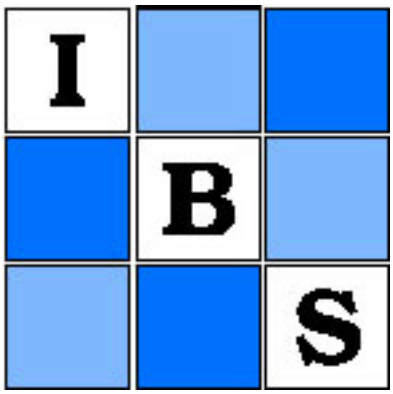

BIOMETRICS 62, 288–296
March 2006

DOI: 10.1111/j.1541-0420.2005.00422.x

Assessment of Agreement under Nonstandard Conditions Using Regression Models for Mean and Variance

Pankaj K. Choudhary

Department of Mathematical Sciences, University of Texas at Dallas, Richardson, Texas 75083-0688, U.S.A.
email: pankaj@utdallas.edu

and

Hon Keung Tony Ng

Department of Statistical Science, Southern Methodist University, Dallas, Texas 75275-0332, U.S.A.
email: ngh@mail.smu.edu

SUMMARY. The total deviation index of Lin (2000, *Statistics in Medicine* **19,** 255–270) and Lin et al. (2002, *Journal of the American Statistical Association* **97,** 257–270) is an intuitive approach for the assessment of agreement between two methods of measurement. It assumes that the differences of the paired measurements are a random sample from a normal distribution and works essentially by constructing a probability content tolerance interval for this distribution. We generalize this approach to the case when differences may not have identical distributions—a common scenario in applications. In particular, we use the regression approach to model the mean and the variance of differences as functions of observed values of the average of the paired measurements, and describe two methods based on asymptotic theory of maximum likelihood estimators for constructing a simultaneous probability content tolerance band. The first method uses bootstrap to approximate the critical point and the second method is an analytical approximation. Simulation shows that the first method works well for sample sizes as small as 30 and the second method is preferable for large sample sizes. We also extend the methodology for the case when the mean function is modeled using penalized splines via a mixed model representation. Two real data applications are presented.

KEY WORDS: Bootstrap; Limits of agreement; Method comparison; Mixed model; Penalized spline; Simultaneous tolerance band; Total deviation index.

1. Introduction

We consider the setting of a method comparison study where a test method of measurement (T) is compared with a reference method of measurement (R) using the paired measurements (T_i, R_i), $i = 1, \ldots, n$. In applications, this R is generally a gold standard, a measurement method that is traditionally considered to be accurate. The goal of the comparison is to assess the agreement of T with R, that is, to determine whether for a subject, a measurement from R can be substituted with a measurement from T without leading to any difference in its clinical interpretation. This issue of agreement arises in medical applications where R is generally expensive or invasive, and T is a convenient alternative.

There are several approaches for evaluating agreement, for example, the concordance correlation of Lin (1989), the total deviation index of Lin (2000), and others—see Lin et al. (2002) and Choudhary and Nagaraja (2004) for reviews of the literature on this topic. We focus on the total deviation index approach and its refinement by P. K. Choudhary and H. N. Nagaraja (unpublished manuscript). The basic idea of this approach is the following: It assumes that the differences $Y_i = R_i - T_i$, $i = 1, \ldots, n$, are a random sample from an $N(\mu, \sigma^2)$ distribution, and constructs a level $(1 - \alpha)$ upper confidence bound for the p_0th percentile of $|Y_i|$, where the cutoff p_0 (>0.5) is specified by the practitioner. Typically, p_0 is chosen from $\{0.80, 0.85, 0.90, 0.95\}$. Let $q = \sigma\{\chi^2_1\ (p_0,\ \mu^2/\sigma^2)\}^{1/2}$ denote this percentile, and U be its upper confidence bound, with $\chi^2_1(p_0,\ \Delta)$ representing the p_0th percentile of a noncentral chi-square-distribution with one degree of freedom and noncentrality parameter Δ. The interval $[-U, U]$ then becomes a tolerance interval with probability content p_0 and confidence $(1 - \alpha)$. In other words, it satisfies

$$\Pr(F(U) - F(-U) \geq p_0) = 1 - \alpha, \tag{1}$$

where $F(\cdot)$ is the cumulative distribution of Y_i's (see P. K. Choudhary and H. N. Nagaraja, unpublished manuscript). One may refer to Guttman (1988) for an introduction to the

concept of tolerance intervals. This interval here essentially estimates the likely range in which most of the differences are expected to lie. The agreement of T with R is inferred as satisfactory if the practitioner believes that all the differences in this interval are equivalent to zero for all clinical purposes. Notice that in this approach, the value of p_0 and the margin of clinically acceptable differences are two subjective choices made by the practitioner.

Frequently, the assumption of identical distribution for the differences does not hold in applications. The dependence of $E(Y_i)$ and $\text{var}(Y_i)$ on the magnitude of measurements through their average, $X_i = (R_i + T_i)/2$, are two common violations (Bland and Altman, 1999; Hawkins, 2002). Consider, for example, the Oestradiol data from Hawkins (2002). Oestradiol is a potent naturally occurring estrogen hormone, which is synthesized to treat estrogen deficiency and menopausal symptoms. The dataset has 142 Oestradiol measurements (in pg/ml) from two assays. Here, for the sake of illustration, we exclude three outlying differences and focus on the remaining $n = 139$ observations. Figure 1a has their scatterplot. The plot of Y_i (assay 1 − assay 2) and $\log(X_i)$ in Figure 1b clearly shows that Y_i's have a nonconstant mean and their variability increases with the magnitude of measurements. These two conditions, namely, the nonconstant mean and the heteroscedasticity, are the nonstandard conditions referred to in the title of this article. In this scenario, a natural approach is to first model $E(Y_i)$ and $\text{var}(Y_i)$ as functions of the observed value x_i of the average X_i, and then construct the tolerance interval as a function of x, hereafter termed a *tolerance band*, that has simultaneous confidence $(1 - \alpha)$ for all x in a given interval $\mathfrak{X}$. The goal of this article is to present the methodology for the construction of this band. Figure 1d shows the resulting band when this methodology is applied to the Oestradiol data with $(p_0, 1 - \alpha) = (0.80, 0.95)$. Also included in Figure 1b and 1d are the graphs of the fitted mean and standard deviation functions (see Section 5.2 for details), and the tolerance interval constructed assuming identical distribution for the differences, as described in P. K. Choudhary and H. N. Nagaraja (unpublished manuscript). Taking into account the dependence of mean and variance of Y on x produces a funnel-shaped band which is much narrower initially than the band assuming constant mean and variance. This shape is consistent with the features of the scatterplot.

In this article, we assume that,

$$Y_i = Y(x_i) = \mu(x_i) + \sigma(x_i)\epsilon_i,$$
$$\epsilon_i \sim \text{independent } N(0,1), \quad i = 1, \ldots, n, \tag{2}$$

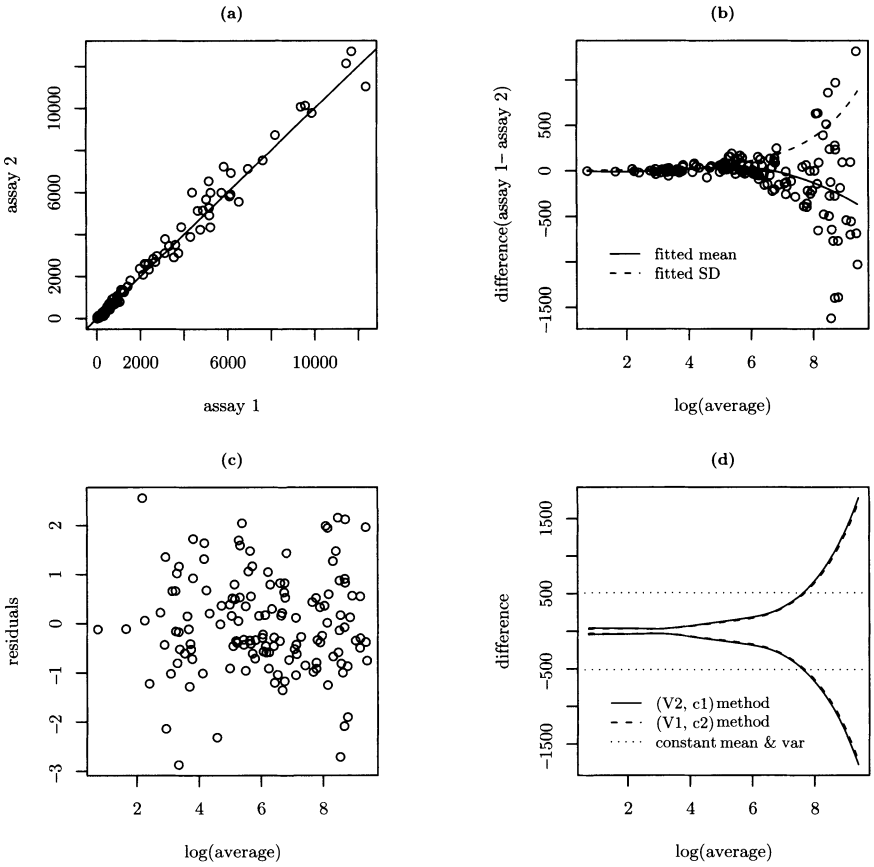

Figure 1. (a) Scatterplot of measurements from two assays for Oestradiol. The solid line represents a 45° line through the origin. (b) Scatterplot of differences and log-averages, superimposed with the fitted mean and standard deviation functions. (c) Residual plot for the fitted model. (d) Approximate 95% tolerance bands with probability content $p_0 = 0.80$. The (V_1, c_2) and (V_2, c_1) methods are described in Section 5.1.

so that the Y_i's follow independent $N(\mu(x_i), \sigma^2(x_i))$ distributions, where

$$\mu(x) = f(x, \beta), \quad \sigma^2(x) = \sigma_e^2\, g(x, \beta, \theta).$$

Here f and g are known fixed functions and $\psi = (\beta, \theta, \sigma_e^2)$ is the $p \times 1$ vector of unknown parameters, with β as a $k \times 1$ vector, and θ also being possibly vector-valued. Throughout this article, a vector is a column vector unless specified otherwise. The forms of f and g are obtained using the standard regression methodology. See Carroll and Ruppert (1988, Chapter 2) for the estimation of the latter. We assume that p_0 (>0.5) is specified. Let $\hat{\psi} = (\hat{\beta}, \hat{\theta}, \hat{\sigma}_e^2)$ be the maximum likelihood estimate (MLE) of ψ. Once we have $\hat{\psi}$, the percentile function

$$q(x) = p_0\text{th percentile of } |Y(x)| = \sigma(x)\left\{\chi_1^2\left(p_0, \frac{\mu^2(x)}{\sigma^2(x)}\right)\right\}^{1/2}, \tag{3}$$

is estimated as

$$\hat{q}(x) = \hat{\sigma}(x)\left\{\chi_1^2\left(p_0, \frac{\hat{\mu}^2(x)}{\hat{\sigma}^2(x)}\right)\right\}^{1/2};$$
$$\hat{\mu}(x) = f(x, \hat{\beta}), \quad \hat{\sigma}^2(x) = \hat{\sigma}_e^2\, g(x, \hat{\beta}, \hat{\theta}). \tag{4}$$

The remainder of this article is organized as follows. Section 2 describes the methodology for the construction of an asymptotic simultaneous upper confidence band (UCB) for $q(x)$, say $U(x)$, which by definition satisfies,

$$\Pr(q(x) \le U(x), \text{ for all } x \in \mathfrak{X}) \approx 1 - \alpha,$$

for large n. In analogy with (1), the region $[-U(x), U(x)]$ then becomes a simultaneous tolerance band for the differences under model (2), that is, we have

$$\Pr(F_x[U(x)] - F_x[-U(x)] \ge p_0, \text{ for all } x \in \mathfrak{X}) \approx 1 - \alpha,$$

where $F_x(\cdot)$ is the cumulative distribution function of $Y(x)$. With the help of this band, the practitioner can find the regions of measurement range where (s)he considers the agreement of T with R to be adequate. In Section 3, we use Monte Carlo simulation to study the small sample performance of the proposed methodology. Section 4 presents an illustration with a real dataset. In Section 5, we generalize the methodology to handle the case when the mean function f is modeled using penalized splines and revisit the Oestradiol data introduced earlier. Section 6 concludes with a discussion. The Appendix, available at the *Biometrics* website `http://www.tibs.org/biometrics`, contains some technical details.

2. Construction of Upper Confidence Band

Let $l(\psi)$ denote the log-likelihood function of model (2),

$$l(\psi) = -(n/2)\log(2\pi) - (1/2)\sum_{i=1}^{n}\log\sigma^2(x_i) - (1/2)\sum_{i=1}^{n}\big(y_i - \mu(x_i)\big)^2/\sigma^2(x_i).$$

The MLE $\hat{\psi}$ is obtained by maximizing $l(\psi)$ with respect to ψ. Pinheiro and Bates (2000, Chapter 2) contains an excellent account of the computational details underlying this maximization process (see also Carroll and Ruppert, 1988, Chapters 2 and 3). It is well known that, for large n, the distribution of $\hat{\psi}$ can be approximated as,

$$\hat{\psi} \approx N(\psi, I^{-1}), \quad \text{where} \quad I = \left(-\frac{\partial^2 l(\psi)}{\partial\psi_i\,\psi_j}; i, j = 1, \ldots, p\right)_{\psi=\hat{\psi}}$$

is the *observed* Fisher information matrix for ψ. From the delta method (see Lehmann, 1998, p. 315), it follows that for a fixed $x \in \mathfrak{X}$ and large n, the approximate distribution of

$$\log\hat{q}(x) \approx N\big(\log q(x), G_x' I^{-1} G_x\big), \quad \text{where}$$
$$G_x = \left(\frac{\partial \log q(x)}{\partial\psi_i}; i = 1, \ldots, p\right)_{\psi=\hat{\psi}} \tag{5}$$

is the $p \times 1$ vector denoting the gradient of $\log q(x)$ evaluated at $\hat{\psi}$, and G_x' is the transpose of G_x. The expression for G_x is given in the Appendix. Consequently,

$$Z_n(x) = (\log\hat{q}(x) - \log q(x))\big/\big(G_x' I^{-1} G_x\big)^{1/2}, \tag{6}$$

is approximately $N(0, 1)$ for every fixed $x \in \mathfrak{X}$. Here we use $\log q(x)$ in place of $q(x)$ as the normal approximation for the former tends to be more accurate.

This result suggests that for the UCB of $q(x)$, we can focus on a curve of the form,

$$U(x, c) = \exp\big\{\log\hat{q}(x) - c\,\big(G_x' I^{-1} G_x\big)^{1/2}\big\}, \quad x \in \mathfrak{X}, \tag{7}$$

where c (<0) in an appropriately chosen critical point. For an approximate *pointwise* $100(1-\alpha)\%$ UCB, we take c as the αth percentile of an $N(0, 1)$ distribution. For a *simultaneous* $100(1-\alpha)\%$ UCB, we must choose c such that

$$1 - \alpha = \Pr(\log q(x) \le \log U(x, c), \text{ for all } x \in \mathfrak{X}) = \Pr\left(\inf_{x\in\mathfrak{X}} Z_n(x) \ge c\right), \tag{8}$$

where $Z_n(x)$ is defined in (6). We now discuss two approximations for this critical point.

2.1 *Numerical Approximation by Bootstrap Method*

The first method is a numerical approximation $c \approx c_1$ obtained using bootstrap (see Efron and Tibshirani, 1993). Let $(x_1^*, \ldots, x_t^*)$ denote a grid of equally spaced x-values in $\mathfrak{X}$. We simulate independent realizations $Y_i^* \sim N(\hat{\mu}(x_i^*), \hat{\sigma}^2(x_i^*)), i = 1, \ldots, t$, where the functions $\hat{\mu}$ and $\hat{\sigma}^2$ defined in (4) are computed using the MLE $\hat{\psi}$ based on the original sample $\{x_1, Y_1\}, \ldots, \{x_n, Y_n\}$. The simulated sample $\{x_1^*, Y_1^*\}, \ldots, \{x_t^*, Y_t^*\}$ serves as a parametric resample of the original sample. This resample is used to compute the MLE, say $\hat{\psi}^*$, the associated observed information matrix, say I^*, the gradient, say $G_{x^*}^*$, and an estimate of $\inf_{x\in\mathfrak{X}} Z_n(x)$, say

$$M = \min_{1\le i\le t}\big(\log\hat{q}^*\big(x_i^*\big) - \log\hat{q}\big(x_i^*\big)\big)\big/\big(G_{x_i^*}^{*\prime} I^{*-1} G_{x_i^*}^*\big)^{1/2}.$$

This process is repeated a large number of times, say $B = 2000$, to simulate B realizations of M. Finally, the αth sample percentile of M is taken as the approximation c_1.

2.2 *Analytical Approximation*

The next is an analytical approximation $c \approx c_2$ obtained by solving the equation

$$\alpha = \Pr(t_\nu \le c_2) + \frac{\kappa_0}{2\pi}\left(1 + \frac{c_2^2}{\nu}\right)^{-\nu/2}, \quad \nu = (n-k), \tag{9}$$

for c_2. Here t_ν follows a t-distribution with ν degrees of freedom; and letting $L_x = I^{-1/2}G_x$ and $\dot{L}_x = I^{-1/2}(\partial/\partial x)G_x$, with the differentiation applied elementwise, κ_0 represents

$$\kappa_0 = \int_{\mathfrak{X}} \frac{1}{(L_x'L_x)}\left((L_x'L_x)(\dot{L}_x'\dot{L}_x) - (L_x'\dot{L}_x)^2\right)^{1/2} dx.$$

It is not difficult to derive a closed-form expression for $\dot{L}_x$ but it gets rather messy. So we avoid presenting it here and will compute it numerically. The motivation for this approximation comes from the methodology used to construct a simultaneous confidence band for the mean function in nonparametric regression (see Loader, 1999, Section 9.2). One can adapt this methodology by taking the regression function as $\log q(x) \approx G_x'\psi$ and the fitted regression function as $\log \hat{q}(x) \approx G_x'\hat{\psi} = L_x'I^{1/2}\hat{\psi}$, where $I^{1/2}\,\hat{\psi}$ serves as the response vector. A more direct asymptotic justification for the approximation appears in the Appendix. In the right-hand side of (9), using $\Pr(t_\nu \le c_2)$ and $(1 + c_2^2/\nu)^{-\nu/2}$ in place of their respective limits $\Phi(c_2)$ and $\exp(-c_2^2/2)$, where $\Phi(\cdot)$ is the cumulative distribution function of an $N(0, 1)$ distribution, leads to a better approximation for c_2 in finite samples.

Both the bootstrap and the analytical approximations described above are asymptotic in nature. So in the next section we evaluate their small-sample performance by estimating the coverage probabilities of the resulting simultaneous UCBs, $U(x, c_1)$ and $U(x, c_2)$, through Monte Carlo simulation.

3. Monte Carlo Simulation Studies

We focus on model (2) with $f(x, \beta) = \beta_0 + \beta_1 x$ and $g(x, \beta, \theta) = x^{2\theta}$, $x > 0$, for this investigation. This model for g is equivalent to modeling $\log\sigma(x)$ as a linear function of $\log x$. We will perform the simulation under three cases of this model:

1. $f(x, \beta) = \beta_0 + \beta_1 x$, $g(x, \beta, \theta) = x^{2\theta}$; so that $\psi = (\beta_0, \beta_1, \theta, \sigma_e^2)$.
2. $f(x, \beta) = \beta_0 + \beta_1 x$, $g(x, \beta, \theta) \equiv 1$ (homoscedastic model); so that $\psi = (\beta_0, \beta_1, \sigma_e^2)$.
3. $f(x, \beta) \equiv \beta_0$, $g(x, \beta, \theta) = x^{2\theta}$; so that $\psi = (\beta_0, \theta, \sigma_e^2)$.

The fourth possibility when $f(x, \beta) \equiv \beta_0$, $g(x, \beta, \theta) \equiv 1$ is not considered here as it refers to the standard case of independently and identically distributed differences that Lin (2000) and P. K. Choudhary and H. N. Nagaraja (unpublished manuscript) discussed. For the simulation, we choose $\mathfrak{X} = (0, 1)$; $p_0 \in \{0.80, 0.90\}$; $\alpha = 5\%$; $B = 2000$; and the parameter values $\beta_0 = 0$, $\beta_1 \in \{0.5, 1, 2\}$, $\theta \in \{0.5, 1\}$, and $\sigma_e^2 = 1$. These settings cover a wide range of possibilities encountered in applications. Finally, we take t, the number of grid points in $\mathfrak{X}$, equal to n. The motivation for this choice comes from our empirical finding that taking t to be substantially smaller than n leads to a somewhat conservative UCB.

Table 1

Estimated coverage probabilities (%) of the 95% confidence level simultaneous UCBs $U(x, c_1)$ and $U(x, c_2)$ for $q(x)$. The parameter vector ψ refers to $(\beta_0, \beta_1, \theta, \sigma_e^2)$ in case 1, $(\beta_0, \beta_1, \sigma_e^2)$ in case 2, and $(\beta_0, \theta, \sigma_e^2)$ in case 3. These estimates are based on 5000 replications and have a standard error of 0.3.

		c_1 ($n = 30$)		c_2 ($n = 30$)		c_2 ($n = 100$)	
		p_0		p_0		p_0	
Case	ψ	0.80	0.90	0.80	0.90	0.80	0.90
1	(0, 0.5, 0.5, 1.0)	95.5	95.7	91.2	90.7	94.4	94.1
	(0, 0.5, 1.0, 1.0)	95.9	96.2	91.6	90.5	94.7	94.5
	(0, 1.0, 0.5, 1.0)	95.5	95.6	92.0	90.6	94.4	94.2
	(0, 1.0, 1.0, 1.0)	95.8	95.6	91.9	89.7	94.9	94.4
	(0, 2.0, 0.5, 1.0)	95.4	95.4	91.9	89.7	94.5	93.7
	(0, 2.0, 1.0, 1.0)	95.5	95.9	90.0	87.7	94.3	93.7
2	(0, 0.5, 1.0)	95.7	95.6	93.3	93.3	93.5	93.6
	(0, 1.0, 1.0)	94.8	95.3	92.4	92.5	94.8	95.0
	(0, 2.0, 1.0)	95.8	95.6	94.3	93.5	95.2	95.0
3	(0, 0.5, 1.0)	95.4	95.2	91.3	91.0	93.5	93.5
	(0, 1.0, 1.0)	95.5	95.7	91.3	90.7	94.1	94.1

To estimate the simultaneous coverage probability of the UCB $U(x, c_1)$ for a given combination of p_0 and ψ, we simulate independent realizations $Y_1, \ldots, Y_n$ from model (2) on a grid of n equally spaced x-values in the range [0.1, 0.99]. Next, we use the data $\{x_1, Y_1\}, \ldots, \{x_n, Y_n\}$ to estimate ψ, compute the critical point c_1, and the UCB $U(x, c_1)$ as described in the previous section. This process is repeated 5000 times and the proportion of times $\{q(x_i) \le U(x_i, c_1)$ for all $i = 1, \ldots, n\}$ is computed, which serves as the estimated simultaneous coverage probability of $U(x, c_1)$. Throughout we use the grid $(x_1^*, \ldots, x_t^*) = (x_1, \ldots, x_n)$ to generate bootstrap resamples. We proceed analogously to estimate the simultaneous coverage probability of $U(x, c_2)$. The computations were performed with sample size $n = 30$ for $U(x, c_1)$ and with $n = 30$, 100 for $U(x, c_2)$ using `FORTRAN` 95 and `R` (R Development Core Team, 2004). Table 1 presents the resulting estimates.

The bootstrap approximation c_1 is remarkably accurate for $n = 30$ as the estimated coverage probabilities are quite close to the target nominal level 95%. In addition, the probabilities in this case do not seem to depend on the model used, p_0, or the true parameter configuration. On the other hand, the approximation c_2 is not good for $n = 30$. It is a liberal approximation (i.e., $U(x, c_2)$ is smaller than what it should be) with probabilities about 3–4% lower than the target. The situation appears worse for $p_0 = 0.90$ than $p_0 = 0.80$, and for $\theta = 1.0$ than $\theta = 0.5$. However, the approximation improves greatly for $n = 100$. The probabilities are now only about 0.5–1% lower. Similar conclusions hold when $p_0 = 0.95$ (results not shown). We also investigate the bootstrap approximation for $n = 60$ with $t \in \{30, 60\}$ and $n = 100$ with $t \in \{30, 50, 100\}$. When $t = n$, our conclusions remain analogous to the $n = 30$ case; but when $t < n$, the coverage probabilities are about 0.5–2% higher than those for $t = n$. Large differences of n and t ($t < n$) produce more conservative results. So, to avoid this situation, taking $t = n$ seems more appropriate than $t < n$, at least when $30 \le n \le 100$. However, if the conservativeness

of the UCB is not a major concern, one may take $t = 30$ to reduce the computational effort. Overall for $0.80 \leq p_0 \leq 0.95$, c_1 is recommended when $n \geq 30$, and c_2 when $n \geq 100$. Notice also that c_1 is more numerically intensive to compute than c_2, particularly for large n.

On further investigation, we discover that $N(0, 1)$ approximation for $Z_n(x)$ is not good for most of $x \in \mathfrak{X}$ when n is not very large. In particular, $E[Z_n(x)] < 0$ and $\text{var}[Z_n(x)] > 1$, that is, both $\log \hat{q}(x)$ and $G_x' I^{-1/2} G_x$ tend to underestimate their respective parameters. Consequently c_2 (<0), whose theory assumes asymptotic $N(0, 1)$ for $Z_n(x)$ for all $x \in \mathfrak{X}$, gets overestimated. However, the bootstrap resampling corrects this problem for c_1.

4. Application to Cyclosporin Data of Hawkins (2002)

Cyclosporin is an immunosuppressant drug widely used to prevent rejection of transplanted organs. This dataset consists of Cyclosporin measurements obtained by assaying an aliquot each of the 56 blood samples from organ transplant recipients using two methods: the high performance liquid chromatography (HPLC) method, which is the standard approved method, and an alternative radio-immunoassay (RIA) method. The question of interest here is whether one can substitute HPLC assays with RIA assays, that is, use the RIA assays as if they were HPLC assays. The scatterplot of these data given in Figure 2a may suggest a reasonable agreement between the methods. We will ignore the observation in the top right corner of this plot and restrict all the subsequent analysis to the measurement range (35–700) with $n = 55$ pairs of measurements.

Our first task is to obtain models for the mean function (f) and the variance function (g) in (2). The plot of difference (Y) against the average (X) of the measurements (popularly known as the Bland–Altman plot after Bland and Altman, 1986) in Figure 2b indicates that $f(x, \beta) \equiv \beta_0$ may be appropriate. The exploratory analysis on the lines of Carroll and Ruppert (1988, Section 2.7) suggests modeling the logarithm of variance of Y as a linear function of $\log x$, so that $g(x, \beta, \theta) = x^{2\theta}$, $x > 0$. Fitting model (2) with these (f, g) produces the MLE of $\psi = (\beta_0, \theta, \sigma_e^2)$ and its estimated asymptotic covariance matrix as

$$\hat{\psi} = \begin{pmatrix} -5.509 \\ 0.802 \\ 0.376 \end{pmatrix}, \quad I^{-1} = \begin{pmatrix} 22.684 & -0.103 & 0.415 \\ & 0.042 & -0.169 \\ & & 0.682 \end{pmatrix}.$$

Further, since the plot of residuals, $r_i = (Y_i - \hat{\beta}_0)/(\hat{\sigma}_e x_i^{\hat{\theta}})$, against x_i in Figure 2c does not show any clear pattern (tests for zero Pearson and Spearman correlations of $(x_i, |r_i|)$ have p-values greater than 0.68), and the Shapiro–Wilk test of normality has a p-value of 0.18, it appears that the above model fits reasonably well. An approximate 95% Wald-type

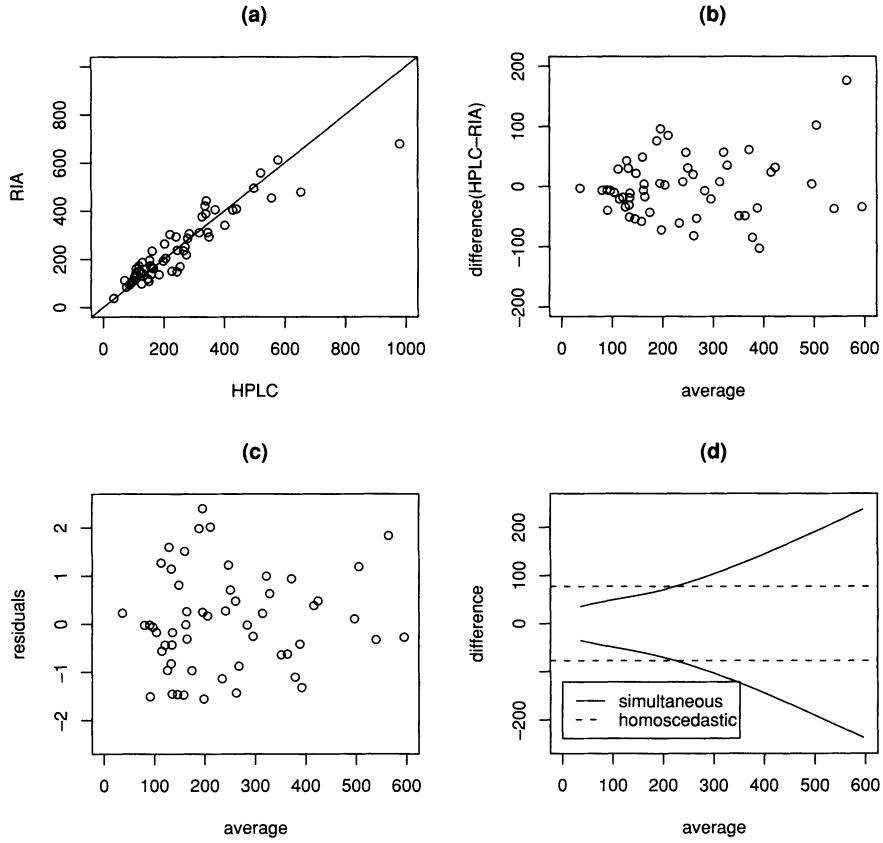

Figure 2. (a) Scatterplot of Cyclosporin measurements from HPLC and RIA assays. The solid line represents a 45° line through the origin. (b) Bland–Altman plot of difference of paired measurements against their averages. (c) Residual plot for the fitted model. (d) Approximate 95% tolerance bands with probability content $p_0 = 0.80$.

confidence interval for θ, $[0.802 \pm 1.96 \ (0.042)^{1/2}] = [0.40, 1.20]$, does not cover zero, thus supporting the need for a heteroscedastic model.

Next, we take $\mathfrak{X} = [\min_i x_i = 36.5, \max_i x_i = 595]$ and obtain 95% simultaneous tolerance band with probability content $p_0 = 0.80$. The approximate critical points are computed as $c_1 = -2.469$ (with $B = 2000$) and $c_2 = -2.264$. From Section 3, we expect c_2 to be a little larger than the more accurate c_1 since the sample size ($n = 55$) here is not very large. Figure 2d presents the simultaneous tolerance band $[-U(x, -2.469), U(x, -2.469)]$, where $U(x, c)$ is computed using (7), along with the band that assumes identical distribution for the differences. Taking heteroscedasticity into account produces a shorter band for low values of Cyclosporin, whereas the converse is true for its high values.

Using this band to judge whether the RIA assays could be substituted for HPLC assays is a matter of subjective judgment for the clinical expert. But since the absolute differences could be as large as 235 when the magnitude of measurements is about 600, it is unlikely that the agreement over the entire measurement range will be considered satisfactory. It may be satisfactory only for low values of Cyclosporin.

5. Extension to Penalized Splines Regression

5.1 *Methodology*

In some situations, the mean function of differences has nonlinear features that are difficult to model parametrically (e.g., a low degree polynomial does not provide a good fit for the Oestradiol data). A popular and relatively straightforward approach to handle this situation is the methodology of penalized splines regression. See Chapters 3–6 of Ruppert, Wand, and Carroll (2003), hereafter termed RWC, for an excellent introduction to this topic.

For simplicity, we assume that the mean function f in (2) can be modeled as

$$f(x, \beta, u) = \beta_0 + \beta_1 x + \beta_2 x^2 + \sum_{j=1}^{m} u_j (x - k_j)_+^2, \qquad (10)$$

where β is the 3×1 vector $(\beta_0, \beta_1, \beta_2)$ and u is the $m \times 1$ vector $(u_1, \ldots, u_m)$. Here f represents a quadratic spline with *knots* at $k_1, \ldots, k_m$; $u_1, \ldots, u_m$ are the respective coefficients of the truncated quadratic basis functions $(x - k_1)_+^2, \ldots, (x - k_m)_+^2$; and $x_+ = \max\{0, x\}$. The use of quadratic basis functions ensures that f does not have any sharp corners and has continuous first derivative. This property is important for reliable numerical computations involved in the construction of a simultaneous UCB. Model (10) is quite flexible in practice, but in principle, a spline function of any degree (≥ 2) may be used to represent f. The number and location of the knots k_j's can be chosen using the guidelines in Chapter 5 of RWC. Also, an appropriate transformation of x may be used in (10) in place of x as the covariate.

Thus in this section, the model for the data $\{x_1, Y(x_1)\}, \ldots, \{x_n, Y(x_n)\}$ is (2) with

$$\mu(x) = f(x, \beta, u), \quad \sigma^2(x) = \sigma_e^2 \, g(x, \beta, \theta). \qquad (11)$$

Substituting these expressions in (3) gives the target percentile function $q(x)$. We use the mixed model representation of the spline f and fit model (2) using maximum likelihood. This approach is equivalent to fitting f using a penalized criterion (see Chapters 3–5 of RWC). The mixed model representation assumes, in addition to the assumptions in (2), that $u_1, \ldots, u_m$ are distributed as independent $N(0, \sigma_u^2)$ random variables, and are mutually independent of the errors $\epsilon_1, \ldots, \epsilon_n$. Thus in this case, the target mean function f is actually an unobservable random variable. This contrasts with the methodology of previous sections in which the target f was fixed. However, when $\sigma_u^2 = 0$, the randomness in f disappears, and the methodology of previous sections applies. When f is random, the percentile function $q(x)$ is also random.

Due to the randomness of u, marginally, $Y(x_1), \ldots, Y(x_n)$ are correlated random variables with $E[Y(x)] = \beta_0 + \beta_1 x + \beta_2 x^2$, $\operatorname{var}[Y(x)] = \sigma_u^2 \sum_{j=1}^{m} (x - k_j)_+^4 + \sigma_e^2 \, g(x, \beta, \theta)$, and $\operatorname{cov}[Y(x), Y(v)] = \sigma_u^2 \sum_{j=1}^{m} (x - k_j)_+^2 (v - k_j)_+^2$ for $x \neq v$. To fit the model, we take the parameter vector as $\psi = (\beta, \theta, \log \sigma_u, \log \sigma_e)$ and maximize the marginal likelihood of $Y(x_1), \ldots, Y(x_n)$ with respect to ψ to obtain the MLE $\hat{\psi}$. We use $(\log \sigma_u, \log \sigma_e)$ in place of (σ_u, σ_e) to get an unconstrained parameterization that leads to more stable and accurate estimates. See Pinheiro and Bates (2000, Chapters 2 and 5) for clever computational strategies to evaluate and maximize the likelihood function associated with a mixed model. Once the model is fitted, $\mu(x)$ and $\sigma^2(x)$ of (11) are estimated as

$$\hat{\mu}(x) = f(x, \hat{\beta}, \hat{u}), \quad \hat{\sigma}^2(x) = \hat{\sigma}_e^2 \, g(x, \hat{\beta}, \hat{\theta}), \qquad (12)$$

and $\hat{q}(x)$ is obtained via (4) using these expressions for $\hat{\mu}(x)$ and $\hat{\sigma}^2(x)$. Here $\hat{u} = (\hat{u}_1, \ldots, \hat{u}_m)$ is the *estimated best linear unbiased predictor* (EBLUP) of u, given as (see RWC, Section 4.6),

$$\hat{u} = \hat{\sigma}_u^2 K' \left[\hat{\sigma}_u^2 K K' + \hat{\sigma}_e^2 \operatorname{diag}(g(x_1, \hat{\beta}, \hat{\theta}), \ldots, g(x_n, \hat{\beta}, \hat{\theta})) \right]^{-1} e,$$

where K is the $n \times m$ design matrix corresponding to the random effect u whose (i, j)th element is $(x_i - k_j)_+^2$, and e is the $n \times 1$ vector with ith element $y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i - \hat{\beta}_2 x_i^2$.

Let ξ denote the stacked column vector $(\beta, u, \theta, \log \sigma_e)$ and $\hat{\xi} = (\hat{\beta}, \hat{u}, \hat{\theta}, \log \hat{\sigma}_e)$ denote its estimate. To obtain the UCB for now random $q(x)$ we must derive the asymptotic distribution of $(\log \hat{q}(x) - \log q(x))$ since the expression given by (5) no longer holds. For a fixed $x \in \mathfrak{X}$, using Taylor series we can write,

$$\log \hat{q}(x) - \log q(x) \approx G_x'(\hat{\xi} - \xi), \quad \text{where}$$
$$G_x = (\partial \log q(x) / \partial \xi)_{\xi = \hat{\xi}}.$$

The gradient G_x can be easily obtained on the lines of its expression in the Appendix for the fixed f case. Assuming asymptotic $N(0, V)$ distribution for $(\hat{\xi} - \xi)$, the delta method suggests

$$\log \hat{q}(x) - \log q(x) \approx N(0, G_x' V G_x). \qquad (13)$$

For the variance approximation here to work well we must also have a small σ_u^2 in addition to a large n. Loosely speaking, this ensures that u behaves like a fixed parameter, so that when n is large, $\hat{u}$ is close to u. This together with the consistency of MLEs implies that $\hat{\xi}$ is close to ξ, and hence $G_x' V G_x$ is close to the true asymptotic variance. Now, as in Section 2, we can use a curve of the form

$$U(x,c) = \exp\left\{\log \hat{q}(x) - c\,(G_x' V G_x)^{1/2}\right\}, \quad x \in \mathfrak{X}, \quad (14)$$

as a simultaneous UCB for $q(x)$. Here the critical point c is such that (8) holds with $Z_n(x) = (\log \hat{q}(x) - \log q(x))/(G_x' V G_x)^{1/2}$. It can be computed using either the bootstrap approximation c_1 of Section 2.1 or the analytical approximation c_2 of Section 2.2.

We next describe two approximations for V, the asymptotic covariance matrix of $(\hat{\xi} - \xi)$, since its direct estimation is largely intractable (see RWC, p. 102). The first is a parametric bootstrap approximation, say V_1, obtained in the following manner:

1. Generate independent $u_1^*, \ldots, u_m^*$ from $N(0, \hat{\sigma}_u^2)$ and define $\xi^* = (\hat{\beta}, u^*, \hat{\theta}, \log \hat{\sigma}_e)$. Here $u^* = (u_1^*, \ldots, u_m^*)$, and $\hat{\psi} = (\hat{\beta}, \hat{\theta}, \log \hat{\sigma}_u, \log \hat{\sigma}_e)$ is the MLE of ψ.
2. Generate $Y_i^*(x_i)$ independently from $N(f(x_i, \hat{\beta}, u^*), \hat{\sigma}_e^2 g(x_i, \hat{\beta}, \hat{\theta}))$, $i = 1, \ldots, n$. The $\{x_i,\ Y^*(x_i)\}$ pairs represent a parametric resample of the original sample.
3. Use this resample to obtain the MLE of ψ, say $\hat{\psi}^*$, EBLUP of u^*, say $\hat{u}^*$, set $\hat{\xi}^* = (\hat{\beta}^*, \hat{u}^*, \hat{\theta}^*, \log \hat{\sigma}_e^*)$, and compute $(\hat{\xi}^* - \xi^*)$.
4. Repeat steps 1–3 a large number of times, say $B = 500$, to obtain B realizations of $(\hat{\xi}^* - \xi^*)$, and use their sample covariance matrix as V_1.

The second is an ad hoc approximation, say V_2, obtained as

$$V_2 = \begin{pmatrix} \widehat{\mathrm{var}}(\hat{\beta}) & \widehat{\mathrm{cov}}(\hat{\beta}, \hat{u} - u) & \widehat{\mathrm{cov}}(\hat{\beta}, \hat{\theta}) & \widehat{\mathrm{cov}}(\hat{\beta}, \log \hat{\sigma}_e) \\ & \widehat{\mathrm{var}}(\hat{u} - u) & 0 & 0 \\ & & \widehat{\mathrm{var}}(\hat{\theta}) & \widehat{\mathrm{cov}}(\hat{\theta}, \log \hat{\sigma}_e) \\ & & & \widehat{\mathrm{var}}(\log \hat{\sigma}_e) \end{pmatrix},$$

where we use the expression given on p. 103 of RWC for

$$\begin{pmatrix} \widehat{\mathrm{var}}(\hat{\beta}) & \widehat{\mathrm{cov}}(\hat{\beta}, \hat{u} - u) \\ & \widehat{\mathrm{var}}(\hat{u} - u) \end{pmatrix} = \hat{\sigma}_e^2 \left(D'D + \left(\hat{\sigma}_e^2 / \hat{\sigma}_u^2\right) \mathrm{diag}(0, 0, 0, 1, \ldots, 1)\right)^{-1},$$

with D denoting the $n \times (m + 3)$ matrix obtained by column-wise combining the $n \times 3$ design matrix for the fixed effect β and the $n \times m$ design matrix for the random effect u, and pre-multiplying the resulting matrix with an $n \times n$ diagonal matrix with diagonal elements $[g(x_i, \hat{\beta}, \hat{\theta})]^{-1/2}, i = 1, \ldots, n$. The remaining covariance matrices in V_2 are the appropriate submatrices of the inverse of the observed Fisher information matrix for ψ. The approximation V_1 is better for V than V_2 because V_2 ignores the potential correlation between the elements of $(\hat{u} - u)$ and $(\hat{\theta}, \log \hat{\sigma}_e)$. This may make V_2 inconsistent for estimating V. But, V_1 is more computationally demanding than V_2, and as we discuss below, V_2 is indeed useful, particularly since our ultimate goal is to construct a simultaneous UCB for $q(x)$.

We next use simulation to evaluate the UCB for $q(x)$, given by (14), computed in two ways. The first is the $(V_1,\ c_2)$ method where V is approximated as V_1 with 500 bootstrap resamples and the critical point c is approximated as c_2 of Section 2.2. The second is the $(V_2,\ c_1)$ method where V is approximated as V_2 and c is approximated as c_1 of Section 2.1 with 2000 bootstrap resamples. We do not pursue $(V_1,\ c_1)$ and $(V_2,\ c_2)$ methods as the former is too computationally demanding to be useful in applications, and the latter does not work well. However, we must add that $(V_1,\ c_1)$ is potentially the best among the four methods.

Table 2

Estimated coverage probabilities (%) of the 95% confidence level simultaneous UCBs for $q(x)$ computed using $(V_1,\ c_2)$ and $(V_2,\ c_1)$ methods with $(n,\ p_0) = (100,\ 0.80)$. Throughout we have $(\beta_0,\ \sigma_e) = (0,\ 1)$. The estimates for $(V_1,\ c_2)$ are based on 2000 replications and have a standard error of 0.5. The $(V_2,\ c_1)$ estimates are based on 1000 replications and have a standard error of 0.7.

	$(V_1,\ c_2)$				$(V_2,\ c_1)$			
	(σ_u/σ_e)				(σ_u/σ_e)			
$(\beta_1,\ \beta_2)$	0.5	1.0	2.0	4.0	0.5	1.0	2.0	4.0
$(-1,\ -1)$	93.9	93.1	91.8	87.4	95.5	95.2	94.1	89.6
$(-1,\ 1)$	92.9	92.9	92.4	88.8	94.5	95.0	94.2	91.7
$(1,\ -1)$	92.6	92.9	91.7	88.8	95.2	94.2	94.9	91.2
$(1,\ 1)$	93.9	93.6	91.4	87.0	96.3	94.5	93.9	90.7

For the simulation study, we focus on the homoscedastic model, that is, model (2) with $\sigma^2(x) \equiv \sigma_e^2$, and consider only $p_0 = 0.80$ to keep the computations manageable. We also take $\mathfrak{X} = (0, 1), \alpha = 5\%$, and $n = 100$. Smaller sample sizes are not considered as the spline regression is typically used for large sample sizes. The number of knots ($m = 25$) and their locations are chosen using the default method of RWC on p. 126. The parameter values are taken as $\beta_0 = 0$, $\beta_1 \in \{-1, 1\}$, $\beta_2 \in \{-1, 1\}$, $\sigma_u \in \{0.5, 1, 2, 4\}$, and $\sigma_e = 1$. The parameter σ_u here actually represents (σ_u/σ_e) since there is no loss of generality in taking $\sigma_e = 1$. We proceed as in Section 3 to estimate the simultaneous coverage probabilities of the two UCBs. The computations were performed in `R` version 1.9.0. Table 2 summarizes the results.

We see that $(V_2,\ c_1)$ outperforms $(V_1,\ c_2)$ in all the cases. The estimated coverage probabilities for $(V_2,\ c_1)$ are close to 95% when $(\sigma_u/\sigma_e) \leq 2$, but when $(\sigma_u/\sigma_e) \geq 4$, the method is liberal. The $(V_1,\ c_2)$ method is liberal throughout. The performance of both methods appears consistent across β values, but worsens as (σ_u/σ_e) increases. The latter is expected as the variance approximation in (13) holds only for small σ_u. Further investigation shows that the normality in (13) is fine throughout, but the variance underestimation becomes more severe as (σ_u/σ_e) increases. These general conclusions also hold for $n = 200$ (results not shown), but both methods become somewhat more liberal here than the $n = 100$ case, particularly for $(\sigma_u/\sigma_e) = 4$. This behavior may be because the effect of σ_u not being small becomes more prominent as n increases. Also, taking t, the number of grid points in $\mathfrak{X}$ for computing c_1, equal to 100 seems adequate for $n \geq 100$.

The estimates of $\mathrm{var}[\log \hat{q}(x) - \log q(x)]$ from V_2 tend to be smaller than those from V_1. However, the $(V_2,\ c_1)$ method corrects for this underestimation by producing a smaller (negative) critical point than the $(V_1,\ c_2)$ method. So, in general, the net effect is that the UCB curve for the former lies above the latter and is more accurate. Due to this property, we prefer $(V_2,\ c_1)$ over $(V_1,\ c_2)$. Finally, we note that most of the

real data applications of penalized spline regression in RWC appear to have estimates of (σ_u/σ_e) that are less than 4. Even for the Oestradiol data this estimate is 2.60 (see below). This indicates that perhaps the liberal behavior of $(V_2,\ c_1)$ for $(\sigma_u/\sigma_e) \geq 4$ may not be a major concern in practice.

5.2 *Application to Oestradiol Data of Hawkins (2002)*

We now return to the Oestradiol data introduced in Section 1. After a preliminary analysis similar to the Cyclosporin data, we decide to model the differences as (2) with

$$f(x,\beta,u) = \beta_0 + \beta_1 \log x + \beta_2 (\log x)^2 + \sum_{j=1}^{m} u_j (\log x - k_j)_+^2, \quad g(x,\theta) = x^{2\theta}; \quad x \in \mathfrak{X},$$

where $\mathfrak{X} = [\min_i x_i = 2, \max_i x_i = 12,201]$. We use the suggestion on p. 126 of RWC to take $m = 34$ knots and their locations as $k_j = ((j+1)/36)$th sample percentile of unique values in $\{\log x_1, \ldots, \log x_n\}$ for $j = 1, \ldots, m$. Fitting this model using R gives the MLE of $\psi = (\beta_0,\ \beta_1,\ \beta_2,\ \theta,\ \log\sigma_u,\ \log\sigma_e)$ as $\hat{\psi} = (7.839, -19.632, 5.496, 0.614, 1.956, 1.001)$. An application of (12) produces $\hat{\mu}(x)$ and $\hat{\sigma}(x)$ whose graphs are given in Figure 1b. The plot of residuals, $r_i = (y_i - \hat{\mu}(x_i))/\hat{\sigma}(x_i)$, in Figure 1c suggests that this model fits well. The normality assumption for the errors also seems reasonable.

Next, we take $(p_0,\ 1-\alpha) = (0.80,\ 0.95)$, and apply the methodology of the previous section to get the simultaneous UCB for $q(x)$. The graphs of the resulting tolerance bands from $(V_1,\ c_2)$ and $(V_2,\ c_1)$ methods are presented in Figure 1d. We used $t = 100$ for computing c_1. The critical points from these methods are -2.623 and -3.508, respectively, and the estimates of $\text{var}[\log \hat{q}(x) - \log q(x)]$ from the latter are smaller than those from the former for most of $x \in \mathfrak{X}$. The two tolerance bands are quite close, but as expected from the simulation study, the $(V_2,\ c_1)$ band is a little wider than the $(V_1,\ c_2)$ band, and is probably more accurate.

As in the Cyclosporin example, using this tolerance band to assess agreement between the two Oestradiol assays is a matter of subjective judgment. But perhaps the agreement here may be inferred as satisfactory over the entire range of measurement since at the widest point of this band, the absolute difference is about 1800, which is only 15% of the average measurement $x = 12{,}200$ at that point.

6. Discussion

In this article, we considered the problem of assessment of agreement between two methods of measurement. We described the methodology of constructing a probability content tolerance band for the distribution of difference of measurements when the mean or the variance of this distribution are modeled as functions of the observed average of the measurements. This modeling approach provides an alternative to the approach that transforms the measurements to achieve identical distribution for the differences on the transformed scale. The log transformation has been shown to work well in some applications (see Bland and Altman, 1999; Hawkins, 2002 for examples). But perhaps only this transformation produces a difference on the transformed scale that has an easy interpretation (as log-ratios) on the original scale. For the Oestradiol data, Hawkins (2002) uses an ad hoc method to come up with a transformation which, he notes, does not have any easy real-world interpretation. Since in the measuring agreement applications there is a clear need to interpret the differences on the original scale, directly modeling the mean and variance functions of the original scale differences provides a more natural approach.

The application of the methodology proposed here involves numerical computations that can be easily implemented in the popular statistical software R. In fact, for the analysis of both datasets, we used R version 1.9.0 installed on a Dell laptop with a 1.8 GHz Pentium 4 processor, 512 MB of RAM and Windows XP operating system. For the Cyclosporin data, the CPU times for computing the tolerance bands were about 8 minutes for c_1 and less than 1 minute for c_2. For the Oestradiol data, these times were about 12 minutes for the $(V_1,\ c_2)$ method and about 35 minutes for the $(V_2,\ c_1)$ method.

The inference procedure described here inherits the nonrobustness properties of the MLEs. To make it more robust one can use the *M*-estimators in place of the MLEs, and adapt the methodology using the asymptotic theory of *M*-estimators described in Carroll and Ruppert (1988, Chapter 7).

Acknowledgements

We express our sincere thanks to the editor, the associate editor, and the two anonymous reviewers for their comments that greatly improved this article. We also thank Prof. D. M. Hawkins for providing the data.

References

Bland, J. M. and Altman, D. G. (1986). Statistical methods for assessing agreement between two methods of clinical measurement. *Lancet* **1,** 307–310.

Bland, J. M. and Altman, D. G. (1999). Measuring agreement in method comparison studies. *Statistical Methods in Medical Research* **8,** 135–160.

Carroll, R. J. and Ruppert, D. (1988). *Transformation and Weighting in Regression.* New York: Chapman and Hall.

Choudhary, P. K. and Nagaraja, H. N. (2004). Measuring agreement in method comparison studies—A review. In *Advances in Ranking and Selection, Multiple Comparisons, and Reliability,* N. Balakrishnan, N. Kannan, and H. N. Nagaraja (eds), 215–244. Boston: Birkhauser.

Efron, B. and Tibshirani, R. J. (1993). *An Introduction to the Bootstrap.* London: Chapman and Hall.

Guttman, I. (1988). Statistical tolerance regions. In *Encyclopedia of Statistical Sciences,* Volume 9, S. Kotz, N. L. Johnson, and C. B. Read (eds), 272–287. New York: John Wiley.

Hawkins, D. M. (2002). Diagnostics for conformity of paired quantitative measurements. *Statistics in Medicine* **21,** 1913–1935.

Lehmann, E. L. (1998). *Elements of Large Sample Theory.* New York: Springer-Verlag.

Lin, L. I. (1989). A concordance correlation coefficient to evaluate reproducibility. *Biometrics* **45,** 255–268 [Corrections: 2000, 56:324–325].

Lin, L. I. (2000). Total deviation index for measuring individual agreement with applications in laboratory performance and bioequivalence. *Statistics in Medicine* **19,** 255–270.

Lin, L. I., Hedayat, A. S., Sinha, B., and Yang, M. (2002). Statistical methods in assessing agreement: Models, issues, and tools. *Journal of the American Statistical Association* **97,** 257–270.

Loader, C. (1999). *Local Regression and Likelihood.* New York: Springer-Verlag.

Pinheiro, J. C. and Bates, D. M. (2000). *Mixed-Effects Models in S and S-PLUS.* New York: Springer-Verlag.

R Development Core Team. (2004). *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria. `http://www.R-project.org.`

Ruppert, D., Wand, M. P., and Carroll, R. J. (2003). *Semiparametric Regression.* New York: Cambridge University Press.

Sun, J. (2001). Multiple comparisons of a large number of parameters. *Biometrical Journal* **43,** 627–643.

Received August 2004. *Revised April* 2005.
Accepted June 2005.