\varepsilon^\perp(\gamma_0))$. If $\gamma_0 \in \Gamma$, then by the assumption (3) for any $\gamma \in \Gamma$ we have

$$R_{\bar{M}}(\varepsilon) = R_{\bar{M}}(\varepsilon^\perp(\gamma))$$

which concludes the proof. ■

Proof of Corollary 1. The result follows from the Schwartz inequality (the first inequality below), Proposition 1 (the second equality below) and convexity of the square root function (the second inequality) as shown below

$$\begin{aligned} E^{\mathcal{F}} \left(\frac{E^{\mathcal{G}_M}(v \cdot \pi \varepsilon_M - v \cdot \pi \varepsilon_M^\perp)^2}{E^{\mathcal{G}_M}(v \cdot \pi \varepsilon_M)^2} \right)^{1/2} \\ &= \sum_K E^{\mathcal{F}} \mathbf{1}_{\{M=K\}} \left(\frac{E^{\mathcal{G}_K}(v \cdot \pi \varepsilon_K - v \cdot \pi \varepsilon_K^\perp)^2}{E^{\mathcal{G}_K}(v \cdot \pi \varepsilon_K)^2} \right)^{1/2} \\ &\leq \sum_K (E^{\mathcal{F}} \mathbf{1}_{\{M=K\}})^{1/2} \left(E^{\mathcal{F}} \frac{E^{\mathcal{G}_K}(v \cdot \pi \varepsilon_K - v \cdot \pi \varepsilon_K^\perp)^2}{E^{\mathcal{G}_K}(v \cdot \pi \varepsilon_K)^2} \right)^{1/2} \\ &= \sum_K (E^{\mathcal{F}} \mathbf{1}_{\{M=K\}})^{1/2} \sqrt{(d-1)/(\text{card}(K) - 1)} \\ &\leq \left(\sum_K E^{\mathcal{F}} \mathbf{1}_{\{M=K\}} \right)^{1/2} \sqrt{(d-1)(m_0 - 1)} \\ &= \sqrt{(d-1)/(m_0 - 1)}. \quad \blacksquare \end{aligned}$$

Proof of Proposition 3. For any random variables Z_1, Z_2 with marginal distribution functions F_1, F_2 and for each $\delta > 0$, $C > 0$ we have that

$$F_2(x - C\delta) - P\left(\frac{Z_2 - Z_1}{C} \geq \delta\right) \leq F_1(x) \leq F_2(x + C\delta) + P\left(\frac{Z_2 - Z_1}{C} \geq \delta\right),$$

which applied to $Z_1 = v \cdot \pi \varepsilon$, $Z_2 = v \cdot \pi \varepsilon^\perp$, $P = P^\varepsilon$, and $C = C_1$ (notice that $Z_1|\varepsilon = v \cdot \varepsilon|\mathcal{F}$, $Z_2|\varepsilon = v \cdot \pi \varepsilon^\perp|Y$ and $C_1 = \sqrt{E^\varepsilon Z_1^2}$), together with the Markov inequality

$$P^\varepsilon\left(\frac{|Z_2 - Z_1|}{C_1} \geq \delta\right) \leq \frac{E^\varepsilon |Z_2 - Z_1|^2}{C_1^2 \delta^{-2}}$$

gives us the second part of Proposition 3.

From the Markov inequality applied to nonnegative $\gamma(\varepsilon)$ we get

$$P(\gamma(\varepsilon) > \delta) \leq E\gamma(\varepsilon)/\delta.$$

Thus the first part of Proposition 3 follows from Proposition 1 since it implies that $E\gamma(\varepsilon) = (d-1)/(n-1)$. ■

Proof of Theorem 1. By Proposition 1 and Lemma 1 we have

$$E^{\mathcal{F}}(v \cdot \pi \varepsilon^\perp - v \cdot \pi \varepsilon)^2 = (d-1) \|v\|^2 \frac{s_\varepsilon^2}{n-1}.$$

Thus by our assumption $Z_n = E^{\mathcal{F}}(v \cdot \pi \varepsilon^\perp - v \cdot \pi \varepsilon)^2$ converges to zero in probability and thus for a given $\delta > 0$ and for sufficient large n we have $P(Z_n > \delta) < \delta$. Now using Markov's inequality in the same manner as in the proof of Proposition 3 we immediately obtain the result. ■

Proof of Corollary 2. By Theorem 2 it is enough to show that the limiting distribution of W_n does not have an atom at zero with probability one. But

$$P\left(\sum_{i=1}^{\infty} v(U_i) \delta_i \Gamma_i^{-1/\alpha} = 0\right) = E\left[P\left(\sum_{i=1}^{\infty} v(U_i) \delta_i \Gamma_i^{-1/\alpha} = 0 \mid (\Gamma, \delta)\right)\right]$$

and $\sum_{i=1}^{\infty} \Gamma_i^{-1/\alpha} \delta_i v(U_i) = \int v dA$ has an α -stable distribution which is continuous. Thus with probability one $P(\sum_{i=1}^{\infty} v(U_i) \delta_i \Gamma_i^{-1/\alpha} = 0 \mid (\Gamma, \delta)) = 0$. ■

Now let us discuss the statements included in Example 1 in Section 2. For a fixed ω belonging to the underlying probability space by $\varepsilon_n^*(\omega) = (\varepsilon_{n,i}^*(\omega))_{i=1}^n$ we will denote a random variable obtained by resampling with replacement from $(\varepsilon_i(\omega))_{i=1}^n$. By the Law of Large Numbers for ω from a set of probability one for $l=1, \dots, k$ we have

$$\lim_{n \rightarrow \infty} E e^{it v_l \cdot \varepsilon_{n,l}^*(\omega)} = \lim_{n \rightarrow \infty} \frac{1}{n} \sum_{j=1}^n e^{it v_l \varepsilon_j(\omega)} = E e^{it v_l \varepsilon_1}.$$

Assume that for our fixed ω the above holds. Denote $\varepsilon_n(\omega) = (\varepsilon_i(\omega))_{i=1}^n$. Then

$$\begin{aligned} & |E(e^{it \mathbf{v}_n \cdot \pi \varepsilon_n(\omega)} - e^{it \mathbf{v}_k \cdot \varepsilon_k})| \\ & \leq |E(e^{it \mathbf{v}_n \cdot \pi \varepsilon_n(\omega)} - e^{it \mathbf{v}_n \cdot \varepsilon_n^*(\omega)})| + |E(e^{it \mathbf{v}_n \cdot \varepsilon_n^*(\omega)} - e^{it \mathbf{v}_k \cdot \varepsilon_k})| \\ & \leq \left(1 - \frac{n!}{(n-k)! n^k}\right) + \left|\prod_{l=1}^k (E e^{it v_l \varepsilon_{n,l}^*(\omega)} - E e^{it v_l \varepsilon_l})\right|, \end{aligned}$$

which implies that $\mathbf{v}_n \cdot \pi \varepsilon_n | \varepsilon_n$ converge weakly to $\mathbf{v}_k \cdot \varepsilon_k$ with probability one.

5. EXAMPLES

The examples in this section have been selected to demonstrate some basic features of the proposed method of resampling. The first one is very simple and somewhat artificial but perfectly illustrates the differences between our approach and other methods, as well as the advantages of using a reduced model.

EXAMPLE 1. Consider an $n \times 2$ design $\mathbf{X}$ consisting of two columns $X_1 = \mathbf{1}$ and X_2 having the first half of coordinates equal to one and the second one equal to minus one (we assume that n is even). Clearly X_2 is

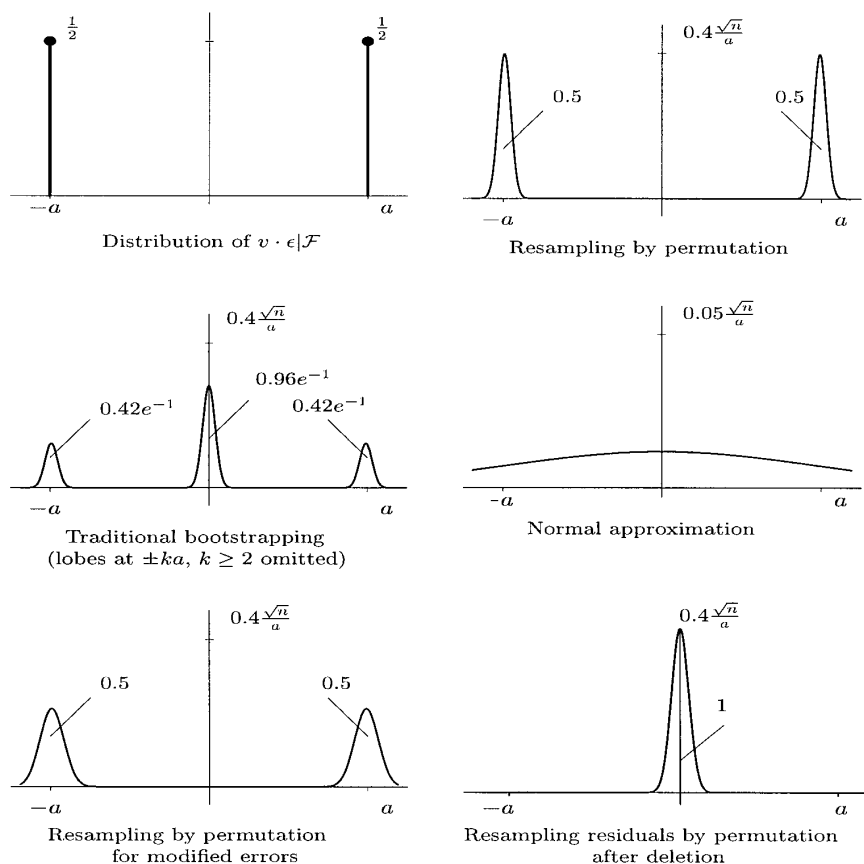
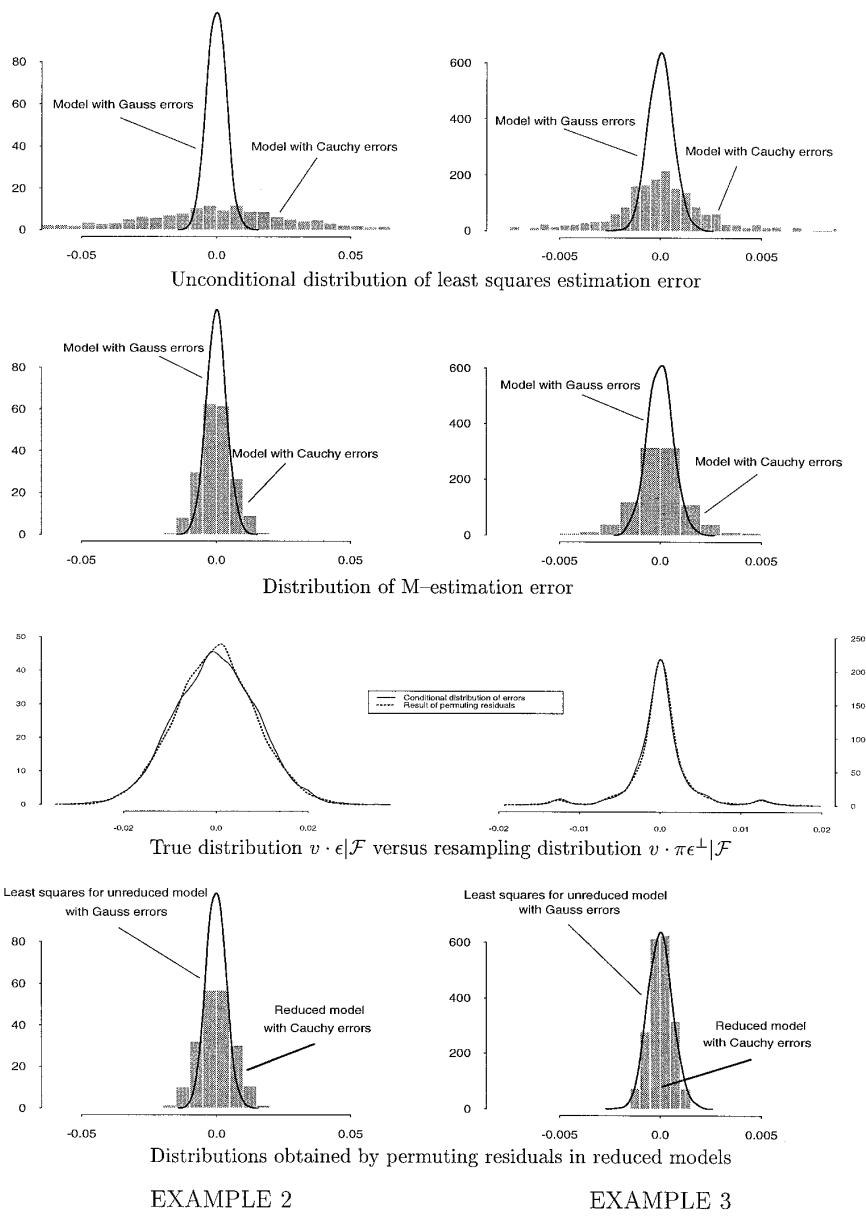


FIG. 3. Resampling permutations versus traditional bootstrap and normal approximation for the model described in Example 1.

a contrast vector and thus orthogonal to X_1 . The vector of errors ε considered here has one value dominating the others and typifies what can happen when the distribution of errors has a long tail. Namely, suppose that our actual errors $\varepsilon = (a, 0, \dots, 0)$ and thus $\varepsilon^\perp = (a - 2a/n, 2a/n, \dots, 2a/n, 0, \dots, 0)$ where the first zero is at the $n/2 + 1$ position. Let us take for our contrast vector v the second column X_2 of the matrix X . One can easily observe that the distribution $v \cdot \varepsilon | \mathcal{F}$ is concentrated on two points a and $-a$ with the equal weights $\frac{1}{2}$. We compare three different methods of estimating this distribution: resampling by permutations of residuals, the traditional bootstrap with replacement, and the normal approximation. On the first two pictures of Fig. 3 we present the distribution of interest versus first-order approximations for $n=625$ of the distribution obtained by resampling residuals. The next two pictures present first-order approximations obtained by applying the traditional bootstrap with replacement and the normal approximation.

If we consider the reduced model (2) for $M = \{n\}$ and if the residual $a - 2a/n$ is large compared to residuals $-2a/n$ we will delete the first row from our model and then we will obtain $\varepsilon_M = \varepsilon_M^\perp = 0$ and thus resampling of residuals gives us a perfect estimate of the conditional distribution of errors. To make this a little bit more realistic let us modify errors by adding to the previous error a sample vector from a Gaussian white noise with a small deviation relative to the value of a . In our numerical simulations we will take it equal to $a/\sqrt{n}$. Let us take as before $M = \{n\}$. The last two pictures of Fig. 3 present recovery of the conditional distributions of error of the contrast in the reduced and unreduced models. This illustrates the advantage of the reduced model. Although approximations in both cases are accurate, in the case of the unreduced model the confidence region consists of two separated intervals around $v \cdot y - a$ and $v \cdot y + a$ while in the reduced model one can consider the more natural confidence interval around $v \cdot y$.

In more natural models in the next examples errors are simulated from the Cauchy distribution. For two different designs we compare performance of the inference based on resampling by permutations and reduced models against M -estimation. Summaries of the numerical studies for both examples are presented in Fig. 4. In these examples we implement one of the possible practical methods of inference based on the results included in this work. All computations were made either in S-Plus[®] statistical package or with a use of our own procedures written in C-language. The latter were used mainly to realize fast procedures for resampling methods. For computing M -estimates we have used the S-Plus[®] function based on an iteratively reweighted least squares fit with Huber's weight function and with a bisquare weight function used for the last two iterations.



EXAMPLE 2

EXAMPLE 3

FIG. 4. Summary of numerical studies of models from Example 2 and Example 3.

EXAMPLE 2. We consider a $2 \times n$ design matrix $\mathbf{X}$ with $n = 101$ having first column $X_1 = \mathbf{1}$ and second column $X_2 = (-50, -49, \dots, 0, \dots, 49, 50)$. We set the actual vector of parameters $\beta = (0, 1)$. For the errors we assume that they are independent and distributed according to the Cauchy law. The following aspects of estimation of slope are considered:

- (a) performance of the least squares estimator and the M -estimator,
- (b) accuracy of recovery of the conditional distribution $v \cdot \varepsilon | \mathcal{F}$ by $v \cdot \pi \varepsilon^\perp | Y$, where v is a contrast corresponding to slope,
- (c) introduction of a statistical procedure based on our theoretical result leading to effective reduced models.

Ad. (a). To examine distributions of the least squares estimator and the M -estimator we have employed Monte Carlo method. By generating independently 1000 vectors of errors we have obtained histograms of 1000 estimation errors for both estimators. These histograms are presented on Fig. 4. To visualize the impact of long tails we have compared these histograms with density estimates of the distribution of estimation errors obtained for “short” tails, i.e., when errors in the model were generated from Gauss distribution. As should be expected the M -estimator convincingly beats the least squares one for Cauchy errors, although performance of both is the same in the Gaussian case.

Ad. (b). To compare $v \cdot \varepsilon | \mathcal{F}$ with $v \cdot \pi \varepsilon^\perp | Y$ we have computed density estimates of both distribution for one sample of Cauchy errors. These estimates were based on 4000 random permutations generated independently and applied to errors and residuals, respectively. The graphs of the densities are presented in Fig. 4. We observe great agreement of these two conditional distributions which confirms our theoretical results.

Ad. (c). We propose here an effective statistical procedure which uses our conditional approach first to determine some extreme ranks of errors and then based on this information to apply resampling by permutations to the reduced model. Consider $n + 1$ regression models consisting of the original one and n models based on “one-at-the-time” deletions of rows. By permuting residuals in all this $n + 1$ models we can recover the conditional distribution of estimation errors and thus construct $n + 1$ confidence intervals on $1 - \alpha$ level. The advantage of the conditional approach is that given fixed errors in the model there is no random relation between the obtained confidence intervals and thus a criterion of a choice of one of them does not involve any additional conditioning. In a sense the obtained confidence intervals are different aspects of the same fixed random event. In our computer studies we have chosen the confidence interval with the smallest width to determine ranks of the errors in the way described in Proposition 2. Then we have examined which group of the ranks of extreme residuals

remains unchanged when the parameters move through the chosen confidence interval. Since the vector of residuals is linear as a function of the parameter and the confidence regions are usually convex sets it is enough to study extreme residuals at vertices of these regions. After determining the set of ranks of extreme residuals which remained the same at the endpoints of the confidence interval we removed the corresponding rows from the model and the reduced model was subjected to resampling by permutations once again. The results of our resampling are presented on Fig. 4. Comparing the obtained histogram with the histogram corresponding to the M -estimator we do not see any significant difference between performance of the estimators. Although the conditional distribution presented here was based on a single run of errors very similar behavior was observed whenever generation of errors was repeated.

The next example shows that our method can actually perform better than M -estimator for reasonable selections of a design matrix.

EXAMPLE 3. Now a $2 \times n$ design matrix $\mathbf{X}$ with $n=43$ is defined by $X_1 = \mathbf{1}$ and $X_2 = (-2^{10}, \dots, -2^{-10}, 0, 2^{-10}, \dots, 2^{10})$. As in the previous example $\beta = (0, 1)$. We carried on the similar analysis as before. Its results are graphically summarized in Fig. 4.

ad. (a). Again the performance of the M -estimator is better than the least squares estimator although the difference is not as drastic as before.

ad. (b). The conditional distributions $v \cdot \varepsilon | \mathcal{F}$ and $v \cdot \pi \varepsilon^\perp | Y$ are nearly identical.

ad. (c). This time we can observe that the distribution obtained by resampling permutations in the reduced model is not only tighter than the sample distribution for M -estimate but also as good as for the model with Gauss errors.

ACKNOWLEDGMENTS

The authors greatly appreciate comments of referees which helped more appropriately accentuate the results of this work.

REFERENCES

1. ARCONES, M. A., AND GINÉ, E. (1989). The bootstrap of the mean with arbitrary bootstrap sample size. *Ann. Inst. H. Poincaré* **25** 456–481.
2. ATHREYA, K. B. (1987). Bootstrap of the mean in the infinite variance case. *Ann. Statist.* **15** 724–731.

3. BICKEL, P. J., AND FREEDMAN, D. A. (1981). Some asymptotic theory for the bootstrap. *Ann. Statist.* **9** 1196–1217.
4. BICKEL, P. J., AND FREEDMAN, D. A. (1984). Asymptotic normality and the bootstrap in stratified sampling. *Ann. Statist.* **12** 470–482.
5. BOX, C. E. P., AND WATSON, G. S. (1962). Robustness to non-normality of regression tests. *Biometrika* **49** 93–106.
6. CHERNOFF, H. (1973). Using faces to represent points in k -dimensional space. *J. Am. Statist. Assoc.* **68** 361–368.
7. EFRON, B. (1979). Bootstrap methods: Another look at the jackknife. *Ann. Statist.* **7** 1–26.
8. FREEDMAN, D. A. (1981). Bootstrapping regression models. *Ann. Statist.* **9** 1218–1228.
9. FREEDMAN, D. A., AND LANE, D. (1983). A nonstochastic interpretation of reported significance levels. *J. Bus. Econ. Statist.* **1** 292–298.
10. HALL, P., AND LEPage, R. (1995). On bootstrap estimation of the distribution of the studentized mean. *Ann. Inst. Statist. Math.*, to appear.
11. KINATEDER, J. (1990). *An Invariance Principle Applicable to the Bootstrap*. Ph.D. thesis, Michigan State University; in *Exploring the Limits of Bootstrap*. (R. LePage and L. Billard, Eds.), Wiley, New York.
12. KUELBS, J. (1973). A representation theorem for symmetric stable process and stable measures. *Z. Wahrsch. Verw. Gebiete* **26** 259–271.
13. LAHIRI, S. N. (1992). Bootstrapping M -estimators of a multiple linear regression parameter. *Ann. Statist.* **20** 1548–1593.
14. LEPage, R. (1980). *Multidimensional Infinitely Divisible Variables and Processes. Part I. Stable Case*. Technical Report 292, Statistics Department, Stanford; in *Lecture Notes in Math.* Vol. 1391, Springer-Verlag, New York, 1989.
15. LEPage, R. (1990). Bootstrapping signs In *Exploring the Limits of Bootstrap* (R. LePage and L. Billard, Eds.), New York.
16. LEPage, R., AND PODGÓRSKI, K. (1994). Giving the boot, block and shuffle to statistics. *Sci. Comput. Automat.* **10** 29–34.
17. LEPage, R., PODGÓRSKI, K., AND RYZNAR, M. (1994). *Strong and Conditional Invariance Principles for Samples Attracted to Stable Laws*. Technical Report No. 425, Center for Stochastic Processes, University of North Carolina; *Probab. Theory Related Field*, submitted.
18. LEPage, R., WOODROOFE, M., AND ZINN, J. (1981). Convergence to a stable distribution via order statistics. *Ann. Probab.* **9** 624–632.
19. ROUSSEEUW, P. J., AND LEROY, A. M. (1987). *Robust Regression and Outlier Detection*. Wiley, New York.
20. SCHILDER, M. (1970). Some structure theorems for the symmetric stable laws. *Ann. Math. Statist.* **41** 522–536.
21. WU, C. F. J. (1986). Jackknife, bootstrap and other resampling methods in regression analysis. *Ann. Statist.* **14** 1261–1295.