Bayesian tolerance regions, with an application to linear mixed models

X. Gregory Chen* and A.W. van der Vaart†

Abstract: We review and contrast frequentist and Bayesian definitions of tolerance regions. We give conditions under which for large samples a Bayesian region also has frequentist validity, and study the latter for smaller samples in a simulation study. We discuss a computational strategy for computing a Bayesian two-sided tolerance interval for a Gaussian future variable, and apply this to the case of possibly unbalanced linear mixed models. We illustrate the method on a quality control experiment from the pharmaceutical industry.

Keywords and phrases: quality control, tolerance, linear mixed model.

1. Introduction

The concept of tolerance region is of central importance in quality control. A tolerance region is a prediction set for a future observation, which takes account of both the random nature of this observation and the uncertainty about its distribution. It incorporates the statistical error of estimation of unknown parameters of this distribution.

Conventional tolerance regions take the uncertainty of estimated parameters into account in one of two ways. Either the region captures the future observation a fraction of $1 - \alpha$ times on average over both future and past observations (the $(1 - \alpha)$-expectation tolerance region), or the region captures the future observation with probability at least $1 - \delta$ with $1 - \alpha$ confidence over past observations (the $(\delta, \alpha)$-tolerance region). (We give precise definitions in Section 2.)

The second way appears to be preferred in the pharmaceutical industry. The “on average” and “confidence” can refer to the sampling distribution of the data in a frequentist sense, but can also refer to a posterior distribution in the Bayesian statistical framework. The main focus on the present paper is the second, but we do relate it to the frequentist setup.

Frequentist tolerance regions have been well studied in the literature. A general reference is the book [8], especially for the situation that the data are i.i.d. For the linear mixed model (LMM), the paper [14] provides an elegant solution to build one- and two-sided $(\delta, \alpha)$-tolerance intervals, and includes a comprehensive review of the literature. One purpose of the present paper is to provide a Bayesian approach for the general LMM.

*Biostatistics and Research Decision Sciences, MSD.
†DIAM, Delft Institute of Applied Mathematics. The research leading to these results has received funding from the European Research Council under ERC Grant Agreement 320637 and is partly financed by the NWO Spinoza prize awarded to A.W. van der Vaart by the Netherlands Organisation for Scientific Research (NWO).
The Bayesian formulation of tolerance regions dates back to at least 1964, but the subsequent literature is relatively small. In his paper [1], Aitchison derived Bayesian \((\delta, \alpha)\)-tolerance regions from a decision-theoretic framework, and contrasted them to the frequentist counterparts. In [2] he extended his discussion to \((1-\alpha)\)-tolerance regions, which are a natural way to build Bayesian prediction intervals.

A one-sided \((\delta, \alpha)\)-tolerance interval for a univariate future observation is usually easy to compute, but two-sided tolerance intervals pose challenges, both conceptually and computationally. There are at least two common approaches: intersecting two one-sided tolerance intervals, or fixing one degree of freedom of the interval (e.g. the midpoint of the interval). The former approach is identical to specifying probability masses in the two tails of the distribution of the future variable separately and gives a valid construction, in view of Bonferroni’s inequality, but it yields longer intervals than necessary. (They are called “anti-conservative” in the pharmaceutical industry in reference to the customers, whereas statisticians use the term “conservative”.) See [5] for an example application. The second approach, fixing one degree of freedom, is the conventional choice, especially in the frequentist framework, but requires untangling the dependence of the interval on the unknown true parameter. Solutions are often not available in analytical form and computationally more challenging. Wolfinger in [20] proposed an algorithm to derive a two-sided Bayesian interval for a future normal variate, which was refined by Krishnamoorthy and Mathew [8]. Their algorithms have been widely adopted in practice, and also in other literature (e.g. [7], [16], [15]).

Other contributions to the literature include the following: [10] used the empirical Bayes method to construct a one-sided tolerance interval given an i.i.d. sample from a normal distribution; [6] derived a Bayesian tolerance interval that contains a proportion of observations with a specified confidence; [4] and [21] focused on the sample size needed to attain a certain accuracy; [16] and [15] allowed data from the unbalanced one-way random effects model and the balanced two-factor nested random effects model; [11] discussed probability matching priors (PMP) in the one-sided case to ensure second-order frequentist validity; [12] extended this to the two-sided case; [13] incorporated it to a balanced one-way random effects model, and evaluated its performance against the frequentist method MLS in [8].

Although the PMP approach has merit when the sample size is small, it is analytically demanding even when data are i.i.d., and it seems difficult to extend to the general LMM setting. The algorithms of Wolfinger [20] and Krishnamoorthy and Mathew [8] can be extended to LMM, but they oversimplify the target function during optimization and may result in less satisfactory performance.

In this paper we propose a computationally efficient solution for the general case that the future observation possesses a normal distribution. We show that this is easy to implement given any data model for which a sample from the posterior distribution is available. We investigate when the shortest interval is centered at the posterior mean of the parameter. We discuss the interval in particular for the linear mixed model, and within this context show its good
performance by simulation. We illustrate the method on an example that is representative for pharmaceutical applications. Finally we also prove that the Bayesian interval has frequentist validity in the case of large samples.

2. Definitions and setup

Given are observed data $X$, with a distribution $P_\theta$ depending on a parameter $\theta$, and future unobserved “data” $Z$, with a distribution $Q_\theta$ depending on the same parameter $\theta$. In both cases the sample space is arbitrary. A tolerance region is a set $R(X)$ in the sample space of $Z$ that captures $Z$ with a “prescribed probability”. It will typically be constructed using the observation $X$ to overcome the problem that $\theta$, and hence the law of $Z$, is unknown. There are various ways to make the “prescribed probability” precise, and these can be divided into frequentist and Bayesian definitions. The probability statement will refer to both $X$ and $Z$, and is fixed by one or two parameters $\alpha$ and $\delta$, which are typically chosen small, e.g. 5%.

The parameter $\theta$ will typically be chosen to identify the distribution of $Z$. The distribution of $X$ may also depend on unknown “nuisance” parameters. For simplicity of notation we do not make this explicit in the following. We shall use the notation $P$ or $P_\theta$ for general probability statements, which may be reduced to $P_\theta$ or $Q_\theta$ if the event involves only $X$ or $Z$.

2.1. Frequentist definitions

The most common frequentist definition is the $(\delta,\alpha)$-tolerance region. For a set $R$, abbreviate $Q_\theta(R) = P_\theta(Z \in R)$. Then $R(X)$ is an $(\delta,\alpha)$-tolerance region if

$$P_\theta\left(x : Q_\theta(R(x)) \geq 1 - \delta \right) \geq 1 - \alpha, \quad \forall \theta. \quad (1)$$

If we let $Q_\theta(R(X))$ denote the probability of $R(X)$ under $Q_\theta$, for $X$ held fixed, then we can also write the display in the shorter form $P_\theta\left(Q_\theta(R(X)) \geq 1 - \delta \right) \geq 1 - \alpha$, where the outer probability $P_\theta$ refers to $X$, and the inequality must hold for all possible values of the parameter $\theta$. The latter reminds us of the definition of confidence sets, and indeed it can be seen that $R(X)$ is a frequentist $(\delta,\alpha)$-tolerance region if and only if the set $\mathcal{C}(X) = \{ \theta : Q_\theta(R(X)) \geq 1 - \delta \}$ is a confidence set for $\theta$ of confidence level $1 - \alpha$.

An alternative is the $\alpha$-expectation tolerance region, which requires that

$$\int Q_\theta(R(x)) \, dP_\theta(x) \geq 1 - \alpha, \quad \forall \theta. \quad (2)$$

With the notational convention as before, the display can be written in the shorter form $E_\theta Q_\theta(R(X)) \geq 1 - \alpha$, which is again required for all possible parameter values.

Both definitions have the form of requiring that $E_\theta \ell\left[Q_\theta(R(X))\right] \geq 1 - \alpha$, for all $\theta$, and some given loss function $\ell$. In the two cases this loss function is given...
by $\ell(q) = 1\{q \geq 1 - \delta\}$ for (1), and $\ell(q) = q$ for (2), respectively, where $1\{A\}$ is the indicator function of a set $A$.

2.2. Bayesian definitions

In the Bayesian setup the parameter $\theta$ is generated from a prior distribution $\Pi$, and the densities $p_\theta$ and $q_\theta$ are the conditional densities of $X$ and $Z$ given $\theta$, respectively. To proceed, it is necessary to make further assumptions that fix the joint law of $(\theta, X, Z)$. The typical assumption is that $X$ and $Z$ are independent given $\theta$.

A natural Bayesian approach is to refer to the predictive distribution of $Z$, and define a tolerance region $R(X)$ to be a set such that $\Pr(Z \in R(X) | X) \geq 1 - \alpha$, i.e. a credible set in the posterior law of $Z$ given $X$. The inequality can be written in terms of the posterior distribution $\Pi(\cdot | X)$ as

$$\int \Pr(Z \in R(X) | X, \theta) \, d\Pi(\theta | X) \geq 1 - \alpha.$$

Under the conditional independence assumption this becomes

$$\int Q_\theta(R(X)) \, d\Pi(\theta | X) \geq 1 - \alpha. \quad (3)$$

This is like a frequentist $\alpha$-expectation tolerance region (2), but with the expectation with respect to $X$ under $P_\theta$ replaced by the expectation with respect to $\theta$ under the posterior distribution.

An alternative, derived from a utility analysis by Aitchison [1], is the Bayesian $(\delta, \alpha)$-tolerance region, which is a set $R(X)$ such that

$$\Pi(\theta : Q_\theta(R(X)) \geq 1 - \delta | X) \geq 1 - \alpha. \quad (4)$$

This may be compared to (1). We can also say that $R(X)$ is a Bayesian $(\delta, \alpha)$-tolerance region if and only if the set $C(X) = \{\theta : Q_\theta(R(X)) \geq 1 - \delta\}$ is a credible set at level $1 - \alpha$.

Both types of Bayesian regions satisfy $\int \ell(Q_\theta(R(X))) \, d\Pi(\theta | X) \geq 1 - \alpha$, for the appropriate loss function $\ell$. Solving the region $R(X)$ from such an equation may seem daunting, but good approximations may be easy to obtain using stochastic simulation. This is true even for complicated data models, as long as one is able to generate a sample from the posterior distribution given $X$, for instance by implementing an MCMC procedure. We make this concrete in Section 3.1 for a Gaussian variable $Z$, and illustrate this in Section 4 for an unbalanced linear mixed model (LMM).

2.3. Comparison

The frequentist and Bayesian definitions differ in the usual way in that the frequentist probabilities in (2) and (1) refer to the possible values of $x$ in the
sample space, whereas the Bayesian probabilities in (3) and (4) condition on the observed value of $X$ and refer to the distribution of the parameter.

As is the case for credible sets versus confidence sets, the Bayesian approach may feel more natural.

An advantage of the Bayesian approach is that while the form of the tolerance set $R(X)$ in (4) is determined by the variable $Z$, through the prediction problem $Q_\theta$, the model for the data $X$ enters only through the posterior distribution $\Pi(\theta \in \cdot | X)$. If in the former the dependence on the parameter $\theta$ is not too complicated, then the problem is solvable for even complicated data models. In contrast, the frequentist problem permits explicit solutions only in very special cases, although approximations and asymptotic expansions may extend their use (see [14]).

Neither the frequentist nor the Bayesian formulation restrains the shape of the region $R(x)$. One may prescribe a fixed form and/or seek to optimize the shape with respect to an additional criterion, such as the volume of the region. The Bayesian formulation is again easier to apply, as the optimization will be given the data $X$. In the case of frequentist region it may be necessary to optimize an expected quantity instead.

In general the two approaches give different tolerance regions, but the difference may disappear in the large sample limit. The requirements of the frequentist and Bayesian tolerance regions $R(X)$ for loss function $\ell$ and level $\alpha$, can be given symmetric formulations, as:

$$E\left(\ell\left(Q_\theta(R(X))\right)|\theta\right) \geq 1 - \alpha, \quad \forall \theta,$$

(5)

$$E\left(\ell\left(Q_\theta(R(X))\right)|X\right) \geq 1 - \alpha.$$

(6)

In the first the expectation is taken with respect to $X$, which gives an integral over $x$ with respect to the density $p_\theta$. In the second the expectation is relative to $\theta$, which leads to an integral relative to the posterior distribution given $X$. The integral of the first relative to the prior is identical to the integral of the second relative to the Bayesian marginal distribution of $X$, but there is no reason that (5) implies (6) or vice versa. In particular, a Bayesian tolerance region need not be a frequentist tolerance region.

However, Bayesian and frequentist inference typically merge if the informativeness in the data tends to a limit. For instance, this is true for regular parametric models in the sense that Bayesian credible sets are frequentist confidence sets, in the limit, with corresponding levels. The prior is then washed out and the Bayesian credible sets are equivalent to confidence sets based on the maximum likelihood estimator. This equivalence extends to tolerance regions, under some conditions. We defer a discussion to Section 5.

2.4. One-sided and two-sided tolerance intervals

For a one-dimensional future variable $Z$ it is natural to choose $R(x)$ an interval in the real line. The endpoints of such an interval are referred to as tolerance
limits.

The single finite tolerance limit of a Bayesian one-sided interval is determined by meeting the \((\delta, \alpha)\)- or \(\alpha\)-tolerance criterion. The pair of tolerance limits of a Bayesian two-sided interval might be optimized to give an interval of minimal length, next to requiring that the tolerance criterion is met.

One-sided tolerance limits possess a straightforward interpretation and implementation. In particular, the \((\delta, \alpha)\)-type has a simple description in terms of confidence intervals and posterior quantiles:

- \((-\infty, U(X)]\) is a frequentist \((\delta, \alpha)\)-tolerance interval if and only if it is a \((1-\alpha)\)-confidence interval for the induced parameter \(Q^{-1}_\theta(1-\delta)\); it is a Bayesian \((\delta, \alpha)\)-tolerance interval if \(U(X)\) is the \((1-\alpha)\)-quantile of the posterior distribution of \(Q^{-1}_\theta(1-\delta)\) given \(X\).
- \([L(X), \infty)\) is a frequentist \((\delta, \alpha)\)-tolerance interval if and only if it is a \((1-\alpha)\)-confidence interval for \(Q^{-1}_\theta(\delta)\); it is a Bayesian \((\delta, \alpha)\)-tolerance interval if \(L(X)\) is the \((1-\alpha)\)-quantile of the posterior distribution of \(Q^{-1}_\theta(\delta)\).

Here \(Q^{-1}_\theta(u) = \inf\{z; Q_\theta(-\infty, z] \geq u\}\) is the usual quantile function of \(Z\), and \(Q^{-1}_\theta_{+}(u)\) is the right-continuous version of this quantile function. (The distinction between the two is usually irrelevant, and linked to the arbitrary convention of including the boundary point in the tolerance intervals.) The assertions follow by inverting the inequality \(Q_\theta(R(X)) \geq 1-\delta\), using the fact that for a cumulative distribution function \(Q\) and its quantile function: \(Q^{-1}(u) \leq x\) if and only if \(u \leq Q(x)\), and \(Q^{-1}_{+}(u) < x\) if and only if \(u < Q(x-)\), for \(u \in [0, 1]\) and \(x \in \mathbb{R}\).

A valid two-sided interval might be constructed as the intersection of two one-sided intervals, each at half of the error rate, but this will be conservative and lead to needlessly wide intervals. It makes good sense to try and minimize the length of the interval. We consider this in the next section for the case that the future observation \(Z\) is univariate Gaussian.

3. Normally distributed future observation

Consider the case that the future observation \(Z\) is univariate Gaussian with mean \(\nu\) and variance \(\tau^2\). Thus the parameter is \(\theta = (\nu, \tau)\), and \(Z|X, \nu, \tau \sim Q_\theta = N(\nu, \tau^2)\). The probability that the future observation is captured within a candidate tolerance interval \([L, U]\) is

\[
Q_\theta[L, U] = \Phi\left(\frac{U - \nu}{\tau}\right) - \Phi\left(\frac{L - \nu}{\tau}\right).
\]

It is convenient to parametrize the interval \([L, U]\) by its midpoint \(A = (L + U)/2\) and half length \(B = (U - L)/2\). For given \([L, U]\), or \(\{A, B\}\), and \(\delta \in (0, 1)\), define the set

\[
G_{A, B, \delta} = \{\theta = (\nu, \tau); Q_\theta[L, U] \geq 1 - \delta\}.
\]

For given \([L, U]\) the set \(G_{A, B, \delta}\) is shaped as in Figure 1. It is symmetric about the vertical line \(\nu = (L+U)/2\) and intersects the horizontal axis \(\tau = 0\) in the interval
Fig 1. The set $G_{A,B,\delta}$ of pairs $(\nu, \tau)$ such that $\Phi((U-\nu)/\tau) - \Phi((L-\nu)/\tau) \geq 1-\delta$, for $A=4$, $B=3$, $\delta=0.1$. The number $A$ is its horizontal point of symmetry and $B$ is the half-length of its base. The base of the set (the line segment at height $\tau = 0$) corresponds to the tolerance interval $[L,U]$.

The set $G_{A,B,\delta}$ moves the set $G_{A,B,\delta}$ horizontally, while changing $B$ changes its shape, with bigger $B$ making the set both wider and taller. Although we use the normal distribution as our example, similarly shaped sets and conclusions would be obtainable for other unimodal symmetric distributions.

It follows that $R(X) = [L(X), U(X)]$ satisfies inequality (1) and hence is a frequentist $(\delta, \alpha)$-tolerance interval for $Z$ if and only if $P_{\theta}(x; \theta \in G_{A(x),B(x),\delta}) \geq 1-\alpha$, $\forall \theta$.

In other words $[L(X), U(X)]$ is an $(\delta, \alpha)$-tolerance interval for $Z$ if and only if $G_{A(X),B(X),\delta}$ is an $(1-\alpha)$-confidence region for $\theta = (\nu, \tau)$. Setting a joint confidence interval for location and dispersion is a familiar problem, but here the shape is restrained to the form $G_{A,B,\delta}$ and the focus will be on minimizing $B = B(X)$ (in some average sense). Solutions will depend on the type of data $X$. Standard solutions are available in closed form for the simplest models, and more generally as approximations.

Similarly $R(X) = [L(X), U(X)]$ satisfies inequality (4) and hence is a $(\delta, \alpha)$-Bayesian tolerance interval for $Z$ if $\Pi(\theta; \theta \in G_{A(X),B(X),\delta}|X) \geq 1-\alpha$.

It is natural to choose $A(X)$ and $B(X)$ to satisfy this inequality in such a way that $B(X)$ is minimal. In the resulting optimization problem the posterior distribution $\Pi(\theta \in \cdot |X)$ is a fixed probability distribution on the upper half plane and optimization entails shifting and scaling the shape shown in Figure 1 in a position such that it captures posterior mass at least $1-\alpha$, meanwhile minimizing its width. In Section 3.1 we show how to achieve this numerically given a large sample from the posterior distribution.

The data $X$ determines the posterior distribution, but does not enter the optimization problem. The parameter $\theta$ may not be the full parameter characterising the distribution of $X$, but our computational strategy will work as long
as \( \theta \) is a function of this full parameter. For instance, if \( X \) follows a linear regression model with predictor “time” and \( Z \) is an observation at time 0, then \( \nu \) will be a function of the regression intercept; if \( X \) follows a random-effects model, then typically \( \nu \) will depend on the fixed effects and \( \tau^2 \) will be a specific linear combination of the variance components, depending on practical interests.

Frequentist methods typically choose \( A(X) \) equal to a standard estimator of \( \nu \). One might guess that the Bayesian solution will be to take \( A(X) \) equal to the posterior mean \( \text{E}(\nu|X) \) of \( \nu \). This would be convenient as it would reduce the optimization of \((A,B)\) to the problem of only optimizing \( B \). However, the posterior mean does not necessarily give the minimal length interval. The following lemma gives a sufficient condition.

Lemma 3.1. Suppose that the conditional distribution of \( \nu \) given \((X, \tau)\) is unimodal and symmetric with decreasing density to the right of its mode and has mean \( \text{E}(\nu|X,\tau) \) that is free of \( \tau \). Then the shortest \((\delta, \alpha)\)-Bayesian tolerance interval \([L, U]\) for a future variable \( Z \) given \( X, \nu, \tau \sim N(\nu, \tau^2) \) is centered at the posterior mean \( \text{E}(\nu|X) \).

Proof. We can decompose the probability on the left side of (7) as

\[ \Pi(G_{A,B,\delta}|X) = \int \Pi(\nu \in (G_{A,B,\delta})|X,\tau) \, d\Pi(\tau|X), \]

where \((G_{A,B,\delta})_\tau = \{\nu: (\nu, \tau) \in G_{A,B,\delta}\}\) is the section of \( G_{A,B,\delta} \) at height \( \tau \). By the unimodality and monotonicity the integrand is maximized over \( A \) for every \( \tau \) and a given \( B \) by choosing \( A = \text{E}(\nu|X,\tau) \). If this does not depend on \( \tau \), then this common maximizer \( A \) will maximize the whole expression. Since we need to determine \( A \) and \( B \) so that the expression is at least \( 1 - \alpha \), maximizing it over \( A \) will give the minimal \( B \). By assumption \( A = \text{E}(\nu|X,\tau) = \text{E}(\nu|X) \).}

The condition of the preceding lemma is not unreasonable, but depends on the prior, as illustrated in the following simple example. In Section 4 we show that for a linear mixed model the condition is approximately satisfied. Then choosing \( A \) equal to the posterior mean is a fast computational shortcut that may perform almost as well as the optimal solution.

Example 3.1. The simplest possible data model is to let \( X = (X_1, \ldots, X_n) \) be a random sample from the \( \mathcal{N}(\nu, \tau^2) \)-distribution. This example was already discussed by Aitchison [1]. Here we highlight the implications of Lemma 3.1.

For the standard priors \( \tau^2 \sim \text{IG}(\alpha_0, \beta_0) \) and \( \nu|\tau \sim \mathcal{N}(a, \tau^2/b) \), the conditional posterior distribution of \( \nu \) given \( \tau \) is normal with mean

\[ \text{E}(\nu|\tau, X) = \frac{ba + n\bar{X}}{b + n}. \]

Since this is independent of \( \tau \), the preceding discussion shows that the shortest \((\delta, \alpha)\)-tolerance interval is centred at the posterior mean of \( \nu \).

Choosing the prior variance \( \text{var}(\nu|\tau) \) proportional to \( \tau^2 \), which is customary, is crucial for this finding. For instance, if we set the prior \( \nu \) to be independent
of $\tau$, say $\nu|\tau \sim N(a,b^{-1})$, then the conditional posterior mean changes to
\[
E(\nu|\tau,X) = \frac{ba\tau^2 + n\bar{X}}{br^2 + n}.
\]
This is $\bar{X}$ if $\tau = 0$ and shrinks to the prior mean $a$ as $\tau \to \infty$. For illustration, let $a = 0$, $b = 0.1$, $\alpha_0 = \beta_0 = 0.01$. Given data with $n = 3$, $\bar{x} = 10$, $s^2 = 1$, we approximated the posterior distribution of $(\nu,\tau)$ given $X$ by a Gibbs sampler, using the full conditional posteriors, where $s^2 = \sum (x_i - \bar{x})^2/(n - 1)$:
\[
\begin{align*}
\nu|\tau,X &\sim N\left(\frac{ba\tau^2 + n\bar{X}}{br^2 + n}, \frac{\tau^2}{br^2 + n}\right), \\
\tau^2|\nu,X &\sim IG\left(\alpha_0 + \frac{n}{2}, \beta_0 + \frac{(n - 1)s^2 + n(\bar{X} - \nu)^2}{2}\right).
\end{align*}
\]

The contour plots of the posterior distribution in the left panel of Figure 2 show dependence between $\nu$ and $\tau$ given $X$, and ensuing functional dependence of $E(\nu|\tau,X)$ on $\tau$. Using Algorithm 2, as explained in Section 3.1, we computed the shortest tolerance interval (i.e. the smallest $\hat{B}$) for every possible location $A$ of the interval, for $A$ in a neighbourhood of the posterior mean $E(\nu|X)$. The right panel of Figure 2 shows $\hat{B}$ as a function of $\hat{A}$. The minimum value is not taken at the location of the posterior mean $E(\nu|X)$, which is indicated by a dashed line.

Admittedly the data in this example has been tweaked to illustrate the principle. Inspection of the vertical scale for $\hat{B}$ shows that the global minimal length of a tolerance interval is only slightly smaller than the length of the minimal interval centred at the posterior mean of $\nu$. In Section 4 we theoretically investigate a similar phenomenon for linear mixed models, and in Section 5 we study this approximation in a large sample context.

### 3.1. Computational strategy

In this section we elaborate on the computation of the two-sided Bayesian $(\delta,\alpha)$-tolerance interval for a normally distributed univariate future variable $Z$, as discussed in Section 3. We also compare our approach to the ones taken in [20] and [8]. We assume given a large sample of values $\theta_j = (\nu_j,\tau_j)$ from the posterior distribution of $\theta = (\nu,\tau)$ given $X$. This could be the result of an MCMC run of a sampler for the posterior distribution, or, depending on the data model, of using an analytic formula for the posterior distribution. We shall use the sample values $(\nu_j,\tau_j)$ to approximate expectations under the posterior distribution, whence they need not be independent, and values from a (burnt-in) MCMC run will qualify. Possible dependence together with sample size will determine the error due to simulation.

The idea is to replace the posterior distribution in (4) or (7) by the empirical distribution of the values $\{\theta_j\}_{j=1}$. For a given interval $R(X) = [L,U]$ we can
(in theory) compute the values $Q_{\theta_j}[L,U]$ and next search for the interval $[L,U]$ of minimal length $U - L$ such that

$$\frac{1}{J} \# \{1 \leq j \leq J : Q_{\theta_j}[L,U] \geq 1 - \delta \} \doteq 1 - \alpha,$$

where $\doteq$ means approximately equal, yielding a slightly conservative or anti-conservative solution in case exact equality is not attainable due to discretization.

Wolfinger [20] proposed the algorithm, summarized as Algorithm 1 in the display, and this was refined (or corrected) by Krishnamoorthy and Mathew [8], Chapter 11. This algorithm has a convenient graphical representation and has been widely adopted in practice. The idea is to compute for every $\theta_j$ the quantiles $L_j = Q_{\theta_j}^{-1}(\delta/2)$ and $U_j = Q_{\theta_j}^{-1}(1 - \delta/2)$, yielding intervals $[L_j, U_j]$ with $Q_{\theta_j}[L_j, U_j] \geq 1 - \delta$, and next setting the tolerance interval $[L, U]$ equal to an interval that is symmetric about the posterior mean and contains a fraction $1 - \alpha$ of the intervals $[L_j, U_j]$ (Krishnamoorthy and Mathew, or is contained in a fraction $\alpha$ of these intervals (Wolfinger). The graphical interpretation is to plot the points $(U_j, L_j)$ in the $x$-$y$-plane and search for a point $(U, L)$ on the line $y + x = 2\nu$, for $\nu$ the posterior mean or some other useful estimator, such that a fraction $1 - \alpha$ of the points are in the left-upper quadrant relative to the point $(L, U)$ (see Figure 11.1 in [8] for an example). This method results in an interval that is more confident than the prescribed level $1 - \alpha$, and appears not to optimize the length of the interval.

Here we propose another algorithm that directly utilizes (4). We seek to minimize $B$ under the constraint that the interval $[L, U] = [A - B, A + B]$ satisfies (4). This takes two steps: for fixed $A$ we optimize over $B$; next we perform a grid search over $A$. Because given $A$, the optimizer over $B$ will yield
Algorithm 1: WKM solution for two-sided tolerance interval

Data: Given $a, \delta, \{(\nu_j, \tau_j)\}_{j=1}^J$

1. Let $\hat{A} = \sum \nu_j / J$ ;
2. Calculate two quantiles sequences: $\{L_j \equiv Q_{\nu_j,\tau_j}(\frac{1}{2})\}_{j=1}^J$ and
   $\{U_j \equiv Q_{\nu_j,\tau_j}(1 - \frac{1}{2})\}_{j=1}^J$ ;
3. Find a point $(\hat{L}, \hat{U})$ such that $\hat{L} + \hat{U} = 2\hat{A}$ satisfying one of the following ;
4. (W) arg min$_{L,U}$$|\frac{S}{J} - \alpha|$, where $S = \{(L_j, U_j): L_j \leq L, U_j \geq U\}$;
5. (KM) arg min$_{L,U}$$|\frac{\# S}{J} - 1 + \alpha|$, where $S = \{(L_j, U_j): L_j \geq \hat{L}, U_j \leq \hat{U}\}$;

Result: two-sided tolerance interval $[\hat{L}, \hat{U}]$

---

Fig 3. Left panel: plot of the curves $g \rightarrow \Phi\left(\frac{\epsilon + g}{\tau}\right) - \Phi\left(\frac{\epsilon - g}{\tau}\right)$ for various settings of $\epsilon$ and $\tau$. Right panel: derivatives of these curves.

equality in (4), $\hat{B}$ will be the solution to

$$\Pi \left[ \Phi\left(\frac{A + B - \nu}{\tau}\right) - \Phi\left(\frac{A - B - \nu}{\tau}\right) \geq 1 - \delta \mid X \right] = 1 - \alpha. \quad (8)$$

The posterior mean $E(\nu \mid X)$ will typically be close to the optimal solution for $A$, and is a good starting point for this parameter. As a fast approximation we may also set $A$ equal to this value and omit the grid search.

In practice, we replace the posterior distribution in equation (8) by an average over the sample values $(\nu_j, \tau_j)$. The posterior mean of $\nu$ can be approximated by the average of the sample values $\nu_j$. Given $\hat{A}$ we approximate $\hat{B}$ by the $(1 - \alpha)$-quantile of the points $g_j$ computed as the solutions to

$$Q_{\nu_j,\tau_j}[\hat{A} - g_j, \hat{A} + g_j] \equiv \Phi\left(\frac{\hat{A} + g_j - \nu_j}{\tau_j}\right) - \Phi\left(\frac{\hat{A} - g_j - \nu_j}{\tau_j}\right) = 1 - \delta. \quad (9)$$

The motivation for this procedure is that $(\nu_j, \tau_j) \in G_{\hat{A}, \hat{B}, \delta}$ if and only if $Q_{\nu_j,\tau_j}[\hat{A} - B, \hat{A} + B] \geq 1 - \delta$, whence precisely the points $(\nu_j, \tau_j)$ with $g_j \leq \hat{B}$ satisfy $Q_{\nu_j,\tau_j}[\hat{A} - B, \hat{A} + B] \geq 1 - \delta$ and hence are inside the set $G_{\hat{A}, \hat{B}, \delta}$, whereas the other points are outside this set. This makes the posterior mass of the set equal to $1 - \alpha$ up to simulation error.
Given $\hat{A}$ and $(\nu_j, \tau_j)$, the function $g \mapsto Q_{\nu_j, \tau_j} [\hat{A} - g, \hat{A} + g]$ in (9) is increasing, from the value 0 when $g = 0$ to 1 as $g \to \infty$ (see Figure 3). The solutions $g_j$ to each equation (9) can be found fast by a Newton-Raphson algorithm, with some caution on choosing the initial value for $g_j$ (the algorithm will diverge if the initial value is chosen in the domain where $Q_{\nu_j, \tau_j} [\hat{A} - g_j, \hat{A} + g_j]$ is very close to 1). An appropriate algorithm is listed in Algorithm 2. Note that the (middle) expression in (9) does not change if $\varepsilon := \hat{A} - \nu_j$ is replaced by $-\varepsilon$.

4. Data from a linear mixed model

In this section we apply the preceding to a model that is representative for practice in pharmaceutical quality control: the linear mixed model (LMM). We assume that the data $X$ are acquired in an LMM design, and that the future variable $Z$ is defined in terms of the same LMM. We concentrate attention to the two-sided ($\delta, \alpha$)-tolerance interval.

In the LMM we observe a vector $X = U\beta + V\gamma + e$, for known (deterministic) matrices $U$ and $V$ of covariates, a vector of fixed effects parameters $\beta$, an unobserved random effect vector $\gamma$, and an error vector $e$. Assume that $\gamma$ and $e$ are independent, with

$$\gamma \sim N(0, D), \quad e \sim N(0, \sigma^2 I).$$  \hfill (10)

Then the data $X$ follows a $N(U\beta, C)$-distribution, for $C = VDV^T + \sigma^2 I$, and the full parameter is $(\beta, D, \sigma^2)$.

Consider predicting a new observation $Z = u^T \beta + v^T \gamma' + e'$ with given fixed and random effects coefficients $u$ and $v$ and newly generated random effect vector $\gamma'$ and error $e'$, with $\gamma' \sim \gamma$ and $e' \sim N(0, \sigma^2 I)$. Thus $\gamma'$ is assumed equal in distribution to, but independent of $\gamma$, and similarly for $e'$. This target for

---

**Algorithm 2: Proposed solution for two-sided tolerance interval**

**Data:** Given $\alpha, \delta, \{((\nu_j, \tau_j))_{j=1}^{J}\}$

1. Let $\tilde{A} = \sum \nu_j / J$
2. for $j = 1, 2, \ldots, J$ do
   3. Solve equation (9) by a Newton-Raphson algorithm as follows;
   4. if $|\tilde{A} - \nu_j| < \tau_j$ then
      5. $g_0 = |\tilde{A} - \nu_j| + \tau_j$
   6. else
      7. $g_0 = |\tilde{A} - \nu_j|$
   8. set initial value for $g_j$ at $g_0$;
   9. while $\omega > 0.00001$ do
      10. let $q_{\nu_j, \tau_j} [\tilde{A} - g_j, \tilde{A} + g_j]$ be the first-order derivative of $Q_{\nu_j, \tau_j} [\tilde{A} - g_j, \tilde{A} + g_j]$;
      11. $g_j = g_j - q_{\nu_j, \tau_j} [\tilde{A} - g_j, \tilde{A} + g_j] - \delta$
      12. $\omega = Q_{\nu_j, \tau_j} [\tilde{A} - g_j, \tilde{A} + g_j] - 1 + \delta$
   13. The above loop results in $(g_j)_{j=1}^{J}$, and let $\hat{B}$ be its $(1 - \alpha)$th sample quantile;

**Result:** two-sided tolerance interval $[\hat{A} - \hat{B}, \hat{A} + \hat{B}]$
prediction is reasonable in many contexts, but sometimes another choice, in particular for $\gamma'$, may be more relevant. Typically $\gamma$ will carry a group structure matched by a block structure in $V$. The vector $v$ will then have nonzero coordinates corresponding to a single group. The target corresponds to setting the distribution $Q_\theta$ of $Z$ equal to $N(\nu,\tau^2)$, with
\[ \nu = u^T \beta, \quad \tau^2 = v^T D v + \sigma^2. \quad (11) \]
The “prediction” parameter $\theta = (\nu,\tau^2)$ is of smaller dimension than the full parameter $$(\beta, D, \sigma^2)$$ governing the distribution of the data $X$, whence part of the latter full parameter can be considered a nuisance parameter. To set a Bayesian tolerance interval we need a posterior distribution of $\theta$ given the data $X$. This will typically be inferred from a posterior distribution of the full parameter, resulting from a prior distribution on $$(\beta, D, \sigma^2)$$.

The (conditional) posterior distribution for a conditional prior $\beta|D,\sigma^2 \sim N(0, \Lambda)$, where $\Lambda$ may depend on $(D, \sigma^2)$, satisfies
\[ \beta|D,\sigma^2, X \sim N \left((U^T C^{-1} U + \Lambda^{-1})^{-1} U^T C^{-1} X, (U^T C^{-1} U + \Lambda^{-1})^{-1}\right). \]
(An alternative expression for the posterior mean is $\Lambda U^T (U \Lambda U^T + C)^{-1} X$.) In general $E(\nu|D, \sigma^2, X) = u^T E(\beta|D, \sigma^2, X)$ will depend on $D$ and $\sigma^2$ (hidden in $C$) and hence typically also on $\tau^2$, in view of (11). Therefore Lemma 3.1 does not apply, and there appears to be no reason that a shortest tolerance interval would be centered at the posterior mean of $\nu$. To obtain the shortest interval, Algorithm 2 should be augmented with a search on possible centerings $A$. As in the standard i.i.d. model in Example 3.1, the dependence on $\sigma^2$ can be removed by choosing the variances $\Lambda$ and $D$ proportional to $\sigma^2$. If $D = \sigma^2 D_0$ and $\Lambda = \sigma^2 \Lambda_0$, then $C$ will be $\sigma^2 (V D_0 V^T + I) =: \sigma^2 C_0$ and the conditional posterior mean of $\beta$ will be $(U^T C_0^{-1} U + \Lambda_0^{-1})^{-1} U^T C_0^{-1} X$. However, the dependence on $D_0$ (through $C_0$) remains, in general.

Letting the prior covariance matrix $\Lambda$ tend to infinity corresponds to a non-informative prior on $\beta$. If all other quantities are fixed and $\Lambda \to \infty$, then
\[ E(\beta|D, \sigma^2, X) \to (U^T C^{-1} U)^{-1} U^T C^{-1} X. \]
The limit is the maximum likelihood estimator of $\beta$ in the model where $C$ is known. Since this is still dependent on $C$ (and hence $D$ and $\sigma^2$), it seems that for both the Bayesian and frequentist tolerance intervals the two parameters $\nu$ and $\tau$ cannot be separated in general. The choice $\Lambda = \lambda (U^T C^{-1} U)^{-1}$ leads to $\lambda/(1 + \lambda)$ times the maximum likelihood estimator, and hence also still depends on $C$.

### 4.1. Approximations to the conditional posterior mean

For special designs the dependence of $E(\beta|D, \sigma^2, X)$ on $(D, \sigma^2)$ is only mild and can be quantified. We discuss three examples of linear mixed models. For clarity we restrict to balanced designs, although its flexibility makes the Bayesian
approach particularly valuable for unbalanced cases, as is illustrated by the numerical examples in Section 4.2.

Example 4.1 (One-way random effects). Suppose $X$ is a vector with coordinates $X_{ik} = \beta + \gamma_i + \epsilon_{ik}$, for $i = 1, \ldots, m$ and $k = 1, \ldots, n$, ordered as $(X_{11}, \ldots, X_{1n}, X_{21}, \ldots, X_{mn})$, where $\beta \in \mathbb{R}$ and $\gamma = (\gamma_1, \ldots, \gamma_m)^T$ with i.i.d. $\gamma_i \sim N(0, d^2)$, so that $D = d^2 I_m$, for $I_m$ the $(m \times m)$-identity matrix. As prior on $\beta$ we choose a one-dimensional normal distribution $N(0, \lambda^2)$.

The matrix $U$ is the $mn$-vector $1_{mn}$ with all coordinates equal to 1, while $V$ is the $(mn \times m)$-matrix with $i^{th}$ column having 1s in rows $(i-1)n + 1$ to $in$ and 0s in the other rows. Then $V^T V = nI_m$, and $U^T V = n1_m$, and it can be verified that $C_{1mn} = (nd^2 + \sigma^2)1_{mn}$ and hence $C^{-1}U = (nd^2 + \sigma^2)^{-1}1_{mn}$. The coefficient vector of $E(\beta | D, \sigma^2, X)$ is

$$
(U^T C^{-1}U + \lambda^{-2})^{-1}U^T C = \left(mn + \frac{nd^2 + \sigma^2}{\lambda^2}\right)^{-1}1_{mn}^T.
$$

For $\lambda = \infty$, this is free of $d^2$ and $\sigma^2$, while for finite, fixed $\lambda$ and $m, n \to \infty$, the coefficient vector is $(mn)^{-1}(1 + O(d^2/(m\lambda^2)) + O(\sigma^2/(mn\lambda^2)))$.

The dependence on $d$ and $\sigma$ can be removed by choosing $d = d_0 \sigma$ and $\lambda = \lambda_0 \sqrt{nd_0^2 + 1}\sigma$.

Example 4.2 (Full random effects). Suppose $X_{ik} = u^T_{ik}\beta + v^T_{ik}\gamma_i + \epsilon_{ik}$, ordered as in the preceding example, but now with observed covariates $u_{ik} \in \mathbb{R}^p$ and $v_{ik} \in \mathbb{R}^q$, fixed effects parameter $\beta \in \mathbb{R}^p$ and i.i.d. random effects $\gamma_i \sim N_q(0, D_q)$, for $i = 1, \ldots, m$. The corresponding matrices $U$ and $V$ are

$$U = \begin{pmatrix} U_1 \\ \vdots \\ U_m \end{pmatrix}, \quad V = \begin{pmatrix} V_1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & V_m \end{pmatrix}, \quad U_i = \begin{pmatrix} u^T_{i1} \\ \vdots \\ u^T_{in} \end{pmatrix}, \quad V_i = \begin{pmatrix} v^T_{i1} \\ \vdots \\ v^T_{in} \end{pmatrix}.$$

Then $C = VDV^T + \sigma^2 I$ is an $(mn \times mn)$-block-diagonal matrix with $m$ blocks $V_i D_q V_i^T + \sigma^2 I_n$, and

$$U^T C^{-1}U = \sum_{i=1}^{m} U_i^T (V_i D_q V_i^T + \sigma^2 I_n)^{-1} U_i,$$

$$U^T C^{-1} = \left(U_1^T (V_1 D_q V_1^T + \sigma^2 I_n)^{-1}, \ldots, U_m^T (V_m D_q V_m^T + \sigma^2 I_n)^{-1}\right).$$

The matrices $U_i^T V_i$ and $V_i^T V_i$ are of dimensions $p \times q$ and $q \times q$, and are sums over the $n$ observations (for $k = 1, \ldots, n$) per group $i$, as defined by the random effect $\gamma_i$. For conventional asymptotics we could view them as $n$ times a matrix
of fixed order. Then, for $D_0 = \sigma^{-2} D_q$, 
\[
(V_i D_0 V_i^T + I)^{-1} = I - V_i(D_0^{-1} + V_i^T V_i)^{-1} V_i^T \\
= I - V_i(V_i^T V_i)^{-1} [D_0^{-1}(V_i^T V_i)^{-1} + I]^{-1} V_i^T \\
= I - V_i(V_i^T V_i)^{-1} \left[ I - D_0^{-1}(V_i^T V_i)^{-1} + (D_0^{-1}(V_i^T V_i)^{-1})^2 + \cdots \right] V_i^T \\
= P_{V_i^\perp} + V_i(V_i^T V_i)^{-1} D_0^{-1}(V_i^T V_i)^{-1} V_i^T \\
- \sum_{k=2}^{\infty} (-1)^k V_i(V_i^T V_i)^{-1}(D_0^{-1}(V_i^T V_i)^{-1})^k V_i^T,
\] 
(12)
where $P_{V_i^\perp}$ is the projection on the orthocomplement of the linear span of the columns of $V$. If $V_i^T V_i$ is large, then the series on the right can be neglected, as its terms contain multiple terms $(V_i^T V_i)^{-1}$. There is then still dependence on $D$ and $\sigma^2$ in the second term, which may dominate the first term.

In a full random effects model, every random effect is matched by a fixed effect with the same covariate vector (supplying a common mean value to the random effects), and hence $u_{ik} = v_{ik}$, for every $(i, k)$, and consequently $U_i = V_i$. Then $U_i^T P_{V_i^\perp} = 0$ and $U_i^T V_i(V_i^T V_i)^{-1} = I_p$. If $n^{-1} V_i^T V_i$ stabilizes as $n \to \infty$, 
\[
\sigma^2 U^T C^{-1} U = \sum_{i=1}^{m} \sum_{k=1}^{\infty} (-1)^{k-1} D_0^{-1} ((V_i^T V_i)^{-1} D_0^{-1})^{k-1} = m \left( D_0^{-1} + O\left(\frac{1}{n}\right) \right),
\]
\[
\sigma^2 U^T C^{-1} = \sum_{k=1}^{\infty} (-1)^{k-1} \left( D_0^{-1}(V_i^T V_i)^{-1} \right)^k V_i^T = \left( D_0^{-1}(V_i^T V_i)^{-1} \left( I + O\left(\frac{1}{n}\right) \right) V_i^T \right)^m.
\]
Thus we find that the coefficient vector of the posterior mean of $\beta$ satisfies 
\[
(U^T C^{-1} U + \Lambda^{-1})^{-1} U^T C^{-1} \\
= \left( \frac{1}{m} \left( I + O\left(\frac{D_0}{n}\right) + \frac{D_0\Lambda^{-1}}{m} \right)^{-1} \left( (V_i^T V_i)^{-1} \left( I + O\left(\frac{1}{n}\right) \right) V_i^T \right) \right)^m.
\]
To first order this is free of $D$ and $\sigma^2$ with relative remainders of the order $D_0/n$ and $D_0\Lambda^{-1}/m$.

**Example 4.3** (Random effects with additional fixed effects). In the preceding example every random effect may be matched by a fixed effect with the same covariate vector (supplying a common mean value to the random effects), but there more fixed than random effects. This corresponds to setting $u_{ik}^T = (v_{ik}^T, u_{ik}^T)$, which implies $U_i = (V_i, \hat{U}_i)$, for an $(n \times (p - q))$-matrix $\hat{U}_i$. The formulas in the preceding example must then be adapted to, where the approximations refer to
cross-comparison against the performance of a standard frequentist solution. We consider an interval with true coverage no less than \(\theta\) and allow for the simulation error, which has a standard error of \(\sigma^2\) up to order 1/n. It follows that again the matrix is free of \(D_{i}^{T}\) resulting in a factor \(1/n\) smaller in order than the matrix \(\bar{U}_{i}^{T}P_{V_{i}}^{+}\). The intra-correlation related to a future observation \(Z\) is expected to be small.

4.2. Numerical examples

We evaluate the small sample performance of our proposed algorithm via a simulation study, with data generated from a one-way random effects model. We observe \((X_{ik} : i = 1, 2, \ldots, m; k = 1, 2, \ldots, n_{i})\), where \(X_{ik}\) is the \(k\)th value of the \(i\)th group and satisfies \(X_{ik} = \nu + \gamma_{i} + e_{ik}\) for i.i.d. \(\gamma_{i} \sim N(0,d^2)\) independent of the i.i.d. error \(e_{ik} \sim N(0,\sigma^2)\). We are interested in the two-sided \((\delta, \alpha)\)-tolerance interval related to a future observation \(Z = \nu + \gamma + e \sim Q_{\theta} = \mathcal{N}(\nu, \tau^2)\), where \(\gamma\) and \(e\) are independent copies of the \(\gamma_{i}\) and \(e_{ik}\). The parameter is \(\theta = (\nu, \tau^2)\), for \(\tau^2 = d^2 + \sigma^2\).

We used 6 different parameter settings. In every setting the overall mean \(\nu = 0\), and the group variance to \(d^2 = 1\). The intra-correlation \(\sigma^2/(d^2 + \sigma^2)\) was chosen equal to the numbers 0.1, 0.3, 0.5, 0.7, 0.9. In every setting the number of groups \(m = 6\) and the group sizes were \((n_{1}, \ldots, n_{6}) = (2, 3, 4, 2, 3, 4)\). This simulation setup is the same as in \([14]\), and facilitates a cross-comparison against the performance of a standard frequentist solution.

We computed the \((\delta = 0.1, \alpha = 0.05)\)-tolerance interval \([L, U]\) by Algorithm 2, for every of \(K = 1000\) replicates of the data in each parameter setting, and computed the true coverage \(Q_{\theta}[L, U]\) of these intervals using the true parameter \(\theta\) of the simulation. We consider an interval with true coverage no less than \(1 - \delta\) as “qualified”, and compared the empirical fraction of qualified intervals out of the \(K\) replicates to the nominal value \(1 - \alpha\). The procedure is considered to perform well in the frequentist sense if this empirical fraction is close to this nominal value. Here we must allow for the simulation error, which has a standard error of \(\sqrt{p(1-p)/K}\), for \(p\) the true coverage, which is unknown, but hopefully close to \(1 - \alpha\).
For each simulated dataset the posterior distribution of \((\nu, d, \sigma^2)\) was approximated by a standard Gibbs sampler (with the vector of random effects \(\gamma_i\) added in as a fourth parameter), before utilizing Algorithm 2. Two setups of priors were deployed, both with independent priors on the three parameters \(\nu, d, \sigma\). The first is the vanilla setup with vague marginal prior distributions \(\nu \sim N(0, 1000)\), \(d^2 \sim IG(0.001, 0.001)\), and \(\sigma^2 \sim IG(0.001, 0.001)\). The second uses the same prior on \(\sigma^2\), but uses a \(t\)-distribution for \(\nu\) given by the hierarchy \(\nu|\sigma_0 \sim N(0, \sigma_0^2)\) and \(\sigma_0^2 \sim IG(0.001, 0.001)\), and the prior on \(d\) given by the structural equation \(d = |\xi|\omega\) for independent \(\xi \sim N(0, 1)\) and \(\omega^2 \sim IG(0.001, 0.001)\). The latter specification can also be understood as over-parameterizing the distribution of the random effect using two parameters instead of one, as \(\gamma|\xi, \omega \sim N(0, \xi^2\omega^2)\). This “parameter expansion” is meant to enhance the mixing rate of the Gibbs sampler, in particular when the number of groups \(m\) is small. See [3] for a comparison of methods (including nonBayesian methods) to fit the LMM.

In all simulation settings the tolerance intervals were constructed both by fixing the center point \(A\) at the posterior mean \(E(\nu|X)\), and by seeking an optimal value of \(A\) to minimize the half length \(B\) of the interval. Thus four tolerance intervals were calculated based on each simulated dataset. To save on computation time the optimization over \(A\) was carried out only approximately. Still for 85% of the simulation cases a shorter interval was obtained than the interval at the posterior mean, in a few cases as much as 20% shorter, but in 75% of the cases no more than a few percentage points. Table 1 reports the quotients of the lengths.

The proportions of intervals that attain true coverage at least \(1 - \delta = 0.9\), are listed in Table 2. They are reasonably close to the nominal value \(1 - \alpha = 0.95\), with deviations in both directions up to several percentage points. The performance seems to depend on the true intra-correlation. This dependence follows a similar pattern as for the high-order asymptotic solution in [14] without correction for imbalance, and is close to their solution that includes correction when the intra-correlation is small or very big. The tolerance intervals centered at the (approximately) optimal value have shorter length and attain lower confidence, but their performance seems to surpass slightly the intervals centered at the posterior mean \(E(\nu|X)\).

The difference in performance of Algorithm 2 between the two prior setups is within the order of the simulation error.

5. Frequentist justification of the Bayesian procedure

In this section we show that Bayesian tolerance regions are often also approximate frequentist tolerance regions, of corresponding levels. We consider an asymptotic setup, with data \(X = X_n\) indexed by a parameter \(n \to \infty\), in which the Bernstein-von Mises theorem holds. The latter theorem (see e.g. [17], Chapter 10) entails that the posterior distribution \(\Pi_n(\cdot|X_n)\) of \(\theta\) can be approximated by a normal distribution with deterministic covariance matrix, centered
Table 1

| intra-corr | under Vanilla setup | under Parameter Expansion setup |
|------------|---------------------|--------------------------------|
|            | Min 0.25th qu. | Median 0.75th qu. | Min 0.25th qu. | Median 0.75th qu. |
| 0.1        | 0.8420          | 0.9958 0.9984 0.9993 | 0.8625          | 0.9969 0.9991 0.9997 |
| 0.3        | 0.8244          | 0.9943 0.9984 0.9994 | 0.8569          | 0.9969 0.9989 0.9997 |
| 0.5        | 0.8192          | 0.9905 0.9981 0.9992 | 0.8374          | 0.9954 0.9985 0.9995 |
| 0.7        | 0.8193          | 0.9897 0.9979 0.9992 | 0.8020          | 0.9947 0.9984 0.9994 |
| 0.9        | 0.8102          | 0.9889 0.9980 0.9993 | 0.8058          | 0.9925 0.9980 0.9993 |

Table 2

Approximated Confidence for (δ = 0.1, α = 0.05)-Bayesian tolerance interval

| intra-corr | under Vanilla setup | under Parameter Expansion setup |
|------------|---------------------|--------------------------------|
|            | A = E(ν|X) | A = Optimal | A = E(ν|X) | A = Optimal |
| 0.1        | 0.972          | 0.960 0.963 | 0.968          | 0.963 |
| 0.3        | 0.964          | 0.955 0.955 | 0.955          | 0.949 |
| 0.5        | 0.936          | 0.921 0.921 | 0.925          | 0.917 |
| 0.7        | 0.925          | 0.907 0.910 | 0.915          | 0.911 |
| 0.9        | 0.952          | 0.941 0.940 | 0.940          | 0.935 |

at an estimator $\hat{\theta}_n = \hat{\theta}_n(X_n)$,

$$\Pi_n(\cdot | X_n) = N\left(\hat{\theta}_n, \frac{1}{n} \Sigma_\theta \right) \Rightarrow 0,$$  \hspace{1cm} (13)

(in total variation norm), where the estimators $\hat{\theta}_n = \hat{\theta}_n(X_n)$ satisfy

$$\sqrt{n}(\hat{\theta}_n - \theta) \xrightarrow{d} N(0, \Sigma_\theta).$$  \hspace{1cm} (14)

Under regularity conditions this is valid for $X_n$ a vector of $n$ i.i.d. observations from a smooth parametric model, with $\hat{\theta}_n$ the maximum likelihood estimator and $\Sigma_\theta$ the inverse Fisher information matrix. More generally, this is true in the case of a local approximation by a Gaussian shift experiment ([9, 19]). The Bernstein-von Mises theorem can be used to show that Bayesian and frequentist inference (testing and confidence sets) merge for large sample sizes. In this section we investigate this for tolerance intervals.

We shall show that Bayesian tolerance regions $R_n(X_n)$ such that the functions

$$h \mapsto Q_{\hat{\theta}_n(X_n) + h / \sqrt{n}}(R_n(X_n)), \quad n = 1, 2, \ldots,$$ \hspace{1cm} (15)

stabilize asymptotically to a deterministic function are asymptotically frequentist tolerance regions, for any given loss function $\ell$ and level $\alpha$. The crux of this stability condition is that the randomness which enters the functions (15) through $X_n$ in $\hat{\theta}_n(X_n)$ asymptotically cancels the randomness which enters through $X_n$ in $R_n(X_n)$: the Bayesian tolerance regions $R_n(X_n)$ should be
“asymptotically pivotal” with respect to the estimators $\hat{\theta}_n$. Some type of stability condition appears to be necessary, because the shape of a Bayesian tolerance region is left free by its definition.

An informal proof of the frequentist validity of Bayesian tolerance regions is as follows. Replacing the posterior distribution in (6) by its normal approximation (13) from the Bernstein-von Mises theorem, we find that

$$\int \ell\left[Q_\theta(R_n(X_n))\right] dN\left(\hat{\theta}_n, \frac{1}{n} \Sigma_\theta\right) \doteq 1 - \alpha.$$  \hspace{1cm} (16)

By the substitution $\theta = \hat{\theta}_n + h/\sqrt{n}$ this can be rewritten in the form

$$\int \ell\left[Q_{\hat{\theta}_n + h/\sqrt{n}}(R_n(X_n))\right] dN(0, \Sigma_\theta)(h) \doteq 1 - \alpha.$$ \hspace{1cm} (17)

By the stability assumption the integrand

$$h \mapsto g_n(h; X_n) = \ell\left[Q_{\hat{\theta}_n + h/\sqrt{n}}(R_n(X_n))\right]$$ \hspace{1cm} (18)

in this expression is asymptotically the same as a deterministic function $h \mapsto g_\infty(h)$. In view of (14) the integral in (17) is then approximately equal to $E_\theta g_\infty(\sqrt{n}(\theta - \hat{\theta}_n))$, which in turn, again by stability, is asymptotically the same as $E_\theta g_n(\sqrt{n}(\theta - \hat{\theta}_n); X_n)$, or

$$E_\theta \ell\left[Q_{\theta + \sqrt{n}(\theta - \hat{\theta}_n)/\sqrt{n}}(R_n(X_n))\right] = E_\theta \ell\left[Q_\theta(R_n(X_n))\right].$$

Thus the final expression, which is the frequentist level of the tolerance region $R_n(X_n)$, is asymptotically equal to $1 - \alpha$.

For an $(\delta, \alpha)$-tolerance region, $\ell(Q_\theta(R_n(X_n)))$ is the indicator of the set $\hat{G}_n = \{\theta: Q_\theta(R_n(X_n)) \geq 1 - \delta\}$ and the function (18) is the indicator of the set

$$\hat{H}_n = \sqrt{n}(\hat{G}_n - \hat{\theta}_n).$$

Thus the stability condition is that the latter sets approximate to a deterministic set, as $n \to \infty$. Condition (17) becomes

$$N(0, \Sigma_\theta)(\hat{H}_n) \doteq 1 - \alpha.$$ \hspace{1cm} (19)

This equality allows to “solve” one aspect of the sets $\hat{H}_n$; in general additional constraints will be imposed to define their shape. As the normal distribution in this display is fixed, it is not unnatural that these constraints would render the sets $\hat{H}_n$ also to become fixed, in the limit: stability is natural.

The following theorem makes the preceding rigorous. We shall verify its conditions for normal prediction variables in the next section.

**Theorem 5.1.** Suppose that (13)–(14) hold, the loss function $\ell$ is bounded, and suppose that there exist (deterministic) functions $f_{n,1}, f_{n,2}: \mathbb{R}^d \to \mathbb{R}$ with the property that $f_{n,i}(h_n) \to f_\infty(h)$ for $i = 1, 2$ and some function $f_\infty$ and
any sequence \( h_n \to h \) with limit \( h \) in a set of probability one under the normal distribution in (14) and such that

\[
f_{n,1}(h) \leq \ell(Q_{\hat{\theta}_n+h/\sqrt{n}}(R_n(X_n))) \leq f_{n,2}(h), \quad h \in \mathbb{R}^d.
\]

Then \( \int \ell(Q_\theta(R_n(X_n))) \ d\Pi(\theta|X_n) \to 1 - \alpha \in (0, 1) \) in probability implies that \( E_\theta \ell(Q_\theta(R_n(X_n))) \to 1 - \alpha \), as \( n \to \infty \), for every \( \theta \).

**Proof.** We may assume without loss of generality that the functions \( f_{n,i} \) are uniformly bounded. Then the condition \( f_{n,i}(h_n) \to f_\infty(h) \) for every sequence \( h_n \to h \) implies that \( E f_{n,i}(Y_n) \to E f_\infty(Y) \), whenever the sequence of random vectors \( Y_n \) tends in distribution to the random vector \( Y \), in view of the extended continuous mapping theorem (see [18], Theorem 1.11.1). Thus by (14), for \( i = 1, 2 \),

\[
E_\theta f_{n,i}(\sqrt{n}(\theta - \hat{\theta}_n)) \to \int f_\infty dN(0, \Sigma_\theta).
\]

By assumption the function \( g_n \) given in (18) is sandwiched between \( f_{n,1} \) and \( f_{n,2} \). Therefore \( E_\theta \ell(Q_\theta(R_n(X_n))) = E_\theta g_n(\sqrt{n}(\theta - \hat{\theta}_n); X_n) \) tends to the same limit.

By (13) and the definition of \( g_n \) (see (16)-(17)), we have

\[
\int \ell(Q_\theta(R_n(X_n))) \ d\Pi(\theta|X_n) = \int g_n(h; X_n) \ dN(0, \Sigma_\theta)(h) + o_P(1).
\]

Again by sandwiching of \( g_n(h; X_n) \) this is asymptotic to \( \int f_{n,i} \ dN(0, \Sigma_\theta) \), and hence tends to \( \int f_\infty \ dN(0, \Sigma_\theta) \). If the left side of the display tends to \( 1 - \alpha \), as assumed, then it follows that \( \int f_\infty \ dN(0, \Sigma_\theta) = 1 - \alpha \). The theorem follows by combining this with the preceding paragraph. \( \blacksquare \)

### 5.1. Normal predictions

An \((\delta, \alpha)\)-tolerance interval \( R_n(X_n) = [A_n - B_n, A_n + B_n] \) for a one-dimensional Gaussian variable \( Z \sim N(\nu, \tau^2) \) is the base (the section at \( \tau = 0 \)) of a set of the form

\[
G_{A,B,\delta} = \{ \theta = (\nu, \tau) : \Phi\left(\frac{A + B - \nu}{\tau}\right) - \Phi\left(\frac{A - B - \nu}{\tau}\right) \geq 1 - \delta \}.
\]

The values \( A_n \) and \( B_n \) are determined so that \( \Pi(G_{A_n,B_n,\delta}|X_n) \geq 1 - \alpha \), for \( \Pi(\cdot|X_n) \) the posterior distribution of \( \theta = (\nu, \tau) \), and so that the length \( 2B_n \) of the interval is minimal.

Under (13) the posterior distribution contracts (at rate \( 1/\sqrt{n} \)) to the Dirac measure at \( \hat{\theta}_n \), which tends to the true value of the parameter (\( \nu, \tau \)) under (14). Hence the equation \( \Pi(G_{A_n,B_n,\delta}|X_n) \geq 1 - \alpha \) forces that any ball of fixed radius around this true value intersects \( G_{A_n,B_n,\delta} \) with probability tending to one. The minimality of \( B_n \) implies that the horizontal locations \( A_n \) of the latter sets must tend to the true value of \( \nu \). Thus \( A_n - \nu_n \to 0 \) in probability and
hence $\Phi(B_n/\hat{\tau}_n) - \Phi(-B_n/\hat{\tau}_n) \to 1 - \delta$, whence $B_n/\hat{\tau}_n \to \xi_\delta/2$, for $\xi_\delta$ the upper $\delta$-quantile of the standard normal distribution.

The function $\theta \mapsto \ell(Q_0([A - B, A + B]))$ corresponding to the $(\delta, \alpha)$-tolerance interval is the indicator of the set $G_{A,B,\Delta}$, and the stability condition (20) is that the (indicator functions) of the sets $\hat{H}_n = \sqrt{n}(G_{A_n,B_n,\theta} - \hat{\theta})$ are asymptotically deterministic. These sets can be written $\hat{H}_n = \{(g,h); \hat{K}_n(g/\sqrt{n},h/\sqrt{n}) \geq 0\}$, for the stochastic processes

$$K_n(g,h) = \Phi\left(\frac{A_n + B_n - \hat{\nu}_n - g}{\hat{\tau}_n + h}\right) - \Phi\left(\frac{A_n - B_n - \hat{\nu}_n - g}{\hat{\tau}_n + h}\right) - (1 - \delta).$$

By a second-order Taylor expansion we see that these processes satisfy the expansion (22) in Lemma 5.2 (below), with

$$\begin{align*}
\hat{a}_n &= K_n(0,0) = \Phi\left(\frac{A_n + B_n - \hat{\nu}_n}{\hat{\tau}_n}\right) - \Phi\left(\frac{A_n - B_n - \hat{\nu}_n}{\hat{\tau}_n}\right) - (1 - \delta), \\
\hat{b}_n &= -\left(\frac{\partial}{\partial h} K_n\right)(0,0) = \psi\left(\frac{A_n + B_n - \hat{\nu}_n}{\hat{\tau}_n}\right) \frac{1}{\hat{\tau}_n} - \psi\left(\frac{A_n - B_n - \hat{\nu}_n}{\hat{\tau}_n}\right) \frac{1}{\hat{\tau}_n}, \\
V(g,h) &= h.
\end{align*}$$

Here $\psi(x) = \phi(x)x = -\phi'(x)$. (Note that the partial derivatives $-\phi((A_n + B_n - \hat{\nu}_n)/\hat{\tau}_n)$ and $\phi((A_n - B_n - \hat{\nu}_n)/\hat{\tau}_n)$ of $K_n$ at $(0,0)$ relative to its first argument $g$ tend to zero.) Since $A_n$, $B_n$, $\hat{\nu}_n$ and $\hat{\tau}_n$ tend in probability to nontrivial limits, the conditions of Lemma 5.2 are satisfied and hence the sets $\hat{H}_n$ are asymptotically sandwiched between pairs of deterministic sets. Functions $f_{n,i}$ as in (20) can be constructed from these sets by letting $\varepsilon \to 0$ and $M \to \infty$ slowly with $n$. Thus the conditions of Theorem 20 are satisfied and we obtain the following corollary.

**Corollary 5.1.** If the posterior distribution of $\theta = (\nu, \tau)$ given $X_n$ satisfies (13)--(14), then the Bayesian $(\delta, \alpha)$-tolerance interval $[A_n - B_n, A_n + B_n]$ of minimal length for a future variable $Z|X_n, \nu, \tau \sim N(\nu, \tau^2)$ is an asymptotic frequentist $(\delta, \alpha)$-tolerance set.

The convergence $A_n - \hat{\nu}_n \to 0$ in probability means that the tolerance intervals are asymptotically centered at the (asymptotic) posterior mean. If the posterior distribution of $\theta$ is exactly normal $N(\hat{\nu}_n, \hat{\tau}_n, \Sigma_\theta)$ with a diagonal covariance matrix, then $E(\nu|X_n, \tau)$ is free of $\tau$ and Lemma 3.1 shows that the tolerance interval is centered exactly at the posterior mean.

For a non-diagonal matrix $\Sigma_\theta$ this is not necessarily true, and in general the normal distribution will also be an approximation only. The approximation $A_n - \hat{\nu}_n \to 0$ can in general be improved to order $n^{-1/4}$. The following lemma also gives an asymptotic expression for the half length $B_n$ of the interval.

**Lemma 5.1.** The centers and half lengths of the Bayesian $(\delta, \alpha)$-tolerance intervals $[A_n - B_n, A_n + B_n]$ of minimal length in Corollary 5.1 satisfy $A_n = \hat{\nu}_n + o_P(n^{-1/4})$ and $B_n = \hat{\tau}_n \xi_{\delta/2} + \xi_\delta/2 \sqrt{\Sigma_{\theta,2,2} n^{-1/2}} + o_P(n^{-1/2})$. 

imsart-generic ver. 2014/10/16 file: BayesTI.tex date: September 20, 2022
Proof. Since \( A_n - \bar{\nu}_n \to 0 \) and \( B_n \to \xi_{\delta/2} \), there exist intervals \( I_n \) and \( J_n \) around \( \nu_0 \) and \( \tau_0 \xi_{\delta/2} \), for \((\nu_0, \tau_0)\) the true value of \( \theta \), that shrink to these points that contain \( A_n \) and \( B_n \) with probability tending to one. Define functions \( F_n, G_n: I_n \times J_n \to \mathbb{R} \)

\[
F_n(A, B) = \Pi(G_{A,B,\delta} \mid X_n),
\]

\[
G_n(A, B) = N(0, \Sigma_\theta) \left\{ (g, h): h \leq \frac{\sqrt{n}K_n(0, 0; A, B)}{2\psi(\xi_{\delta/2})/\hat{\tau}_n} \right\},
\]

Here \( K_n(h, g; A, B) \) is the expression on the right side of (21), but with \((A_n, B_n)\) replaced by a generic \((A, B)\), and \( \psi(x) = x\phi(x) \). The values \((A_n, B_n)\) are determined so that \( B_n \) is the smallest value so that there exists \( A_n \) such that \( F_n(A_n, B_n) \geq 1 - \alpha \). In other words \( B_n \) is the minimum of \( \{ B: \sup_A F_n(A, B) \geq 1 - \alpha \} \) and \( A_n \) is the point of maximum of \( A \mapsto F_n(A, B_n) \). We shall show that the functions \( F_n \) and \( G_n \) satisfy the conditions of Lemma 5.3 with \( c_n = \hat{\tau}_n \xi_{\delta/2} \) and \( \xi(\alpha) = \xi_n \xi_{\delta/2} / \sqrt{\Sigma_{\theta,2,2}} \), whence the lemma follows from that lemma.

In view of (13) we have \( \sup_{A,B} |F_n(A, B) - N(0, \Sigma_\theta)(\hat{H}_{A,B,\delta})| \to 0 \), for \( \hat{H}_{A,B,\delta} = \{ (g, h): K_n(g/\sqrt{n}, h/\sqrt{n}; A, B) \geq 0 \} \). The supremum here and in the following is taken over \((A, B) \in I_n \times J_n \). By a second-order Taylor expansion, for functions \( \eta_{n,1} \) and \( \eta_{n,2} \) with \( \sup_{A,B} \eta_{n,i}(A, B) \to 0 \) and a constant \( b \) independent of \((A, B)\),

\[
\frac{\sqrt{n}}{\hat{\tau}_n} K_n \left( \frac{g}{\sqrt{n}}, \frac{h}{\sqrt{n}}; A, B \right) - K_n(0, 0; A, B) - \eta_{n,1}(A, B) - \frac{\sqrt{n}}{\hat{\tau}_n} \left( -\frac{2\psi(\xi_{\delta/2})}{\hat{\tau}_n} + \eta_{n,2}(A, B) \right) \frac{h}{\sqrt{n}} \leq b g^2 + h^2.
\]

By a sandwiching argument as in the proof of Lemma 5.2, this shows that, with \( C_n = \{ (g, h): g^2 + h^2 \leq L_n \} \) and \( L_n \to \infty \) sufficiently slowly

\[
\sup_{A,B} |N(0, \Sigma_\theta)(\hat{H}_{A,B,\delta}\cap C_n) - N(0, \Sigma_\theta) \{ (g, h) \in C_n: h \leq \frac{\sqrt{n}K_n(0, 0; A, B)}{2\psi(\xi_{\delta/2})/\hat{\tau}_n} \}| \to 0.
\]

Because \( N(0, \Sigma_\theta)(C_n^{\infty}) \to 0 \), this is then also true without intersecting the sets by \( C_n \). This finishes the proof that \( \sup_{A,B} |F_n(A, B) - G_n(A, B)| \to 0 \).

By the unimodality and symmetry of the normal distribution, the map \( A \mapsto K_n(0, 0; A, B) \) has a maximum at \( A = \bar{\nu}_n \), for every \( B \), and hence the same is true for \( A \mapsto G_n(A, B) \). It is elementary to verify by several Taylor expansions that \( G_n(\bar{\nu}_n, B) \to 1 - \alpha \), for \( \sqrt{n}(\Phi(B/\hat{\tau}_n) - 1 + \delta/2) \hat{\tau}_n \to \xi_n \psi(\xi_{\delta/2}) \sqrt{\Sigma_{\theta,2,2}} \), or \( \hat{B} = \hat{\tau}_n \xi_{\delta/2} + \xi_n \xi_{\delta/2} / \sqrt{\Sigma_{\theta,2,2}} \).

Finally, since \( \frac{\partial}{\partial A} K_n(0, 0; \bar{\nu}_n, B) = 0 \), again by Taylor expansion there exist functions \( \eta_{n,3} \) with the property \( \sup_A \eta_{n,3}(A, \hat{B}) \to 0 \) such that

\[
K_n(0, 0; A, B) \leq K_n(0, 0; \bar{\nu}_n, \hat{B}) + (A - \bar{\nu}_n)^2 (\phi'(B/\hat{\tau}_n)/\hat{\tau}_n^2 + \eta_{n,3}(A, \hat{B})).
\]

As \( \phi'(x) < 0 \), for \( x > 0 \), we obtain with probability tending to one that \( \sqrt{n}K_n(0, 0; A, B) \) is smaller than \( \sqrt{n}K_n(0, 0; \bar{\nu}_n, \hat{B}) - \varepsilon \) for some \( \varepsilon > 0 \) if \( \sqrt{n}(A-\bar{\nu}_n)^2 \phi''(x) \to 0 \).
\[ (\hat{\nu}_n)^2 > \delta \text{ for some } \delta > 0. \] 

On this event

\[ G_n(A, \hat{B}) = N(0, \Sigma_o) \left\{ (\delta, h) : h \leq \frac{\sqrt{n}\xi_n(0, 0; A, \hat{B})}{2\psi(\xi_n/2)/\tau_n} \right\} \]

is strictly smaller than its asymptotic value 1 − \alpha at \( A = \hat{\nu}_n \). This verifies the last displayed condition of Lemma 5.3.

**Lemma 5.2.** Suppose that for every \( M > 0 \) the stochastic processes \( (K_n(h) : h \in \mathbb{R}^d) \) satisfy

\[
\sup_{\|h\| \leq M/\sqrt{n}} \sqrt{n}|K_n(h) - \hat{\alpha}_n + \hat{b}_n Vh| \xrightarrow{P} 0,
\]

for random variables \( \hat{\alpha}_n \) and \( \hat{b}_n > 0 \) such that \( \hat{b}_n^{-1} \) is bounded in probability, and a linear map \( V: \mathbb{R}^d \to \mathbb{R} \). If the sets \( \hat{H}_n = \{ h \in \mathbb{R}^d : K_n(h/\sqrt{n}) \geq 0 \} \) satisfy \( N(0, \Sigma)(\hat{H}_n) \to 1 - \alpha \in (0, 1) \), then \( \sqrt{n}\hat{\alpha}_n/\hat{b}_n \to \xi_\alpha\sqrt{\Sigma V\Sigma^T} \) and for every \( \varepsilon, M > 0 \), with probability tending to 1,

\[
\{ h \in \mathbb{R}^d : \|h\| \leq M, Vh \leq \xi_\alpha\sqrt{\Sigma V\Sigma^T} - \varepsilon \} \subset \hat{H}_n \subset \{ h \in \mathbb{R}^d : Vh \leq \xi_\alpha\sqrt{\Sigma V\Sigma^T} + \varepsilon \text{ or } \|h\| > M \}.
\]

**Proof.** Define \( \hat{\nu}_n(h) = \sqrt{n}(K_n(h) - \hat{\alpha}_n + \hat{b}_n Vh) \) and set \( \hat{\varepsilon}_n = \sup_{\|h\| \leq M/\sqrt{n}} |\hat{\nu}_n(h)| \), for given \( M \). Then by assumption \( \hat{\varepsilon}_n \to 0 \) in probability, and \( |K_n(h/\sqrt{n}) - \hat{\alpha}_n + \hat{b}_n Vh/\sqrt{n}| \leq \hat{\varepsilon}_n \), for every \( h \) with \( \|h\| \leq M \). From the latter inequality we find that

\[
\|h\| \leq M, K_n(h/\sqrt{n}) \geq 0 \Rightarrow \hat{b}_n Vh \leq \sqrt{n}\hat{\alpha}_n + \hat{\varepsilon}_n, \\
\|h\| \leq M, \hat{b}_n Vh \leq \sqrt{n}\hat{\alpha}_n - \hat{\varepsilon}_n \Rightarrow K_n(h/\sqrt{n}) \geq 0.
\]

This implies that

\[
\{ \|h\| \leq M, Vh \leq (\sqrt{n}\hat{\alpha}_n - \hat{\varepsilon}_n)/\hat{b}_n \} \subset \hat{H}_n \subset \{ Vh \leq (\sqrt{n}\hat{\alpha}_n + \hat{\varepsilon}_n)/\hat{b}_n \text{ or } \|h\| > M \}.
\]

Combining this with the fact that \( N(0, \Sigma)(\hat{H}_n) \to 1 - \alpha \in (0, 1) \) and the fact that \( Vh \sim N(0, V\Sigma V^T) \) if \( h \sim N(0, \Sigma) \), we conclude that there exists \( \delta_M > 0 \) such that \( \delta_M \to 0 \) as \( M \to \infty \) such that \( (\sqrt{n}\hat{\alpha}_n - \hat{\varepsilon}_n)/\hat{b}_n \geq \xi_\alpha - \delta_M \sqrt{\Sigma V\Sigma^T} + o_P(1) \) and \( (\sqrt{n}\hat{\alpha}_n + \hat{\varepsilon}_n)/\hat{b}_n \geq \xi_\alpha + \delta_M \sqrt{\Sigma V\Sigma^T} + o_P(1) \), for every \( M \). Since \( \hat{b}_n^{-1} = O_P(1) \) by assumption, we have \( \hat{\varepsilon}_n/\hat{b}_n \to 0 \) in probability, and hence \( \sqrt{n}\hat{\alpha}_n/\hat{b}_n = \xi_\alpha\sqrt{\Sigma V\Sigma^T} + o_P(1) \). We substitute this in the last display to obtain the result of the lemma.

**Lemma 5.3.** Let \( F_n, G_n : I_n \times J_n \to \mathbb{R} \) be stochastic processes indexed by rectangles \( I_n \times J_n \subset \mathbb{R}^2 \) that are nondecreasing in their second argument, such that

\[
\sup_{A, B} |F_n(A, B) - G_n(A, B)| \xrightarrow{P} 0,
\]
and such that for numbers $c_n \in J_n$ and continuous functions $\xi: (0, 1) \to \mathbb{R}$, and every $\alpha \in (0, 1)$,

$$\sup_A G_n(A, c_n + \xi(\alpha)/\sqrt{n}) = G_n(0, c_n + \xi(\alpha)/\sqrt{n}) \xrightarrow{P} 1 - \alpha,$$

$$\Pr \left( \sup_{A:|A| > \delta_n n^{-1/4}} G_n(A, c_n + \xi(\alpha)/\sqrt{n}) < 1 - \alpha, \right) \to 1, \quad \text{some } \delta_n \to 0.$$

Then $B_n(\alpha) := \inf(B: \sup_A F_n(A, B) \leq 1 - \alpha)$ satisfies $B_n(\alpha) = c_n + \xi(\alpha)/\sqrt{n} + o_P(n^{-1/2})$ and $\arg\max_A F_n(A, B_n) = o_P(n^{-1/4}).$

Proof. The functions $F_n$ and $G_n$ defined by $F_n(A, B) = \sup_A F_n(A, B)$ and similarly for $G_n$ satisfy $\sup_B |\tilde{F}_n(B) - F_n(B)| \xrightarrow{P} 0$. Combined with the first displayed assumption on $G_n$, this gives that $F_n(c_n + \xi(\alpha)/\sqrt{n}) \xrightarrow{P} 1 - \alpha$, for every $\alpha \in (0, 1)$. The definition of $B_n(\alpha)$ and monotonicity of $F_n$ now readily give that $c_n + \xi(\alpha_2)/\sqrt{n} \leq B_n(\alpha) \leq c_n + \xi(\alpha_1)/\sqrt{n}$ for every $\alpha_1 < \alpha < \alpha_2$, eventually with probability tending to one, or equivalently $\xi(\alpha_2) \leq \sqrt{n}(B_n(\alpha) - c_n) \leq \xi(\alpha_1)$, eventually. By the continuity of $\xi$ it follows that $\sqrt{n}(B_n(\alpha) - c_n) \xrightarrow{P} \xi(\alpha)$.

By the uniform approximation of $F_n$ by $G_n$ and the second displayed assumption on $G_n$, we have that

$$\sup_{|A| > \delta_n n^{-1/4}} F_n(A, B_n(\alpha)) = \sup_{|A| > \delta_n n^{-1/4}} G_n(A, B_n(\alpha)) + o_P(1).$$

By monotonicity $G_n(A, B_n(\alpha)) \leq G_n(A, c_n + \xi(\alpha_1)/\sqrt{n})$, for every $\alpha_1 < \alpha$, for every $A$. Thus the right side of the display is strictly smaller than $1 - \alpha$, eventually, by assumption, which is strictly smaller than $1 - \alpha$, for $\alpha_1$ close enough to $\alpha$. Similarly also $F_n(0, B_n(\alpha)) = G_n(0, B_n(\alpha)) + o_P(1) \to 1 - \alpha$. It follows that the maximum of $A \mapsto F_n(A, B_n(\alpha))$ is taken on the interval $[-\delta_n n^{-1/4}, \delta_n n^{-1/4}].$

References

[1] J. Aitchison. Two papers on the comparison of bayesian and frequentist approaches to statistical problems of prediction: Bayesian tolerance regions. *Journal of the Royal Statistical Society. Series B (Methodological)*, 26(2):161–175, 1964. [2], [4], [8]

[2] J. Aitchison. Expected-cover and linear-utility tolerance intervals. *Journal of the Royal Statistical Society. Series B (Methodological)*, 28(1):57–62, 1966. [2]

[3] William J. Browne and David Draper. A comparison of bayesian and likelihood-based methods for fitting multilevel models. *Bayesian Analysis*, 1(3):473–514, 09 2006. [17]

[4] Robert G. Easterling and David L. Weeks. An accuracy criterion for bayesian tolerance intervals. *Journal of the Royal Statistical Society. Series B (Methodological)*, 32(2):236–240, 1970. [2]
[5] Michael Hamada. Bayesian tolerance interval control limits for attributes. *Quality and Reliability Engineering International*, 18(1):45–52, 2002. 2

[6] Michael Hamada, Valen Johnson, Leslie M. Moore, and Joanne Wendelberger. Bayesian prediction intervals and their relationship to tolerance intervals. *Technometrics*, 46(4):452–459, 2004. 2

[7] Hormuzd A. Katki, Eric A. Engels, and Philip S. Rosenberg. Assessing uncertainty in reference intervals via tolerance intervals: application to a mixed model describing hiv infection. *Statistics in Medicine*, 24(20):3185–3198, 2005. 2

[8] K. Krishnamoorthy and T. Mathew. *Statistical Tolerance Regions: Theory, Applications, and Computation*. Wiley Series in Probability and Statistics. Wiley, 2009. 1, 2, 9, 10

[9] L. Le Cam. Limits of experiments. In *Proceedings of the Sixth Berkeley Symposium on Mathematical Statistics and Probability (Univ. California, Berkeley, Calif., 1970/1971), Vol. I: Theory of statistics*, pages 245–261, 1972. 18

[10] Robert W. Miller. Parametric empirical bayes tolerance intervals. *Technometrics*, 31(4):449–459, 1989. 2

[11] Rahul Mukerjee and N. Reid. Second-order probability matching priors for a parametric function with application to bayesian tolerance limits. *Biometrika*, 88(2):587–592, 2001. 2

[12] Dharini Pathmanathan, Rahul Mukerjee, and S. H. Ong. Two-sided bayesian and frequentist tolerance intervals: general asymptotic results with applications. *Statistics*, 48(3):524–538, 2014. 2

[13] Dharini Pathmanathan and S. H. Ong. A Monte Carlo simulation study of two-sided tolerance intervals in balanced one-way random effects model for non-normal errors. *Journal of Statistical Computation and Simulation*, 84(11):2329–2344, 2014. 2

[14] Gaurav Sharma and Thomas Mathew. One-sided and two-sided tolerance intervals in general mixed and random effects models using small sample asymptotics. *Journal of the American Statistical Association*, 107(497):258–267, 2012. 1, 5, 16, 17

[15] A.J. Van der Merwe and Johan Hugo. Bayesian tolerance intervals for the balanced two-factor nested random effects model. *TEST*, 16(3):598–612, 2007. 2

[16] A.J. Van der Merwe, Albertus L. Pertorius, and J. H. Meyer. Bayesian tolerance intervals for the unbalanced one-way random effects model. *Journal of Quality Technology*, 38(3):280–293, 2006. 2

[17] A. van der Vaart. *Asymptotic Statistics*, volume 3 of *Cambridge Series in Statistical and Probabilistic Mathematics*. Cambridge University Press, Cambridge, 1998. 17

[18] A. van der Vaart and J. Wellner. *Weak Convergence and Empirical Processes*. Springer Series in Statistics. Springer-Verlag, New York, 1996. With applications to statistics. 20

[19] Aad van der Vaart. An asymptotic representation theorem. *International Statistical Review / Revue Internationale de Statistique*, 59(1):97–
[20] Russell D. Wolfinger. Tolerance intervals for variance component models using bayesian simulation. *Journal of quality technology*, 30(1):18–32, 1998.

[21] Derek S. Young, Charles M. Gordon, Shihong Zhu, and Bryan D. Olin. Sample size determination strategies for normal tolerance intervals using historical data. *Quality Engineering*, 28(3):337–351, 2016.