Semiparametric Regression for Assessing Agreement Using Tolerance Bands

Pankaj K. Choudhary[1]

Department of Mathematical Sciences, University of Texas at Dallas

Abstract

This article describes a Bayesian semiparametric approach for assessing agreement between two methods for measuring a continuous variable using tolerance bands. A tolerance band quantifies the extent of agreement in methods as a function of a covariate by estimating the range of their differences in a specified large proportion of population. The mean function of differences is modelled using a penalized spline through its mixed model representation. The covariance matrix of the errors may also depend on a covariate. The Bayesian approach is straightforward to implement using the Markov chain Monte Carlo methodology. It provides an alternative to the rather ad hoc frequentist likelihood-based approaches that do not work well in general. Simulation for two commonly used models and their special cases suggests that the proposed Bayesian method has reasonably good frequentist coverage. Two real data sets are used for illustration, and the Bayesian and the frequentist inferences are compared.

Key Words: Limits of agreement; Method comparison; Mixed model; Penalized spline; Tolerance interval; Total deviation index.

[1]Address: Department of Mathematical Sciences EC 35, University of Texas at Dallas, PO Box 830688, Richardson, TX 75083-0688, USA. Email: pankaj@utdallas.edu, Tel: (972) 883-4436, Fax: (972) 883-6622.

1 Introduction

In this article, we discuss inference procedures for the following problem: We have a scalar response y_x, which conditional on a scalar continuous covariate $x \in \mathfrak{X}$, follows a $\mathcal{N}(\mu_x \equiv f(x, \beta, b), \sigma_x^2)$ distribution. The mean function f is modelled nonparametrically via penalized splines regression. A p-th degree spline model is,

$$f(x, \beta, b) = \beta_0 + \beta_1 x + \ldots + \beta_p x^p + \sum_{k=1}^{K} b_k (x - c_k)_+^p = X_x \beta + Z_x b, \tag{1}$$

where β is the $(p+1) \times 1$ vector $(\beta_0, \ldots, \beta_p)$; K is the number of knots; b is the $K \times 1$ vector $(b_1, \ldots, b_K)$; $c_1 < \ldots < c_K$ are the knot locations; $b_1, \ldots, b_K$ are the coefficients of the truncated polynomial basis functions $(x - c_1)_+^p, \ldots, (x - c_K)_+^p$; $(x - c)_+ = \max\{0, x - c\}$; X_x is the $1 \times (p+1)$ vector $(1, x, \ldots, x^p)$; and Z_x is the $1 \times K$ vector $\big((x - c_1)_+^p, \ldots, (x - c_K)_+^p\big)$. See Ruppert, Wand and Carroll (2003) for an excellent introduction to the methodology of penalized splines regression. We use the mixed model representation of the spline f, where the coefficients $b_1, \ldots, b_K$ are treated as independently distributed $\mathcal{N}(0, \sigma_b^2)$ variables. In the mixed model terminology, β is called a fixed-effect and b is called a random-effect. Our main interest lies in the parameter function q_x — the p_0-th quantile of $|y_x|$ for a given large probability p_0. It is defined as,

$$q_x = \sigma_x \big\{ \chi_1^2(p_0, \mu_x^2/\sigma_x^2) \big\}^{1/2}, \tag{2}$$

where $\chi_1^2(p_0, \Delta)$ denotes the p_0-th quantile of a noncentral chisquare distribution with one degree of freedom and noncentrality parameter Δ. This function is random under the mixed model representation of (1). Our goal is to obtain a simultaneous upper bound U_x such that

$$Pr(q_x \leq U_x, x \in \mathfrak{X}) = 1 - \alpha. \tag{3}$$

In practice, $\mathfrak{X}$ is a finite interval representing the range of observed values of x.

We are interested in the application of this methodology to assess agreement between two methods of measuring a continuous response. In a typical method comparison study, one method serves as a test method and the other serves as a reference. Generally the test method provides a cheaper or less invasive alternative to the reference method. Sometimes the test method may well be more accurate and precise than the reference method. The goal of their comparison is to evaluate the extent of their agreement and judge whether it is high enough to warrant their interchangeable use in practice. In this context, x is a covariate and y_x represents the population of differences in paired measurements from the two methods at x. The quantile q_x measures the extent of agreement between the methods. Its small value implies a good agreement at x. With U_x defined by (3), the interval $[-U_x, U_x]$, $x \in \mathfrak{X}$, becomes a p_0 probability content simultaneous tolerance band for the distribution of y_x over $\mathfrak{X}$ in the sense that

$$Pr\big\{F_x(U_x) - F_x(-U_x) \geq p_0, \text{ for all } x \in \mathfrak{X}\big\} = 1 - \alpha,$$

where F_x is the cumulative distribution function of y_x. This band estimates the range of p_0 proportion of population differences as a function of x. The practitioner uses it to infer regions of $\mathfrak{X}$ where the differences within the band are clinically unimportant. The agreement in these regions is considered good enough for interchangeable use of the two methods. This approach for agreement evaluation was introduced in Lin (2000), Lin et al. (2002) and Choudhary and Nagaraja (2007) for the case when the differences are independently and identically distributed. The agreement measure here — the p_0-th quantile of absolute differences, is called the "total deviation index" in Lin (2000). Choudhary and Ng (2006) extended this approach for the case when the distribution of differences depends on x, and Choudhary (2006) generalized it to incorporate repeated measurements data. We now introduce

two real data applications.

Oestradiol data: In this example from Hawkins (2002), the interest lies in comparing two assays for Oestradiol — a naturally occurring female hormone synthesized to treat estrogen deficiency. The data consist of pairs of measurements of Oestradiol concentration (in pg/ml) from the two assays. Here we take the difference (assay 1 − assay 2) in concentrations as the response y_x and the average concentration as the covariate x. This average serves as a proxy for the magnitude of the true concentration. Its choice as the covariate is motivated by an exploratory analysis of the data. Let (x_i, y_i) be the value of (x, y_x) on the i-th unit in the sample, $i = 1, \ldots m = 139$. We take $\mathfrak{X} = [\min x_i = 2, \max x_i = 12, 201]$. The scatterplot of $(\log x, y_x)$ in Figure 1(a) reveals that the mean response and the variability in response depend on x. For these data, Choudhary and Ng (2006) consider a model of the form:

$$y_i = f(x_i, \beta, b) + h^{1/2}(x_i, \beta, \lambda)\, \epsilon_i, \tag{4}$$

where the mean function f is given by (1); the errors ϵ_i's follow independent $\mathcal{N}(0, \sigma_e^2)$ distributions and are mutually independent of the random-effects b_k's; and h is a variance function for modelling heteroscedasticity. It follows that the response $y_x \sim \mathcal{N}\big(\mu_x = f(x, \beta, b), \sigma_x^2 = \sigma_e^2\, h(x, \beta, \lambda)\big)$. Based upon this model, the authors describe a likelihood-based methodology for computing U_x to satisfy (3). We analyze these data in Section 4.

[Figure 1 about here.]

Body fat data: In this Young Women's Health Study example from Chinchilli et al. (1996), we are interested in comparing two methods for measuring percentage body fat — skinfold calipers and dual energy x-ray absorptiometry (DEXA). The data consist of paired body fat measurements from the two methods taken over a course of about five years on a cohort of

$m = 91$ adolescent girls. There were nine visits of the girls roughly six months apart with the first visit around age twelve. DEXA measurements are not available for the first visit. We have between four to eight complete pairs of measurements on each girl yielding a total of 654 pairs after excluding three outliers. Figure 2 presents the scatterplots of these data for visits two through nine. The methods do not seem to be highly correlated. Here we take the covariate x as (age in years at the time of visit $- 12$) $\in \mathfrak{X} = [-0.80, 5.30]$. Let y_{ij} be the difference (calipers $-$ DEXA) in the body fat measurements of the i-th girl on the j-th visit, $j = 1, \ldots, n_i$, $i = 1, \ldots, m$, and x_{ij} be the value x at this visit. These longitudinal data are modelled in Choudhary (2006) as

$$y_{ij} = v_i + f(x_{ij}, \beta, b) + \epsilon_{ij}, \tag{5}$$

where the v_i's denote unit-specific random intercepts following independent $\mathcal{N}(0, \sigma_v^2)$ distributions; the errors ϵ_{ij}'s follow $\mathcal{N}(0, \sigma_e^2)$ distributions with autocovariance $cov(\epsilon_{ij}, \epsilon_{il}) = \sigma_e^2 \phi^{|x_{ij} - x_{il}|}$, $0 < \phi < 1$, and are independent for different i; and v_i's, b_k's and ϵ_{ij}'s are mutually independent. This model implies that the response $y_x \sim \mathcal{N}\big(\mu_x = f(x, \beta, b), \sigma_x^2 = \sigma_v^2 + \sigma_e^2\big)$. The methodology for computing U_x by Choudhary and Ng (2006) is adapted in Choudhary (2006) for this model. We return to these data in Section 5.

[Figure 2 about here.]

Both the models (4) and (5) are special cases of a mixed model. Let θ denote the column vector of all model parameters — the fixed-effects, the random-effects and the variance-covariance parameters. Also let $\hat{\theta}$ be an estimator of θ. In the frequentist mixed model framework, the fixed-effects and the variance-covariance parameters are generally estimated by their maximum likelihood estimators (MLE's) and the random-effects are estimated using their estimated best linear unbiased predictors (see, e.g., ch 4, Ruppert et al., 2003). Next

let $\hat{q}_x$ be an estimator of the quantile function q_x obtained by substituting $\hat{\theta}$ for θ in (2). In the frequentist approach, a key issue involved in computing an upper bound U_x for q_x that satisfies (3) is how to obtain the standard error of $(\hat{q}_x - q_x)$. Deriving an analytical expression for it is not so straightforward since q_x is a random function involving θ and the standard technique of delta method is not directly applicable in this case. Choudhary and Ng (2006) do employ a delta method type approximation, but it is expected to work well only when σ_b is small and it requires an estimate of $var(\hat{\theta} - \theta)$, whose direct estimation is largely intractable. However, one can use parametric bootstrap for this estimation. Section 3 gives further details of this approach. There the simulations also show that the resulting U_x does not have good confidence coverage in general for $(\sigma_b/\sigma_e) \geq 1.0$. Another alternative is to simulate a joint bootstrap distribution of $\hat{q}_x$ on a fine grid of x values in $\mathfrak{X}$ and use the percentile method to compute the desired simultaneous upper bound. However, the simulations in Section 3 demonstrate that this procedure also does not have acceptable accuracy for sample sizes typically found in applications.

In this article, we describe a Bayesian approach for computing U_x in a general mixed model setting. It overcomes the above drawbacks of the frequentist approach and is easy to implement using Markov chain Monte Carlo (MCMC) techniques. Our Bayesian approach to semiparametric regression is on the lines of Ruppert et al. (2003, ch 16). Their key point is to first cast the problem of fitting a penalized spline regression model into a problem of fitting a mixed model and then use the Bayesian machinery for performing inference. The main difference between this work and the vast literature on Bayesian smoothing (see, e.g., Lang and Brezger, 2004, and the references therein) is that the parameter of our main interest is a quantile function, which is a nonlinear function of both the mean and the variance functions, instead of the usual mean function.

This article is organized as follows: In Section 2, we describe a Bayesian procedure for computing U_x. Simulations in Section 3 show that it has reasonably good frequentist confidence coverage. In Sections 4 and 5, we revisit the two data sets introduced earlier and compare the frequentist and the Bayesian inferences. Section 6 concludes with a discussion.

2 Methodology for computing U_x

Suppose that the observed data on (x, y_x) for the i-th unit can be modelled using a linear mixed model of the form,

$$y_i = X_i\beta + Z_i b + W_i v + \epsilon_i, \quad i = 1, \ldots, m,$$

where y_i is the $n_i \times 1$ response vector; β is the $(p+1) \times 1$ vector of fixed coefficients in the spline function (1); b is the $K \times 1$ vector of random coefficients in this spline function; v is the $r \times 1$ vector of random-effects of the units; X_i, Z_i and W_i are the design matrices associated with β, b and v, respectively; and ϵ_i is the $n_i \times 1$ vector of within-unit errors. The j-th rows of X_i and Z_i are $(1, x_{ij}, \ldots, x_{ij}^p)$ and $\big((x_{ij} - c_1)_+^p, \ldots, (x_{ij} - c_K)_+^p\big)$, respectively, $j = 1, \ldots, n_i$. We assume that $b|\sigma_b^2 \sim \mathcal{N}(0, \sigma_b^2 I_K)$, with I_K denoting a $K \times K$ identity matrix; $v|(\sigma_v^2, G) \sim \mathcal{N}(0, \sigma_v^2 G)$; $\epsilon_i|(\sigma_e^2, R_i) \sim$ independent $\mathcal{N}(0, \sigma_e^2 R_i)$; and three vectors are conditionally independent, $i = 1, \ldots, m$. Here G, a $r \times r$ matrix, and R_i, a $n_i \times n_i$ matrix, are positive definite; and R_i may depend on x, for example to model heteroscedasticity or autocorrelation. Let $N = \sum_{i=1}^m n_i$, and define

$$y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_m \end{bmatrix}, \; X = \begin{bmatrix} X_1 \\ X_2 \\ \vdots \\ X_m \end{bmatrix}, \; Z = \begin{bmatrix} Z_1 \\ Z_2 \\ \vdots \\ Z_m \end{bmatrix}, \; W = \begin{bmatrix} W_1 \\ W_2 \\ \vdots \\ W_m \end{bmatrix}, \; R = \begin{bmatrix} R_1 & 0 & \cdots & 0 \\ 0 & R_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & R_m \end{bmatrix}.$$

We also assume that X, Z and W have full ranks. This mixed model can now be written in a hierarchical fashion as

$$y|(\beta, b, v, \sigma_e^2, R) \sim \mathcal{N}(X\beta + Zb + Wv,\ \sigma_e^2 R),\ b|\sigma_b^2 \sim \mathcal{N}(0, \sigma_b^2 I_K),\ v|(\sigma_v^2, G) \sim \mathcal{N}(0, \sigma_v^2 G), \quad (6)$$

with last two vectors being conditionally independent.

Consider the vectors X_x and Z_x defined in (1) that represent a row of the design matrices X_i and Z_i, respectively, corresponding to a single covariate value $x \in \mathfrak{X}$. Let W_x denote a similar $1 \times r$ vector associated with the design matrix W_i. Also let R_x be a scalar so that $\sigma_e^2 R_x$ is the error variance at x. From (6), it follows that the distribution of y_x, which represents the population of the scalar response at x, after integrating out the unit random-effects v, is

$$y_x|(\mu_x, \sigma_x^2) \sim \mathcal{N}(\mu_x, \sigma_x^2), \quad \mu_x = X_x\beta + Z_x b,\ \sigma_x^2 = \sigma_v^2 W_x G W_x' + \sigma_e^2 R_x, \quad (7)$$

and W_x' is the transpose of W_x. The parameter of our main interest — the quantile function q_x, is given by (2) upon substituting these expressions for (μ_x, σ_x^2).

To complete the Bayesian specification of model (6), we need priors on $(\beta, \sigma_b^2, \sigma_e^2, \sigma_v^2)$, and also on (R, G) when they are unknowns. Throughout we use "$[\cdot]$" to denote a probability density. For $(\beta, \sigma_b^2, \sigma_e^2)$, we adopt the following priors,

$$[\beta] \equiv 1,\ \sigma_b^2 \sim \mathrm{IG}(A_b, B_b),\ \sigma_e^2 \sim \mathrm{IG}(A_e, B_e),\ \sigma_v^2 \sim \mathrm{IG}(A_v, B_v), \quad (8)$$

where $\mathrm{IG}(A, B)$ represents an inverse gamma distribution, i.e., its reciprocal follows a gamma distribution with mean A/B and variance A/B^2. The four hyperparameters are positive and must be specified. Choosing their values close to zero leads to largely noninformative priors (see ch 16, Ruppert et al., 2003). For β, an alternative is to use a mean zero multivariate normal prior with independent components and a common variance. A large variance will ensure a proper but an essentially noninformative distribution. Gelman et al. (2003, ch 19)

give general advice for priors when (R, G) are unknown arbitrary matrices. Frequently these matrices are either known or are specified in terms of a small number of parameters. One then specifies priors for these parameters. We give some examples in Section 3 for the models (4) and (5). Following Searle, Casella and McCulloch (1992, sec 9.2), it is easily seen that the joint posterior distribution of parameters is proper despite their improper joint prior, at least when (R, G) are known or assume proper priors. However, this posterior is not available in a closed-form and we need to use MCMC techniques to sample from this distribution. The Appendix contains some examples. The computations are easy to program in softwares such as `R` (R Development Core Team, 2006) or `WinBUGS` (Spiegelhalter, Thomas and Best, 2004). See also Crainiceanu, Ruppert and Wand (2005) for examples of how to use `WinBUGS` for penalized splines regression. We have used `R` for all the computations here.

We continue to refer to (6) as a "mixed model" despite the fact that all unknowns are random in a Bayesian context. This somewhat confusing term is fairly common in the Bayesian lexicon (see, e.g., p. 391, Gelman at al., 2003). It simply means that the components of β in this model have infinite or infinitely large prior variance.

We now describe how to compute an upper credible bound (UCB) U_x for the quantile function q_x that approximately satisfies (3). Consider a fine grid $\{x_1^*, \ldots x_g^*\}$ of $\mathfrak{X}$ with say $g = 100$ equally spaced points. Recall that θ is the vector of all model parameters. Let $\{\theta_1, \ldots, \theta_L\}$ be a large number of posterior draws of θ; and $\left\{q_{x_i^*}^1, \ldots, q_{x_i^*}^L\right\}$ be the corresponding draws from the posterior of $q_{x_i^*}$ obtained by applying (2) and (7), $i = 1, \ldots, g$. Further, let $\left\{q_{x_i^*}^{(1)}, \ldots, q_{x_i^*}^{(L)}\right\}$ denote the ordered values of $\left\{q_{x_i^*}^1, \ldots, q_{x_i^*}^L\right\}$, and $\left\{r_i^1, \ldots, r_i^L\right\}$ be their ranks. Define l_α as the $(1-\alpha)$-th sample quantile of $\left\{\max_{i=1,\ldots,g} r_i^j, j = 1, \ldots, L\right\}$. From Besag et al. (1995), the empirical posterior probability of the simultaneous event $\left\{q_{x_i^*} \leq q_{x_i^*}^{(l_\alpha)}, i = 1, \ldots, g\right\}$ is approximately $(1-\alpha)$. Hence we can take $\left\{U_{x_i^*} \equiv q_{x_i^*}^{(l_\alpha)}, i = 1, \ldots, g\right\}$ as the desired UCB

with $(1-\alpha)$ simultaneous credible probability since

$$Pr(q_x \leq U_x,\ x \in \mathfrak{X} \,|\, y) \ \approx\ Pr(q_{x_i^*} \leq U_{x_i^*},\ i = 1, \ldots, g \,|\, y) \ \approx\ 1 - \alpha.$$

The second approximation above is due to the finiteness of the number of posterior draws L. This inference is "exact" with respect to the sample size. The band $\{[-U_{x_i^*}, U_{x_i^*}], i = 1, \ldots, g\}$ now is a Bayesian tolerance band. It has the property that

$$Pr\big\{F_{x_i^*}(U_{x_i^*}) - F_{x_i^*}(-U_{x_i^*}) \geq p_0,\ i = 1, \ldots, g \,|\, y\big\} \ \approx\ 1 - \alpha,$$

where F_x is the cumulative distribution function of y_x in (7).

The method of Besag et al. (1995) is slightly conservative for a finite L as the true empirical probability of the simultaneous event $\big\{q_{x_i^*} \leq q_{x_i^*}^{(l_\alpha)}, i = 1, \ldots, g\big\}$ is more than $(1-\alpha)$ due to the ties in the set $\{\max_{i=1,\ldots,g} r_i^j,\ j = 1, \ldots, L\}$. This problem goes away as $L \to \infty$. Even for a finite L, it is not likely to be a serious concern in practice since L tends to be quite large in comparison with g. Held (2004) provides an alternative to Besag et al.'s approach for making simultaneous posterior probability statements that is appropriate when the primary interest lies in determining how well a certain point in the parameter space is supported by the posterior distribution.

3 Simulation study

It is well-known that, under mild regularity conditions, the confidence coverage of a $(1-\alpha)$ Bayesian credible interval approaches $(1-\alpha)$ as the sample size increases to infinity (see, e.g., ch 4, Gelman et al., 2003). In this section, we use simulation to investigate the simultaneous coverage probability of the UCB $\{U_{x_i^*}, i = 1, \ldots, g\}$ for finite N. We will only consider $N \geq 100$ as the semiparametric regression is typically not used for small sample sizes. Our

investigation will concentrate on the following four models. The Appendix describes how we use MCMC to draw samples from the posteriors in these cases.

Model 1A: Consider model (6) with $n_i \equiv 1$ so that $N = m$; $R = I_m$; $v = 0$; and prior (8). Thus this model is the homoscedastic case of (4), and is given as

$$y|(\beta, b, \sigma_e^2) \sim \mathcal{N}(X\beta + Zb,\ \sigma_e^2 I_m),\ \ b|\sigma_b^2 \sim \mathcal{N}(0, \sigma_b^2 I_K),$$

$$[\beta] \equiv 1,\ \ \sigma_b^2 \sim \mathrm{IG}(A_b, B_b),\ \ \sigma_e^2 \sim \mathrm{IG}(A_e, B_e).$$

In this case, $y_x \sim \mathcal{N}(\mu_x = f(x, \beta, b), \sigma_x^2 = \sigma_e^2)$.

Model 1B: Same as Model 1A except $R = \mathrm{diag}\{x_1^{2\lambda}, \ldots, x_m^{2\lambda}\}$, where $\lambda > 0$; and we take Uniform$(0, A_l)$, $A_l > 0$, as the prior for λ. This is also model (4) with $h(x, \beta, \lambda) = x^{2\lambda}$. We have $y_x \sim \mathcal{N}(\mu_x = f(x, \beta, b), \sigma_x^2 = \sigma_e^2 x^{2\lambda})$ for this model.

Model 2A: Consider model (6) with $v = (v_1, \ldots, v_m)$ as the $m \times 1$ vector of unit-specific random intercepts; $R_i = I_N$; $G = I_m$; and prior (8). In this case, $r = m$; and W_i is a $n_i \times m$ matrix with ones in its i-th column and zeros elsewhere, $i = 1, \ldots, m$. This is model (5) with no autocorrelation in errors. Here $y_x \sim \mathcal{N}(\mu_x = f(x, \beta, b), \sigma_x^2 = \sigma_v^2 + \sigma_e^2)$.

Model 2B: Same as Model 2A except R_i now is a matrix with (j, l)-th element $\phi^{|x_{ij} - x_{il}|}$, $0 < \phi < 1$, $j, l = 1, \ldots, n_i$, $i = 1, \ldots, m$. This is our model (5). We take Uniform$(0, 1)$ as the prior for ϕ. In this case also, $y_x \sim \mathcal{N}(\mu_x = f(x, \beta, b), \sigma_x^2 = \sigma_v^2 + \sigma_e^2)$.

We consider the following settings for the simulation: grid size $g = 100$; credible probability $(1 - \alpha) = 0.95$; degree of spline $p = 2$; $K = 25$ knots; k-th knot location $c_k = \big((k + 1)/(K+2)\big)$-th sample percentile of x; parameter values $(\beta_0, \sigma_e) = (0, 1)$, $\sigma_b \in \{0.5, 1, 2, 4, 8\}$; 10^{-3} as the common variance hyperparameter value; and 10,000 as the number of MCMC iterations with a burn-in period of 1,000, so that the number of posterior draws L =9,000.

In case of Model 1A, we additionally take: covariate range $\mathfrak{X} = [0.1, 0.99]$; number of

units $m = 100$; $(x_1, \ldots, x_m)$ as equally spaced points in $\mathfrak{X}$; probability $p_0 \in \{0.80, 0.95\}$; and parameters $\beta_1, \beta_2 \in \{-1, 1\}$. The parameters (β, σ_b) actually represent $(\beta/\sigma_e, \sigma_b/\sigma_e)$ since $\sigma_e = 1$ without any loss of generality. Thus we study a total of 40 settings covering a variety of scenarios. At each setting, we perform the following steps:

1. Draw $b \sim \mathcal{N}(0, \sigma_b^2 I_K)$. Then compute (μ_x, σ_x^2) using their expressions given with the model definition, and apply (2) to obtain q_x for each $x \in (x_1^*, \ldots, x_g^*)$.

2. Draw $y \sim \mathcal{N}(X\beta + Zb, \sigma_e^2 I_m)$.

3. Simulate from the posterior taking $(x_1, y_1), \ldots, (x_m, y_m)$ as the sample data, and compute the UCB $\{U_{x_i^*}, i = 1, \ldots, g\}$ as described in the previous section. Check whether $\{q_{x_i^*} \leq U_{x_i^*}$, for all $i = 1, \ldots, g\}$.

4. Repeat the above steps 1,000 times. Take the proportion of times the quantile function remains below its UCB as the estimated confidence coverage probability of the UCB.

Table 1 displays the resulting estimates. They are about 2-2.5% higher than the target 95% when $(\sigma_b/\sigma_e) \leq 2$, and seem to decrease towards the target as (σ_b/σ_e) increases. The situation is slightly better for $p_0 = 0.95$ than for $p_0 = 0.80$, and the estimates for different (β_1, β_2) configurations are practically the same. Thus from a frequentist viewpoint, the Bayesian UCB appears conservative, particularly when (σ_b/σ_e) is small. The coverage probability estimates in case of other models also do not seem to depend on (β_1, β_2). So from now on we will present the results only for $(\beta_1, \beta_2) = (1, 1)$. For Model 1B, we consider the same settings as Model 1A, and fix the heteroscedasticity parameter $\lambda \in \{0.25, 0.75\}$ and its hyperparameter $A_l = 1$. The estimated probabilities, obtained similarly as the previous model, are also given in Table 1. They are fairly close to 95% in all cases. We also repeated

this study at several settings with $m = 200$ and $\sigma_b = 16$. We found, as expected, that the estimates are more accurate when m is larger. Further, their values at $\sigma_b = 16$ are close to 95% and are comparable with those at $\sigma_b = 8$.

For Models 2A-B, we simulate data on the lines of the body fat data introduced in Section 1. In particular, we randomly select $m = 25$ units from this data set and take their x values as the settings of our longitudinal covariate x. There are between 6 to 8 observations per unit totalling to $N = 184$. In this case, $\mathfrak{X} = [-0.42, 5.30]$. Moreover, we take $p_0 = 0.80$; $(\beta_1, \beta_2) = (1, 1)$; $\sigma_v \in \{0.5, 1, 2, 4\}$, its hypeparameters equal to 10^{-3}; and autocorrelation $\phi = 0.5$ (for 2B only). The simulation for these models proceeds on the lines of Model 1A with the exception that we also draw $v \sim \mathcal{N}(0, \sigma_v^2 I_m)$ in step 1, and in step 2, $y \sim \mathcal{N}(X\beta + Zb + Wv, \sigma_e^2 R)$ is drawn using the appropriate choice of R. Table 2 shows that the resulting coverage probability estimates are about 96% irrespective of (σ_b, σ_v).

We now focus on the two frequentist alternatives for computing U_x described in the introduction. For simplicity, we consider only the Models 1A-B. The first procedure is from Choudhary and Ng (2006, sec 5) who argue that, when m is large and σ_b^2 is small, $\log \hat{q}_x - \log q_x \approx \mathcal{N}(0, H_x' V H_x)$, where $H_x = (\partial \log q_x / \partial \theta)_{\theta=\hat{\theta}}$ and V is the asymptotic covariance matrix of $(\hat{\theta} - \theta)$. Based on this result, they suggest $U_x = \exp\big(\log \hat{q}_x - c\,(H_x' V H_x)^{1/2}\big)$, $x \in \mathfrak{X}$, as the simultaneous upper bound where the critical point $c\,(< 0)$ is computed to satisfy (3). They consider two methods for approximating (V, c). The first is to estimate V using parametric bootstrap with B resamples, and compute c by numerically solving

$$\alpha = Pr(t_{n-3} \leq c) + \frac{\kappa_0}{2\pi}\left(1 + \frac{c^2}{n-3}\right)^{\frac{n-3}{2}},$$

where t_ν has a t-distribution with ν degrees of freedom, and

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

with $L_x = V^{1/2}H_x$ and $\dot{L}_x = V^{1/2}(\partial H_x/\partial x)$. The second is to use an ad hoc approximation of V obtained by setting the covariances between estimates of the fixed- and random-effects and the variance-covariance parameters to zero, and c is estimated using parametric bootstrap on a fine grid in $\mathfrak{X}$. We, however, study only the first method (with $B = 500$) here as the estimate of V in the second method may be inconsistent. The second frequentist alternative for computing U_x is motivated by our Bayesian approach. It is also based on bootstrap and consists of the following steps after a model is fitted to observed data using ML:

1. Draw $b^* \sim \mathcal{N}(0, \hat{\sigma}_b^2 I_K)$ and $y^* \sim \mathcal{N}(X\hat{\beta} + Zb^*, \hat{\sigma}_e^2 \hat{R})$, where $\hat{R}$ represents I_m for Model 1A and diag$\{x_1^{2\hat{\lambda}}, \ldots, x_m^{2\hat{\lambda}}\}$ for 1B.

2. Use ML to fit model to the resampled data $(x_1, y_1^*), \ldots, (x_m, y_m^*)$. Let $\hat{\beta}^*$, $\hat{\sigma}_e^*$ and $\hat{\lambda}^*$ (for 1B only) respectively be the resulting MLE's of β, σ_e and λ; and $\hat{b}^*$ be the estimated best linear unbiased predictor of b.

3. Compute $\hat{\mu}_x^* = f(x, \hat{\beta}^*, \hat{b}^*)$ and $\hat{\sigma}_x^*$ as $\hat{\sigma}_e$ (for 1A) and as $\hat{\sigma}_e x^{\hat{\lambda}^*}$ (for 1B). Apply (2) to obtain $\hat{q}_x^*$ for each $x \in (x_1^*, \ldots, x_g^*)$.

4. Repeat steps 1-3 a large number of times (say 2,000) to obtain a bootstrap distribution of $\hat{q}_x$ for each $x \in (x_1^*, \ldots, x_g^*)$.

5. To compute the simultaneous upper confidence bound $\{U_{x_i^*}, i = 1, \ldots, g\}$, use the Besag et al. (1995) method described in the previous section with bootstrap resamples of $\hat{q}_x$ in place of the posterior draws of q_x. It can be thought of as an extension of the bootstrap-percentile method for computing upper bounds for multiple parameters.

We refer to the first method above as the "(V, c) method" and the second method as the "bootstrap-percentile method". Our simulation study to investigate the true simultaneous

confidence coverage probability of these frequentist procedures proceeds on the lines of the Bayesian procedure. The simulations are performed at the same settings as before. We use the `nlme` package in `R` for fitting a mixed model. Table 1 contains the probability estimates based on 1,000 replications for the (V, c) method with $N = 100$. They are close to the nominal level of 95% only when $\sigma_b = 0.5$ in case of Model 1A. The method is quite liberal at all other settings. Further, the liberal behavior progressively becomes more severe as σ_b increases. The performance of the bootstrap-percentile method is much worse than the (V, c) method at every setting, so we do not present these results separately. This finding contrasts with that of the Bayesian procedure, which tends to be conservative for small values of σ_b and becomes more accurate as σ_b increases. We also performed simulations for $N = 200$. These results (not shown) indicate that the behavior of the bootstrap-percentile method improves slightly compared with $N = 100$. However, the results are still liberal — about 92% when $\sigma_b = 0.5$ and quite worse for higher σ_b values. In addition, for the (V, c) method, the results for $N = 200$ are similar to $N = 100$ case when $\sigma_b = 0.5$, but are more liberal at all other settings. This may be because the effect of σ_b not being small becomes more prominent at a larger N leading to substantial underestimation in the variability of the estimated q_x.

To summarize, the proposed Bayesian UCB has reasonably good confidence coverage in general for the models investigated with $N \geq 100$. Although it may be somewhat conservative particularly in case of Model 1 with a small (σ_b/σ_e). Among the frequentist procedures, only the (V, c) method is recommended, that too when (σ_b/σ_e) is near 0.5 or smaller.

In all the simulations so far we generated data from the underlying true model. However, as a reviewer pointed out, the mean curve under the true model, $f(x, \beta, b) = X_x\beta + Z_x b$, may become wiggly when σ_b^2 is high. On this reviewer's suggestion, we also investigate the

case when the true mean is a sine function. In particular, we simulate data from the model,

$$y_i = \sin(x_i) + \epsilon_i, \quad i = 1, \ldots, m,$$

where $\epsilon_i | \sigma_e^2 \sim$ independent $\mathcal{N}(0, \sigma_e^2)$. Thus, in this case, $y_x \sim \mathcal{N}(\mu_x = \sin(x), \sigma_x^2 = \sigma_e^2)$. We fit Model 1A to the simulated data and proceed as before to compute the Bayesian and the frequentist bounds, and compare them with the true quantile function. The simulation settings are also the same as before except that $\mathfrak{X} = [-3, 3]$, $\sigma_e \in \{0.25, 0.5, 1\}$ and we only consider $(m, p_0, 1-\alpha) = (100, 0.80, 0.95)$. In the Bayesian case, we additionally thin the draws so that only 1,000 posterior draws are saved. The estimated coverage probabilities of various bounds at the three σ_e settings are as follows: 0.964, 0.967 and 0.961 for the Bayesian method; 0.712, 0.793 and 0.721 for the (V, c) method; and 0.998, 0.996 and 0.899 for the bootstrap-percentile method. These estimates are based on 1,000 replications. A similar investigation with models studied in Lang and Brezger (2004) for simulating data confirms the above finding that the frequentist procedures do not tend to be accurate and their behavior is quite erratic. On the other hand, the Bayesian method tends to be more stable, robust and accurate than its frequentist counterparts.

4 Illustration: Oestradiol data

4.1 Model fitting

In this case, x and y_x respectively represent the average and the population difference of Oestradiol concentrations from the two assays. We model these data using Model 1B of Section 3 with mean function f as a quadratic spline (1) in $\log x$, and variance function h as $x^{2\lambda}$, $\lambda > 0$. The number of knots $K = 34$ and their locations $c_k = \big((k+1)/(K+2)\big)$-th

sample quantile of the unique observed values of $x \in \mathfrak{X}$, $k = 1, \ldots, K$, are chosen using Ruppert et al. (2003, p. 126). We set the four variance hyperparameters equal to 10^{-3} and take $A_l = 2$. These specify a noninformative joint prior. We have a total of 40 parameters in this model (3 for β, 34 for b, and σ_b^2, σ_e^2 and λ).

We apply the MCMC algorithm described in the Appendix for posterior simulation. We run four parallel chains for 10,000 iterations each. Overdispersed starting points for the chains are obtained by randomly drawing values for $\left(\log \sigma_e^2, \log \sigma_v^2, \text{logit}\,(\lambda/A_l)\right)$ from a trivariate t_4-distribution centered on the mode of their marginal posterior density and the scale as the inverse information matrix of this density at the mode. The first 1,000 iterations of each chain are discarded as burn-in, and every eighth iteration is saved to thin the chains. Thus the posterior summaries are based on 1,125 $\times 4 =$ 4,500 draws. We assess the convergence of these chains using the Gelman-Rubin scale reduction factor that compares the within- and between-chain variation for each scalar parameter. They should be close to one if the chains are near their target distribution (see ch 11, Gelman et al., 2003). The 95-th quantile of this factor for each of the 40 parameters is less than 1.01. This finding together with the time series plots of the draws and their cumulative quantile plots (not shown) seem to suggest that an acceptable degree of convergence has been reached. Table 3 displays the MLE's and the posterior summaries for selected parameters. The MLE's and the posterior means differ to some extent except for $\log \sigma_e^2$ and $\text{logit}(\lambda/A_l)$. However, we will see later that the two sets of estimated functions for μ_x and $\log q_x$ are largely identical.

The posterior predictive checks suggested in Gelman et al. (2003, ch 6) verify that the assumed model fits well to these data. Further, the deviance information criterion (DIC) of Spiegelhalter et al. (2003) and graphical checks confirm the need for the spline component in this model. This conclusion was also obtained in Choudhary and Ng (2006) using frequen-

tist criteria. Finally, to assess the sensitivity of posterior to the chosen common variance hyperparameter value of 10^{-3}, we refit this model twice — once with 10^{-1} and again with 10^{-6} as the common value. The resulting posteriors are practically identical, indicating that the hyperparameter choice is unimportant provided they are noniformative.

4.2 Agreement evaluation and comparison

The population difference y_x for the assumed model follows a $\mathcal{N}(\mu_x = f(x, \beta, b), \sigma_x^2 = \sigma_e^2 x^{2\lambda})$ distribution, $x \in \mathfrak{X} = [2, 12201]$. For the agreement evaluation, we take $p_0 = 0.80$ as the cutoff probability and $1-\alpha = 0.95$ as the simultaneous credible probability. Figure 1 presents the posterior mean functions of μ_x and $\log \sigma_x$, and the simultaneous Bayesian tolerance band for the distribution of y_x, $x \in \mathfrak{X}$. The band is computed as described in Section 2 using an equally spaced grid of $g = 100$ points in $\mathfrak{X}$. It gives the estimated range of differences in concentration for 80% of population as a function of the average concentration x. Also included in Figure 1 are the MLE's of μ_x and $\log \sigma_x$, and a simultaneous 95% confidence frequentist tolerance band using the (V, c) method. The bootstrap-percentile band is not included as it is much wider than the two bands. The two curves for μ_x and $\log q_x$ are largely identical. This is not surprising, however, since the posterior means and the MLE's are asymptotically equivalent. The Bayesian band is generally slightly shorter than the frequentist band. This finding contrasts with the simulation results since the MLE and the Bayesian estimate of (σ_b/σ_e) are 2.60 and 3.32, respectively — they fall in the region of its values where the frequentist band tends to be liberal and the Bayesian band tends to be accurate. Their funnel shapes are consistent with the scatterplot of the data. Since the extent of differences appear small compared to the magnitude of measurements, the agreement between the two assays may be considered satisfactory by the practitioner.

5 Illustration: Body fat data

5.1 Model fitting

In this case, x is the age (minus 12) of a girl at the time of measurement and y_x is the population difference in her percent body fat measurements at this time from the calipers and DEXA methods. As in Choudhary (2006), we use Model 2B of Section 3 for these data with f as a quadratic spline (1) in x with $K = 34$ knots located at $c_k = \big((k+1)/(K+2)\big)$-th sample quantile of the unique observed values of $x \in \mathfrak{X}$, $k = 1, \ldots, K$. In addition, we set the six hyperparameters of the variances equal to 10^{-3}. Thus in essence we have a noninformative joint prior distribution. We have a total of 132 parameters $\big((p = 3) + (m = 91) + (K = 34) + 4$ for $(\sigma_b^2, \sigma_e^2, \sigma_v^2, \phi)\big)$ in this case.

We apply the MCMC algorithm described in the Appendix to simulate from the posterior. We run four parallel chains for 5,000 iterations each. The overdispersed starting points for the chains are generated in the same way as the Oestradiol data. The first 500 iterations of each chain are discarded as burn-in and the chains are thinned by saving every fourth iteration. Thus the posterior summaries are based on 1,125 $\times 4 =$ 4,500 draws. The 95-th quantile of Gelman-Rubin scale reduction factor for each of the 132 parameters is less than 1.02. This finding and other graphical diagnostics (not shown) suggest that the chains have converged to an acceptable degree. Table 3 presents the MLE's and the posterior summaries for selected parameters. The MLE's and the posterior means are quite close except for $\log \sigma_v^2$.

Here also the posterior predictive checks verify that the assumed model fits well to these data, and DIC confirms the need for the spline term in this model. These conclusions were also obtained in Choudhary (2006) through a frequentist evaluation. Next, to assess the sensitivity of posterior to the chosen common variance hyperparameter value 10^{-3}, we refit

this model twice — once with 10^{-1} and next with 10^{-6} as the common value. The resulting posteriors are virtually unchanged. It demonstrates that the choice for hyperparameter values is unimportant provided they are noninformative.

5.2 Agreement evaluation and comparison

For the assumed model, the population difference $y_x \sim \mathcal{N}(\mu_x = f(x, \beta, b), \sigma_x^2 = \sigma_e^2 + \sigma_v^2)$, $x \in \mathfrak{X} = [-0.80, 5.30]$. As for the Oestradiol data, we take $p_0 = 0.80$, $1 - \alpha = 0.95$ and $g = 100$. Figure 3 presents the posterior mean functions of μ_x and $\log q_x$, and the simultaneous Bayesian tolerance band $[-U_x, U_x]$ over $\mathfrak{X}$. It also displays the MLE's of μ_x and $\log q_x$, and a simultaneous 95% frequentist tolerance band using an adaptation of the (V, c) method by Choudhary (2006). Here also the two curves for μ_x and for $\log q_x$ are virtually identical. The Bayesian and the frequentist bands have similar shapes. They roughly coincide for ages 14.5 or more. For lower ages, however, the Bayesian band tends to be shorter than the frequentist one. These bands demonstrate how the extent of agreement between calipers and DEXA methods varies with age. The agreement is best between ages 14.5 and 16.5, and is worst around age 13.5. In the best case, 80% of the differences in measurements are estimated to lie within $\pm 4\%$, whereas in the worst case, this interval widens to $\pm 7\%$. The magnitude of body fat measurements in these regions is about 25%. Since a 4% difference in percent body fat measurements is usually considered important by practitioners, the agreement between the two methods does not seem enough for their interchangeable use.

6 Discussion

In this article, we described a Bayesian mixed model methodology for computing a simultaneous tolerance band for normally distributed differences when their mean function is modelled using a penalized spline with polynomial basis functions. This band is used for assessing agreement between two methods of continuous measurement. Simulations suggest that the frequentist ML procedures based on bootstrap do not work well in general for the sample sizes typically encountered in this application.

Although we have focused only on the polynomial basis functions for the spline, the methodology can be easily adapted for splines with radial basis functions. These bases lead to similar inferences with an adequate number of knots. However, as Crainiceanu et al. (2005) note in the context of `WinBUGS`, their choice does affect the mixing of the Markov chain. A rapidly mixing chain converges quickly. These authors have experimented with several basis functions, including the polynomial bases, and conclude that the low-rank thin-plate splines (a special case of the radial bases) is best suited for MCMC sampling from `WinBUGS` as it leads to reduced correlation in the draws. This finding may also be relevant for the MCMC scheme used here as we have also observed high autocorrelation in the variance parameters. One may also thin the chains to reduce the autocorrelation.

The semiparametric model (6) of concern here is a Bayesian mixed model. On the lines of Choudhary (2006), it is straightforward to generalize it to incorporate other continuous and categorical covariates, and also to directly model the measurements from the two methods instead of modelling their differences (as we do here). In addition, several special cases of (6) may also be interest, including the regression model with only the fixed-effects, and the identical distribution case. Some of these issues are currently under investigation.

Appendix: Posterior simulation

Let C be the $N \times (p+1+K+r)$ matrix $[X, Z, W]$ and D be the $(p+1+K+r) \times (p+1+K+r)$ block diagonal matrix $\text{diag}\{0, (1/\sigma_b^2)I_K, (1/\sigma_v^2)G^{-1}\}$. For the model (6) with priors (8), the full conditionals — the conditional posterior of a parameter given all other parameters, are:

$$
\begin{aligned}
&(\beta, b, v)|(\text{others}, y) \sim \mathcal{N}\left(\left(C'R^{-1}C + \sigma_e^2 D\right)^{-1} C'R^{-1}y,\ \sigma_e^2\left(C'R^{-1}C + \sigma_e^2 D\right)^{-1}\right),\\
&\sigma_b^2|(\text{others}, y) \sim \text{IG}\big(A_b + K/2, B_b + b'b/2\big),\\
&\sigma_e^2|(\text{others}, y) \sim \text{IG}\big(A_e + N/2, B_e + (y - X\beta - Zb - Wv)'R^{-1}(y - X\beta - Zb - Wv)/2\big),\\
&\sigma_v^2|(\text{others}, y) \sim \text{IG}\left(A_v + r/2, B_v + v'G^{-1}v/2\right). \qquad (9)
\end{aligned}
$$

The full conditionals for σ_b^2, σ_e^2 and σ_v^2 are independent. When (G, R) are known, this model is conditionally conjugate. So one can simulate a Markov chain using a Gibbs sampler that cycles through the following steps in each iteration until convergence:

1. Sample (β, b, v) from their normal distribution in (9).

2. Sample $(\sigma_b^2, \sigma_e^2, \sigma_v^2)$ from their inverse gamma distributions in (9).

This is the strategy we use for Models 1A and 2A. The terms involving (v, σ_v^2, G) are neglected in the first case as they are not a part of the model. For Model 1B, R depends on λ, and its full conditional distribution is not a standard distribution. In this case, instead of simulating λ alone, we simulate $\left(\log\sigma_e^2, \log\sigma_b^2, \text{logit}\,(\lambda/A_l)\right)$ jointly from their full conditional,

$$
\begin{aligned}
\left[\left(\log\sigma_e^2, \log\sigma_b^2, \text{logit}\,(\lambda/A_l)\right) | (\text{others}, y)\right] \ \propto\ & \det(R)^{-1/2}\,(\lambda/A_l)(A_l - \lambda) \times\\
& (\sigma_e^2)^{-\{(m/2)+A_e\}} \exp\left\{-\big(B_e + (y - X\beta - Zb)'R^{-1}(y - X\beta - Zb)/2\big)/\sigma_e^2\right\} \times\\
& (\sigma_b^2)^{-\{(K/2)+A_b\}} \exp\left\{-\big(B_b + b'b/2\big)/\sigma_b^2\right\},
\end{aligned}
$$

and use an MCMC algorithm that cycles through the following steps in each iteration:

1. Sample $\left(\log \sigma_e^2, \log \sigma_b^2, \text{logit}\left(\lambda/A_l\right)\right)$ from the above distribution using a Metropolis-Hastings algorithm.

2. Sample (β, b) from their normal distribution in (9) omitting (v, σ_v^2, G) terms.

In case of Model 2B, R depends on ϕ and its full conditional also is not available in a closed form. We sample logit(ϕ) from its full conditional,

$$[\text{logit}(\phi)|(\text{others}, y)] \propto \det(R)^{-1/2}\phi(1-\phi)\exp\left\{-(y - X\beta - Zb)'R^{-1}(y - X\beta - Zb)/(2\sigma_e^2)\right\},$$

and use an MCMC algorithm that cycles through the following steps in each iteration:

1. Sample logit(ϕ) from the above distribution using a Metropolis-Hastings algorithm.

2. Sample $(\sigma_b^2, \sigma_e^2, \sigma_v^2)$ from their inverse gamma distributions in (9).

3. Sample (β, b, v) from their normal distribution in (9).

In the Metropolis-Hastings algorithms, we use a normal proposal distribution with mean (vector) equal to the parameter value in the previous iteration and a fixed variance (matrix). This variance (matrix) is obtained from a likelihood-based analysis and is scaled from trial and error to ensure that the acceptance rate of the proposals is about 0.44 for a univariate parameter and about 0.23 for a multivariate parameter (see ch 11, Gelman et al., 2003). Also, to improve the approximation, this algorithm is implemented on a normalizing transformation scale. The starting points for simulation are obtained from a frequentist mixed model fit. Finally, the matrix inverse in (9) is computed using a QR decomposition.

Acknowledgement

We thank the reviewers for a thorough review of the manuscript. Their comments have led to substantial improvements in this work.

References

Besag, J., Green, P., Higdon, D., and Mengersen, K. (1995). Bayesian computation and stochastic Systems. *Statistical Science* **10**, 3–41.

Chinchilli, V. M., Martel, J. K., Kumanyika, S. and Lloyd, T. (1996). A weighted concordance correlation coefficient for repeated measurement designs. *Biometrics* **52**, 341–353.

Choudhary, P. K. (2006). A tolerance interval approach for assessment of agreement in method comparison studies with repeated measurements. Submitted.

Choudhary, P. K. and Nagaraja, H. N. (2007). Tests for assessment of agreement using probability criteria. *Journal of Statistical Planning and Inference* **137**, 279–290.

Choudhary, P. K. and Ng, H. K. T. (2006). Assessment of agreement under non-standard conditions using regression models for mean and variance. *Biometrics* **62**, 288–296.

Crainiceanu, C., Ruppert, D. and Wand, M. P. (2005). Bayesian analysis for penalized spline regression using WinBUGS. *Journal of Statistical Software* **14**, Issue 14.

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

Held, L. (2004). Simultaneous posterior probability statements from Monte Carlo output. *Journal of Computational and Graphical Statistics* **13**, 20–35.

Lang, S. and Brezger, A. (2004). Bayesian P-splines. *Journal of Computational and Graphical Statistics* **13**, 183–212.

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.

R Development Core Team (2006). *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*. Cambridge University Press, New York.

Searle, S. S., Casella, G. and McCulloch, C. E. (1992). *Variance Components*. John Wiley, New York.

Spiegelhalter, D. J., Best, N. G., Carlin, B. P. and van der Linde, A. (2003). Bayesian measures of model complexity and fit (with discussion). *Journal of the Royal Statistical Society, Series B* **64**, 583–616.

Spiegelhalter, D. J., Thomas, A. and Best, N. G. (2004). *WinBUGS Version 1.4 User Manual*. Medical Research Council Biostatistics Unit. Cambridge, UK. `http://www.mrc-bsu.cam.ac.uk/bugs`.

		$p_0 = 0.80$					$p_0 = 0.95$				
				σ_b					σ_b		
Model	Parameter	0.5	1.0	2.0	4.0	8.0	0.5	1.0	2.0	4.0	8.0
		Bayesian method									
1A	$(-1,-1)$	97.2	97.4	97.2	95.7	95.0	96.8	97.2	97.2	95.5	94.9
1A	$(-1,1)$	98.4	97.6	96.8	96.1	95.1	97.9	97.1	96.3	95.7	94.7
1A	$(1,-1)$	98.1	97.6	96.8	95.9	94.5	97.9	97.3	96.4	95.5	95.3
1A	$(1,1)$	97.7	97.5	96.8	95.7	94.4	96.8	96.8	96.0	96.1	94.2
1B	0.25	95.5	95.7	94.8	94.9	94.3	95.6	95.8	95.1	95.0	94.0
1B	0.75	97.7	96.1	94.7	93.5	93.7	96.7	94.6	94.3	95.1	93.9
		Frequentist (V, c) method									
1A	$(-1,-1)$	94.9	93.5	92.3	89.1	85.7	95.0	93.0	93.2	89.6	86.9
1A	$(-1,1)$	94.1	93.9	93.4	89.2	86.8	94.5	93.8	93.4	89.8	88.1
1A	$(1,-1)$	94.4	93.5	93.4	88.7	86.1	95.0	93.6	93.2	89.8	88.5
1A	$(1,1)$	95.6	93.3	92.1	88.2	84.6	95.2	92.9	92.2	87.8	86.3
1B	0.25	92.7	91.3	88.8	86.1	88.1	92.8	93.2	88.9	87.7	84.3
1B	0.75	93.7	90.1	86.7	81.8	80.0	92.8	90.1	88.2	85.1	84.3

Table 1: Estimated simultaneous confidence coverage (in %) of 95% upper bounds for the quantile function q_x in case of model (4) with $N = 100$. The parameter refers to (β_1, β_2) for Model 1A and to λ for Model 1B.

	Model 2A					Model 2B				
			σ_b					σ_b		
σ_v	0.5	1.0	2.0	4.0	8.0	0.5	1.0	2.0	4.0	8.0
0.5	95.5	95.7	96.0	96.9	96.4	95.2	96.0	95.5	95.8	96.1
1.0	96.0	97.0	96.8	97.1	96.3	96.2	95.7	96.3	95.8	96.9
2.0	96.8	96.4	96.1	96.7	96.8	95.7	96.4	96.4	96.2	96.0
4.0	96.1	95.8	96.6	96.1	96.2	96.8	95.4	96.4	95.8	96.0

Table 2: Estimated simultaneous confidence coverage (in %) of 95% Bayesian UCB for the quantile function q_x in case of model (5) with $p_0 = 0.80$ and $N = 184$.

		Oestradiol data				
parameter	MLE	mean	sd	2.5%	50%	97.5%
β_0	7.84	5.84	11.66	-17.49	6.00	28.78
β_1	-19.63	-16.42	14.31	-43.58	-16.84	12.79
β_2	5.50	4.59	3.58	-2.97	4.73	11.20
$\log \sigma_e^2$	2.00	2.12	0.36	1.46	2.12	2.84
$\log \sigma_b^2$	3.91	4.53	1.43	1.88	4.50	7.32
$\text{logit}\,(\lambda/A_l)$	-0.81	-0.83	0.06	-0.96	-0.83	-0.71
		Body fat data				
parameter	MLE	mean	sd	2.5%	50%	97.5%
β_0	2.03	2.01	0.70	0.65	2.01	3.43
β_1	1.59	1.58	1.63	-1.72	1.59	4.77
β_2	-0.36	-0.56	1.49	-3.63	-0.51	2.25
$\log \sigma_e^2$	1.28	1.30	0.10	1.12	1.30	1.53
$\log \sigma_b^2$	1.48	1.46	0.20	1.06	1.47	1.85
$\log \sigma_v^2$	-0.92	-0.43	0.74	-1.82	-0.46	1.13
$\text{logit}\,(\phi)$	-1.32	-1.29	0.31	-1.95	-1.29	-0.69

Table 3: MLE's and posterior summaries of selected parameters for the two data sets.

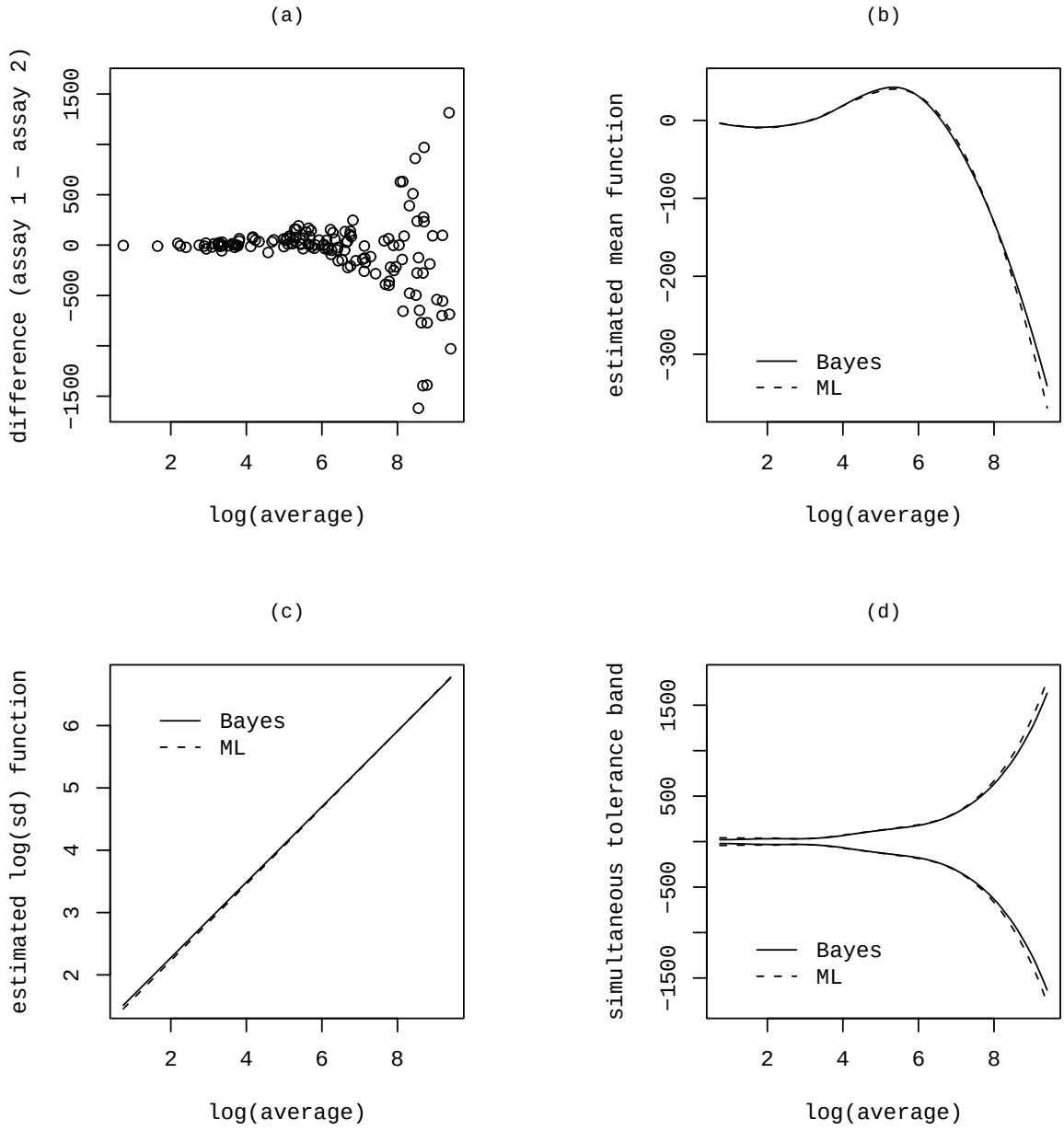

Figure 1: (a) Scatterplot of differences and log-averages for the Oestradiol data. (b) MLE and posterior mean of μ_x. (c) MLE and posterior mean of $\log(\sigma_x)$. (d) Tolerance bands with 80% probability content for the distribution of y_x with simultaneous 95% credible probability in the Bayesian case and simultaneous 95% confidence in the frequentist case.

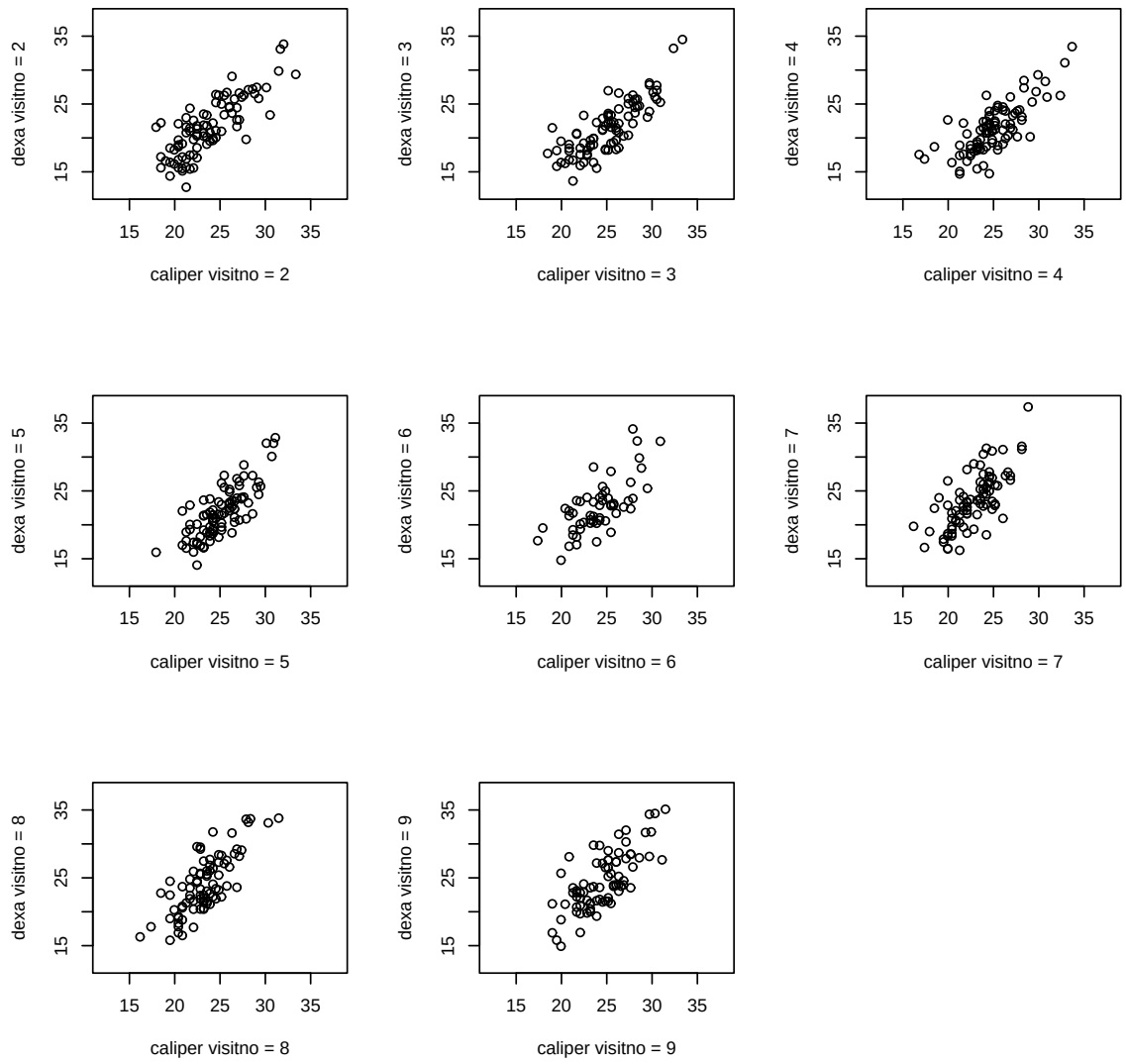

Figure 2: Scatterplots of percent body fat measurements from skinfold calipers and DEXA methods for visit numbers two through nine.

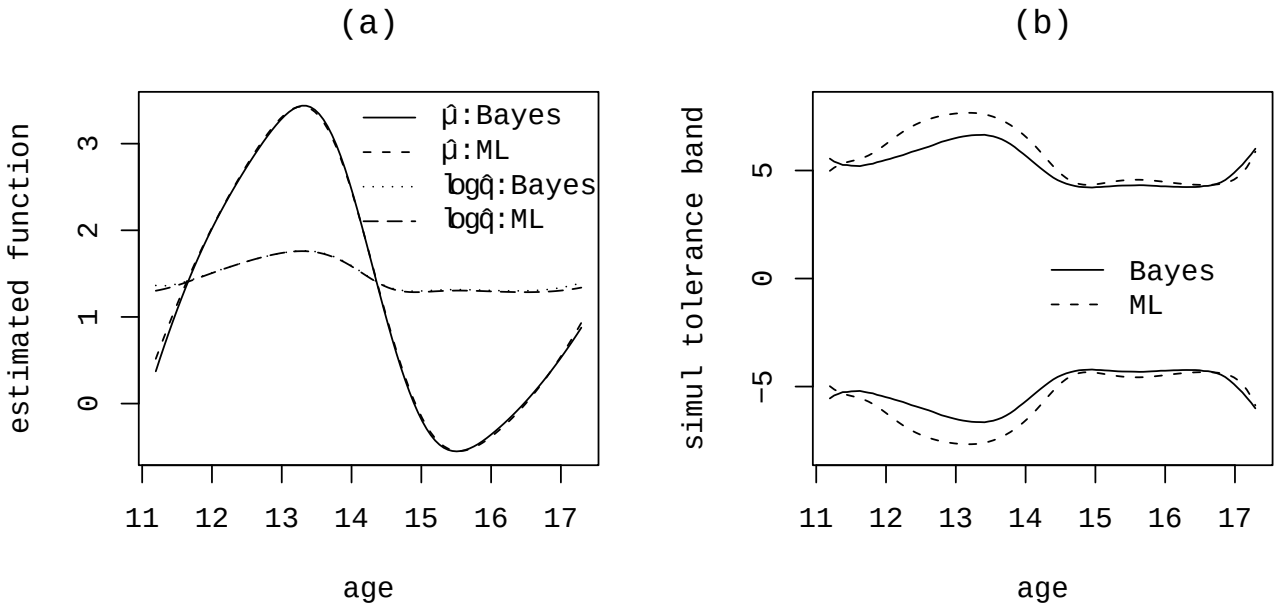

Figure 3: (a) MLE's and posterior means of parameter functions μ_x and $\log q_x$ with $p_0 = 0.80$. (b) Tolerance bands with 80% probability content for the distribution of population difference y_x with simultaneous 95% credible probability in the Bayesian case and simultaneous 95% confidence in the frequentist case.

List of Tables

List of Figures