CONTRASTS UNDER LONG-RANGE CORRELATIONS, OR: WHY STATISTICS SOMETIMES WORKS

by

H. Künsch, ETH Zürich

J. Beran, Texas A&M University

F. Hampel, ETH Zürich

Abstract:

The background of the paper is the empirical fact that the fundamental assumption of *independence* in statistics is generally *false* and has to be replaced by long-range dependence. Since this is far from generally known, some evidence of it is reviewed. It can easily be seen that long-range dependence has *disastrous* effects on the naive statistical treatment of *absolute constants*. The natural question arising from this is how the ubiquitous long-range dependence affects the statistics for *contrasts* . The main answer given in this paper is twofold.

(i) If the experimental conditions are well mixed as provided by randomization, the *levels* of tests and confidence intervals derived under the independence assumption are still *correct* , asymptotically and usually in good approximation for finite samples.

(ii) Even under randomly mixed designs there are typically large unnoticed *power* and *efficiency losses* due to the long-range dependence. They can be greatly reduced without estimating the correlations by a simple *blocking* device.

These results can be viewed as a late theoretical justification for the relative successes of statistical methodologies like regression and analysis of variance as opposed to the classical standard error for absolute constants so often dismissed by further experiments. Furthermore our results corroborate the good applied practice of randomization and blocking.

Key Words and Phrases: long-range dependence, ordinary and generalized least squares, regression with correlated errors, robustness of validity, robustness of efficiency, randomization, blocking.

1. Introduction:

This paper grew out of an attempt to understand why some parts of statistical applications are successful (according to folklore in applications), despite the virtually ubiquitous occurence of long-range correlations (see below) which are not considered in the analysis. The answer, put briefly, is that the principal effects of long-range correlations, so disastrous for absolute constants, cancel out for contrasts if the treatments are applied in random order. These results provide a late justification for a major part of regression and analysis of variance methodology, in which the data normally are treated as if they were independent. (Intercepts and grand means are another matter.) On the other hand, it turned out that with the usual methods for linear models there are efficiency and power losses due to long-range correlations which can be considerable. These losses can be partly avoided not only by cumbersome estimation and explicit inclusion of the long-range correlations into the inference process, but also simply by the use of randomized block designs (again corroborating a good applied practice from another theoretical angle).

Put differently, the levels of tests etc. for effects and regression slopes are still approximately correct if (as usual) long-range correlated data are treated as independent, causing statisticians to make correct positive claims on effects with the controlled error probability of the first kind; but there are – partly considerable – unnoticed power and efficiency losses, causing statisticians to discover too few true effects.

This situation is markedly different from that for absolute constants (grand means, intercepts) where the effects of long-range correlations are disastrous already for the level even in moderately small samples. This is probably a major reason for the disdain many physicists have towards statistics for their constants (cf. Jeffreys, (1939) 1961, p.301; p.270f), and it turns out to be a remarkably wise custom not to include the grand mean into the analysis of variance table.

For simplicity we restrict ourselves to serial correlations assuming time ordered observations, although more general structures might also be of interest. The standard models for serial dependence are ARMA-processes. They are well suited for describing an arbitrary behavior of the small lag correlations, but asymptotically their correlations decay exponentially fast and their spectrum is bounded and continuous. By standard results this implies that the variance of the arithmetic mean eventually decays like n^{-1}. Such properties are in striking contrast to the empirical behavior of a large number of observed time series from all subject areas including high quality measurement series. There the spectrum seems to be unbounded at the origin and the variance of the arithmetic mean decays like $n^{-\alpha}$ with some $\alpha < 1$ even for quite large $n's$. Stationary processes whose correlations decay so slowly that their sum is infinite can model such a behavior. This is what we mean by long-range correlations. Since not all readers may be familiar with this topic and some may doubt its practical importance, we give in Section 2 a survey where we collect empirical evidence for long-range correlation. In view of this evidence it seems crucial to us to work with long-range correlated errors although this imposes some mathematical difficulties, for instance Rosenblatt's strong mixing property does not hold.

There is already a considerable amount of mathematical results on design and analysis of experiments with serially correlated error structures. The special role played by contrasts has been realized and attention was mostly restricted to these. The difference between these papers and our work is two-fold. First the correlation structures in the earlier work allowed only short term correlations (e.g. Williams 1952, Jenkins and Chanmugan 1962, Berenblut and Webb 1974, Kiefer and Wynn 1981 and 1984, Cheng 1983, Martin 1986, Morgan and Chakravarti 1988). Second the goal of many of these papers is to construct in some sense optimal designs for prespecified correlation structures. By contrast we consider a broad and realistic class of correlation structures which are close to a nonstationary behavior and we investigate what actually happens to common statistics, tests and confidence intervals with designs which are in good approximation frequently used.

The problems considered here belong to the area of regression with correlated errors. In the case of short-range correlations some of our results can be derived from the classical theorems of Grenander and Rosenblatt (1957, Section 7.3). Still we have not seen these results stated explicitly in the literature despite their importance for statistical practice. Moreover for long-range dependence the theorems of Grenander and Rosenblatt do no longer apply since the spectrum has a singularity at zero. So far no general theory for regression with long-range correlated errors has been developed. Yajima (1988, 1989) covers the case of deterministic regressors, but it is an open problem to verify his conditions in our case.

We use here the special structure of the regressors to construct a martingale and then apply the martingale convergence theorem.

Our results can be considered as another justification of randomization and blocking. We show that randomization chooses with high probability a design for which the usual i.i.d. inference about contrasts is approximately correct. Hence randomization is also important for an analysis given the design simply because it leads to a good design. Moreover blocking reduces the variance even though in our model different blocks are not independent and the block effect is not constant within a block.

Our main results are asymptotic, but in a situation where dependence does not become asymptotically negligible. Moreover we give also results on the speed of convergence and on the bias of the estimated variance. It turns out that for any finite sample size the effect of correlated errors on the variance of a contrast is compensated by the bias of the estimated variance. This is a further explanation why usual inference under i.i.d. assumptions works well for contrasts. Finally in Section 7 we give some simulation studies which complement our theoretical results.

2. Evidence for long-range correlations:

Long-range correlations are serial correlations between observations which may be small individually, but whose cumulative effect is such that even the distant past becomes important for the distant future. As mentioned in the Introduction, there are three nearly equivalent mathematical formalizations of this vague description: the power spectrum has a pole at zero, the serial correlations decay so slowly (e.g., hyperbolically) that their sum becomes infinite, and the variance of the mean of n consecutive observations does not go to zero at rate $1/n$, but slower. Because of the strong dependence typical sample paths look nonstationary to the untrained eye and invite to search for trends and "hidden periodicities" which in this case are only spurious (cf. the simulated paths in Mandelbrot & Wallis 1969a, Graf et al. 1984, Hampel et al. 1986, p.392, or Hampel 1987, where still more literature is cited, among others).

Although long-range correlations look rather pathological at first glance and have only recently attracted the attention of mathematicians, eminent applied statisticians have been aware of their common existence in real data at least qualitatively for almost 100 years. The astronomer Newcomb (1895, p.103 ff) gave a rather detailed description of their qualitative genesis by overlapping errors sources of strongly varying duration in astronomical data (yielding intermediates between "pure" independent random errors and constant systematic errors, which he called "semi-systematic errors"). Karl Pearson (1902) simulated long series of astronomical observations experimentally, with various striking results in this direction. The chemist Gosset ("Student" 1927) gave several series of chemical measurements, all exhibiting what he termed "semi-constant errors". He remarked that they were typical for all chemical measurements he knew, and provided also a thorough analysis of the composition of the error of a chemical measurement. The geophysicist Jeffreys ((1939)1961, Ch. 5.6, 5.7, cf. esp. p.298) reanalyzed K. Pearson's series, among others, and concluded that what he termed "internal correlation" was the normal thing in real data, and that the assumption of independence and the $\sigma/\sqrt{n}$-rule for the standard error should never be used unless they had been checked under special circumstances which made them somewhat more plausible.

Further evidence against independence in "ordinary" data sets, though not so directly and specifically for long-range dependence, can be found in Berkson (1938), Daniel and Wood (1971), and in chapter headings of Box et al. (1978) ("The Declaration of Independence") and of Mosteller and Tukey (1977) ("Hunting out the Real Uncertainty. How $\sigma/\sqrt{n}$ can mislead.").

While the early researchers had no usable mathematical model to describe long-range correlations, the first such model became the increment processes of (usually Gaussian) self-similar processes invented in an abstract setting by Kolmogorov (1940). For these processes, the variance of the mean of n consecutive observations decreases exactly like $n^{-\alpha}$ for α anywhere between 0 and 1, a law that had been found purely empirically by Smith (1938) in his analysis of 40 uniformity trials in agriculture. After a specific modelling application to textile engineering data by D.R.Cox (1954; Cox and Ingham 1950), it was B. B. Mandelbrot who applied these processes and made them popular with data from geophysics (and a few other fields like economics and linguistics). Some references are Mandelbrot and Wallis (1968), Mandelbrot and van Ness (1968), Mandelbrot and Taqqu (1979), further references are in Hampel et al. (1986) and in Hampel (1987). Mandelbrot and Wallis (1969b) contains a list of about 70 geophysical data sets such as river flows in hydrology, rainfall data, earthquake frequencies, tree-ring indices, thickness of geological layers, and also sunspot numbers. The exponent α in $n^{-\alpha}$ ranges fairly evenly distributed over the whole interval between 0 and 1 and is in most cases clearly below 1.

The starting point for Mandelbrot were the floods of the river Nile which were known for about 800 years and were analyzed by H.E.Hurst before the construction of the Aswan dam. (Mandelbrot named the parameter $H = 1 - \alpha/2$ in honor of Hurst.) Mandelbrot was the first to describe the data satisfactorily by a stationary model using self-similar processes, while all attempts to use short-term correlated processes (like ARMA-processes) by a number of top probabilists failed in that they could only describe the 800 years as transient behaviour of a globally very different process with very many parameters.

While by now many (though not all) researchers working with geophysical data sets are convinced of the importance and necessity to use self-similar or other long-range correlated processes as models, it is only the long-range and not the correlation as such which is surprising for these data. But much more surprising – and still too little known – is the evidence that practically all pure measurement series in physics, chemistry, and astronomy which should be prime examples for "independent identically distributed observations" are not only correlated, but even long-term correlated.

For one thing, after the development of asymptotically (nearly) efficient estimation and testing procedures for the parameter H (or α) of long-range correlation in the model of (near-)Gaussian self-similar processes by Graf (1983; Graf et al. 1984), some of the old evidence for pure measurement series was quantitatively corroborated. Thus, for a "typical" series of 135 chemical measurements by Student (1927), the long-range correlation, tested against independence, was highly significant, and the actual variance of the mean was about twenty times the variance derived under the independence assumption (Graf et al. 1984)! (With a longer record of these data, the factor would be still larger.)

It was also tried to find new series of high-quality measurements of the same constant which were long enough to test for long-range correlations. Such series are extremely rare; either they are done (like Pearson's and Student's series) to check upon the random structure of the measuring process itself, or sometimes they are done for a highly important constant.

Data series of the former kind play the same role for "typical" physical, chemical etc. measurements as do uniformity trials for agricultural experiments. They allow to check the usual assumptions when it is known that no effect is present. It is much harder and often impossible to check for long-range correlations in actual regression or analysis of variance data sets: even if the overall sample size is large and the time sequence of the data known, there are usually too many parameters being estimated, and these estimates are partly confounded with the irregular fluctuations caused by long-range dependence. If the levels occur only over short stretches of time, their estimates will be almost completely confounded with the effects of long-range dependence, that is, the latter will be hidden in partly misleading parameter estimates. We do not know large experiments with many dozens of replicates for each level spread (hopefully regularly) throughout the duration of the experiment, but we have no doubt that they would behave on each level like the "pure" measurement series.

A recent example of such a series are eleven hundred repeated measurements of the U.S. 1 kg check standard weight by the National Bureau of Standards in Washington, D.C., done with the greatest care, under conditions as uniform as possible, with relative random errors of about 10^{-10}, to yield series as independently and identically distributed as any series of high-quality physical measurements can be. Yet many parts of the series looked "out of statistical control" already with the naked eye; and even the longest homogeneous subseries of 289 observations, which for the untrained eye looks very much like an i.i.d. series, shows fairly weak $(\alpha = 0.8)$ but highly significant long-range correlations (Graf 1983, Graf et al. 1984, Hampel et al. 1986, Ch. 8.1a, Hampel 1987). (By the way, a test of the lag 1 serial correlation is not even weakly significant.)

An example where one actually wanted to measure an unknown constant with utmost accuracy were the almost 3000 measurements of the velocity of light by Michelson et al. (1935). This, or part of this data set, was probably the data W. Shewhart, according to Daniel (1976, p.22f), considered to be in as good a "state of statistical control" as any he had seen. But also here we find clear deviations from an i.i.d. pattern, and even the last 294 measurements, which are one of the most homogeneous looking subseries and profit from the accumulated experience of the experiments, contain quite strong $(\alpha = 0.6)$ and highly significant long-range correlations (Hampel 1987).

There is also evidence for long-range correlations in sample surveys (the wireworm data in Yates (1949) 1981, Sec. 7.20) and in quality control (360 data chosen from a fully automated production line analysed by Graf, see Hampel 1987 for more details), besides the broad fields mentioned above. In the frequency analysis of time series, '$1/f$ noises', i.e. processes whose spectrum has a singularity at the origin, have been found quite frequently and in widely divergent fields of application, for example in fluctuations of atomic clocks (Percival 1977), fluctuations of freeway traffic (Musha and Higuchi 1976), pitch fluctuations in music (Voss and Clarke 1975), electrical resistance deviations (Voss 1979), voltage fluctuations across cell membranes (Holden 1976), sand density fluctuations (Schick and Verveen 1974). For a discussion of examples in economics, cf. Granger (1966) and Mandelbrot (1973), for linguistics cf. Damerau and Mandelbrot (1973).

In all cases, one could try in principle to fit the data by transient parts of short-term correlated processes (such as ARMA-processes) with very many parameters, and/or take out trends, "hidden periodicities" and the like and thus declare them to be nonstationary. But nonstationary data, for example with linear trends, prevent us from estimating an absolute constant (unless, of course, we can find *and* eliminate effects of reasonable explanatory variables), and the many parameters of ARMA-processes are not interpretable either. In addition, all evidence points in the direction that in the "very long run" we do not have a change of behaviour towards the relative stability of an ARMA-process, but rather even bigger fluctuations as is typical for long-range processes.

The best-known models for long-range correlated processes are the increments of self-similar processes as mentioned above which are called fractional noise, and fractional ARIMA(p,d,q)-processes, where d is between 0 and $\frac{1}{2}$ (Granger and Joyeux 1980, Hosking 1981). For near-Gaussian fractional noise, statistical methods have been worked out and programmed, including confidence and prediction intervals for the mean ("one-sample t-test for long-range correlated data", see Beran 1986, 1989). For a more detailed survey of these statistical methods see Beran (1988), Künsch (1987).

It can be asked why the widespread occurrence of long-range correlations in statistical data is still so little known (cf. also Heyde 1987 "... persuasive case for long-range dependence. The recognition of and modelling of long-range dependene will be a topic of the utmost scientific importance."). First of all, there is of course the usual reluctance to abandon any traditional paradigm, especially if it is as simple and basic as the "i.i.d." paradigm. But there are also more specific reasons. One is that there were no usable alternative mathematical models available more than about fifty years ago; the first model – self-similar processes – became known in small subfields of statistics only since about 25 years ago; and most of the few statistical methods available for these processes were developed during the last ten years. Another reason is that both in small samples and in highly structured large samples the effects of long-range correlations tend to get confounded with the parametric effects of the model (i.e., in small samples with the grand mean) and hence pass unnoticed (but give partly distorted results). In contrast to such data sets, long sequences (say, $n > 100$) of supposedly i.i.d. "pure" measurements are extremely rare and hence unknown to most statisticians. (In geophysical and other "time series", some correlation structure is assumed a priori, see the discussion above.)

But the main reason that it was possible to ignore long-range correlations without too bad effects for at least part of statistics seems to be that for *contrasts* (differences between means, effects in the analysis of variance, slopes in regression etc.), as opposed to absolute constants, the first-order effects of long-range correlations cancel out. In how much they do so, what (unnoticed) efficiency losses still remain, and how these losses can be reduced by simple devices, is analysed in the following parts of this paper.

3. Notations and definitions

3.1 Contrasts. We consider in this paper a one-way layout with p treatments. Let β_j denote the mean under the j-th treatment. A linear combination of treatment means $\sum_{j=1}^{p} \lambda_j \beta_j$ with $\sum_{j=1}^{p} \lambda_j = 0$ is called a contrast. Obviously the set of all contrasts forms a $(p-1)-$dimensional vector space. Multiway layouts can be reduced to this case in the following way. If we have k factors on $p_1, \ldots, p_k$ levels respectively, we put $p = \Pi p_j$ and take each possible level combination as one treatment. All main effects and interactions in the original layout become then contrasts in the one-way layout. In our asymptotics the number of treatments p will be fixed whereas the number of observations increases to infinity. If we want to include fractional factorials, we have to let p increase at the same speed as n. This is a problem for further research.

3.2 Randomization. The p treatments are applied sequentially in time or along a transect in the plane. Denote the treatment allocation at time t by $a_t \epsilon \{1, 2, \ldots, p\}$. Hence the design matrix $X_n = (x_{t,j}; 1 \le t \le n; 1 \le j \le p)$ is given by

$$x_{t,j} = \begin{cases} 1 & \text{if } a_t = j \\ 0 & \text{otherwise} \end{cases} \tag{3.1}$$

and the model becomes

$$y_t = \sum_{j=1}^{p} \beta_j x_{t,j} + \varepsilon_t. \tag{3.2}$$

The treatment allocations a_t are supposed to be random. We will consider three different randomization schemes. The first one is <u>complete randomization</u> which means that

$$(a_t) \text{ is i.i.d. with } P[a_t = j] = \pi_j \quad (j = 1, \ldots, p). \tag{3.3}$$

This makes the number of observations with a given treatment, $n_j = \sum_{t=1}^{n} x_{t,j}$ $(j = 1, \ldots, p)$, random. With <u>restricted randomization</u> we fix n_j in advance and use the conditional distribution given the n_j's :

$$\begin{array}{ll} \text{All} & n!/(n_1! \ldots n_p!) \text{ allocations } (a_1, \ldots, a_n) \\ \text{with} & \displaystyle\sum_{t=1}^{n} x_{t,j} = n_j \quad (j = 1, \ldots, p) \text{ are equally likely.} \end{array} \tag{3.4}$$

For large n the difference between (3.3) and (3.4) with $n_j = n\pi_j$ will be small. The first one is somewhat easier to analyze because we can keep the allocations already chosen as n increases.

The third randomization scheme to be considered is <u>blockwise randomization</u>. Here we form blocks $B_k = \{(k-1)\ell + 1, (k-1)\ell + 2, \ldots, k\ell\}$ of length ℓ $(k = 1, \ldots, b = n/\ell)$. Then we require:

$$\begin{array}{l} \text{The allocations in different blocks are independent} \\ \text{and in each block } B_k \text{ all allocations with} \\ \displaystyle\sum_{t \varepsilon B_k} x_{t,j} = \ell_j \ge 1 \quad (j = 1, \ldots, p) \text{ are equally likely.} \end{array} \tag{3.5}$$

Obviously $\sum_{j=1}^{p} \ell_j = \ell$. The most common case is $\ell = p$, $\ell_j = 1$ for all j, i.e. each treatment occurs exactly once in each block.

3.3 Correlations of the errors. The errors ε_t in the model (3.2) are supposed to have expectation zero, constant variance σ^2 and long-range correlations. We assume that either

$$|\rho(t,s)| = |\operatorname{Corr}(\varepsilon_t,\varepsilon_s)| \le \text{ const.} \,|t-s|^{2H-2} \quad (t \neq s) \text{ with } H\epsilon(\frac{1}{2},1) \tag{3.6}$$

or in the stationary case $\rho(t,s)=\rho(t-s)$ with

$$\begin{aligned} \rho(t-s) &= \int_{-\pi}^{\pi} e^{i\omega(t-s)} f(\omega) d\omega \quad \text{where} \quad \int_{-\pi}^{\pi} f(\omega)d\omega = 1 \text{ and} \\ f(\omega) &= |\omega|^{1-2H} f_0(\omega) \quad \text{with} \quad H\epsilon(\frac{1}{2},1) \end{aligned} \tag{3.7}$$

and $f_0(\omega)$ continuous, of bounded variation and bounded away from zero. Condition (3.7) implies (3.6), see Zygmund (1959, Chap. V.2). In (3.6), stationarity is not required. The two most common models with long-range correlations, fractional noise (Mandelbrot and van Ness, 1968) and fractional ARIMA (p,d,q)-processes (Hosking, 1981), satisfy (3.7), see Sinai (1976) in the former and Hosking (1981) in the latter case.

If the treatments are allocated in space as for instance in agriculture, the index t becomes multidimensional, i.e. the ε_t's are then a random field. We expect our results to generalize to this case.

3.4 Estimators. The ordinary least squares estimator (OLSE) in the model (3.1)-(3.2) is

$$\hat\beta(n) = (X_n^T X_n)^{-1}(X_n^T y) = \Big(\sum_t y_t x_{t,1}/n_1, \ldots, \sum_t y_t x_{t,p}/n_p\Big)^T. \tag{3.8}$$

Here we are interested in the variance of a contrast $\sum \lambda_j \hat\beta_j(n)$ given the design X_n because the design is ancillary. It is sufficient to consider the variances $V_n(j,k \mid X_n)$ of $\hat\beta_j(n)-\hat\beta_k(n)$ because $\sum\lambda_j\hat\beta_j(n) = \sum\lambda_j(\hat\beta_j(n)-\hat\beta_1(n))$ and $\operatorname{Cov}(\hat\beta_j(n)-\hat\beta_1(n),\ \hat\beta_k(n)-\hat\beta_1(n)) = \frac{1}{2}(\operatorname{Var}(\hat\beta_j(n)-\hat\beta_1(n)) + \operatorname{Var}(\hat\beta_k(n)-\hat\beta_1(n)) - \operatorname{Var}(\hat\beta_j(n)-\hat\beta_k(n)))$. From (3.8) it follows immediately that

$$V_n(j,k \mid X_n) = \sigma^2 \sum_{t,s=1}^{n} (x_{t,j}/n_j - x_{t,k}/n_k)\rho(t,s)(x_{s,j}/n_j - x_{s,k}/n_k) \tag{3.9}$$

If the ε_t's were uncorrelated, $V_n(j,k \mid X_n)$ would be $\sigma^2(n_j^{-1}+n_k^{-1})$. Hence if one is not aware of the correlations, one estimates $V_n(j,k \mid X_n)$ by $\hat\sigma^2(n_j^{-1}+n_k^{-1})$ where

$$\begin{aligned} \hat\sigma^2 &= (n-p)^{-1}\sum_{t=1}^{n}\Big(y_t - \sum_t \hat\beta_j(n)x_{t,j}\Big)^2 \\ &= (n-p)^{-1}\Big(\sum_t \varepsilon_t^2 - \sum_j n_j^{-1}\big(\sum_t \varepsilon_t x_{t,j}\big)^2\Big). \end{aligned} \tag{3.10}$$

In Section 4 we are investigating the difference $\hat{\sigma}^2(n_j^{-1}+n_k^{-1}) - V_n(j,k \mid X_n)$ under the assumptions (3.4) and (3.6). This is relevant for the question of robustness of the level of tests and confidence intervals (robustness of validity) under deviations from independence. In Section 5 the same question is investigated under assumptions (3.3) and (3.7).

If we are interested in the robustness of power (robustness of efficiency) we have to study the efficiency of the ordinary least squares estimator. In Section 5 we will compare $V_n(j,k \mid X_n)$ with the conditional variance of the best linear unbiased estimator (BLUE) under assumptions (3.3) and (3.7). The best linear unbiased estimator $\tilde{\beta}(n)$ is given by

$$\tilde{\beta}(n) = (X_n^T \Sigma_n^{-1} X_n)^{-1}(X_n^T \Sigma_n^{-1} y) \tag{3.11}$$

where Σ_n is the covariance matrix of $(\varepsilon_1, \ldots, \varepsilon_n)$. The variance of $\tilde{\beta}_j(n) - \tilde{\beta}_k(n)$ given the design X_n is then

$$V_n^*(j,k \mid X_n) = \mathrm{Var}[\tilde{\beta}_j(n) - \tilde{\beta}_k(n) \mid X_n] = \lambda^T(X_n^T \Sigma_n^{-1} X_n)^{-1}\lambda \tag{3.12}$$

with $\lambda = (\lambda_1, \ldots, \lambda_p)^T$ such that $\sum \lambda_i \beta_i = \beta_j - \beta_k$.

Finally in Section 6 we will compare $V_n(j,k \mid X_n)$ under the two different randomizations (3.3) and (3.5), i.e. we will see how much we gain by changing the randomization instead of changing the method of estimation. With randomized blocks one will take the estimated block effects $\hat{\mu}_k = \ell^{-1}\sum_{t\epsilon B_k}(y_t - \sum_j \hat{\beta}_j x_{t,j})$ into account for the estimation of the variance of the errors. We thus consider instead of (3.10)

$$\begin{aligned}\hat{\sigma}^2_{\text{Block}} &= (n-p-b+1)^{-1}\sum_{t=1}^{n}(y_t - \sum_j \hat{\beta}_j x_{t,j} - \sum_k \hat{\mu}_k 1_{[t\epsilon B_k]})^2 \\ &= (n-p-b+1)^{-1}(\sum_t \varepsilon_t^2 - \sum_j n_j^{-1}(\sum_t \varepsilon_t x_{t,j})^2 \\ &\quad - \ell^{-1}\sum_k(\sum_{t\epsilon B_k}\varepsilon_t)^2 + n^{-1}(\sum_t \varepsilon_t)^2).\end{aligned} \tag{3.13}$$

4. Restricted randomization with ordinary least squares

4.1 Asymptotic behavior of the variance. The variance $V_n(j,k|X_n)$ of a contrast depends on the random design matrix X_n and is thus a random variable. Its expectation and variance under the assumption (3.4) is given in the following:

Proposition 4.1.

i) $E[V_n(j,k \mid X_n)] = (n_j^{-1} + n_k^{-1})\sigma^2\big(1 - n^{-1}(n-1)^{-1}\sum_{t\neq s}\rho(t,s)\big)$.

ii) If $n_j/n \to \pi_j\epsilon(0,1)$ $(j=1,\ldots,p)$, then $\mathrm{Var}[V_n(j,k \mid X_n)] \sim$
$(n_j^{-1}+n_k^{-1})^2\sigma^4\big(2n^{-2}\sum'\rho(t,s)^2 - 4n^{-3}\sum'\rho(t,s)\rho(s,u) + 2n^{-4}\sum'\rho(t,s)\rho(u,v)\big)$
where $\sum'$ denotes summation over indices which are all different.

The proof is given in the Appendix. From this we can obtain easily the asymptotic behavior of $V_n(j,k \mid X_n)$.

Theorem 4.1. If $n_j/n \to \pi_j\epsilon(0,1)$ $(j=1,\ldots,p)$ and $\rho(t,s)$ satisfies (2.6), then

$$E[n(V_n(j,k \mid X_n) - (n_j^{-1}+n_k^{-1})\sigma^2)] = O(n^{2H-2}),$$

$$\mathrm{Var}[nV_n(j,k \mid X_n)] = O(c_n) \quad \text{where} \quad c_n = n^{-1} \quad (H<\frac{3}{4}),\ \ c_n = \log(n)/n \quad (H=\frac{3}{4})$$

$$\text{and} \quad c_n = n^{4H-4} \quad (H>\frac{3}{4}).$$

In particular by Chebyshev's inequality $nV_n(j,k \mid X_n) \xrightarrow{P} (\pi_j^{-1}+\pi_k^{-1})\sigma^2$. The proof is again given in the Appendix. This says that for n large randomization will pick with high probability a design such that the variance of a contrast is almost the same as in the i.i.d. case. The correlations of the errors do however affect the speed of convergence in this asymptotic result: The larger H, the slower it is. Moreover for $H>\frac{3}{4}$ the bias is of the same order as the standard deviation.

4.2 Asymptotic behavior of $\hat\sigma^2$. First we show that $\hat\sigma^2$ is consistent. The effect of using the residuals instead of the true errors is by (3.10) $-(n-p)^{-1}\sum_j n_j^{-1}(\sum_t \varepsilon_t x_{t,j})^2$. The following Lemma shows that this is asymptotically negligible.

Lemma 4.1. If (3.6) holds, $E[(n-p)^{-1}\sum_j n_j^{-1}(\sum_t \varepsilon_t x_{t,j})^2 \mid X_n] = O(n^{2H-2})$ uniformly for all X_n.

Proof: see Appendix.

By the Chebyshev inequality this implies that $(n-p)^{-1}\sum n_j^{-1}(\sum \varepsilon_t x_{t,j})^2$ is $O_p(n^{2H-2})$ both conditionally on the design X_n and unconditionally. As a direct consequence we have

Theorem 4.2 If (ε_t) is stationary and ergodic, $\hat{\sigma}^2 \xrightarrow{P} \sigma$ conditionally on the design and unconditionally. Hence $n(V_n(j,k \mid X_n) - \hat{\sigma}^2(n_j^{-1} + n_k^{-1})) \xrightarrow{P} 0$, i.e. asymptotically the correlations of the errors can be neglected when estimating the variance of a contrast.

In order to get an idea about the effect of the correlations for finite n we look at the bias of $\hat{\sigma}^2$ more closely. By (3.10) we obtain the conditional expectation of $\hat{\sigma}^2$ given the design

$$E[\hat{\sigma}^2 \mid X_n] = \sigma^2\big(1 - (n-p)^{-1} \sum_{t \neq s} \rho(t,s) \sum_j n_j^{-1} x_{t,j} x_{s,j}\big) \tag{4.1}$$

In the Appendix we prove

Proposition 4.2.

i) $E\big[E[\hat{\sigma}^2 \mid X_n]\big] = \sigma^2\big(1 - n^{-1}(n-1)^{-1} \sum_{t \neq s} \rho(t,s)\big)$.

ii) If (3.6) holds and $n_j/n \to \pi_j \epsilon (0,1)$ $(j = 1, \ldots, p)$, $\operatorname{Var}\big[E[\hat{\sigma}^2 \mid X_n]\big] = O(c_n n^{-2})$ with c_n defined in Theorem 4.1.

If we consider Propositions 4.1 and 4.2, we see that the correlations of (ε_t) have two effects: if they are positive, the variance of contrasts tends to be smaller than under independence and we underestimate the variance of the errors. Surprisingly on the average these two effects cancel each other exactly for any n and arbitrary correlations.

For a complete discussion we also have to look at the variance of $\hat{\sigma}^2$. This depends on the fourth moments of (ε_t). The most important case is the one with Gaussian errors.

Proposition 4.3. If (ε_t) is Gaussian, (3.6) holds and $n_j/n \to \pi_j \epsilon (0,1)$, then $\operatorname{Var}[\hat{\sigma}^2 \mid X_n] = O(c_n)$ uniformly in X_n where c_n has been defined in Theorem 4.1.

Proof: see Appendix

Hence when using the classical t- and F-tests for contrasts derived under independence the true level is presumably a little larger than we assume because the effective number of degrees of freedom is smaller. For a conclusive statement one would have to look also at the correlation between $\hat{\beta}_j(n) - \hat{\beta}_k(n)$ and $\hat{\sigma}^2$. In any case, asymptotically the level is correct.

5. Complete randomization

If (ε_t) is stationary, the model (3.2) is a special case of the general regression model with correlated errors. Moreover $n^{-1}\sum x_{t,j}x_{t+h,k}$ converges a.s. to $\pi_j\pi_k$ for $h \neq 0$ and to $\pi_j\delta_{jk}$ for $h = 0$ respectively. Hence the following theorem is an easy consequence of Section 7.3 of Grenander and Rosenblatt (1957).

Theorem 5.1. If the spectrum $f(\omega)$ of (ε_t) is piecewise continuous and $0 < c \leq f(\omega) \leq C$, then $nV_n(j,k \mid X_n)$ converges a.s. to $\sigma^2(\pi_j^{-1} + \pi_k^{-1})$ and $nV_n^*(j,k \mid X_n)$ converges a.s. to $\sigma^2((2\pi)^{-2}\int f(\omega)^{-1}d\omega)^{-1}(\pi_j^{-1} + \pi_k^{-1})$.

However if (ε_t) has long-range correlations, the spectrum is of the form (3.7) and thus the condition of Theorem 5.1 does not hold. Nevertheless we have

Theorem 5.2. If (3.7) holds, then the conclusions of Theorem 5.1 continue to hold.

The proof which uses a completely different approach and exploits the independence of the x_t's, is given in the Appendix.

Hence asymptotically $V_n(j,k \mid X_n)$ has the same behavior under complete and under restricted randomization. We conjecture that this is true also for $V_n^*(j,k \mid X_n)$.

It is well known (see e.g., Rozanov, 1967, Formula (10.40), p.104) that $\sigma^2\left((2\pi)^{-2}\int f(\omega)^{-1}d\omega\right)^{-1}$ is the variance of the best linear interpolation of ε_t based on ε_s, $s \neq t$. If the dependence is strong, this will be much smaller than $\mathrm{Var}[\varepsilon_t]$, i.e. ordinary least squares can loose quite some efficiency. Some numerical values are given in Table 1 for the $\mathrm{ARIMA}(0, H-\frac{1}{2}, 0)$-model where the efficiency can be expressed with the gamma-function, see Theorem 1 of Hosking (1981). In practice the correlations of (ε_t) are unknown, so one cannot expect to really achieve $V_n^*(j,k \mid X_n)$. Still, it should be possible to improve the ordinary least squares estimator if n and H are large.

6. Randomized blocks with ordinary least squares

6.1 Asymptotic behavior of the variance. First we give the exact value of the expectation and the order of magnitude of the variance of $V_n(j,k \mid X_n)$.

Proposition 6.1. Assume that $\rho(t,s)$ satisfies (3.7). Under (3.5) we have:

i) $E[V_n(j,k \mid X_n)] = (n_j^{-1} + n_k^{-1})\sigma^2\big(1 - 2(\ell-1)^{-1}\sum_{t=1}^{\ell-1}\rho(t)\,(1-t/\ell)\big),$
$= (n_j^{-1} + n_k^{-1})\sigma_\ell^2$ say.

ii) $\mathrm{Var}[nV_n(j,k \mid X_n)] = O(c_n)$ where c_n has been defined in Theorem 4.1.

Proof: see Appendix.

Proposition 6.1 implies the convergence of $nV_n(j,k \mid X_n)$ in probability. In fact we have even almost sure convergence.

Theorem 6.1. If (3.7) and (3.5) hold, $nV_n(j,k \mid X_n)$ converges almost surely to $(\pi_j^{-1} + \pi_k^{-1})\sigma_\ell^2$ where $\pi_j = P[x_{t,j} = 1] = \ell_j/\ell$.

The proof is given in the Appendix. Hence with blockwise randomization the variance of a contrast depends on the correlations of the errors. If $\rho(t)$ is positive for all t, blocking decreases the variance. Moreover the decrease is larger for small blocks if the correlations decay monotonically.

Lemma 6.1. If $\rho(t) > \rho(s) > 0 \quad (0 < t < s)$, then σ_ℓ^2 is strictly increasing for $\ell \geq 2$. If $\lim_{t\to\infty}\rho(t) = 0$, σ_ℓ^2 converges to σ^2 as ℓ goes to infinity.

Proof: see Appendix.

Therefore blocking with large block size does not improve much over complete randomization. However, for small blocks the gain can be substantial as Table 2 shows. Comparing Tables 1 and 2 we see that ordinary least squares with small blocks is better than generalized least squares with complete randomization for the $\mathrm{ARIMA}(0, H-\frac{1}{2}, 0)$-model. This is however not true for any model with correlations decaying to zero monotonically. For instance, for a MA(1)-model with parameter α and $p = \ell = 2$, the efficiency of ordinary least squares with randomized block versus generalized least squares with complete randomization turns out to be $(1-\alpha^2)/(1+\alpha^2-\alpha)$. For α going to one, this tends to zero.

One may ask at this point how much more efficiency can be gained by using generalized least squares and/or different designs. With the same methods as used for Theorems 4.2 and 6.1 we can obtain the efficiency of generalized versus ordinary least squares for randomized blocks. For the ARIMA $(0, H-\frac{1}{2}, 0)$-model the results are given in Table 3. It shows that ordinary least squares are sufficient for all practical purposes. If $p=2, n_1=n_2=n/2$ and the spectral density $f(\omega)$ satisfies the conditions of Theorem 5.1 and takes its minimum at $\omega=\pi$, one can also show that the most efficient design alternates the two treatments. Moreover, ordinary and generalized least squares are equivalent for this design. However, this design is very sensitive to the presence of a monotone trend and thus should be avoided in general. We, therefore, do not pursue the search for optimal designs any further.

6.2 Asymptotic behavior of $\hat{\sigma}^2_{Block}$. Since we have exactly the same behavior as for $\hat{\sigma}^2$ in Section 4.1, we give only the main results. By generalizing Lemma 4.1 we obtain first the consistency of $\hat{\sigma}^2_{Block}$.

Theorem 6.2. If (ε_t) is stationary and ergodic and (2.7) holds, $\hat{\sigma}^2_{Block} \xrightarrow{P} \sigma^2_\ell$ both conditionally on the design and unconditionally.

Hence also with randomized blocks the correlations of the errors can be neglected asymptotically when estimating the variance of a contrast because the inclusion of block effects makes the correct adjustment for the estimated variance. Looking at finite n's we have by (3.13)

$$\begin{aligned} E[\hat{\sigma}^2_{Block} \mid X_n] = \sigma^2 - \sigma^2(n-p-b+1)^{-1}\ (2n/\ell \sum_{t=1}^{\ell-1} \rho(t)\ (1-t/\ell) \\ + \sum_{t\neq s} \rho(t-s)\ (\sum_j n_j^{-1} x_{t,j} x_{s,j} - n^{-1}) \end{aligned} \tag{6.1}$$

In the Appendix we prove

Proposition 6.2. $E\big[E[\hat{\sigma}^2_{Block} \mid X_n]\big] = \sigma^2_\ell$ and $\mathrm{Var}\big[E[\hat{\sigma}^2_{Block} \mid X_n]\big] = O(n^{-2} c_n)$.

In particular $\hat{\sigma}^2_{Block}(n_j^{-1}+n_k^{-1})$ is also here unbiased for $V_n(j,k|X_n)$ on the average.

7. Simulation results

We simulated 400 replicates of the restricted randomization design (3.4) for sample sizes $n=16$ and $n=64$, and number of treatments $p=2$ and $p=8$. For the random errors ϵ_t we took zero mean, variance one and correlations $\rho(t,s) = (|t-s+1|^{2H} - 2|t-s|^{2H} + |t-s-1|^{2H})/2$, $H=0.7$ and 0.9. In the Gaussian case this defines standardized fractional Gaussian noise. We then studied the conditional variance of the ordinary least squares estimator (OLSE) and the best linear unbiased estimator (BLUE) given the random design X_n. The results are summarized in Tables 4 to 8.

In the case $p = 8$ there is a large number of possible contrasts to look at. Fortunately, by a simple symmetry argument the distributions of $V_n(j,k \mid X_n)$ or $V_n^*(j,k \mid X_n)$ are the same for any $j \neq k$. Hence we estimated the moments of $V_n(j,k \mid X_n)$ or $V_n^*(j,k \mid X_n)$ for a fixed $j \neq k$ by taking averages over all j and k. Besides these pairwise differences we looked also at orthogonal contrasts $\gamma_j = \sum_{i=1}^{8} \lambda_{ji}\beta_i$ $(j = 1,\ldots,7)$ where the vectors $(\lambda_{j1},\ldots,\lambda_{j8})^T$ are orthogonal and have elements ± 1. This corresponds to main effects and interactions in a 2^3 factorial. Again the distribution of $\mathrm{Var}(\hat{\gamma}_j(n) \mid X_n)$ and $\mathrm{Cov}(\hat{\gamma}_j(n), \hat{\gamma}_k(n) \mid X_n)$ are the same for all j and $j \neq k$ respectively.

The asymptotic values of $nV_n^*(j,k \mid X_n)/(1/n_j + 1/n_k)$, given in Theorem 5.1, are equal to 0.810 for $H = 0.7$ and 0.317 for $H = 0.9$. Tables 4, 5 and 6 illustrate that already for short series the expected value of the variance for the BLUE is very close to the asymptotic limit. Also the standard deviation of $nV_n^*(j,k \mid X_n)$ decreases fast, probably with the rate $n^{-1/2}$. On the other hand $E(nV_n(j,k \mid X_n))/(1/n_j + 1/n_k)$ is quite far from the asymptotic value of 1, as predicted by Proposition 4.1 i). In particular for $H = 0.9$ the bias is large. Also the standard deviation decreases very slowly. Note furthermore that the skewness of the BLUE is decreasing with increasing sample size. Less clear is the behavior of the simulated skewness for the variance of the OLSE. The s.d. of the empirical skewness of 400 Gaussian random variables is 0.123, i.e. most skewness values in Tables 4-8 are significant. It is an open problem if $V_n(j,k \mid X_n)$, normalized by a factor of the order $c_n^{1/2}$ given by Theorem 4.1, is asymptotically normal or if it has another limit distribution. In principle, $V_n(j,k \mid X_n)$ belongs to the class of randomly associated arrays considered by Barbour and Eagleson (1986), but we were unable to verify any of their conditions for asymptotic distributions.

The smaller variability of the conditional variance for the BLUE can also be seen by comparing $\Delta_n^* = max_{j \neq k} V_n^*(j,k \mid X_n) - min_{j \neq k} V_n^*(j,k \mid X_n)$ with $\Delta_n = max_{j \neq k} V_n(j,k \mid X_n) - min_{j \neq k} V_n(j,k \mid X_n)$ (Table 5). For the BLUE, increasing the sample size from $n = 16$ to $n = 64$ leads to a decrease of the mean of this difference by an approximate factor of 1/2. This suggests a decay of the order $n^{-1/2}$. On the other hand, for the OLSE and $H = 0.7$ the decay is much slower. For $H = 0.9$ the mean of Δ_n is even approximately the same for n=16 and n=64. This means that for relatively small samples the conditional variance of the OLSE of a contrast varies considerably, depending on which design matrix we obtained from the randomization. For the BLUE the variance is less sensitive to the (randomized) choice of the treatment assignments. It is an open question if one can find reasonable classes of design matrices and randomization designs for which the conditional LS-variance $V_n(j,k)$ is more stable when n is not very large.

Table 7 illustrates how the biases of $V_n(j,k \mid X_n)$ and of $E[\hat{\sigma}^2 \mid X_n]$ compensate each other, cf. Propositions 4.1 and 4.2. The mean of the ratio of the two quantities is practically equal to one even for $n = 16$. The value of H or the number of treatments does not seem to play any role here. Table 8 shows that the situation is completely different for the arithmetic mean. There the increase in the variance due to the dependence of the errors and the bias of $E[\hat{\sigma} \mid X_n]$ reinforce each other. Hence the ratio of the two quantities is huge even for $n = 16$ and $H = 0.7$. This emphasizes once more the fundamental difference between contrasts and absolute constants.

Appendix: Proofs

Without loss of generality we assume $\sigma^2 = 1$ throughout.

Proof of Proposition 4.1: We put $\xi_t = x_{t,j}/n_j - x_{t,k}/n_k$. Because $\sum \xi_t^2 = (n_j^{-1} + n_k^{-1})$, it follows from (3.9) that $V_n(j,k \mid X_n) = (n_j^{-1} + n_k^{-1}) + \sum_{t \neq s} \rho(t,s)\xi_t\xi_s$. The randomization (3.4) implies that for $t \neq s$

$$\begin{aligned} E[\xi_t\xi_s] &= n^{-1}(n-1)^{-1}\{n_j^{-2}n_j(n_j-1) + n_k^{-2}n_k(n_k-1) - 2n_j^{-1}n_k^{-1}n_jn_k\} \\ &= -n^{-1}(n-1)^{-1}(n_j^{-1} + n_k^{-1}). \end{aligned}$$

This proves (i). For (ii) we observe that

$$\begin{aligned} \mathrm{Var}[V_n(j,k \mid X_n)] = &\sum{}' \rho(t,s)^2 \mathrm{Var}[\xi_t\xi_s] \\ &+ 4\sum{}' \rho(t,s)\rho(s,u)\mathrm{Cov}[\xi_t\xi_s, \xi_s\xi_u] \\ &+ \sum{}' \rho(t,s)\rho(u,v)\mathrm{Cov}[\xi_t\xi_s, \xi_u\xi_v]. \end{aligned}$$

Now a lengthy but straightforward calculation shows that for t, s, u, v all different

$$\mathrm{Var}[\xi_t \xi_s] = n^{-2}(n_j^{-1} + n_k^{-1})^2 + O(n^{-5}),$$

$$\mathrm{Cov}[\xi_t \xi_s, \xi_s \xi_u] = -n^{-3}(n_j^{-1} + n_k^{-1})^2 + O(n^{-6}),$$

$$\mathrm{Cov}[\xi_t \xi_s, \xi_u \xi_v] = 2n^{-4}(n_j^{-1} + n_k^{-1})^2 + O(n^{-7}),$$

Together this proves (ii).

<u>Proof of Theorem 4.1</u> By (3.6) $|\sum' \rho(t,s)| \leq$ const. $n \sum_{t=1}^{n-1} t^{2H-2} \leq$ const. n^{2H}, and

$$\begin{aligned}\sum{}' \rho(t,s)^2 \leq \text{const. } n \sum_{t=1}^{n-1} t^{4H-4} &\leq \text{const. } n \text{ if } H < \frac{3}{4}, \\ &\leq \text{const. } n \log n \text{ if } H = \frac{3}{4}, \\ &\leq \text{const. } n^{4H-2} \text{ if } H > \frac{3}{4}.\end{aligned}$$

Similarly $|\sum' \rho(t,s)\rho(s,u)| \leq$ const. $\sum' |\rho(t,s)| \sum_{u=1}^{n-1} u^{2H-2} \leq$ const. $n(\sum_{u=1}^{n-1} u^{2H-2})^2 \leq$ const. n^{4H-1}. Finally $|\sum' \rho(t,s)\rho(u,v)| \leq$ const. $(n \sum_{t=1}^{n-1} t^{2H-2})^2$ $\leq$ const. n^{4H}.

Hence Theorem 4.1 follows from Proposition 4.1.

<u>Proof of Lemma 4.1:</u> Because $x_{t,j} \epsilon \{0, 1\}$, $E[(n-p)^{-1} \sum_j n_j^{-1} (\sum_t \varepsilon_t x_{t,j})^2 \mid X_n]$ $\leq (n-p)^{-1} \sum_j n_j^{-1} \sum_{t,s} |\rho(t,s)| \leq$ const. $n^{2H-1} \sum_j n_j^{-1} = O(n^{2H-2})$.

<u>Proof of Proposition 4.2:</u> We put $\eta_{t,s} = (n-p)^{-1} \sum n_j^{-1} x_{t,j} x_{s,j}$. Then $E[\eta_{t,s}] = (n-p)^{-1} n^{-1} (n-1)^{-1} \sum n_j^{-1} n_j (n_j - 1) = n^{-1}(n-1)^{-1} (t \neq s)$. This proves i). By a lengthy but straightforward calculation we find for t, s, u, v all different

$$\mathrm{Var}[\eta_{t,s}] = n^{-4}(p-1) + O(n^{-5})$$

$$\mathrm{Cov}[\eta_{t,s}, \eta_{s,u}] = -n^{-5}(p-1) + O(n^{-6})$$

$$\mathrm{Cov}[\eta_{t,s}, \eta_{u,v}] = 2(p-1)n^{-6} + O(n^{-7}).$$

The proof is then completed with the arguments used above for Proposition 4.1 and Theorem 4.1.

Proof of Proposition 4.3: By the well known formula for the fourth moment of a Gaussian random variable we have

$$
\begin{aligned}
&\mathrm{Var}[\sum_t \varepsilon_t^2] = 2\sum_{t,s} \rho(t,s)^2, \\
&\mathrm{Var}[(\sum_t \varepsilon_t x_{t,j})^2 \mid X_n] = 2\,\mathrm{Var}[\sum_t \varepsilon_t x_{t,j} \mid X_n]^2 = 2\big(\sum_{t,s} x_{t,j} x_{s,j} \rho(t,s)\big)^2 \\
&\quad \leq 2\big(\sum \mid \rho(t,s) \mid\big)^2, \\
&\mid \mathrm{Cov}[\sum_t \varepsilon_t^2, (\sum_t \varepsilon_t x_{t,j})^2 \mid X_n] \mid = \mid 2 \sum_{t,s,u} x_{t,j} x_{s,j} \rho(t,u)\rho(s,u) \mid \leq 2 \sum_{t,s,u} \mid \rho(t,u) \mid\mid \rho(s,u) \mid .
\end{aligned}
$$

Now we apply again the arguments of the proof of Theorem 4.1.

The proof of Theorem 5.2 is more difficult than the previous ones. We break it up into several parts.

Lemma A.1: $\lim n\, V_n(j,k \mid X_n) = (\pi_j^{-1} + \pi_k^{-1})$.

Proof: As above we introduce

$$\xi_t = x_{t,j}/n_j - x_{t,k}/n_k.$$

Because n_j is random, it is much easier if we can replace ξ_t by

$$\theta_t = n^{-1}(x_{t,j}/\pi_j - x_{t,k}/\pi_k).$$

Because the θ_t's are independent with expectation zero and variance $n^{-2}(\pi_j^{-1} + \pi_k^{-1})$,

$$M_n = n^2 \sum_{t,s=1}^{n} \theta_t \rho(t-s) \theta_s - n(\pi_j^{-1} + \pi_k^{-1})$$

is a martingale with respect to $\mathcal{F}_n = \sigma(\theta_1, \ldots, \theta_n)$. Moreover

$$E[(M_n - M_{n-1})^2] = \mathrm{Var}[n^2\theta_n^2] + 4E[n^2\theta_n^2]^2 \sum_{t=1}^{n-1} \rho(n-t)^2 \leq \text{const. } n\, c_n.$$

Therefore by monotone integration

$$E[\sum n^{-2} E[(M_n - M_{n-1})^2 \mid \mathcal{F}_{n-1}] = \sum n^{-2} E[(M_n - M_{n-1})^2] < \infty.$$

The strong law of large numbers for martingales (see e.g. Stout (1974), Th. 3.3.1) implies therefore that $n^{-1}M_n = n\sum_{t,s=1}^{n} \theta_t \rho(t-s)\theta_s - (\pi_j^{-1} + \pi_k^{-1})$ converges to zero a.s.

Finally we have to show that

$$V_n(j,k \mid X_n) - \sum_{t,s=1}^{n} \theta_t \rho(t-s)\theta_s = \xi^T \Sigma_n \xi - \theta^T \Sigma_n \theta = o(n^{-1}) a.s..$$

Observe that

$$\begin{aligned} \mid \xi^T \Sigma_n \xi - \theta^T \Sigma_n \theta \mid &\le \big((\xi-\theta)^T \Sigma_n (\xi-\theta)\big)^{1/2} \big((\xi^T \Sigma_n \xi)^{1/2} + (\theta^T \Sigma_n \theta)^{1/2}\big) \\ &\le 2\big((\xi-\theta)^T \Sigma_n (\xi-\theta)\big)^{1/2} (\theta^T \Sigma_n \theta)^{1/2} + (\xi-\theta)^T \Sigma_n (\xi-\theta). \end{aligned}$$

By the argument above $\theta^T \Sigma_n \theta = O(n^{-1})$ a.s. By (3.6) the largest eigenvalue of Σ_n is bounded by const. n^{2H-1}. Hence

$$\begin{aligned} (\xi-\theta)^T \Sigma_n (\xi-\theta) &\le \text{const. } n^{2H-1} (\xi-\theta)^T (\xi-\theta) \\ &\le \text{const. } n^{2H} \big((n_j^{-1} - (n\pi_j)^{-1})^2 + (n_k^{-1} - (n\pi_k)^{-1})^2\big) \end{aligned}$$

which is $O\Big(n^{2H-3} (\log \log (n))^2\Big) = o(n^{-1})$ by the law of the iterated logarithm. This completes the proof of Lemma A.1.

<u>Lemma A.2:</u> $\liminf n\, V_n^*(j,k \mid X_n) \ge \big((2\pi)^{-2} \int f(\omega)^{-1} d\omega)^{-1} (\pi_j^{-1} + \pi_k^{-1})$.

<u>Proof:</u> Let (g_m) be the sequence of spectral densities $g_m(\omega) = \min\big(f(\omega), m\big)$. By $\Sigma_{m,n}$ we denote the corresponding covariance matrix of n consecutive observations. Now $g_m \le f$ implies $\Sigma_{m,n} \le \Sigma_n$, the last inequality being in the positive definite sense. By standard arguments it follows that $(X_n^T \Sigma_{m,n}^{-1} X_n)^{-1} \le (X_n^T \Sigma_n^{-1} X_n)^{-1}$ whence $\liminf n\, V_n^*(j,k \mid X_n) \ge \lim n\, \lambda^T (X_n \Sigma_{m,n}^{-1} X_n)^{-1} \lambda$. This limit exists a.s. by Theorem 5.1 and equals $\big((2\pi)^{-2} \int g_m(\omega)^{-1} d\omega\big)^{-1} (\pi_j^{-1} + \pi_k^{-1})$. Letting m tend to infinity completes the proof.

<u>Lemma A.3:</u> $\limsup n\, V_n^*(j,k \mid X_n) \le \big((2\pi)^{-2} \int f(\omega)^{-1} d\omega\big)^{-1} (\pi_j^{-1} + \pi_k^{-1})$:

<u>Proof:</u> This is the hard part because we cannot approximate f from above. We are going to approximate the best linear unbiased estimator $\hat{\beta}(n)$ by $\hat{\beta}^W(n) = (X_n^T W_n X_n)^{-1} (X_n^T W_n y)$ where $W_n = \{w(t-s); 1 \le t,s \le n\}$ with $w(t) = w(-t)$, $w(t) = 0$ $(\mid t \mid > q)$, $\hat{w}(\omega) = (2\pi)^{-1} \Sigma_t w(t) e^{it\omega} > 0$. This way we avoid approximating Σ_n^{-1} which is very delicate, cf. Bleher (1981).

Because $\tilde{\beta}(n)$ is optimal, we have

$$\begin{aligned} n\, V_n^*(j,k \mid X_n) &\le n\, Var\big(\tilde{\beta}_j^W(n) - \tilde{\beta}_k^W(n) \mid X_n\big) \\ &= n\lambda^T (X_n^T W_n X_n)^{-1}\, X_n^T W_n \Sigma_n W_n X_n (X_n^T W_n X_n)^{-1} \lambda. \end{aligned}$$

Using Lemma A.4 below and approximating $f(\omega)^{-1}$ by a sequence of positive trigonometric polynomials $\hat{w}(\omega)$ completes the proof.

Lemmas A.1-A.3 imply Theorem 5.2. Finally we show

Lemma A.4: $n\,\lambda^T(X_n^T W_n X_n)^{-1} X_n^T W_n \Sigma_n W_n X_n (X_n^T W_n X_n)^{-1}\lambda$ converges a.s. to $w(0)^{-2}(2\pi)^2 \int \hat{w}(\omega)^2 f(\omega) d\omega (\pi_j^{-1} + \pi_k^{-1})$. ($W_n$, w and $\hat{w}$ as defined above).

Proof: This is basically the same proof as the convergence of $n\,V_n(j,k \mid X_n)$. Let γ_t be defined by

$$\gamma_t = \big(X_n(X_n^T W_n X_n)^{-1}\lambda\big)_t w(0).$$

and let θ_t be as above. Again $M_n = n^2\theta^T W_n \Sigma_n W_n \theta - tr(W_n \Sigma_n W_n)\ (\pi_j^{-1} + \pi_k^{-1})$ is a martingale. Because

$$\mid (W_n \Sigma_n W_n)_{t,s} \mid \leq (\sum_u \mid w(u) \mid)^2 \max_{-2q \leq j \leq 2q} \mid \rho(t-s+j) \mid \leq \text{const.} \quad \mid t-s \mid^{2H-2},$$

$n^{-1}M_n$ converges a.s. to zero. Moreover for $q < t < n-q$

$$(W_n \Sigma_n W_n)_{t,t} = \sum_{s,u} w(s)w(u)\rho(s-u) = (2\pi)^2 \int \hat{w}(\omega)^2 f(\omega) d\omega.$$

Hence $n^{-1} tr(W_n \Sigma_n W_n)$ converges to $(2\pi)^2 \int \hat{w}(\omega)^2 f(\omega) d\omega$.

Finally we note that the largest eigenvalue of $W_n \Sigma_n W_n$ is of the same order as the one of Σ_n. Hence we only need to show that $\sum_t (\gamma_t - \theta_t)^2 = o(n^{-2H})$ a.s. But

$$\begin{aligned}
(X_n^T W_n X_n)_{jk} &= w(0)\sum_{t=1}^{n} x_{t,j}x_{t,k} + \sum_{s=1}^{q} w(s) \sum_{t=1}^{n-s} (x_{t,j}x_{t+s,k} + x_{t,k}x_{t+s,j}) \\
&= n\{w(0)\delta_{jk}\pi_j + 2\sum w(s)\pi_j\pi_k\} + O\big(n^{1/2}\log\log(n)\big) \\
&= n\,A_{jk} + O\big(n^{1/2}\text{loglog}(n)\big)\text{say}.
\end{aligned}$$

It is straightforward to check that $(A^{-1}\lambda)_i = w(0)^{-1}(\delta_{ij}\pi_j^{-1} - \delta_{ik}\pi_k^{-1})$. Together this implies $\mid \gamma_t - \theta_t \mid = O\big(n^{-3/2}\log\log(n)\big)$ a.s. (uniformly in t) which completes the proof of Lemma A.4.

Proof of Proposition 6.1 and Theorem 6.1: Again let $\xi_t = x_{t,j}/n_j - x_{t,k}/n_k = b^{-1}(x_{t,j}/\ell_j - x_{t,k}/\ell_k)$ and

$$Z(i_1, i_2) = \sum_{t\epsilon B_{i_1}} \sum_{s\epsilon B_{i_2}} \rho(t-s)\xi_t\xi_s \quad (1 \leq i_1, i_2 \leq b)$$

By the independence of different blocks $E[Z(i_1,i_2)]=0$ for $i_1\neq i_2$. If $t\neq s$ are in the same block

$$E[\xi_t\xi_s]=-n^{-1}(\ell-1)^{-1}(n_j^{-1}+n_k^{-1}).$$

Hence $E[Z(i,i)]=(n_j^{-1}+n_k^{-1})b^{-1}\big(1-2(\ell-1)^{-1}\sum_{t=1}^{\ell-1}\rho(t)(1-t/\ell)\big)$.

Because $V_n(j,k\mid X_n)=\sum_{i_1,i_2}Z(i_1,i_2)$, this proves i) of Proposition 6.1.

Again by the independence of different blocks $\mathrm{Var}[V_n(j,k\mid X_n)]$ $=\sum_{i_1,i_2}\mathrm{Var}\big(Z(i_1,i_2)\big)$. Because $\mid\xi_t\mid\leq$ const. b^{-1}, we have by (3.6)

$$\mathrm{Var}\big(Z(i_1,i_2)\big)\leq \text{const. } b^{-4}\mid i_1-i_2\mid^{4H-4}$$

From this ii) of Proposition 6.1 follows easily. Finally, for Theorem 6.1 we note that

$$M_b=b^2\sum_{i_1=1}^{b}\sum_{i_2=1}^{b}Z(i_1,i_2)-\ell^{-1}\sigma_\ell^2(\pi_j^{-1}+\pi_k^{-1})$$

is a martingale with respect to $\mathcal{F}_b=\sigma(x_1,\ldots,x_{\ell b})$. The condition for the strong law of large numbers is easily checked. Details are left to the reader.

Proof of Lemma 6.1: For any $1\leq t<(\ell+1)/2$ we have

$$\begin{aligned}&\rho(t)\{(\ell-1)^{-1}(1-t/\ell)-\ell^{-1}(1-t/(\ell+1))\}\\&=\rho(t)(\ell+1-2t)(\ell^3-\ell)^{-1}>\rho(\ell+1-t)(\ell+1-2t)(\ell^3-\ell)^{-1}\\&=\rho(\ell+1-t)\{\ell^{-1}\big(1-(\ell+1-t)/(\ell+1)\big)-(\ell-1)^{-1}\big(1-(\ell+1-t)/\ell\big)\}.\end{aligned}$$

Summing this inequality over all $1\leq t\leq(\ell+1)/2$ gives $\sigma_\ell^2<\sigma_{\ell+1}^2$. The convergence of σ_ℓ^2 to σ^2 is an easy consequence of (3.6).

Proof of Proposition 6.2:

Denote $\eta_{t,s}=\sum_j n_j^{-1}x_{t,j}x_{s,j}-n^{-1}$ $(t\neq s)$, $\eta_{t,t}=0$, and introduce

$$Z(i_1,i_2)=\sum_{t\epsilon B_{i_1}}\sum_{s\epsilon B_{i_2}}\rho(t-s)\eta_{t,s}.$$

For t and s in different blocks, we have $E[\eta_{t,s}\mid x_t]=\sum_j x_{t,j}/n-\ n^{-1}=0$. Hence $E[Z(i_1,i_2)]=0$ for $i_1\neq i_2$. On the other hand for t and s in the same block

$$E[\eta_{t,s}]=(\sum_j(\ell_j-1)/(\ell-1)-1)n^{-1}=-(p-1)/\big((\ell-1)n\big).$$

Inserting this into (6.1) gives $E\big[E[\hat\sigma^2_{Block} \mid X_n]\big] = \sigma^2_\ell$. Finally

$$\mathrm{Var}\big[E[\hat\sigma^2_{Block} \mid X_n]\big] \sim n^{-2}\ \mathrm{Var}\big(\sum_{i_1=1}^{b}\sum_{i_2=1}^{b} Z(i_1,i_2)\big) = n^{-2}\sum_{i_1=1}^{b}\sum_{i_2=1}^{b} \mathrm{Var}\ \big(Z(i_1,i_2)\big)$$

by conditioning on the variables in one block. Because $\mid \eta_{t,s} \mid \leq$ const. n^{-1} we can complete the proof with the same argument as for Proposition 6.1.

REFERENCES

Barbour, A. D., and Eagleson, G. K. (1986). Random association of symmetric arrays. Stoch. Anal. Appl. **4**, 239-281.

Beran, J. (1986). Estimation, testing and prediction for self-similar and related processes. Ph.D. thesis, ETH Zürich.

Beran, J. (1989). A test of location for data with slowly decaying serial correlations. Biometrika **76**, 261-269.

Bleher, P.M. (1981). Inversion of Toeplitz matrices. Trans. Moscow Math. Soc. **40**, 201-229.

Graf, H. P. (1983). Long-range correlations and estimation of the self-similarity parameter. Ph.D. thesis, ETH Zürich.

Grenander, U., and Rosenblatt, M. (1957). *Analysis of Stationary Time Series.* Wiley, New York.

Hampel, F. R. (1987). Data analysis and self-similar processes. In Proc. 46th Session of the ISI, **LII**, Book 4, 235-254.

Hampel, F. R., Ronchetti, E. M., Rousseeuw, P. J. and Stahel, W. A. (1986). *Robust Statistics: The Approach based on Influence Functions.* Wiley, New York.

Hosking, J. R. M. (1981). Fractional differencing. Biometrika **68**, 165-176.

Künsch, H. R. (1987). Statistical aspects of self-similar processes. In Proc. 1st World Congress Bernoulli Soc. (Yu. Prohorov and V.V. Sazonov, eds.) **1**, 67-74. VNU Science Press, Utrecht.

Mandelbrot, B. B., and van Ness, J. W. (1968). Fractional Brownian motions, fractional noises and applications. SIAM Rev. **10**, 422-437.

Mandelbrot, B. B., and Wallis, J. R. (1968). Noah, Joseph and operational hydrology. Water Resources Research **4**, 909-918.

Rozanov, Yu. A. (1967). *Stationary Random Processes.* Holden-Day, San Francisco.

Sinai, Ya. G. (1976). Self-similar probability distributions. Theory Prob. Appl. **21**, 64-80.

Stout, W. F. (1974). *Almost Sure Convergence.* Academic Press, New York.

Yajima, Y. (1988). On estimation of a regression model with long-memory stationary errors. Ann. Stat. **16**, 791-807.

Yajima, Y. ((1989).). Asymptotic properties of the LSE in a regression model with long-memory stationary errors. Preprint.

Zygmund, A. (1959). *Trigonometric Series.* Cambridge University Press, London.

Table 1:

Asymptotic efficiency of generalized versus ordinary least squares under complete randomization in the $ARIMA(0, H - \frac{1}{2}, 0)$ -model.

H	0.6	0.7	0.8	0.9
efficiency	1.03	1.16	1.46	2.45

Table 2:

Asymptotic efficiency of blockwise versus complete randomization with ordinary least squares in the $ARIMA(0, H - \frac{1}{2}, 0)$ -model. Block size $= l$.

H	0.6	0.7	0.8	0.9
eff for $l = 2$	1.12	1.33	1.75	3.00
eff for $l = 4$	1.09	1.25	1.59	2.61

Table 3:

Asymptotic efficiency of generalized versus ordinary least square under blockwise randomization in the $ARIMA(0, H - \frac{1}{2}, 0)$-model.

H	0.6	0.7	0.8	0.9
eff for $l = 2$	1.003	1.012	1.027	1.050
eff for $l = 4$	1.005	1.023	1.053	1.099

Table 4:

Mean (upper figure), standard deviation (middle figure) and skewness (lower figure) of $Var(\hat{\beta}_1(n) - \hat{\beta}_2(n)|X_n)/(n_1^{-1} + n_2^{-1})$ for $p = 2$, $n_1 = n_2 = n/2$. The errors are discrete fractional noise with unit variance. Each value is estimated from a simulation of the design X_n under restricted randomization with 400 replicates.

		$n = 16$	$n = 64$
	$H = 0.7$	0.857	0.934
OLSE		0.112	0.094
		0.85	0.11
	$H = 0.9$	0.460	0.564
		0.142	0.125
		1.46	0.95
	$H = 0.7$	0.789	0.806
BLUE		0.081	0.047
		1.37	0.54
	$H = 0.9$	0.320	0.315
		0.066	0.030
		1.46	0.83

Table 5:

Mean, standard deviation and skewness of the variance of treatment differences given the design X_n for $p = 8$, $n_j = n/p$. Besides a fixed pair of treatments also the minimal and maximal variances among all pairs of treatments are given. Notation: $V_n^{st}(j,k) = Var(\hat{\beta}_n(j) - \hat{\beta}_n(k)|X_n)/(n_j^{-1} + n_k^{-1})$, $V_n^{max} = max_{j \neq k} V_n^{st}(j,k)$ and $V_n^{min} = min_{j \neq k} V_n^{st}(j,k)$.

		$V_n^{st}(j,k)$ for fixed $j \neq k$		V_n^{max}		V_n^{min}	
		$n = 16$	$n = 64$	$n = 16$	$n = 64$	$n = 16$	$n = 64$
	$H = 0.7$	0.865	0.931	1.06	1.09	0.696	0.801
OLSE		0.100	0.079	0.072	0.069	0.030	0.027
		0.52	0.67	-0.22	0.70	0.38	-0.18
	$H = 0.9$	0.455	0.572	0.702	0.844	0.276	0.393
		0.117	0.125	0.087	0.140	0.028	0.029
		0.79	1.11	-0.38	0.85	0.56	0.10
	$H = 0.7$	0.819	0.811	0.971	0.899	0.676	0.729
BLUE		0.087	0.046	0.080	0.036	0.022	0.017
		0.57	0.43	0.39	0.43	0.41	-0.23
	$H = 0.9$	0.359	0.326	0.498	0.389	0.243	0.271
		0.079	0.033	0.084	0.030	0.015	0.012
		0.89	0.68	0.29	0.76	0.39	0.07

Table 6:

Mean, standard deviation and skewness of the variance and covariance of orthogonal contrasts given the design X_n with $p = 8$, $n_j = n/p$ $(j = 1, ..., p)$.

		$Var(\hat{\gamma_n}(j)\|X_n)p/n_j$ for j fixed		$Cov(\hat{\gamma_n}(j), \hat{\gamma_n}(k)\|X_n)p/(n_j n_k)^{\frac{1}{2}}$ for $j \neq k$ fixed	
		$n = 16$	$n = 64$	$n = 16$	$n = 64$
	$H = 0.7$	0.863	0.933	0.000	0.001
OLSE		0.110	0.083	0.076	0.057
		0.80	0.72	-0.02	0.04
	$H = 0.9$	0.453	0.574	0.001	0.001
		0.126	0.127	0.090	0.091
		1.05	1.02	0.02	0.00
	$H = 0.7$	0.816	0.813	0.000	0.001
BLUE		0.093	0.049	0.064	0.034
		0.86	0.56	-0.04	0.06
	$H = 0.9$	0.358	0.326	0.001	0.000
		0.082	0.033	0.056	0.023
		1.18	0.58	0.05	-0.04

Table 7:

Mean, standard deviation and skewness of $Var(\hat{\beta}_n(j) - \hat{\beta}_n(k)|X_n)/[(n_j^{-1} + n_j^{-1})E(\hat{\sigma}^2|X_n)]$ for $j \neq k$ fixed, $n_j = n/p$, $(j = 1, .., p)$. The estimator used is OLSE.

	p=2		p=8	
	$n = 16$	$n = 64$	$n = 16$	$n = 64$
$H = 0.7$	1.01	1.01	1.00	1.00
	0.143	0.098	0.067	0.034
	0.97	0.89	0.51	0.28
$H = 0.9$	1.00	1.02	1.01	1.00
	0.313	0.231	0.150	0.091
	1.45	1.15	0.58	0.57

Table 8:

Mean, standard deviation and skewness of $Var(n^{-1}\sum_{t=1}^{n} y_t)/n^{-1}E(\hat{\sigma}^2|X_n)$ for $n_j = n/p$, $(j = 1, .., p)$. The estimator used is OLSE.

	p=2		p=8	
	$n = 16$	$n = 64$	$n = 16$	$n = 64$
$H = 0.7$	3.51	5.66	3.51	5.66
	0.033	0.009	0.109	0.021
	0.99	0.94	0.51	0.30
$H = 0.9$	20.2	48.6	20.4	48.5
	0.423	0.178	1.42	0.494
	1.45	1.19	0.59	0.58