Calibration for Simultaneity: (Re)Sampling Methods for Simultaneous Inference with Applications to Function Estimation and Functional Data

ANDREAS BUJA [1] and WOLFGANG ROLKE [2]

We survey and illustrate a Monte Carlo technique for carrying out simple simultaneous inference with arbitrarily many statistics. Special cases of the technique have appeared in the literature, but there exists widespread unawareness of the simplicity and broad applicability of this solution to simultaneous inference.

The technique, here called "calibration for simultaneity" or CfS , consists of 1) limiting the search for coverage regions to a one-parameter family of nested regions, and 2) selecting from the family that region whose estimated coverage probability has the desired value. Natural one-parameter families are almost always available.

CfS applies whenever inference is based on a single distribution, for example: 1) fixed distributions such as Gaussians when diagnosing distributional assumptions, 2) conditional null distributions in exact tests with Neyman structure, in particular permutation tests, 3) bootstrap distributions for bootstrap standard error bands, 4) Bayesian posterior distributions for high-dimensional posterior probability regions, or 5) predictive distributions for multiple prediction intervals.

CfS is particularly useful for estimation of any type of function, such as empirical Q-Q curves, empirical CDFs, density estimates, smooths, generally any type of fit, and functions estimated from functional data.

A special case of CfS is equivalent to p-value adjustment (Westfall and Young, 1993). Conversely, the notion of a p-value can be extended to any simultaneous coverage problem that is solved with a one-parameter family of coverage regions.

Key Words: Permutation tests, randomization tests, bootstrap confidence regions, credible regions, posterior probability regions, predictive regions, Q-Q plots, p-values, p-value adjustment, multiple comparison.

1 Introduction

The following are typical examples of problems that are often in need of simultaneous inference:

[1] Andreas Buja is Professor, The Wharton School, University of Pennsylvania.
Web page: http://www-stat.wharton.upenn.edu/~buja/

[2] Wolfgang Rolke is Professor, Department of Mathematics, University of Puerto Rico – Mayaguez.
Web page: http://math.uprm.edu/~wrolke

- When comparing a univariate sample with a theoretical distribution, it is standard practice to examine Q-Q plots where empirical quantiles $\hat{q}(\alpha)$ are plotted against their theoretical counterparts $q(\alpha)$. If the values of the variable at hand are in fact samples from the theoretical distribution, the Q-Q plot is expected to be *near* the line $\hat{q} = q$. The problem is assessing how near is "near", preferably with some type of simultaneous coverage bands.
- In the two-sample problem of comparing two empirical distributions, it is standard to examine Q-Q plots where empirical quantiles $\hat{q}_2(\alpha)$ of the second sample are plotted against the corresponding empirical quantiles $\hat{q}_1(\alpha)$ of the first sample. If the two samples arise from the same population, the Q-Q plot is expected to be *near* the line $\hat{q}_1 = \hat{q}_2$. Again, the problem is assessing how near is "near".
- When fitting smooths, additive models or ACE regressions to predictor-response data, one likes to provide standard error or confidence bands with the estimated curves. The bands tend to have pointwise coverage probabilities of, say, 95%. It would be desirable to show bands that have simultaneous 95% coverage across a dense grid of predictor values and across all predictors.
- Functional data are a form of multivariate data in which the n'th case is thought of as the discretized realization of a stochastic process $x_n(t)$. The parameter t ranges in a finite subset T of a domain such as time, frequency, spatial location. If such functional data are further structured by a one-way classification into K groups, one may be interested in comparing the group means $m_k(t) = \text{ave}_{\{n \in group\ k\}}\, x_n(t)$. One would want to establish whether some of these process means are significantly different from the others, and one would like to see significance in a simultaneous sense across the whole domain T.

These are just a few examples, selected for no other reason than being the authors' starting point for this article. These examples and others will be illustrated in Section 5.

The search for simultaneous coverage bands often ties statisticians' minds in knots: it seems infeasible to perform a search among coverage bands because all pointwise intervals that make up the band can be widened and shifted independently, imposing on us a search space of dimension $2K$, if K is the number of locations at which we require simultaneous coverage.

One can cut through this knot of problems with a simplifying assumption: limit the search to a one-parameter family of nested bands. An example would be a band consisting of intervals $f(t) \pm s \cdot \sigma(t)$, where s is chosen for 95% pointwise coverage. By playing with the parameter s, however, we can adjust the width of the band such that its approximate simultaneous coverage probability is a prescribed value $1 - \alpha$. In other words, we calibrate the bands for desired simultaneous coverage (hence the proposed name). It may, for example, be found that a band with 98% pointwise coverage yields 90% simultaneous coverage. The essential step is to consider the collection of bands as a one-parameter family that can be searched for the desired simultaneous coverage. Limiting oneself to a one-parameter family of bands is often natural, and it has the crucial advantage of reducing the search for simultaneous coverage to a calibration problem. Simultaneous coverage probabilities can often be estimated with sampling or resampling, depending on the context; it is then natural to perform CfS by searching the estimated simultaneous coverage probabilities.

CfS has been used in a couple of contexts, in particular in the bootstrap literature. Examples are the "wild bootstrap" of Härdle and Marron (1991) and proposals by Faraway

(1990) and Hall and Pittelkow (1990). They consist of bootstrapping residuals in nonparametric and parametric regression and using CfS for confidence bands. Another published example is by Tibshirani (1992) who used bootstrap-based CfS for confidence bands in the analysis of medical data. CfS has also been proposed for Bayesian inference by Besag, Green, Higdon and Menger (1995). Their particular version of CfS turns out to be equivalent to frequentist p-value adjustment (Section 7).

The present article is a partial survey, a systematic account and an illustration of CfS , as well as an extension to frequentist null bands. Because we have observed considerable unawareness of this simple technique, we think this survey may be useful. In the following sections we give motivations and some history (Section 2), describe some controversies (Section 3), we illustrate the generality of the problem (Section 4), demonstrate a few worked applications (Section 5), outline computations in a general framework (Section 6), and describe the connection with p-value adjustment (Section 7). Among the latter are the one-sample problem of distributional assessment with normal Q-Q plots (Section 5.1), the two-sample problem of comparing two groups of observations with empirical Q-Q plots (Section 5.2), a one-way ANOVA problem involving functional data (Section 5.3), bootstrap standard error bands for the transformations of ACE regression (Section 5.4), and simultaneous bootstrap intervals for correlations (Section 5.5).

The first author has had a long-standing interest in the relation between inference and exploratory data analysis (EDA), with an emphasis on infusing a measure of inferential honesty into the visual tools of EDA (Buja et al. 1988, Section 5, pp. 292-295). The same can be said about the present article: null bands, standard error or confidence bands, and posterior bands are often inferential crutches propped onto exploratory graphical tools. By lifting coverage properties from pointwise to simultaneous, one may obtain a greater degree of trustworthiness when facing the question “is what we see real?”

2 Motivation

The motivation for this article came from the first of the above examples, the problem of assessing Q-Q plots for comparing univariate data with a theoretical distribution. The importance of this problem for normal Q-Q plots of regression residuals was recognized early on by Daniel and Wood (1980, 1999) who proposed and practiced the following informal method: Generate extensive series of normal Q-Q plots of normal pseudo-random samples in order to convey a sense of the variability that should be expected for a given sample size when the null hypothesis of normality is in fact true. Many pages of such “null plots” are contained in Daniel and Wood’s book, a testament to the fact that it was published before the advent of ubiquitous computing.

While Daniel and Wood proposed plotting many null curves *one at a time*, it is equally plausible to *overplot* many null curves in a single plot, often called a “spaghetti plot”. Overplotted null curves create the visual impression of a null band in which the observed curve should fall if it is compatible with the null hypothesis. If the observed curve reaches clearly outside the null band anywhere on its domain, significant deviation from the null hypothesis is inferred, and the exact shape of the excursion outside the band gives clues to the nature of the deviation. Figure 1 (top right) shows a band of overplotted null curves, but the curves are only plotted as series of points at the abscissae of interest. Actual spaghetti plots are shown in Figure 9 in a different context.

A next step was taken by Ripley (1981, chap. 8, p. 175) in the analysis of spatial data. He plots envelopes for cumulative distribution functions by computing the minimum and the maximum of 50 cdf's simulated under a null assumption. Note again that these are not confidence bands but null bands for cdf's.

Subsequently, Atkinson (1981, 1987) proposed the following inferential twist, for Q-Q plots again: simulate 19 Q-Q curves under the assumption of normal data; then the minimum and maximum envelopes can be interpreted as having 90% pointwise coverage, because a 20'th curve following the assumption would fall outside the extrema at a given point with probability 2*(1/20)=10%. This recipe was adopted by Cook and Weisberg (1982, 1999) and Venables and Ripley (2002).

Atkinson's proposal, intended for linear regression, has a second twist: it involves simulation of curves from artificial residuals obtained by regressing normal pseudo-random responses, the idea being to mimic the covariance structure of the actual residuals. Together with the inferential twist, the procedure amounts to a parametric bootstrap test.

Ripley's and Atkinson's envelopes worked well at a time when computing was difficult and slow. Being computationally less constrained now, there is no reason to favor computational shortcuts over conceptual and practical needs. An example of such a need is greater precision in estimating the lower 5% and upper 95% pointwise quantiles, which is what Atkinson's extrema of 19 simulations are doing. Another need is to move from pointwise to simultaneous coverage. A third need is to clarify the choices we have in selecting shapes of coverage bands. Here is a discussion of these issues:

- **Shapes of coverage bands:** Atkinson's inferential twist points in the direction of using bands formed by collections of intervals that have a specified pointwise coverage such as 90% or 95%. This choice has been universally adopted for confidence bands in nonparametric regression. For null bands in testing, this may be a lesser known choice, probably because there does not exist a theory of power for the resulting tests. Still, in the absence of such theory, we adopt the principle for all coverage problems, including confidence bands, null bands, posterior bands, predictive bands: the bands we consider have pointwise constant coverage probabilities, and the intervals that make up a band are formed by fixed marginal quantiles.

  This choice solves the problem of selecting one-parameter families of nested bands. There is obviously no deeper principle behind this choice other than the desire for a uniform pointwise treatment (see Härdle and Marron (1991), p. 783), but there is an added convenience: the natural parameter for the resulting one-parameter families of bands is the pointwise coverage probability, which permits a direct translation of pointwise to simultaneous coverage, and nesting is automatic because a band with lower pointwise coverage is necessarily contained in a band with higher pointwise coverage.

- **Approximation of pointwise quantiles:** Since we have decided to generally use pointwise quantiles to define coverage intervals, there is a question of how to compute them. Here are three typical cases:

  - Marginal quantiles can be theoretically known. An example are Q-Q curves for testing a fixed distribution: its order statistics can be mapped to uniform order statistics whose marginals are known to be certain Beta distributions.
  - If theoretical quantiles are not available, MC approximation may be used in one of two ways:

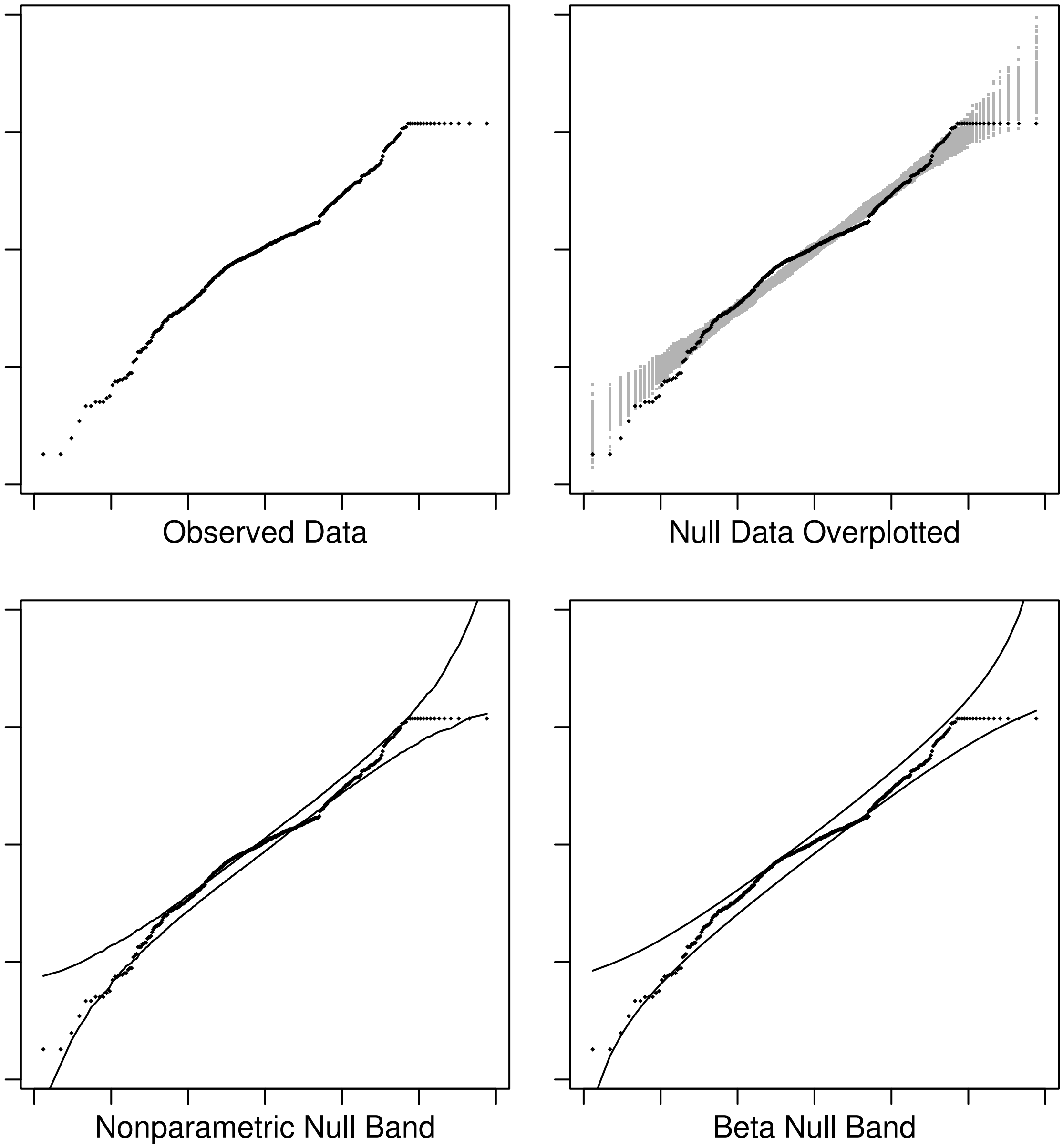

Figure 1: *Q-Q Plots for One-Sample Comparison with the Gaussian. Top left: the observed curve; top right: 100 overplotted null curves (shown only at quantile locations); bottom left: a 90% simultaneous null band based on pointwise quantile estimates; bottom right: a 90% simultaneous null band based on the Beta distribution of uniform order statistics.*

  - ∗ Quantiles may be approximated with simulated order statistics. To gain greater precision for estimating lower and upper 5% quantiles, for example, Atkinson's extrema of 19 simulations would be replaced with the 500'th and the 9,500'th order statistics of 9,999 simulations.
  - ∗ If a normal approximation is reasonable, it may be more efficient to estimate marginal means and variances from simulations and use the corresponding normal quantiles.

- **MC approximation of simultaneous coverage:** For any given band, one can approximate its simultaneous coverage probability as the fraction of simulated curves that are simultaneously contained in all intervals that make up the band. As with all simulations, this approximation can be made arbitrarily precise by increasing the number of simulations.

- **Search for bands with prescribed simultaneous coverage, or CfS :** Assuming that a one-parameter family of nested bands is given, and also assuming that MC sampling or resampling is used for calibration, here is how an implementation may proceed: We start by observing that it is often easy to find the minimal band that fully contains a given curve. Therefore, in an MC simulation of curves, it is sufficient to collect for each curve the parameter of the minimal containing band. After the simulation of, say, 9,999 curves, one has as many minimal parameter values; their upper 5% quantile (= order statistic 9,500) determines a band with an approximate simultaneous coverage probability of 95%. This procedure is similar to an algorithm described by Besag et al. (1995) for the construction of simultaneous posterior credible regions.

  Figure 1 (bottom plots) shows two examples of null bands that have been calibrated for an approximate simultaneous null coverage probability of 95%. One band is obtained by CfS of theoretical quantile bands, the other is obtained by CfS of pointwise MC estimates of quantiles. For more details see Section 5.1.

A case of CfS for Q-Q plots is in Davison and Hinkley's book on bootstrap (1997, section 4.2.5), where they interpret Q-Q plots as graphical parametric bootstrap tests. As far as they give computational details, they seem to be similar to those of Besag et al. (1995) in their re-use of simulations for constructing a one-parameter family of bands and estimating their simultaneous coverage properties; their description of the calibration step is left somewhat informal. In spite of its potential, this use of CfS seems isolated and tentative in their book.

CfS for confidence bands based on bootstrapping regression residuals has been proposed multiple times in different versions (Faraway 1990; Härdle and Marron 1991; Hall and Pittelkow 1990; Tibshirani 1992). Curiously this is the most problematic application area of CfS , for two reasons: (1) Even with a full enumeration of bootstrap samples, that is, the true bootstrap distribution, one obtains only an estimate of an intended error distribution, not the error distribution itself, whereas in frequentist testing the null distributions and in Bayesian approaches the posterior distributions are not approximations to anything other than themselves. (2) Confidence intervals and bands in regression always face difficult bias problems; see Loader's (1993) criticisms, but also Sun and Loader's (1994) work. The bias problem can be done away with by redefining and down-scaling the goal: instead of confidence bands, estimate standard error bands that attempt to capture the expectation of the estimates, as opposed to the underlying truth. [This would be in the spirit of Chaudhuri and Marron's (1999) SiZer if done for multiple bandwidths.] The former problem, though, that of approximation due to estimation, is inherent in the bootstrap and cannot be defined away. Yet, applying CfS with bootstrap is simply irresistible, and we show a couple of bootstrap applications as well (Sections 5.4 and 5.5).

3 Controversies and Objections

There are objections to some types of sampling/resampling-based simultaneous inference. For example, C. Loader (2001) argued in characteristically strong language against bootstrapping nonparametric regression residuals for simultaneous confidence bands:

> ... one should beware of using bootstrap methods for this type of problem, since they are extremely unreliable, and just an unnecessary addition to computational expense. The basic problem is that one works far out in the tail of marginal distributions, and the combination of bootstrap approximation with simulations cannot accurately model these tails.

We answer as follows: Whether one is far out in the tails of the marginal distributions can be readily learned from sampling/resampling itself. If, for example, a simulation shows 99% pointwise intervals to have only 40% simultaneous coverage, we will conclude that it is fruitless and even dangerous to strive for 99% simultaneous coverage, because this will drive us extremely far out in the pointwise quantiles. As a corollary, the goal of constructing bands with pre-specified stringent simultaneity properties is sometimes unattainable; instead, one will have to make do with estimating the effective simultaneity properties of realistic bands that are not too extreme in terms of pointwise quantiles. This is in fact the idea behind frequentist p-value adjustment (Westfall and Young 1993), see Section 7.

If simultaneous inference based on (re)sampling is to be useful, it is best to be in a situation where the Bonferroni adjustment is too conservative because of strong correlations among the multiple statistics for which simultaneous coverage intervals are sought. Independence provides a benchmark for extreme cleavage between pointwise and simultaneous coverage properties; for example, 100 intervals with 99% pointwise coverage will have only 36.6% simultaneous coverage, and in order to obtain 99% simultaneous coverage, one needs 99.99% pointwise coverage. By contrast, strong correlations will prevent pointwise and simultaneous coverage properties from becoming unbridgeable. Obvious examples are function estimates whose evaluations at nearby locations have strong correlations by construction, and this is indeed where we find the majority of applications of CfS .

In practice we like to compute the savings achieved by CfS over Bonferroni by forming the ratio $\alpha_{sim}/\alpha_{ptwise}$, where the α's are the complements of the coverage probabilities of either type for a given band; we call this ratio the "equivalent Bonferroni ratio." It describes approximately for how many independent variables this would be the Bonferroni adjustment. In one example below (Section 5.4) the coverage probabilities are 10% and 0.02% for simultaneous and pointwise, respectively. The equivalent Bonferroni ratio is 10/.06≈167. In other words, this would be the Bonferroni adjustment for 167 independent variables, which compares favorably with the fact that this example involved 1,006 variables.

A further objection against simultaneous inference says that the resulting intervals, bands, regions, are often too conservative. S. Marron (1997) formulates it as follows:

> ... "too conservative", i.e. you don't find features that are actually discernible by other means.

The "other means" he has in mind is his powerful SiZer methodology (Chaudhuri and Marron 1999). This argument points out the simple fact that it matters to what type of statistics one applies simultaneous inference. For example, evaluations of raw function estimates may form a rather powerless set of statistics when it comes to establishing the reality of qualitative features such as modes; it may be necessary to use a set of evaluations of estimated derivatives instead. Thus, "other means" might just be another set of statistics, whereas simultaneous inference as such is not the issue. The most elementary lesson of the theory of testing statistical hypotheses should not be lost: for detecting certain features (= rejecting certain null hypotheses in favor of certain alternatives), some statistics are more powerful than others. In this article we do not address the question of power and choice of statistics; our only focus is the question of simultaneity given a set of statistics.

This article focuses mostly on simultaneous coverage bands, but CfS can also be applied to one-parameter families of coverage regions that are not bands. An alternative are high-density coverage regions, which make natural nested one-parameter families by raising and lowering the contour threshold of a density function.

One may wonder whether coverage bands/boxes aren't generally inferior to high-density coverage areas. This is indeed what Knorr-Held (2003) maintains against Besag et al.'s posterior credible boxes. Note that the objection is general and not peculiar to Bayesian inference. Knorr-Held argues that boxes often contain large areas with very little posterior mass, in particular when the estimates are highly correlated under the posterior. This is illustrated in Figure 2 (a).

While this is true, it seems that the argument applies only in relatively low-dimensional settings and when the estimates are not curves. When they are curves, a very different viewpoint is needed. Counter-intuitively, it is in very high-dimensional situations with very strong correlations where Knorr-Held's criticism applies the least.

For one thing, there is a question of feasibility: it is infeasible and meaningless to attempt density estimation, for example, on the distribution of a 100-dimensional vector of a nonparametric fit evaluated at 100 locations.

The less intuitive aspect is that of high correlation that also works against Knorr-Held's argument. We use the following stylized situation as an "intuition pump" (the philosopher Daniel Dennett's term): We assume the estimates are variables $X(t)$ that are correlated in such a way that they have only two independent degrees of freedom, such as $X(t) = \cos(t)Y + \sin(t)Z$, where Y and Z are two latent independent variables with zero mean and equal variance, and t is in a finite subset of $\mathbb{R}$. We note that there is no interest in the fact that a large part of a box $\{x| - c \leq x(t) \leq c \ \forall \ t\}$ has low probability density because in this example we know about the trivial correlations for nearby t's, just as we know about correlations between nearby evaluations of a nonparametric fit. It is of interest, however, that the same high-dimensional box can give a pretty good approximation to a high-density coverage region in the plane spanned by the latent variables Y and Z. This is illustrated in Figure 2(b) where the meaning of the ten intervals is shown in the space of Y and Z: the intervals that define the simultaneous coverage box approximate a circular disk. Indeed, if Y and Z are normal, this circular disk is a simultaneous highest-density coverage region. The bands therefore represent a coverage region that is more meaningful in the space of latent variables Y and Z than in the space of observables $X(t)$.

4 Generality of the Simultaneous Coverage Problem

Bands with simultaneous coverage properties have generality in at least two dimensions: 1) the type of object for which simultaneous inference is sought, and 2) the type of distribu-

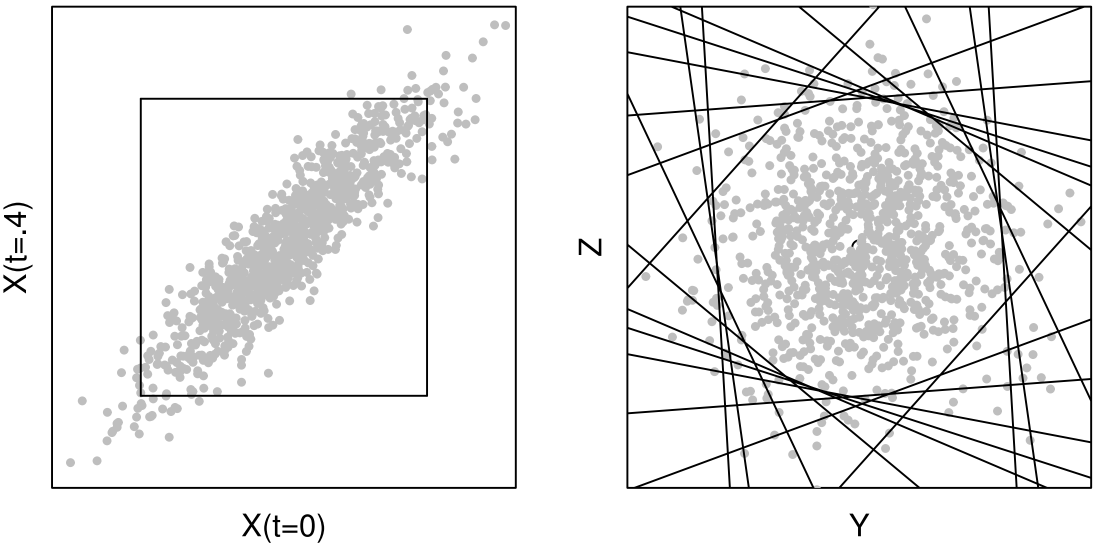

Figure 2: *Coverage Boxes/Bands: (a) Two variates with correlation* $\cos(.4) = 0.92$*; the north-west and south-east have low probability. (b) A simultaneous band on ten variables with only two latent dimensions,* Y *and* Z*, shown in the* Y*-*Z *plane; the ten intervals in ten directions approximate a circular disk. Each point represents a curve* $X(t) = \cos(t)Y + \sin(t)Z$*.*

tion that gives rise to simultaneous coverage probabilities.

We start with distributional situations where simultaneous inference is possible. The simple but essential assumption is that there exists a single probability distribution for which simultaneous coverage probabilities need to be calculated.

- **Null regions for testing null hypotheses:** In many testing situations significance levels and hence null coverage probabilities are obtained from a single distribution.
  - *Simple null hypotheses:* An example is testing a normal distribution with a specific mean and variance.
  - *Composite null hypotheses with pivotal structure:* An example is testing for normality, with unknown mean and variance. Pivotality is achieved by standardizing the data, that is, subtracting a location estimate and dividing by a scale estimate. The distribution of the standardized data is independent of the true mean and variance. Although this joint distribution may not be analytically tractable under most location and scale estimates, it can be easily simulated.
  - *Exact tests with Neyman structure:* In many (though far from all) composite null hypotheses, nuisance parameters can be eliminated by conditioning on a statistic that is sufficient under the null hypothesis but not under alternatives. The result are "conditional" or "similar" or "exact" tests (Lehman 1986).

    The best-known examples are permutation tests (Good 2000, Mielke and Berry 2001) of null hypotheses of independence or, more generally, exchangeability. They can be illustrated with two-sample tests, in which a quantitative variable is assumed to be independent of a binary variable that labels the two samples. In

  Section 5.2 we give an example where an empirical Q-Q plot of the two groups is assessed in terms of a null band whose simultaneous coverage probability is obtained from the permutation distribution of the data. In Section 5.3 we give another example where a permutation distribution is applied to a 1-way ANOVA situation.

  The class of exact tests comprises many examples in addition to permutation tests and they are probably insufficiently exploited. A companion paper (Buja 2003) looks into these issues and proposes another type of exact test that may be called "rotation test".

  - *Nonparametric Bootstrap tests:* Bootstrap tests are unlike the usual non-parametric bootstrap used for standard error estimation. In bootstrap tests one resamples the data after modifying them to agree with the null hypothesis (Efron and Tibshirani 1993, Davison and Hinkley 1997). For example, the null hypothesis may state that the medians of two or more groups are equal. In a bootstrap test, one would subtract from each observation the median of its group and resample the resulting data.

    Bootstrap tests are more often applicable than permutation tests, but the latter are exact, the former are not. When they both apply, they differ in the resampling mode: permutation tests use sampling without replacement, bootstrap tests use sampling with replacement. For example, a null hypothesis of independence of two variables is tested as follows: a permutation test matches permutations of the values of the first variable with permutations of the values of the second variable (the second permutation is vacuous); a bootstrap test matches i.i.d. resamples of the first variable with i.i.d. resamples of the second variable (the second resampling is essential).

  - *Parametric Bootstrap tests:* Parametric bootstrap relies on a parametric model. The test version of this type of bootstrap consists of 1) fitting a model under the null hypothesis and 2) calculating significance levels or null coverage probabilities under the fitted model.

- **Bootstrap standard error regions:** These are typically constructed with pointwise coverage properties, but we have seen that CfS has been proposed more than once in the context of bootstrapping regression residuals. The distribution used to estimate simultaneous coverage is an appropriate type of bootstrap distribution. For examples see Sections 5.4 and 5.5.

- **Bayesian posterior regions:** Bayesian posteriors are usually analyzed one parameter at a time, which means using pointwise rather than simultaneous posterior coverage probabilities. Bayes procedures, however, have simultaneity problems also, and the revolution in Bayesian computing has produced a wealth of opportunities for using Markov chain Monte Carlo techniques to simulate draws from posterior distributions in complex models and using them for simultaneous inference. We already mentioned two proposals: "simultaneous credible regions based on order statistics" (Besag et al. 1995) which amount to a type of simultaneous posterior bands or boxes, and "posterior contour areas" (Korn-Held 2003) which are high-density regions with regard to the posterior, or an approximation thereof.

- **Predictive regions:** When the statistical problem is not about interpretation or prediction of parameters, it may be about prediction of future observations instead. When the predicted observations are multidimensional, it may be necessary to account for multiplicity, just like in any other type of inference. Given a predictive distribution, be it frequentist or posterior Bayesian, CfS may be the construction of choice for predictive simultaneous coverage regions when analytics are not available.

The other direction in which the reach of simulation-based simultaneous inference can be extended is in terms of the objects for which inference is sought. Fundamentally, all objects have the same structure: they are finite collections of statistics for which coverage intervals are to be constructed in such a way that the joint coverage probability has a prescribed size $1 - \alpha$. Just the same, it is worthwhile pointing out the variety of special cases that are covered by this simple setup. Here is an incomplete list:

- **Multiple parameter estimates:** This is the conventional case of simultaneous inference. In a typical situation one would have estimates of several regression coefficients or ANOVA effects, for which one needs simultaneous confidence intervals or non-rejection regions. Other examples are the entries in a correlation matrix or eigenvalue profiles in principal components analysis (Buja and Eyuboglu 1993). Situations with more heterogeneous sets of parameter estimates are easy to conceive, for example two-sample problems for which one needs simultaneous comparisons of means and variances.

  General multiple parameter estimates can be given a graphical device that parallels the idea of plotting curves with coverage bands. The idea is to use a parallel coordinate display for the observed and simulated values of the estimates as well as their coverage intervals, as in Figure 11 of Section 5.5. This representation is formally equivalent to the more familiar scatterplot display (Figure 10), but it gives us a way to think of coverage boxes as coverage bands, even when the estimates are not curves but heterogeneous types of variables.

- **Diagnostics curves for distributional assessment and two-sample comparisons:** Both one- and two-sample problems are often approached with Q-Q plots for which we now know how to construct simultaneous null bands. See Sections 5.1 and 5.2 for worked examples.

  Null plots and null bands are not restricted to Q-Q plots. In the two-sample problem, for example, they can be equally applied to other types of plots such as paired histograms, paired density plots, or paired percentile plots. More recently, Ghosh (1996) proposed a graphical tool called T_3 plot for the one-sample problem, and Ghosh and Beran (2000) proposed a variant of T_3 plots for the two-sample problem.

- **Smooths and non-parametric transformations of data:** Smooths for simple x-y data, generalized additive models based on smooths, and ACE regression with smooth transformation of the response are all amenable to CfS . Smooths can be augmented with permutation null bands for the overall null hypothesis, or with bootstrap standard error bands. The type of smoother is immaterial: kernel, local polynomial, smoothing spline, regression spline smooths can all be subjected to the same simultaneous inference procedures. Nonparametric regression seems to stretch the framework of CfS due to the potentially large number of variables involved. Indeed, see Section 5.4 for an

ACE fit involving fourteen variables with simultaneous inference for a total of over 1,000 locations, executed in a high-level language on a small computer. Computing was not a limiting factor.

- **Surfaces in spatial/temporal modeling:** Models can be used to simulate possible data, either in a parametric bootstrap fashion by using the fitted parameters, or in a Bayesian fashion by drawing parameters from a posterior. Either way, simulation of spatial or temporal processes can be used to obtain simultaneous confidence or posterior intervals for surfaces or simultaneous predictive intervals for future data at locations or times of interest.

- **Curves that arise as functional data:** Functional data are multivariate data in which the collection of variables is thought of as a discretized random function or random signal. Typically the number of variables is large and all variables are on the same scale. Estimation for functional data usually leads to parameter vectors that are also thought of as discretized dual processes. One therefore interprets a set of confidence intervals for the parameters as a confidence band because the parameters are thought of as representing a function defined on a domain such as space, time or frequency. A worked example involving a functional one-way ANOVA problem is in Section 5.3.

In all examples we assume that the functions are discretized at finitely many locations, and we consider only simultaneous inference at these locations (as in Härdle and Marron 1991, p. 783). We ignore the problem of simultaneous inference at all or infinitely many locations; in practice this is more often an aesthetic than an essential problem. In many situations one can choose the discretization fine enough that the original function estimates and their bounds can be replaced by linear interpolants, so that simultaneous inference at the discretization locations is equivalent to simultaneous inference on the whole function domain.

5 Some Applications

5.1 Null Bands for a One-Sample Comparison Based on Q-Q Plots

Here are details for the plots in Figure 1 for testing normality with Q-Q curves. The data are the response variable "Median Housing Value" (MEDV0 from the well-known Boston Housing data, in all a set of 506 values, here denoted $\{Y_t\}_{t=1..506}$. The values were sorted and standardized (studentized): $X(t) = (Y_{(t)} - \hat{\mu})/\hat{\sigma}$. Because the distribution of these values is independent of the true mean and variance, we can use sorted studentized values from any normal distribution (in particular N(0,1)) in the null simulation.

For the plot in the bottom left of Figure 1, we estimated pointwise quantiles from 9,999 normal pseudo-random samples. Another simulation of the same size was used to estimate simultaneous coverage for the 506 test statistics. Figure 3 shows the correspondence between pointwise and simultaneous significance levels. One reads off that in order to achieve a 10% simultaneous significance (= 1-coverage), one needs a 0.12% pointwise significance, for an equivalent Bonferroni factor of 10/0.12≈83, which compares favorably with the number of variables, 506. In the bottom right of Figure 1 we used a parametric family of bands derived from the known Beta distribution of uniform order statistics (Section 6.3), which strictly

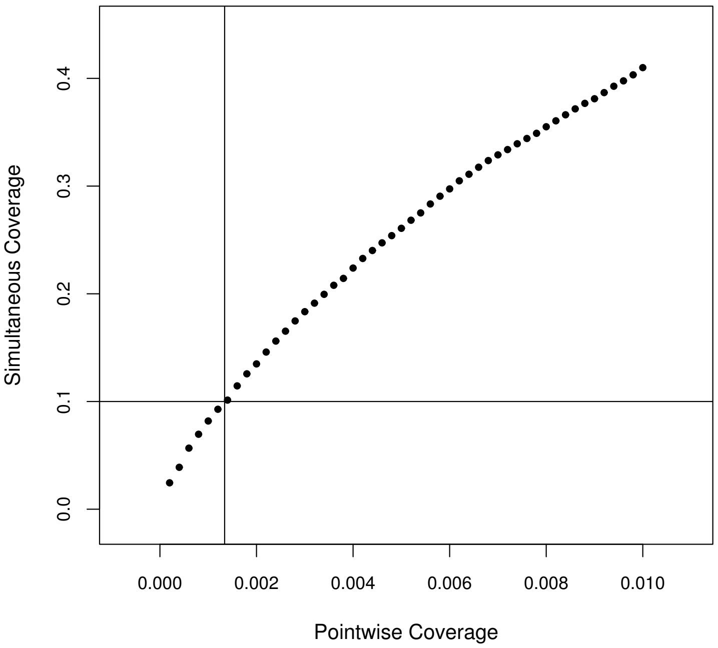

Figure 3: *Simultaneous versus Pointwise Significance Levels for the One-Sample Problem. For example, in order to achieve a simultaneous significance level of 0.10, a pointwise significance level of approximately 0.0012 is required.*

speaking is not quite proper because it does not account for the standardization. CfS was again achieved with a simulation of the same size. Both bands are significant at the 10% level due to 90% simultaneous coverage. The nonparametric band on the left is somewhat more powerful than the parametric band on the right: the former has 181 out of 506 excursions of the actual data outside the band, the latter only 129.

It is one of the virtues of simulation approaches that the distributions of exactly the desired statistics can be obtained, without resorting to approximations such as that of a t_n with a Gaussian. If the statistics arise for example as studentized residuals from a regression, one can account for the correlations among the residuals by regressing normal null data on the predictors and forming studentized null residuals that mimic the exact computation of the actual studentized residuals (Atkinson 1981). Furthermore, the regression does not need to be linear OLS: It can be robust, model-searched, nonlinear, nonparametric, cross-validated, ...

5.2 Permutation Null Bands for Two-Sample Comparisons with Q-Q Plots

The following example uses the New Jersey lottery payoff data from the "New S" book by Becker, Chambers and Wilks (1988). The payoffs are given in dollars for each number, but we ignore the dependence on the numbers. Instead, we consider the dependence on time periods: We compare the payoffs of the 1975-76 period and the 1980-81 period.

Retracing the steps for the one-sample comparison in the previous section, we show a

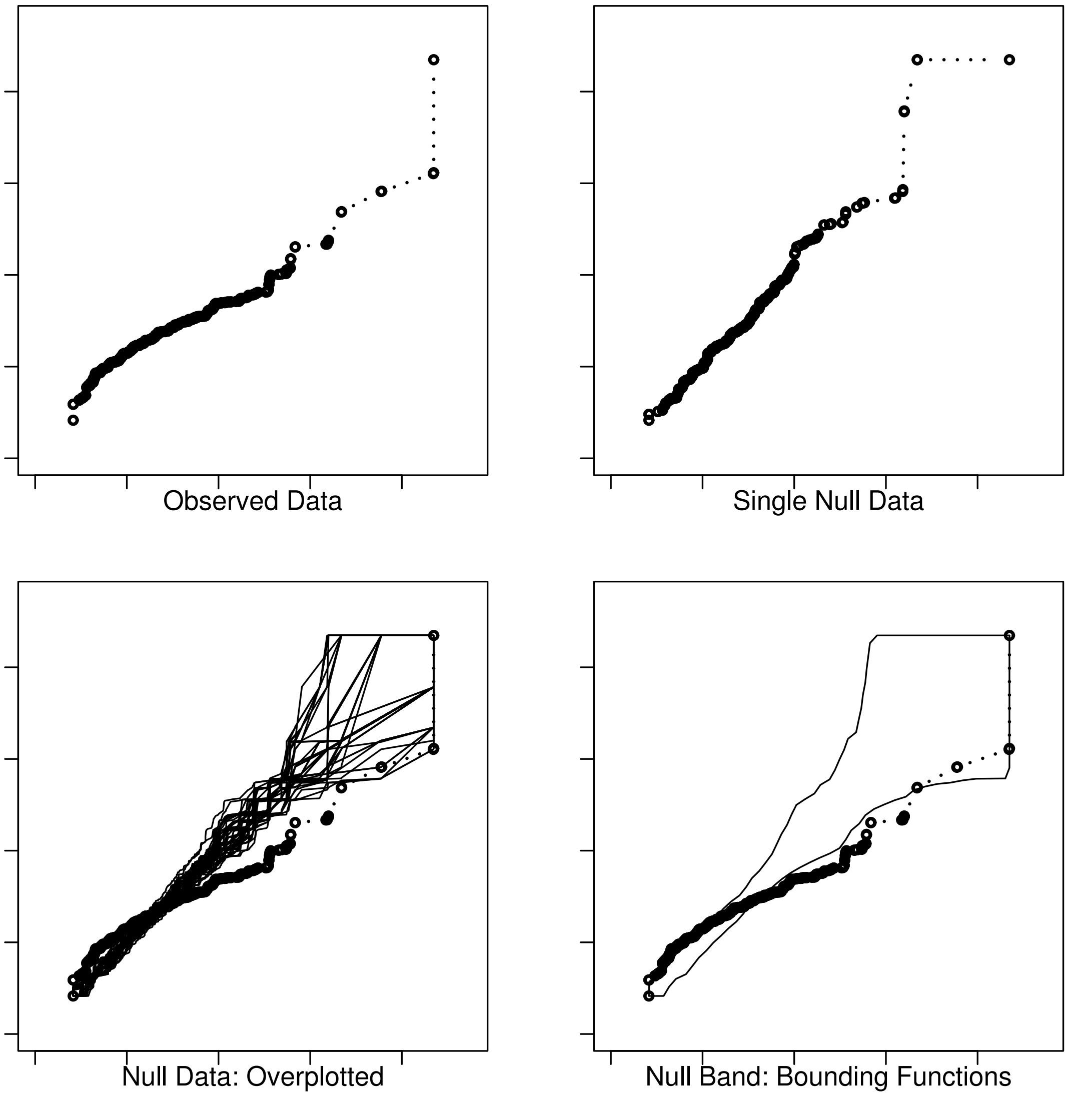

Figure 4: *Q-Q Plots for the Two-Sample Problem of Testing Equality of Two Populations. Top left: a single null curve; top right: the observed curve; bottom left: a null band represented by overplotted null curves; bottom right: a 95% simultaneous null band represented by lower and upper bounding curves.*

Q-Q plot of the actual two payoff distributions against each other in the top left panel of Figure 4. The null assumption of identical payoff distributions is tested with a permutation test whereby null data are generated by randomly permuting the period labels "1975-76" and "1980-81" against the payoff values. The remaining panels of Figure 4 show the results: The top right shows an example of a Q-Q plot for one such set of permutation null data. The bottom left shows 100 overplotted Q-Q curves of null data, together with the Q-Q curve of the actual data. Finally, the bottom right shows a simultaneous permutation null band for the Q-Q curves at the 0.05 level, together with Q-Q curve of the actual data. CfS required widening pointwise null bands to the 0.002 level (0.998 coverage) to achieve a simultaneous 0.05 level, for an equivalent Bonferroni factor of 0.05/0.002=25.

As the null band shows, the 1975-76 period had significantly greater variability than would be expected under the null assumption of no difference: The high payoffs were significantly larger and the low payoffs significantly lower in 1975-76 than 1980-81.

A point of interest are the computations of the null bands for two-sample Q-Q curves: The symmetry between the two samples suggests that the null band should be constructed symmetrically. Note that for the one-sample comparison in the previous subsection we treated the theoretical and the empirical quantiles asymmetrically by choosing the theoretical quantile axis for parametrization of the Q-Q curves and their null band. For the present two-sample comparison by contrast, we chose a parametrization in terms of the 45 degree diagonal to achieve a symmetric treatment of the two samples. We constructed 50 test statistics in terms of 50 equispaced locations on the 45 degree diagonal. The number 50 was chosen for computational expediency. The bands are drawn by computing the orthogonal distances of the upper and lower bounds from the 45 degree diagonal.

5.3 Permutation Null Bands for Multivariate/Functional One-Way ANOVA

The following "pizza commercial data" form a sample of curves that fall into several groups and should hence be analyzed as a functional one-way ANOVA. The data are courtesy of Jianqing Fan who in turn obtained them from D. Hudge and N.M. Didow at the University of North Carolina at Chapel Hill. Fan and Lin (1998) applied what they call "adaptive Neyman tests on Fourier transformed data" in order to detect group differences. Our aim is to use simultaneous confidence bands on the more or less raw data to the same end. Fan and Lin (1998) give the following background description:

> In evaluating business advertisements, evaluators are asked to dynamically assign scores to a commercial as they are watching. The resulting observations are a collection of curves: the score of the i'th subject assigned at time t of the commercial advertisement. Figure 1 presents this kind of data on a pizza commercial. The commercial was played at studios at six different time slots and assessed by different evaluators. Of interest is to test if there is any significant time effect.

The curve of each evaluator is given at 200 *time points*. The 6 *time slots* (not to be confused with the 200 time points) constitute the six groups of curves. The groups are unbalanced and their sizes can be read from Figure 5 where the raw data are shown. The evaluation values observed at each time point range between 0 and 100; because the evaluation device was initialized at 50, one observes many 50s in the beginning of each time slot. We take the last sentence of Fan and Lin's description as meaning that there is an interest in significant effects of *time slots* because they presumably correspond to controlled experimental conditions.

Because the design does not use repeated measures, one will expect considerable noise caused by the presence of different evaluators in each group. As Figure 5 shows, the noise is indeed considerable and no clear group differences are evident at this point.

Fan and Lin mention a need for dimension reduction which they perform with Fourier analysis. For our techniques, there is in principle no need to reduce the dimensionality: Null bands for 200 test statistics (one for each time point) do not pose conceptual problems; to the opposite, staying close to the raw data is advantageous for plotting and interpretation.

Just the same, we made a small concession to dimension reduction to speed up the computations: we reduce the number of abscissa by using averages of four consecutive time points, thus reducing the effective number of time points to 50.

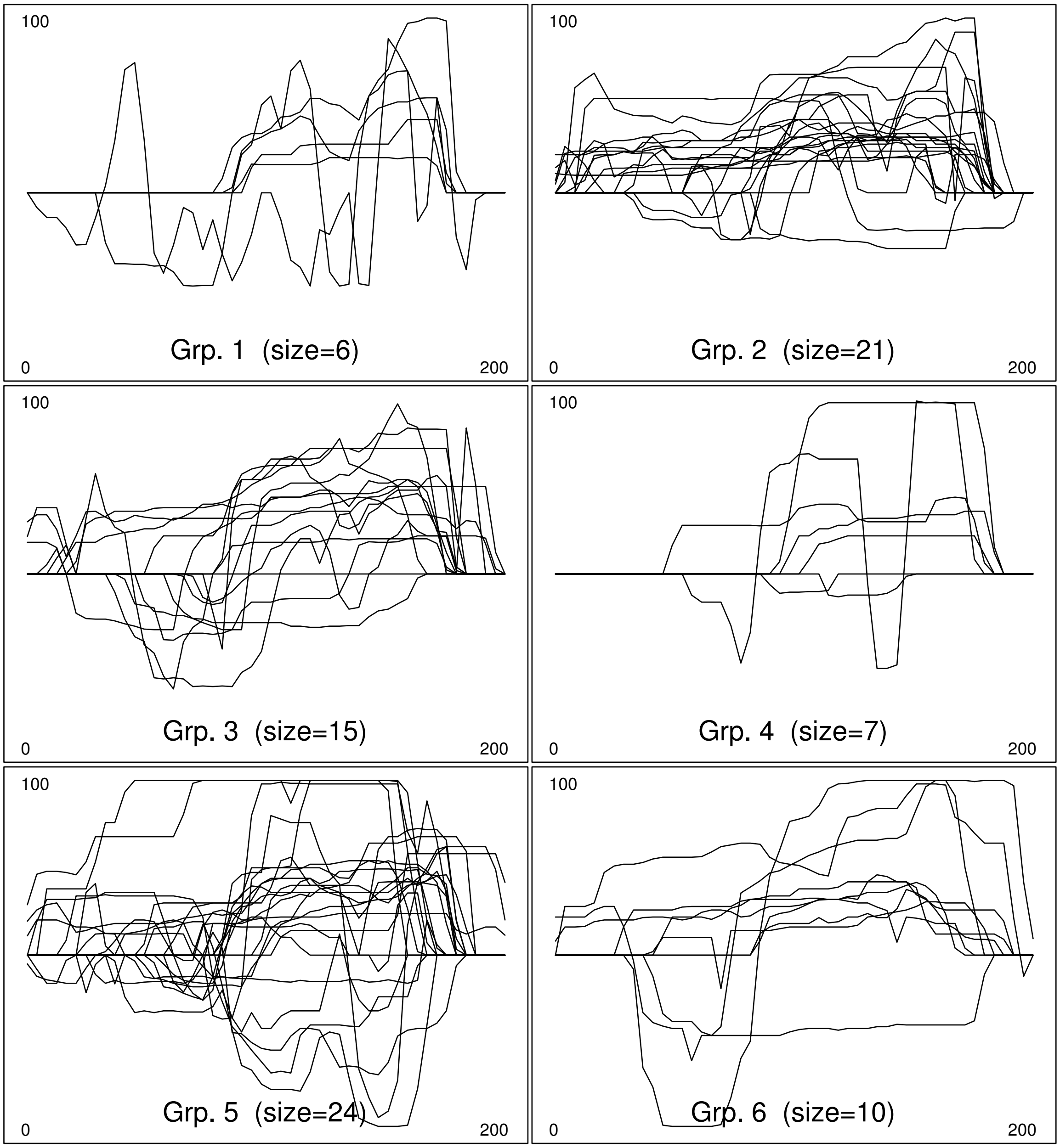

Figure 5: *Raw Pizza Commercial Data.*

Although the data look quite badly behaved, it was assumed that differences between groups should be formulated in terms of the mean curves of the groups. Thus, one is interested in simultaneous null bands for the six mean curves under the null assumption of absent group differences. The test statistics are effectively the group means at the 50 reduced time points in all 6 groups, amounting to 300 test statistics all together.

The null assumption of identical distributions of the curves in all groups suggests the use of a permutation distribution whereby the curves are randomly assigned group memberships. The permutation distribution was approximated by 10,000 randomly permuted datasets. Figure 6 shows mean curves for 10 of these random datasets. Note that the mean curves of the smaller groups have greater variability, as would be expected. Consequently, null bands should be expected to be wider for small groups. It becomes apparent that most likely there

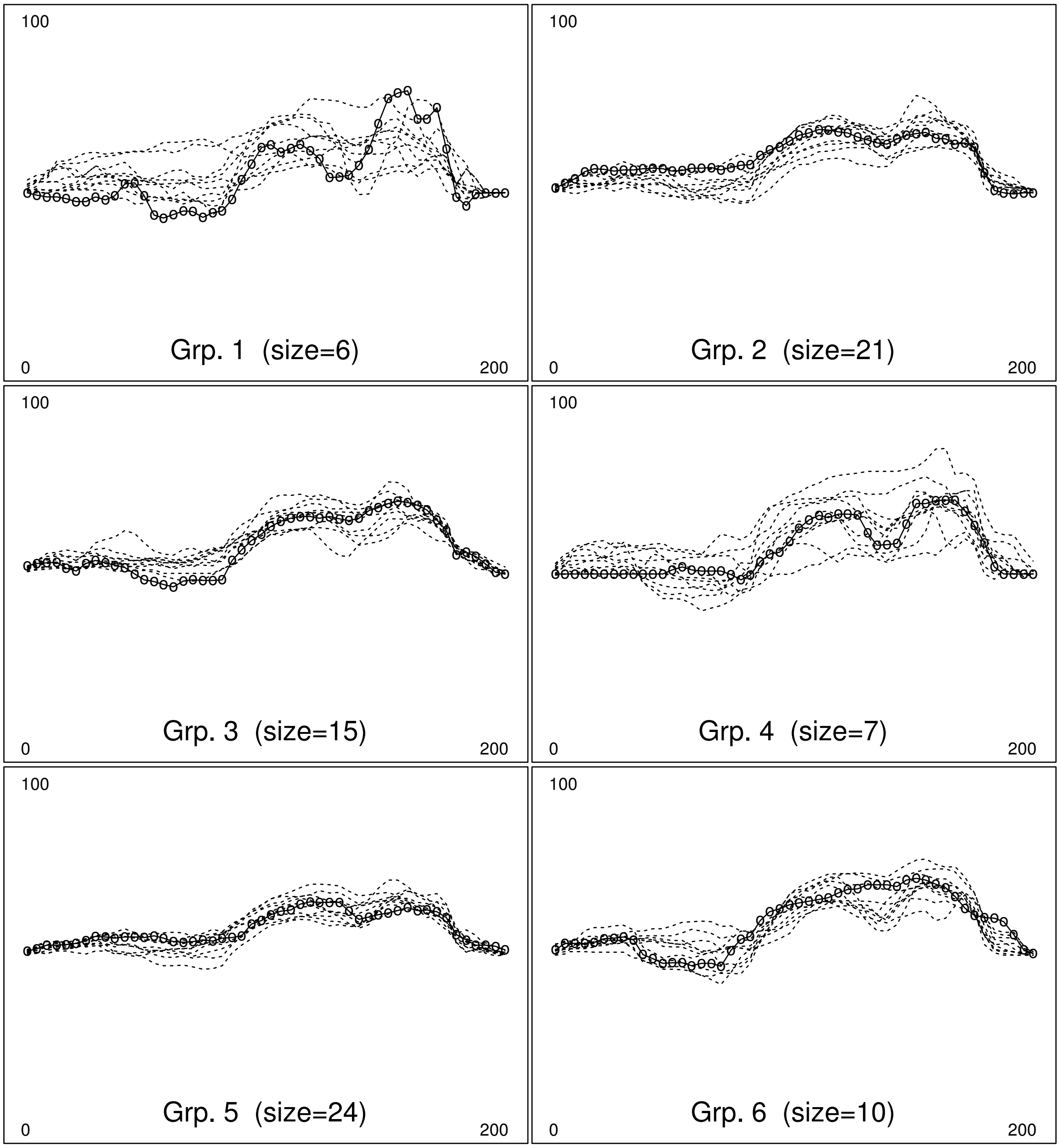

Figure 6: *Pizza Commercial Data: Mean curves of the actual data (solid with circles) and mean curves of ten randomly permuted datasets (dotted)*

is not much significance to be achieved for group differences among the mean curves.

From the 10,000 null datasets, estimated pointwise quantiles were obtained for each of the 300 means (by sorting the 10,000 mean values at each time point). The result are 6 pointwise quantile curves for any quantile level. The 10,000 null datasets are then re-used to estimate the simultaneous coverage for each pointwise quantile level. The desired simultaneous null coverage can thus be found by a simple search over the outer pointwise quantile levels.

Figure 7 shows three types of null bands at the 10% simultaneous level: a two-sided null band (the outer band) and an upper and a lower one-sided null bound. It appears that no mean curve achieves significance in the two-sided mode, and only few curves come close to significance in the considerably more lenient one-sided mode, most notably the early time

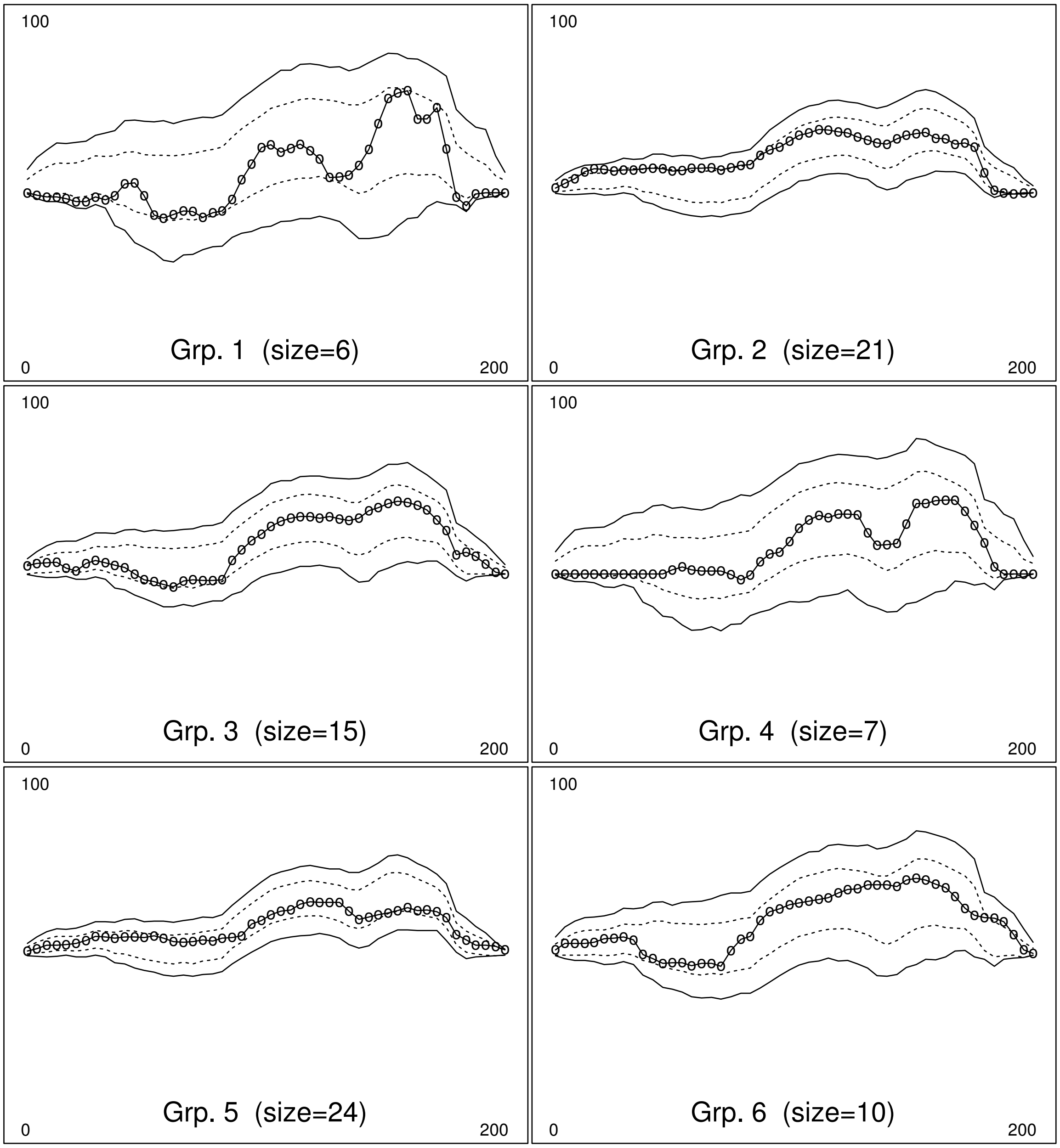

Figure 7: *Pizza Commercial Data: Mean curves and 90% simultaneous null bands. Solid: two-sided, dotted: one-sided.*

points of group 2.

Although it should be obvious, we note that the null bands are *not* confidence bands; it is only the absence of clear group differences that gives the visual appearance of bands accompanying the mean curves. One would have hoped that at least some of the observed mean curves would reach clearly outside the null band.

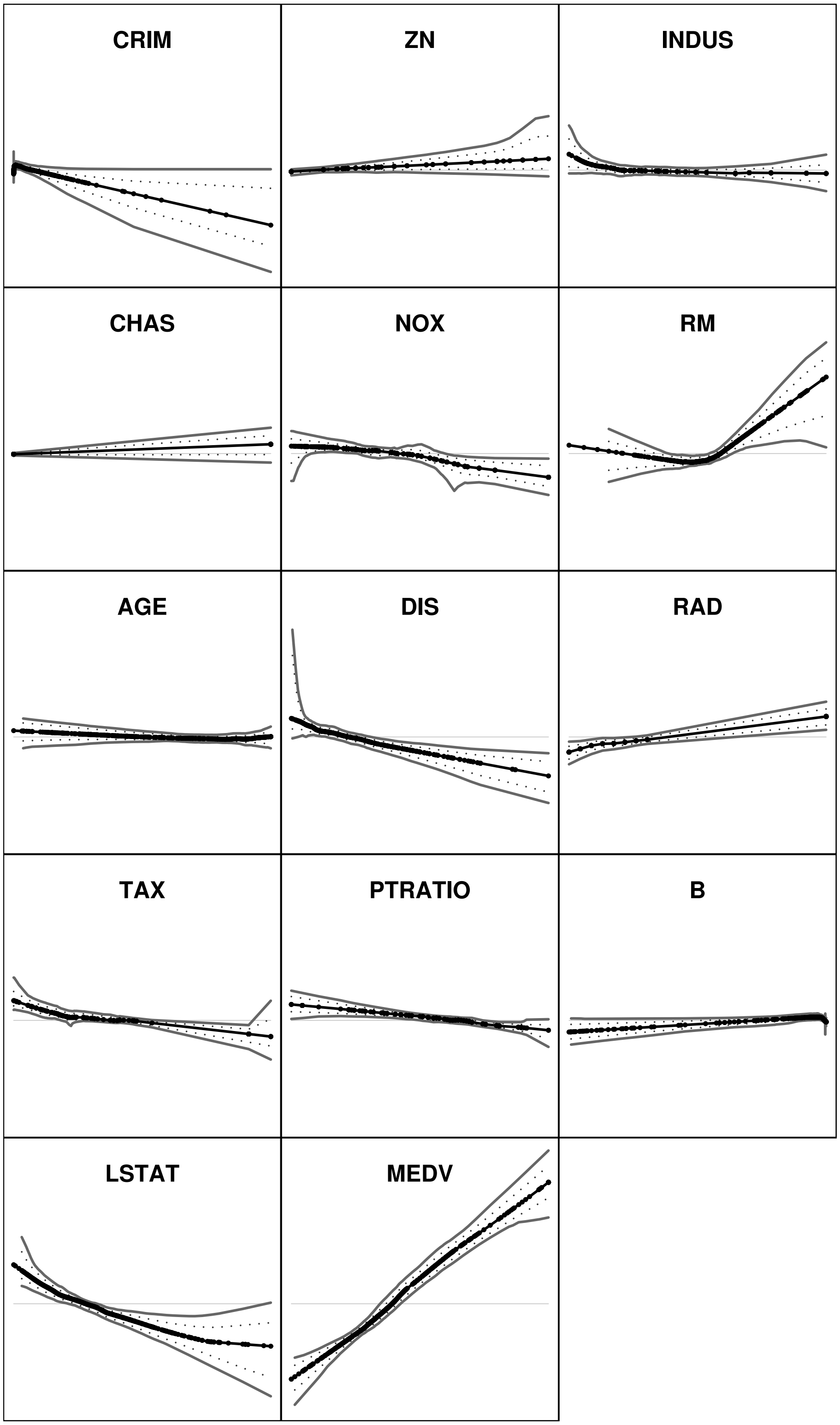

Figure 8: *Boston Housing Data: Bootstrap standard error bands for ACE transformations. The inner bands (dotted) have 95% pointwise coverage; the outer bands (solid gray) have 90% simultaneous coverage. The transformations are shown on identical vertical scales in order to show their importance graphically.*

5.4 Bootstrap Standard Error Bands for Smooths, Additive Fits, ACE Regression

Function estimates (often called "smooths" when smoothness is assumed) are usually shown with confidence bands or standard error bands or posterior bands. Which of the three terms applies depends on the coverage claims: confidence bands for the true underlying function, standard error bands for the expectation of the function estimates (thus bypassing the bias problem), posterior bands for Bayesian posterior uncertainty. For most major estimation methods (kernel smooths, smoothing splines, regression splines,...) pointwise coverage bands are quite easily derived (for a concise and very readable reference see Loader 1993; more complete is Sun and Loader 1994). Some smoothers, however, have no bands with known coverage properties, an example being Friedman and Stuetzle's (1982) supersmoother. But even for fixed-bandwidth smoothers coverage properties of bands are invalidated when subjected to bandwidth selection based on cross-validation. The situation complicates further when smoothers are used as building blocks of additive models, although here again coverage bands exist when fixed-bandwidth smoothers are used.

We illustrate CfS with bootstrap standard error bands for Friedman's original implementation of ACE regression (Breiman and Friedman 1985), which we treat as a black-box. Two complications make CfS particularly appealing for ACE: first, Friedman's implementation of ACE rests on the supersmoother, for which coverage bands may be difficult to justify by any other means; second, ACE is strictly speaking not a regression but a canonical correlation method, which puts it outside the analytical approaches suitable for regression methods.

Breiman and Friedman applied ACE to the well-known Boston Housing data, and so will we, although many objections can be raised against i.i.d.-based inference about data that is aggregate (census tracts), spatial (greater Boston), and not a sample (complete enumeration). Just the same, the point is to demonstrate feasibility of simultaneous bootstrap inference on a dataset that involves 14 data transformations on 506 objects.

Bootstrap has seen several proposals for CfS for nonparametric regression (Sections 1 and 2), and they all are cases of conditional bootstrap that conditions on the predictors and resamples residuals. By contrast, we favor when applicable the observational-data bootstrap that resamples cases with their full predictor-response vectors. Resampling cases mimics observational data better than resampling residuals, which ignores for example inhomogeneous variances or capped response values, which we find in the Boston Housing data where the response (median housing values, MEDV) seems to be capped at 50 thousand with a tie of 16 census tracts.

As always we replaced an uncountable number of function values with a finite number, here as follows: for the seven variables that take on fewer than 100 values, we computed intervals for the transformations at each observed value; for the other seven variables we subselected 100 values (for example, about every fifth order statistic if there are about 500 values). All together we were left with 1,006 different variable values at which we obtained standard error intervals for the variable transformations. We thus computed simultaneous coverage properties across the transformations of the response (MEDV) and all 13 predictors, and across evaluations of the transformations at up to 100 locations per variable

The construction of the simultaneous bootstrap standard error bands was as follows: 1) We used ACE fits on 9,999 bootstrap samples in order to approximate extreme pointwise quantiles at the 1,006 variable values. 2) We used ACE fits on another 9,999 bootstrap

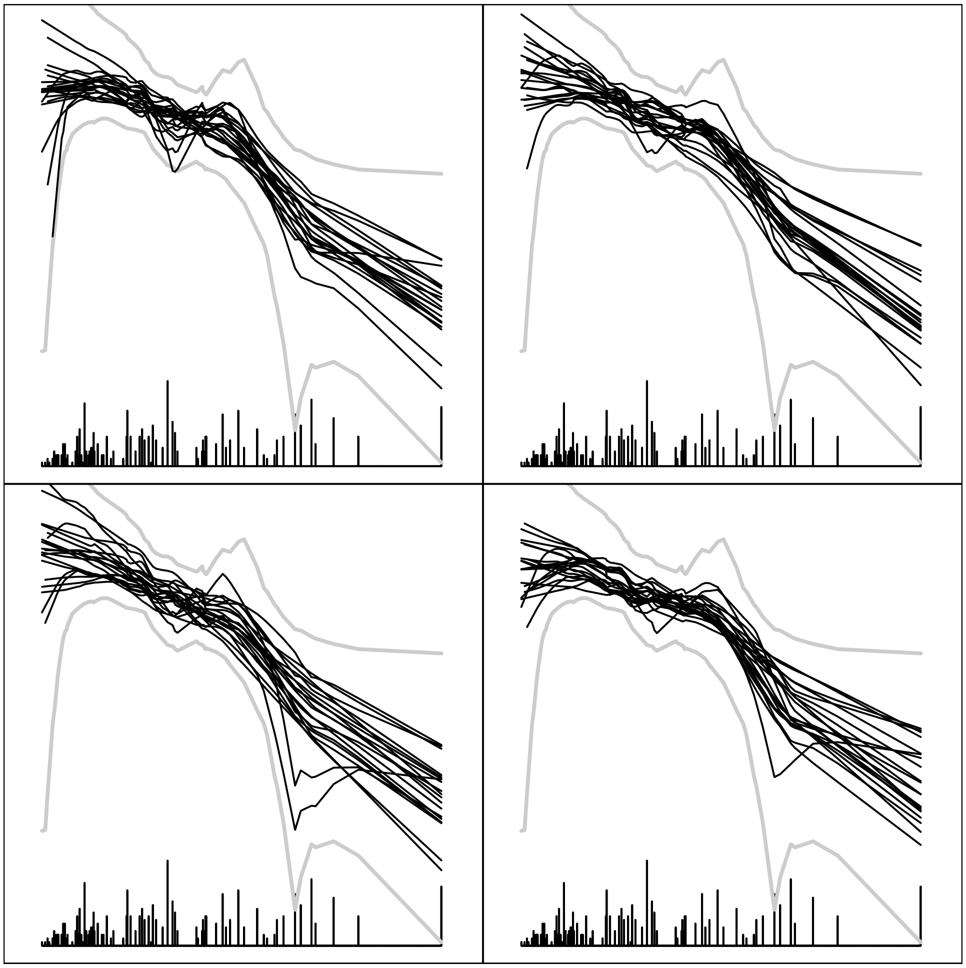

Figure 9: *Boston Housing Data: Four "spaghetti plots" of ACE transformations of the variable NOX. Each frame shows 20 bootstrap replications. Also shown are: the 90% simultaneous coverage band (gray lines), and a fine-grained histogram of the variable NOX, which takes on only 81 values for 506 cases. The highest spike corresponds to a tie of 23 cases.*

samples in order to obtain simultaneous coverage properties of the pointwise intervals. It turned out that a pointwise coverage of 99.94% was required in order to achieve an approximate 90% simultaneous coverage at the 1,006 variable values, corresponding to an equivalent Bonferroni factor of 167. With 99.94% pointwise coverage, Loader's (2001) warning against going too far out in the tails of the pointwise distributions may have some validity here; yet it seems preferable to use this imperfect crutch than relying on purely pointwise intervals.

In Figure 8 we show both the pointwise 95% and the simultaneous 90% standard error bands. In contrast to Breiman and Friedman (1985), we retained the full set of variables whereas they performed a stepwise backward variable selection procedure. A couple of observations: It appears that the response MEDV (Median Housing Values) does not require a nonlinear transformation. It is gratifying to see that the descending stretch of the transformation of RM (average number of rooms) is not significant, but some kind of kink in the center of the transformation seems to hold up. The band for NOX (an air pollution variable of primary interest) shows two cusps that are somewhat implausible; they are, however, "real" in the sense that the bootstrap transformations of NOX exhibit the most erratic behavior among all variables, in particular at the locations of the cusps which may have to do with the many ties present in this variable. The "spaghetti plots" in Figure 9 give an idea of the qualitative variety of transformations that are typical under bootstrap sampling. In our experience, spaghetti plots retain their usefulness and can never be fully replaced by coverage bands. They may indicate in this case that the real significance may exist in the first derivatives more so than in the raw function values.

Finally, note that we used the basic percentile bootstrap as opposed to, for example, the adjusted percentile bootstrap which is known to generally have better properties (Efron and Tibshirani 1993, Davison and Hinkley 1997). The reason is simply a desire to limit the complexity of computations. One should also keep in mind that we choose the width of the bootstrap intervals with CfS , which means that the pointwise percentile bootstrap is used for choosing the relative shapes of the intervals at different locations, not for selecting their actual widths. CfS may amount to a small degree of "borrowing strength" across locations by relying on all locations for selecting the ultimate widths in the one-parameter families of pointwise bootstrap percentile intervals.

5.5 Simultaneous Bootstrap Confidence Regions for Correlations

The data example in this section arose from a phone usage study at a large telecommunications firm. For a sample of 500 customers, three variables were recorded: the number of incoming and outgoing calls as well as the average monthly bill. The three variables had sizeable correlations as expected: 0.66 0.58 0.79 for the variable pairs (1,2), (1,3) and (2,3).

We used bootstrap to construct a confidence box with simultaneous 90% coverage for the three correlations. From 10,000 bootstrap samples we approximated marginal lower and upper quantiles. We then searched all boxes formed from intervals spanned by marginal lower and upper α-quantiles. It turned out that simultaneous 90% coverage was achieved by marginal $\alpha = 1.9\%$ quantiles, as opposed to 5% quantiles that yield the marginal 90% coverage intervals. The equivalent Bonferroni factor is 10/3.8≈2.6, which is not a huge savings compared to a Bonferroni adjustment with 3.

The result is depicted in Figure 10: The pairwise correlations of 10,000 bootstrap samples are plotted as three pointclouds on the same axes. The narrower marginal 90% intervals are shown as well as the projections of the wider simultaneous 90% coverage box. Figure 11 shows the same but in a parallel coordinate representation which is reminiscent of confidence bands for function estimates. The differences to function plots are that 1) there are only three abscissae, and 2) the abscissae do not form the domain of a function (they are just representations of some statistics).

Although a drop from the marginal 5% quantiles to the 1.9% quantiles to achieve simultaneity might seem like a substantial difference, in terms of absolute increase in size of the intervals on the correlation scale there is relatively little difference. This is not too surprising because the simultaneity problem among three variables is not as great as it is in problems with 20 or even 1,000 variables.

A point of possible confusion should be addressed: The correlated shapes of the pointclouds in Figure 10 express a different kind of correlation. The statistics considered here are the estimates of pairwise correlation. These statistics are correlated as they should be when for example the estimate of cor(1,2) and the estimate of cor(1,3) share variable 1. In summary: each point in Figure 10 expresses a bootstrapped correlation estimated from the data, while the correlated shape of the pointclouds expresses the correlation between the correlation estimates...

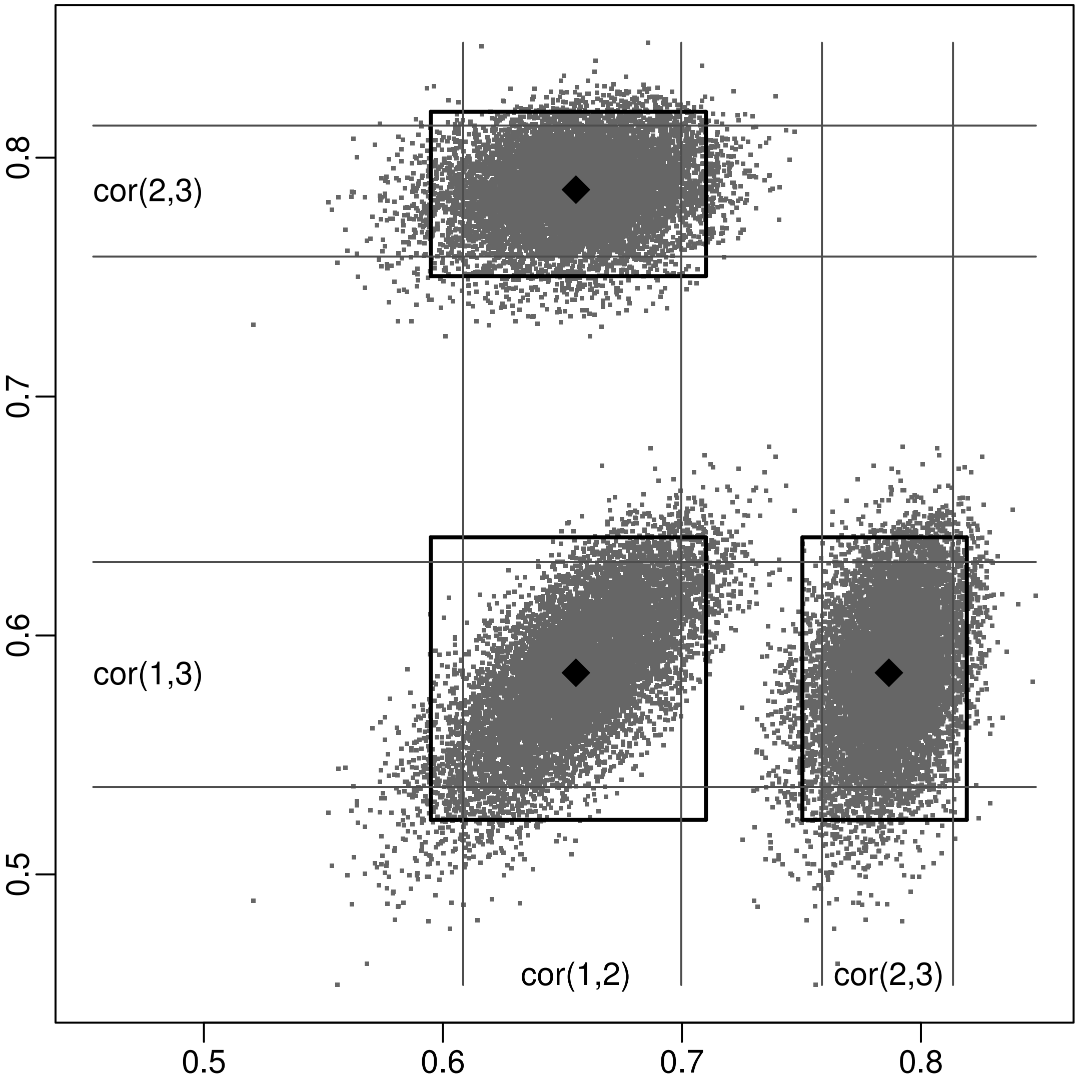

Figure 10: *Marketing data: pointwise 90% confidence intervals and simultaneous 90% confidence box for the correlations of three variables. The three pointclouds represent the pairwise views of the correlations of variable pairs (1,2), (1,3) and (2,3) in 10,000 bootstrap replications. The correlation estimates from the full data are shown as diamonds. Note that not all points inside the simultaneous confidence rectangles are inside the simultaneous confidence cube in 3-D; some points may fall outside in the third dimension.*

The final Figure 12 illustrates a different type of simultaneous confidence region that is not based on intervals: it shows a three-dimensional confidence ellipsoid in all three coordinate projections. The ellipsoid is computed from a principal component analysis of the three-dimensional bootstrap resamples, and it is calibrated for 90% coverage. The ellipsoid is adapted to the shape of the bootstrap resampling distribution in 3-D. In effect we may have constructed an approximate high-density coverage region (compare Knorr-Held, 2003) without density estimation.

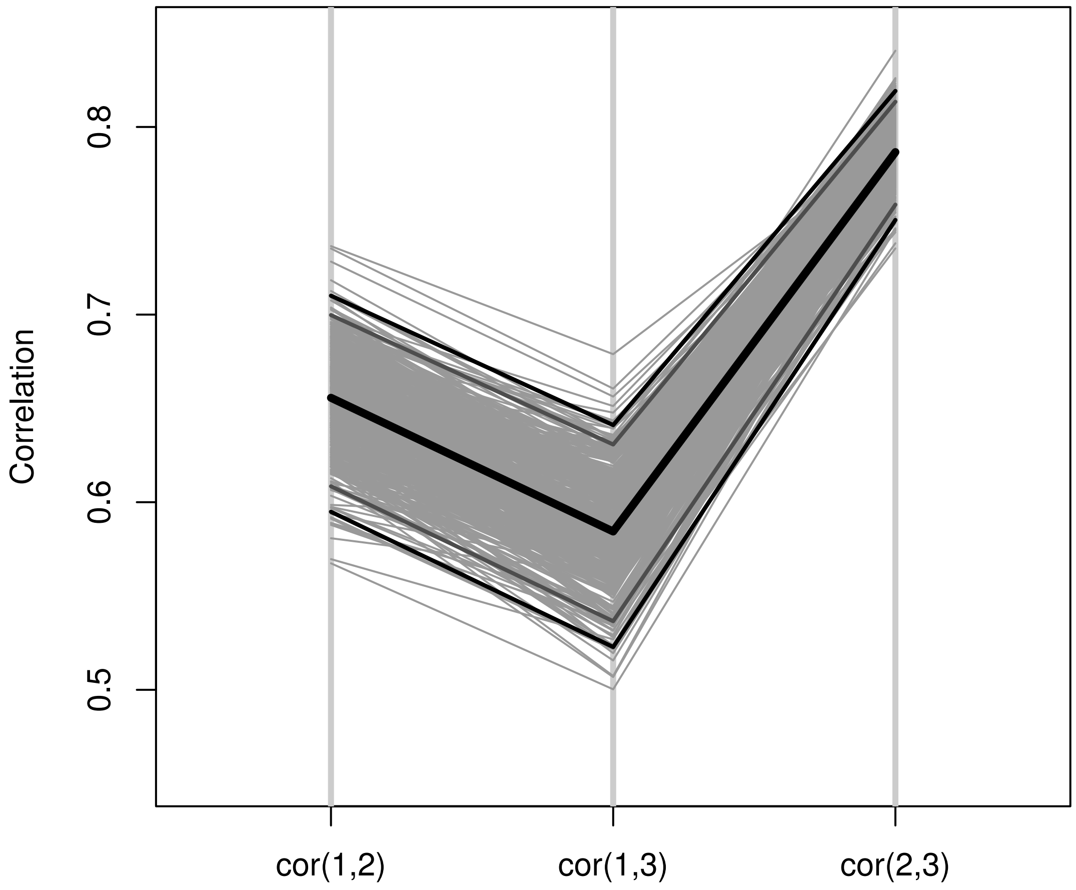

Figure 11: *Same as Figure 10 but in parallel coordinate representation.*

6 A Framework for Simultaneous Coverage Problems

6.1 Nested One-Parameter Families of Coverage Regions

The examples of Section 4 share these commonalities: In each case, one is given a random vector or random function, $X(t)$, with a *known* probability distribution. In order to avoid technicalities that play no role in practice, we assume a finite domain $T(\ni t)$ for $X(t)$. The random function $X = (X(t))_{t \in T}$ is therefore just a random vector with values in $\mathbb{R}^T$. A realization of X is written lower-case $x = (x(t))_{t \in T}$, and if it is an observed realization we write it as $x_o = (x_o(t))_{t \in T}$.

The general task is to find coverage regions $C \subset \mathbb{R}^T$ for which $Pr[X \in C] = 1 - \alpha$. We confine the search to one-parameter families $\{C_s\}_s$ of nested sets:

$$C_{s'} \subset C_{s"} \quad \text{for} \quad s' \leq s" \in S \ ,$$

where we assume that all C_s are closed subsets of $\mathbb{R}^T$, that the parameter set $S \subset \mathbb{R}$ is

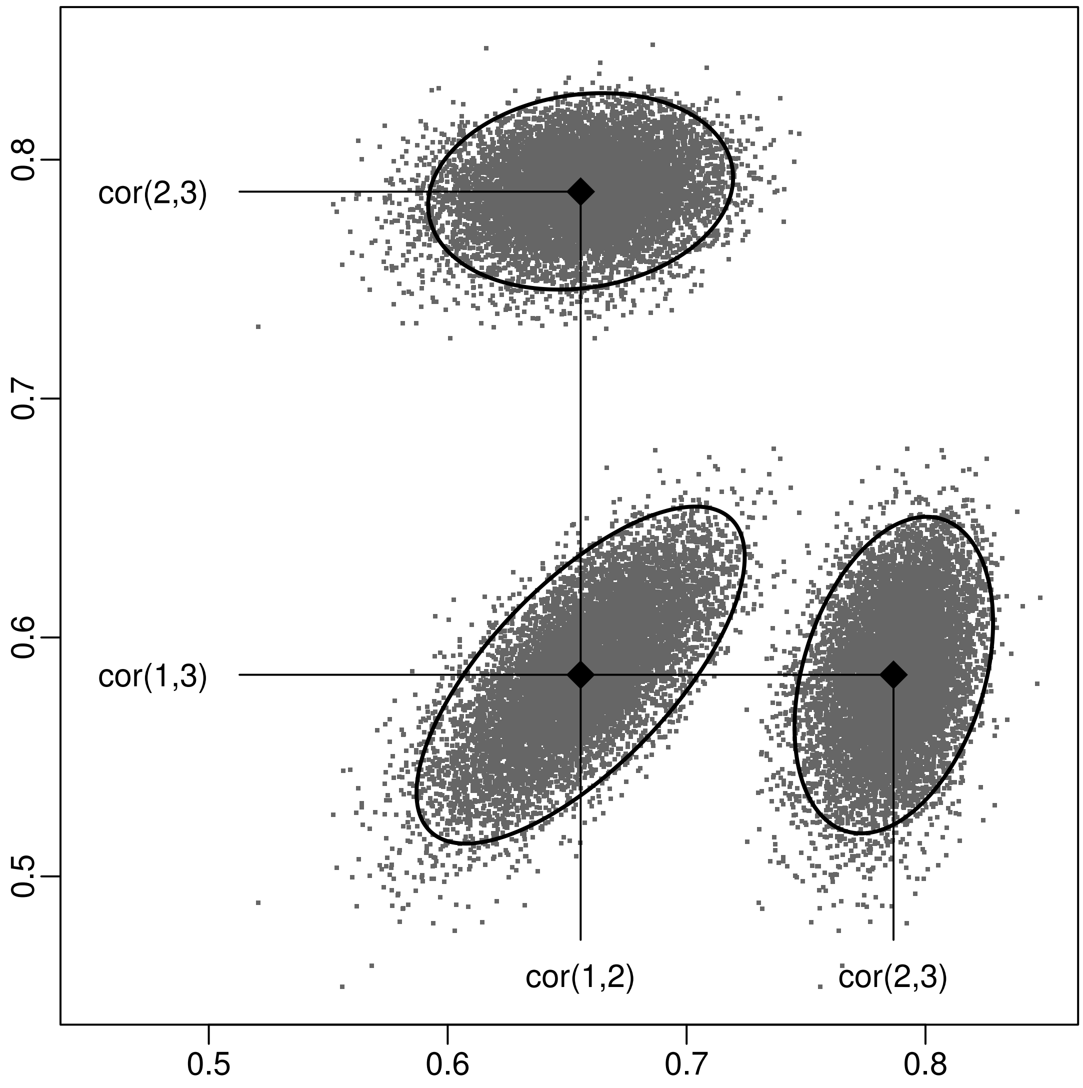

Figure 12: *Projections of a three-dimensional confidence ellipsoid with 90% coverage. The ellipsoid is adapted to the 3-D pointscatter by way of a principal component analysis.*

closed, and that the family is downward continuous:

$$C_s \;=\; \bigcap_{s'>s,\in S} C_{s'} \; .$$

Coverage probabilities will then also be downward continuous:

$$Pr[C_s] \;=\; \inf_{s'>s,\in S} \; Pr[C_{s'}] \; .$$

Such regularity conditions are practically irrelevant because the probability distributions P are estimated by discrete empirical measures obtained from MC or MCMC sampling, and the values of s are often confined to a finite grid. Regularity conditions would be relevant in proofs of consistency.

Now the calibrated choice for coverage $1-\alpha$ is

$$s_\alpha \;=\; \inf \; \{\, s' \mid Pr[C_{s'}] \geq 1-\alpha \,\} \; .$$

This does not guarantee exact coverage $1-\alpha$, but it is the closest possible approximation from above, which is conservative in the sense that the coverage probability never drops below the target value $1-\alpha$.

It is important to note that in practice the construction of the sets C_s is based on the observed $x_o(t)$ as well as the known distribution of $X(t)$: $C = C_{x_o,P}$. A simple example is in Section 5.5 where we construct a family of nested bootstrap confidence ellipsoids in 3-space. They can be interpreted as balls $C_t = \{\, x \,|\, \mathrm{dist}(x, x_o) \leq t\}$ in the Euclidean metric on the sphered bootstrap resamples. The resamples are not only used for calibration, but for shaping the neighborhoods as well.

Often, the sets are constructed as upper level sets of a function on $\mathrm{I\!R}^T$:

$$C_s \;=\; \{x \,|\, f(x) \geq \sup f - s\,\} \;.$$

[We assumed f bounded to use this parametrization which maintains our ordering convention. We also assumed f upper semi-continuous so C_s is closed.] An example are highest-density regions, where $f(x)$ is a density function, usually a density estimate of the distribution of X. An example in Bayesian inference are highest posterior density regions, investigated by Knorr-Held (2003).

Given a nested family C_s as above, one can construct a lower semi-continuous function $f(x)$ whose lower level sets are the sets C_s: If we let $f(x) = \min\{s \,|\, x \in C_s\,\}$, then $\{x \,|\, f(x) \leq s\,\} = C_s$. [The construction could be for upper level sets by reversing the parametrization.]

6.2 Nested One-Parameter Families of Coverage Bands and Boxes

We speak of coverage bands when $X(t)$ is a interpreted as a curve, and of coverage boxes when $X(t)$ is an arbitrary collection of variables. Bands and boxes are described by lower and upper bounding functions, $l(t)$ and $u(t)$, where $l(t) \leq u(t)$ $\forall t \in T$:

$$C \;=\; \{x \,|\, l(t) \leq x(t) \leq u(t),\; \forall t \in T\,\} \;.$$

The one-sided cases of upper or lower coverage regions are trivially accommodated by setting either $l \equiv -\infty$ or $u \equiv +\infty$. A t-like version of writing C is as

$$C \;=\; \{x \,|\; \max_t \frac{|x(t) - m(t)|}{b(t)} \leq 1\,\} \;,$$

where $m(t) = (u(t) + l(t))/2$ and $b(t) = (u(t) - l(t))/2$.

Nested one-parameter families of coverage bands/boxes are specified by parametrizing the bounding functions, $l_s(t)$ and $u_s(t)$, such that

$$l_{s"}(t) \leq l_{s'}(t) \leq u_{s'}(t) \leq u_{s"}(t) \qquad \forall s' \leq s" \;.$$

Then

$$C_s \;=\; \{\, x \,|\, l_s(t) \leq x(t) \leq u_s(t) \;\; \forall t\,\}$$

represents a particular form of nested family of coverage regions. [We may assume $l_s(t)$ upper and $u_s(t)$ lower semi-continuous in s to satisfy the regularity conditions of the previous section.] Calibration for (upper approximate) coverage probability $1-\alpha$ is given by

$$s_\alpha \;=\; \min\{\, s \,|\;\; \Pr[\, l_s(t) \leq X(t) \leq u_s(t),\; \forall t] \geq 1-\alpha\,\} \;.$$

Again, the coverage is not necessarily exactly $1-\alpha$, but it is the lowest attainable conservative coverage $\geq 1-\alpha$.

When estimating s_α in practice, one simulates a large number n_{sim} of realizations $X_\nu(t)$, $\nu = 1, ..., n_{sim}$, and uses the obvious Monte Carlo estimate $\hat{\Pr}$ of Pr:

$$\hat{\Pr}[\, l_s(t) \leq X(t) \leq u_s(t),\ \forall t \in T] \ = \ \frac{1}{n_{sim}} \#\{\nu |\ l_s(t) \leq X_\nu(t) \leq u_s(t),\ \forall t \in T\ \}$$

The estimate for s_α becomes

$$\hat{s}_\alpha \ = \ \min\{s \in \mathbb{R} \mid \hat{\Pr}[\, l_s(t) \leq X(t) \leq u_s(t),\ \forall t \in T] \ \geq \ 1-\alpha\ \}$$

By construction, the set $\{x \in \mathbb{R}^T |\ l_{\hat{s}_\alpha} \leq x(t) \leq u_{\hat{s}_\alpha}(t),\ \forall t \in T\ \}$ is a band with simultaneous coverage probability $1-\alpha$, up to simulation error.

6.3 Types of One-Parameter Families of Bounding Functions and Their Construction

The simplest example of a one-parameter family of bounding functions are constants such as

$$l_s(t) = 0,\ u_s(t) = s\ ,$$

in its upper one-sided version. This has been used by Ghosh (1996) and Ghosh and Beran (2000) in their T_3 method for one and two sample problems as alternatives to Q-Q plots.

More commonly, though, the bands are at least conceptually based on marginal or pointwise quantiles:

- The bounding functions $l_s(t)$ and $u_s(t)$ are the lower and upper pointwise s-quantiles, respectively, of the distribution of $X(t)$:

  $$l_s(t) \ = \ q_{X(t)}(s)\ , \quad u_s(t) \ = \ q_{X(t)}(1-s)\ , \quad 0 \leq s \leq 1/2\ , \tag{1}$$

  where the pointwise quantile functions $q_{X(t)}(s)$ are implicitly defined by $\Pr[X(t) \leq q_{X(t)}(s)] = s$ and assumed continuous and strictly monotone.

- If approximate marginal normality holds, one may let

  $$l_s(t) = \mu(t) - s \cdot \sigma(t)\ , \quad u_s(t) = \mu(t) + s \cdot \sigma(t) \quad (s \geq 0)\ ,$$

  where $\mu(t)$ is the mean function and $\sigma(t)$ the standard deviation function. This is equivalent to the pointwise quantile approach with $q_{X(t)}(s) = \mu(t) + \sigma(t)\Phi^{-1}(1/2+s)$.

The more important distinctions between types of coverage bands are in terms of construction and estimation:

- **Analytic bands:** In some cases families of bands can be derived analytically. Important examples are the following:

– Q-Q plots for a fixed distribution (1): In Q-Q plots the variables $X(t) = Y_{(t)}$ are the order statistics of univariate data. Their marginal distributions are known if the data are i.i.d. with fixed known cdf F because they can be reduced to the distribution of order statistics of a uniform distribution. The marginal distribution of the t'th order statistic of a uniform sample of size N in $[0, 1]$ is a Beta distribution $B(t, N - t + 1)$. Therefore

$$q_{X(t)}(s) = F^{-1}(F^{-1}_{B(t,N-t+1)}(s)) \ .$$

– Q-Q plots for a fixed distribution (2): Chambers et al. (1983, p. 229ff) gave approximate standard errors for empirical quantiles, to be used in normal null bands:

$$\mu(t) = q_t \ , \qquad \sigma(t) = \frac{\sigma_Y}{f(q_t)} \left(\frac{p_t(1-p_t)}{N} \right)^{1/2} \ ,$$

where $t = 1, ..., N$, $p_t = (t-1/2)/N$, $q_t = F^{-1}(p_t)$, and $f(q) = F'(q)$ is the density of the distribution. The standard deviation of the data, σ_Y, is irrelevant because the multiple will be chosen through CfS from the family

$$l_s(t), \ u_s(t) \ = \ \mu(t) \pm s \frac{(p_t(1-p_t))^{1/2}}{N^{1/2} \, f(q_t)} \ .$$

– Linear estimation in parametric and nonparametric regression: Estimates of linear model coefficients and many nonparametric regression fits with fixed bandwidths are linear in the responses. Under the assumption of normal errors with constant variance, one can then construct normal quantile-based standard error bands (or even confidence bands if one is willing to tackle bias). The reason is that $\sigma(t)$ can be derived up to a proportionality factor: if

$$X(t) = \sum_j \beta_j(t) Y_j = \langle \boldsymbol{\beta}(t), \boldsymbol{Y} \rangle$$

is a nonparametric fit at location t, then

$$\sigma^2(t) = \sum_j \beta_j^2(t) \ \sigma_Y^2 = \|\boldsymbol{\beta}(t)\|^2 \ \sigma_Y^2 \ ,$$

where σ_Y^2 is the same for all j. The specific value of σ_Y is irrelevant because we determine the width of the band with CfS , as in the preceding example.

Often one uses these analytic bands even in situations where they do not strictly apply. For example, the above Beta-based null bands for Q-Q plots are used when the data are standardized to zero mean and unit variance, even though standardization invalidates the derivation of these analytical forms. Similarly in nonparametric regression, one uses the above bands even if the bandwidth is selected with cross-validation, for example, which also invalidates their derivation.

The reason for stretching the reach of analytic bands is that they obviate the need for individual pointwise approximation of quantiles, which carries a burden of approximation error that may produce unaesthetic results in cases where smooth dependence on t across the underlying domain T is assumed.

- **Pointwise estimated bands:** We distinguish two cases, a direct case in which quantiles are estimated based on order statistics of an MC simulation, and a normal case in which quantiles are those of normal distributions for which means and variances are estimated pointwise.

  - **Direct quantile bands:** One approximates $l_s(t)$ and $u_s(t)$ by empirical quantiles of MC samples $X_\nu(t)$ ($\nu = 1$, ..., n_{sim}). These empirical quantiles are simply the pointwise sorted values or order statistics, $(X_{(\nu)}(t))_{\nu=1...n_{sim}}$, sorted separately for each $t \in T$.
  - **Normal quantile bands:** One approximates the mean profile $\mu_{X(t)}$ and variance profile $\sigma^2_{X(t)}$ by empirical means and standard deviations of MC samples, for each $t \in T$ separately.

  For the direct quantile bands there are only a finite number of quantile estimates available: $q_{s=\nu/(n_{sim}+1)}(t) = X_{(\nu)}(t)$. If $n_{sim} = 9,999$, for example, one has pointwise estimates of quantiles for probabilities that are multiples of 1/10,000. For any one of these 9,999 quantiles, one has a pointwise quantile curve which can be used for the construction of coverage bands: $l_s(t) = q_s(t)$ and $u_s(t) = q_{1-s}(t)$, for $s \in \{1/(n_{sim}+1), 2/(n_{sim}+1), ..., 1/2\}$. [This parametrization follows the reverse convention.]

Direct quantile bands as well as pointwise normal bands can suffer from unsightly roughness due to pointwise approximation. If the set T is a subset of a domain such as time or space, one might gain by slightly smoothing $\hat{\mu}_{X(t)}$, $\hat{\sigma}_{X(t)}$, or $\hat{q}_s(t)$ over T, rather than completely relying on separate approximations for each $t \in T$. In the worked examples below, though, we did not take advantage of smoothing.

Simple early versions of direct quantile bands are Ripley (1981, for cdf's) and Atkinson (1981, for Q-Q plots of regression residuals). Their envelopes formed from minima and maxima of small numbers of MC simulations (Atkinson $n_{sim} = 19$) are crude estimates of tail quantiles. Direct quantile bands were also proposed by Landwehr, Pregibon and Shoemaker (1984) with $n_{sim} = 25$ for P-P plots as a diagnostic tool for logistic regression. With simulations so small, it is impossible to even entertain the idea of CfS . Two later examples where CfS on direct quantile bands were used are Härdle and Marron (1991) for confidence bands based on resampling nonparametric regression residuals, and Besag et al. (1995) for simultaneous posterior credible regions in Bayesian inference.

6.4 Computation of Direct Quantile Bands

Direct quantile bands can be computationally intensive, in particular when the number $|T|$ of quantities $X(t)$ is large. In one of our examples, bootstrap standard error bands for ACE regression, $|T|$ is in excess of 1,000, for which we performed over 20,000 ACE runs using a high-level language (R, see http://www.r-project.org/). At the time of writing this took in the order of an hour or two to compute.

Depending on the size of T, we use different ways of computing coverage bands: a memory-intensive method for small $|T|$, a memory-saving method for larger $|T|$. They have different advantages: using more memory is not only faster but permits the computation of greater detail; using relatively less memory permits the computation of considerably larger problems.

In what follows we denote the number of simulation runs by n_{sim}. At typical number would be $n_{sim} = 9,999$. The reason for the odd number is that n_{sim} order statistics make $n_{sim} + 1$ equally likely spacings and hence are estimates of quantiles that are multiples of $1/(n_{sim} + 1)$.

- Memory-intensive method: We store equally spaced quantiles across the whole range of each variable. To this end, allocate a matrix of size $n_{sim} \times |T|$, so that each column stands for one variable $X(t)$. Fill each row with the results of one simulation. Then sort the columns. This destroys the relationships across variables but yields for each variable a set of order statistics that are estimates of marginal or pointwise quantiles. The ν'th row will contain the estimates for the $\nu/(n_{sim} + 1)$-quantiles. This is part of the algorithm described by Besag et al. (1995, sec. 6.3).

- Memory-saving method: We store order statistics only for the extremes of each variable. Hence a decision has to be made how many extreme order statistics should be kept. For $n_{sim} = 9,999$, for example, we may limit ourselves to order statistics corresponding to quantiles below 1.5% and above 98.5%. This may seem excessively little, but given that we look for simultaneous coverage it is realistic to expect the desired pointwise quantiles to be somewhat extreme. Continuing for concreteness with a figure of 1.5%, the number of order statistics on each side will be $n_{ord} = 150$. The algorithm processes small batches of simulations one at a time, hence we also have to choose a batch size, such as $n_{batch} = 100$. We allocate a matrix with the number of rows equal to $n_{ord} + n_{batch} + n_{ord} = 150 + 100 + 150 = 400$, and the number of columns equal to the number of variables $|T|$.The algorithm is as follows:

  - Initialize the 400 rows of the matrix with simulation results, one simulation per row.
  - Sort each column of the matrix.
  - Let $n_{rep} = \text{ceiling}((n_{sim} - 2n_{ord} - n_{batch})/n_{batch}) = 96$. This is the number of batches to be processed. Also let $n_{rem} = (n_{sim} - 2n_{ord} - 1) \mod n_{batch} + 1 = 99$. This is the size of a remainder batch in case the numbers n_{sim}, n_{batch} and n_{ord} do not make a whole number of batches (which is almost always the case).
  - Repeat $n_{rep} = 96$ times:
    - Fill the $n_{batch} = 100$ center rows (here 151 to 250) with as many simulation results.
    - Sort each column of the matrix. This moves very large values in the current batch to rows 251...400, and very small values to rows 1...150.

    In one repetition fill only a remainder batch of $n_{rem} = 99$ rows out of $n_{batch} = 100$ to achieve exactly $n_{sim} = 9,999$ simulations.

  The first and last 150 rows of the matrix will contain the extreme order statistics of the variables. The working memory for the batches can be discarded at the end. We may name this algorithm "RSS", for "**R**epeatedly **S**imulate and **S**ort."

Strictly speaking, the memory-intensive method is a special case of RSS, whenever $2n_{ord} + n_{batch} = n_{sim}$. A comparison of the two methods shows the following:

- With the given numbers, the RSS algorithm requires a factor of 25 (≈400/9,999) less memory. This permits us to compute problems with 25 times more variables $X(t)$. For example, the first method may accommodate 40 variables, the second 1,000 variables with the same memory usage.

- RSS computations are somewhat more time-consuming, but the repeated sorts are actually not as laborious as it may seem because the top and bottom n_{ord} entries of each column are in sorted order at all times, which helps the sorting algorithm.

- The RSS algorithm can be restarted. For example, $n_{sim} = 9,999$ can be increased to $n_{sim} = 19,999$ by running another 10,000 simulations in 100 batches. However, the quantile range shrinks from $n_{ord}/(n_{sim} + 1) = 1.5\%$ to $.75\%$ because the number n_{ord} of order statistics on each side cannot be increased.

- The major advantage of the memory-intensive method is to permit simultaneous (=adjusted) p-value computations (Section 7) of all sizes because it stores all quantile estimates. The RSS method does so only for the chosen extreme ranges on either side.

6.5 Computation of CfS

At this point we are given a one-parameter family of bounding functions. The way it may be given is either as computable functions in case of analytical bands and normal bands, or as two matrices of upper and lower bounding values in case of direct quantile bands. These matrices are the result of the computations described in the previous subsection. They contain $l_s(t)$ and $u_s(t)$, respectively, for a selection of parameter values s. For direct quantile bands computed with RSS they are of size $n_{ord} \times |T|$. Even if the the bands are analytic, though, it may be more efficient to form an l- and a u-matrix with bounding vectors for a grid of quantiles such as the outermost 150 quantiles on each side for probabilities that are multiples of 1/10,000. Greater precision is rarely meaningful. If l and u are stored in a matrix, the parameter to be searched is effectively the row number.

One needs next a set of simulations for calibration. Two situations are possible:

- Simulations are relatively cheap and a new set of simulation runs for calibration can be afforded.

- Simulations are expensive and have been stored in a matrix of size $n_{sim} \times |T|$. This may be the same set of simulations that were used to construct direct quantile bands by sorting the columns according to the memory-intensive method.

In either situation, we have a series of simulations of length denoted again by n_{sim}, and results denoted again by $X_\nu(t)$ ($\nu = 1...n_{sim},\ t \in T$).

Calibration can now be computed as follows:

- For each simulated X_ν determine the minimal parameter $s = s_\nu$ for which $l_s(t) \le X_\nu(t) \le u_s(t)$ simultaneously for all $t \in T$. This can be done efficiently with bisection which requires searching $\log_2(n_{ord})$ parameter values. For $n_{ord} = 150$, for example, this means searching about 7 values.

- For the collection of parameter values $(s_\nu)_{\nu=1...n_{sim}}$ determine the upper $1-\alpha$ quantile. This will be the estimate $\hat{s}_\alpha$ for a band with coverage probability minimally $\ge 1-\alpha$: $[\,l(t),\, u(t)\,] = [\,l_{\hat{s}_\alpha}(t),\, u_{\hat{s}_\alpha}(t)\,]$.

Besag et al. (1995, sec. 6.3) give the following alternative computation in case the simulated data is completely stored and used both for the construction of direct quantile bands and calibration: They compute the rank transformations of the columns of the matrix at the same time that they sort them. Denoting the rank of $X_\nu(t)$ in the t'th column by $R_\nu(t)$, they form the set of values

$$\{ \max(\max_t R_\nu(t),\, n_{sim} + 1 - \min_t R_\nu(t) \mid \nu = 1, ..., n_{sim} \} \ ,$$

which is equivalent to our set $(s_\nu)_{\nu=1...n_{sim}}$. If n^* is the α-quantile of this set, then $[l(t),\, u(t)] = [X_{(n_{sim}+1-n^*)}(t),\, X_{(n^*)}(t)]$ is the same band found above.

7 Simultaneous P-Values for One-Parameter Families of Coverage Bands

Coverage bands in the context of testing amount to non-rejection ("acceptance") regions. Non-rejection and rejection regions, however, are in practice the less frequent way of expressing testing results. More often, one uses p-values, because they express test results at all possible significance levels. P-values work well for pointwise testing, but for simultaneous testing they are problematic. The problem of adjusting p-values for simultaneous inference is the topic of a book by Westfall and Young (1993). While their book is concerned with many cases of simultaneous inference, including multiple nested and non-nested null hypotheses, we are concerned here with only one case: multiple test statistics under a single null hypothesis. For us, however, testing is just one of several simultaneous coverage problems.

It turns out that one can define adjusted p-values not only for null bands in testing problems, but for all types of simultaneous coverage problems, including confidence bands and posterior bands. [We will say more on Bayesian p-values at the end of this section.] These bands require a test profile which we denote by $X_o = (X_o(t))_{t \in T}$. A p-value with regard to a set of coverage bands will indicate how extreme X_o is with regard to the underlying (null, bootstrap, posterior, ...) distribution of $X = (X(t))_{t \in T}$. To make this work, it is again essential to limit oneself to a one-parameter family of nested bands. We describe pointwise, simultaneous, and adjusted p-values in turn:

- The usual pointwise p-value at t is defined as follows: Let $s_t(X)$ be the parameter for the smallest interval $[l_s(t),\, u_s(t)]$ that contains $X(t)$:

  $$s_t(X) \;=\; \min\{s \mid l_s(t) \le X(t) \le u_s(t) \} \ .$$

  Upper and lower semicontinuity assumptions on $s \mapsto l_s(t)$ and $s \mapsto u_s(t)$, respectively, grant the existence of the minimum. The *pointwise p-value at t* is

  $$\mathrm{pval}_t(X_o) \;=\; \Pr[s_t(X) \ge s_t(X_o)] \;=\; 1 - \Pr[\, l_{s_t(X_o)} \le X(t) \le u_{s_t(X_o)}] \ .$$

- The simultaneous p-value for all of T is defined as follows: Let $s_{sim}(X)$ be the parameter for the smallest band $[l_s,\, u_s]$ that contains all of X:

  $$s_{sim}(X) \;=\; \min\{s \mid l_s(t) \le X(t) \le u_s(t),\ \forall t \in T \} \ .$$

In particular, the band corresponding to $s_{sim}(X)$ covers all the variables. More generally one has

$$s_{sim}(X) \leq s \quad \Longleftrightarrow \quad l_s(t) \leq X(t) \leq u_s(t), \ \forall t \in T \ .$$

The *simultaneous p-value* is obtained for $s = s_{sim}(X_o)$:

$$\text{pval}_{sim}(X_o) \ = \ \Pr[s_{sim}(X) > s_{sim}(X_o)] \ = \ 1{-}\Pr[\, l_{s_{sim}(X_o)} \leq X(t) \leq u_{s_{sim}(X_o)}, \ \forall t \in T] \, .$$

where X_o is fixed, and $Pr[...]$ refers to the random variable X which follows the underlying distribution.

- It may seem a contradiction in terms, but adjusted p-values are pointwise yet simultaneous. As for pointwise p-values, let $s = s_t(X_o)$ be the parameter corresponding to the smallest interval at t that contains $X_o(t)$. While the pointwise p-value is the complement of the marginal coverage probability at t alone, the *adjusted p-value* at t is the complement of the simultaneous coverage probability of the whole band $[l_s, \, u_s]$, not just the interval $[l_s(t), \, u_s(t)]$ at t:

  $$\text{pval}_{t,adj}(X_o) \ = \ 1 - \Pr[\, l_{s_t(X_o)} \leq X(\tau) \leq u_{s_t(X_o)}, \ \forall \tau \in T] \ .$$

  Thus the adjusted p-value at t adjusts the pointwise p-value for the presence of the other statistics. This agrees precisely with the notion of p-value adjustment in the literature (Westfall and Young 1993, Sec. 1.3).

Adjusted p-values are always less than (or equal to) pointwise p-values, and the simultaneous p-value is the smallest of the adjusted p-values:

$$\text{pval}_{t,adj}(X_o) \geq \text{pval}_t(X_o) \ , \qquad \text{pval}_{sim}(X_o) \ = \ \min_{t \in T} \ \text{pval}_{t,adj}(X_o) \ .$$

We see that the role of one-parameter families in p-value calculations is two-fold:

1. One-parameter families of nested intervals provide a notion of extremeness: in order to measure how extreme $X_o(t)$ is, one obtains the narrowest interval $[l_s(t), \, u_s(t]$ that contains $X_o(t)$ and calculates either the coverage probability of this interval or its complement, which is the pointwise p-value.
2. One-parameter familes of nested bands provide a link between coverage intervals across variables by imposing the simultaneous use of intervals $[l_s(t), \, u_s(t)]$ and $[l_s(t'), \, u_s(t')]$ at locations t and t'. Therefore, if one asks in p-value fashion what the probability of observing something more extreme than $X_o(t)$ is, one needs to consider not only the narrowest interval $[l_s(t), \, u_s(t]$ at t, but all intervals $[l_s(t'), \, u_s(t']$ with that parameter s at all locations t' and calculate the simultaneous coverage or its complement, which is the adjusted p-value.

Obviously some one-parameter families of coverage bands are more plausible than others. With p-value technology in mind, quantile bands seem particularly attractive because they link intervals with identical pointwise coverage across all locations.

Approximations of p-values can be obtained as always by approximating probabilities Pr with relative frequencies $\hat{\Pr}$ from MC sampling.

The interpretation of p-values in the context of frequentist testing is clear, but less clear is their meaning for confidence bands or posterior bands. In fact, their interpretation is quite different. The proper interpretation follows from basic observations:

- The bootstrap estimates the distribution of parameter estimates based on data.
- The posterior distribution is weighted evidence about the parameters after observing data.

Therefore, in both cases $X_o(t)$ is a possible parameter value at t, as opposed to the usual observed test statistic at t. The use of $X_o(t)$ as parameter values of interest in bootstrap and posterior inference is practical and useful, particularly in model selection: submodels can often be expressed by zeroing specific model parameters, which means in current notation: $X_o(t) = 0$ for a specific t. It is then natural to ask how extreme such a parameter is, and it is natural to answer with a bootstrap or posterior p-value. In either case one asks how wide an interval has to be to cover $X_o(t) = 0$. For the interval with the required width one can calculate a bootstrap or posterior coverage probability or its complement, a bootstrap or posterior p-value.

Simultaneous inference enters when many parameters are assessed with a joint bootstrap or posterior distribution, in which case a pointwise p-value for $X(t)$ has to be adjusted for the presence of the other $X(t')$ that could also have produced small pointwise p-values. For example, a smooth curve fitted at locations $t \in T$ might be essentially straight; it might permit a straight line (represented by X_o) to run simultaneously through all intervals with one or two exceptions. The significance of those exceptions can obviously not be assessed without taking into account the whole range of fitted t's that could also have produced exceptions.

[We conclude with remarks on *Bayesian p-values.* We start by noting that the posterior p-values defined here are *not* identical with the posterior predictive p-values found in the literature (see Gelman et al. (1995) or Carlin and Louis (1996), and the references therein). The short version is that our posterior p-values are for testing submodels against the current model, whereas posterior predictive p-values are for testing the current model against an unspecified larger model. Here is why:

Posterior predictive p-values are based on a test quantity $T(y, \theta)$, a function of the data and the model parameters. Extremeness of the observed data y_o in relation to a model $p(y|\theta)$ and prior $p(\theta)$ is judged in terms of $\Pr[T(y,\theta) > T(y_o,\theta) \mid y_o\,]$, where the probability is w.r.t. the posterior of θ and the corresponding predictive distribution of y — hence the name "posterior predictive p-value". This type of p-value yields a test of the present model against unspecified larger models that are hinted at by $T(y, \theta)$. Gelman et al. (p. 170f) illustrate this with binary data assumed i.i.d. Bernoulli and T equal to the number of runs (-1), so that the posterior predictive p-value measures evidence of correlation among the binary observations, which is the unspecified larger model. The test quantity T is then not even a function of θ.

By contrast, the present posterior p-values are based solely on the parameters and their posterior distribution, without the predictive distribution of the data. Recall that a one-parameter family of bands defines implicitly a test quantity, namely: $T(\theta) = \inf\ \{\ s \mid l_s(t) \leq \theta(t) \leq u_s(t)\ \}$. Our posterior p-value is $\Pr[\ T(\theta) > T(\theta_o)\ ]$, which defines a test of a hypothetical parameter θ_o within the current model and in light of the posterior distribution.

We finally note a possible wrinkle in posterior predictive model checks: they also run into simultaneity issues if they use multiple test quantities, $X(t) = T_t(y, \theta)$. Naturally, CfS applies here as well, and the resulting p-values will indeed be pointwise, simultaneous and adjusted posterior predictive p-values.]

8 Conclusions

We surveyed a method for simultaneous inference based on one-parameter families of coverage regions and calibration for simultaneity (CfS). The method is of great generality, both in terms of types of inference and types of objects to which it applies. CfS has appeared many times in the literature, but generally with a particular application in mind; pointing out its generality and reach is one of the main purposes of this article. Computations of coverage regions with almost no parametric assumptions is currently possible for more then 1,000 variables. While some caution is needed about the fact that simultaneous regions tend to reach far out in marginal distributions, the method can be used in two ways, either for finding regions with given simultaneous coverage or, if this is unrealistic, for estimating the simultaneous coverage of a given realistic region. CfS is another example for the advantages of sampling/resampling-based methods: most simultaneity problems are theoretically intractable, but easily solved with simulations.

References

[1] Atkinson, A.C. (1981), "Two Graphical Displays for Outlying and Influential Observations in Regression," *Biometrika*, **68**, pp. 13-20.

[2] Atkinson, A.C. (1987), *Plots, Transformations, and Regression - An Introduction to Graphical Methods of Diagnostic Regression Analysis* Oxford, UK: Oxford University Press.

[3] Becker, R.A., Chambers, J.M., Wilks, A.R. (1988), *The New S Language*, Pacific Grove, CA: Wadsworth & Brooks.

[4] Besag, J.E., Green, P.J., Higdon, D.M., and Mengersen, K.L. (1995), "Bayesian Computation and Stochastic Systems" (with discussion), *Statistical Science*, **10**, pp. 3-66.

[5] Breiman, L., and Friedman, J.H. (1985), "Estimating Optimal Transformation for Multiple Regressin and Correlation," *J. of the American Statistical Association*, **80** (391) pp. 580-598.

[6] Buja, A. (2003), "Testing Sufficiency: Exact Tests as Generalized Randomization Tests," see www-stat.wharton.upenn.edu/˜buja/ for a preprint.

[7] Buja, A. and Eyuboglu, N. (1993), "Remarks on Parallel Analysis," *Multivariate Behavioral Research*, **27** 4, pp. 509-540.

[8] Buja, A., Asimov, D., Hurley, C., and McDonald, J.A. (1988), "Elements of a viewing pipeline for data analysis," in *Dynamic Graphics for Statistics*, eds. W.S. Cleveland and M.E. McGill, Belmont, CA: Wadsworth, pp. 277-308.

[9] Carlin, B.P., and Louis, T.A. (1996), *Bayesian and Empirical Bayes Methods of Data Analysis*, New York: Chapman & Hall.

[10] Chambers, J.M., Cleveland, W.S., Kleiner, B., Tukey, P.A. (1983), *Graphical Methods for Data Analysis*, Boston: Duxbury Press.

[11] Chaudhuri, P., and Marron, J.S. (1999), "SiZer for Exploration of Structure in Curves," *Journal of the American Statistical Association*, **94**, pp. 807-823.

[12] Cook, D.R., and Weisberg, S. (1982). *Residuals and Influence in Regression*, New York and London: Chapman and Hall.

[13] Cook, D.R., and Weisberg, S. (1999). *Applied Regression Including Computing and Graphics*, New York and London: Wiley.

[14] Daniel, C., and Wood, F.C. (1980), *Fitting Equations to Data: Computer Analysis of Multifactor Data*, republished 1999 as part of Wiley Classics Library, New York: Wiley.

[15] Davison, A.C., and Hinkley, D.V. (1997), *Bootstrap Methods and their Application*, Cambridge: Cambridge University Press.

[16] Efron, B. (1983), "Estimating the Error-Rate of a Prediction Rule: Some Improvements on Cross-Validation." *Journal of the American Statistical Association*, **78**, pp. 316-331.

[17] Efron, B., and Tibshirani, R. (1993), *An Introduction to the Bootstrap*, New York and London: Chapman and Hall.

[18] Faraway, J. (1990), "Bootstrap Selection of Bandwidth and Confidence Bads for Nonparametric Regression," *Journal of Statistical Computation and Simulation*, **37**, pp. 37-44.

[19] Friedman, J.H., and Stuetzle, W. (1982), "Supersmoother," Technical Report, Dept. of Statistics, Stanford Univ.

[20] Fan, J., and Lin, Sh.-K. (1998), "Test of Significance when Data are Curves," *Journal of the American Statistical Association*, **93**, pp. 1007-1021.

[21] Gelman, A., Carlin, J.B., Stern, H.S., and Rubin, D.B. (1995), it Bayesian Data Analysis, New York: Chapman & Hall.

[22] Ghosh, S. (1996), "A new graphical tool to detect non-normality," *Journal of the Royal Statistical Society*, Series B, **37**, pp. 37-44

[23] Ghosh, S., and Beran, J. (2000), "Two Sample T_3-plot: A Graphical Comparison of Two Distributions," *Journal of Computational and Graphical Statistics*, **9**, pp. 167-179.

[24] Good, P. (2000). *Permutation Tests*, 2nd ed., New York: Springer.

[25] Hall, P., and Pittelkow, Y.E. (1990), "Simultaneous Bootstrap Confidence Bands in Regression," *Journal of Statistical Computation and Simulation*, **37**, pp. 99-113.

[26] Härdle, W., and Marron, J.S. (1991), "Bootstrap simultaneous error bars for nonparametric regression," *The Annals of Statistics*, **19**, pp. 778-796.

[27] Hsu, J. C. (1996), *Multiple Comparisons, Theory and Methods*, New York: Chapman & Hall.

[28] Knorr-Held, L. (2003), "Simultaneous Posterior Probability Statements from Monte Carlo Output," *Journal of Computational and Graphical Statistics* (to appear).

[29] Landwehr, J.M., Pregibon, D., and Shoemaker, A.C. (1984), "Graphical Methods for Assessing Logistic Regression Models," *Journal of the American Statistical Association*, **79**, pp. 61-71.

[30] Lehman, E. L. (1986), *Testing Statistical Hypotheses*, 2nd edition, New York: Wiley.

[31] Loader, C. (1993), "Nonparametric Regression, Confidence Bands and Bias Correction," *Computing Science and Statistics*, 25, pp. 131-136.

[32] Loader, C. (2001), "Re: simultaneous bands," posting to the smoothing mailing list at smoothing@stat.unc.edu, dated "14 Aug 2001".

[33] Marron, J.S. (1997), "Re: "the confidence bands"," posting to the smoothing mailing list at smoothing@stat.unc.edu, dated "7 Oct 1997".

[34] Mielke, P.W. Jr., and Berry, K.J. (2001), *Permutation Methods*, New York: Springer.

[35] Rosenkrantz, W.A. (2000), "Confidence Bands for Quantile Functions: A Parametric and Graphic Alternative for Testing Goodness of Fit," *The American Statistician*, **54**, pp. 185-190.

[36] Ripley, B.D. (1981). *Spatial statistics*, New York: Wiley.

[37] Sun, J., and Loader, C.R. (1994), "Simultaneous Confidence Bands for Linear Regression and Smoothing," *The Annals of Statistics*, **22**, pp. 1328-1345.

[38] Tibshirani, R. (1992), "Some Applications of the Bootstrap in Complex Problems," in *Exploring the Limits of the Bootstrap*, ed. R. LePage. New York: Wiley.

[39] Venables, W.N., and Ripley, B.D. (2002), *Modern Applied Statistics with S-PLUS,* 4th ed., New York: Springer

[40] Westfall, P. H., and Young, S. S. (1993), *Resampling-Based Multiple Testing*, New York: Wiley.