Journal of the Royal Statistical Society Series C: Applied Statistics, 2024, **00**, 1–25
https://doi.org/10.1093/jrsssc/qlae077

ROYAL STATISTICAL SOCIETY
DATA | EVIDENCE | DECISIONS
Journal of the Statistics Society
Series C
Applied Statistics

Original Article

Tree models for assessing covariate-dependent method agreement with an application to physical activity measurements

Siranush Karapetyan[1], Achim Zeileis[2], André Henriksen[3] and Alexander Hapfelmeier[1,4]

[1]Institute of General Practice and Health Services Research, Technical University of Munich, Munich, Germany
[2]Faculty of Economics and Statistics, University of Innsbruck, Innsbruck, Austria
[3]Department of Computer Science, The Arctic University of Norway, Tromsø, Norway
[4]Institute of AI and Informatics in Medicine, Technical University of Munich, Munich, Germany

Address for correspondence: Siranush Karapetyan, Institute of General Practice and Health Services Research, Technical University of Munich, Munich, Germany. Email: siranush.karapetyan@mri.tum.de

Abstract

Method comparison studies assess agreement between different measurement methods. In the present work, we are interested in comparing physical activity measurements using two different accelerometers. However, a potential issue arises with the popular Bland–Altman analysis, as it assumes that differences between measurements are identically distributed across all observational units. In the case of the physical activity measurements, agreement might depend on sex, height, weight, or age of the person wearing the accelerometers, among others. To capture this potential dependency, we introduce the concept of conditional method agreement, which defines subgroups with heterogeneous agreement in dependence of covariates. We propose several tree-based models that can detect such a dependency and incorporate it into the model by splitting the data into subgroups, showing that the agreement of the activity measurements is conditional on the participant's age. Simulation studies also showed that all models were able to detect subgroups with high accuracy as the sample size increased. We call the proposed modelling approach conditional method agreement trees and make them publicly available through the R package `coat`.

Keywords: Bland-Altman analysis, hypothesis testing, method agreement, recursive partitioning, subgroup analysis

1 Introduction

Method comparison studies are relevant in all scientific fields whenever the agreement of continuously scaled measurements made by two or more methods is to be investigated. However, they have found particular application in medical research, for example in laboratory research (Chhapola et al., 2015; Giavarina, 2015), anaesthesiology (Abu-Arafeh et al., 2016), ophthalmology (Bunce, 2009), and pathology (Jensen & Kjelgaard-Hansen, 2006) among many others. In the field of epidemiological research, particularly in the measurement of physical activity, a variety of wearable devices are available for objective measurement. However, taking measurements with some of these devices can be complex, time-consuming, and expensive for applicants. Therefore, it is of importance to conduct method agreement studies to compare the performance of different devices in order to be able to use these devices interchangeably.

Received: October 10, 2023. **Revised:** October 28, 2024. **Accepted:** December 2, 2024

© The Royal Statistical Society 2025.
This is an Open Access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted reuse, distribution, and reproduction in any medium, provided the original work is properly cited.

Downloaded from https://academic.oup.com/jrsssc/advance-article/doi/10.1093/jrsssc/qlae077/7942358 by UiT The Arctic University of Norway user on 06 January 2025

1.1 Measuring agreement

A well-established methodology for analysis was developed by Bland and Altman and is known as the Bland–Altman analysis or plot (Altman & Bland, 1983). In its most basic form, it illustrates the differences against the mean values of paired measurements made by two methods. Here, two quantities of interest are the mean difference, referred to as 'bias', and the standard deviation of the differences, which is used to determine the width of the so-called 'Limits of Agreement' (LoA) (Hanneman, 2008). The bias is a measure of the overall deviation of the methods but has limited interpretability, since large positive and negative deviations can still add up to a small overall bias. Therefore, Bland and Altman proposed to estimate the LoA, that is a prediction interval in which about 95% of individual differences between the measurements of the two methods are expected to lie. The mean and standard deviation of differences can be calculated directly from the observed data, but it has also been suggested to use regression modelling under the assumption of normally distributed residuals (Carstensen, 2010, 2011).

Proper planning, conduct, interpretation, and reporting of method comparison studies has been the subject of ongoing research and recommendations have been provided in respective publications and through reviews of the relevant literature (Abu-Arafeh et al., 2016; Bunce, 2009; Chhapola et al., 2015; Francq & Govaerts, 2016; Gerke, 2020; Giavarina, 2015; Hanneman, 2008; Hapfelmeier et al., 2016; Jensen & Kjelgaard-Hansen, 2006; Stöckl et al., 2004). These works are also concerned with the data description, processing, and analysis, the plotting of results, the (pre)-specification of acceptable agreement, the precision of estimation, the repeatability of measurements, and the investigation of homoscedastic variances and trends. Regarding the latter two, Bland and Altman already discussed early the question whether the agreement between the methods depends on the magnitude of the measured values, that is whether there is a relationship between the differences and the means of paired values (Altman & Bland, 1983; Bland & Altman, 1986). In that case, they suggested either transforming (e.g. log-transforming) the differences to remove the dependency or modelling the differences with mean values as explanatory variable in a linear regression model. Recent developments have proposed to address the problem of heteroscedasticity through a heteroscedastic mixed effects model (Nawarathna & Choudhary, 2013, 2015; Taffé, 2018, 2020).

In the present work, we are interested in agreement between physical activity measurements made by two different accelerometers, one worn with a belt on the hip and the other attached to the skin on the chest. We suppose that the underlying assumption of a Bland–Altman analysis, that is that the agreement of methods, i.e. accelerometers is identically distributed for all observational units or subjects, may not be valid in that case. The basic idea is that the methods' measurements, in particular the differences between accelerometers' measurements, can be affected by internal and external factors, such as the subjects' characteristics and measurement settings, with direct implications on the agreement of methods. Previous studies have used heuristic approaches to address this issue, for example through the post-hoc fitting of additional regression models and subgroup analyses (Haghayegh et al., 2020; Huber et al., 2014). An early example is the regression of mean values on differences as originally suggested by Bland and Altman and outlined above (Altman & Bland, 1983; Bland & Altman, 1986).

Here, we introduce a unifying framework and analysis approach for conditional method agreement in case of single measurements per subject or observational unit. Recursive partitioning is used to simultaneously explore relations between covariates and agreement and to define corresponding subgroups with heterogeneous agreement in terms of bias and/or the width of LoA, taking advantage of the fact that a Bland–Altman analysis can be parametrized accordingly (Carstensen, 2010, 2011; Möller et al., 2021). We consider three different modelling approaches, that is conditional inference trees with an appropriate transformation of the outcome (CTreeTrafo) (Hothorn, Hornik, et al., 2006), distributional regression trees (DistTree) (Schlosser et al., 2019), and model-based trees (MOB) (Zeileis, Hothorn, et al., 2008). We call the proposed modelling approach conditional method agreement trees (COAT) and demonstrate its relevance to epidemiological research through the application to a data of accelerometer measurements made by different wearable devices. The ability of these approaches to control the type-I error probability at a nominal level, the power to detect given subgroups, and the ability to accurately define these subgroups is investigated in simulation studies. In addition, we propose a two-sample test of differences in method agreement suitable for exploratory or confirmatory hypothesis testing of differences in agreement between two (pre)defined subgroups.

Downloaded from https://academic.oup.com/jrsssc/advance-article/doi/10.1093/jrsssc/qlae077/7942358 by UiT The Arctic University of Norway user on 06 January 2025

1.2 Physical activity measurements

Lack of physical activity is one of the main risk factors contributing to global mortality (WHO, 2022). Tracking physical activity is therefore important in research. Accelerometers are widely used to objectively measure physical activity. These are compact, lightweight, wearable devices designed to measure acceleration in one or more axis. They provide data on the frequency, duration, and intensity of physical activity per unit of time (Butte et al., 2012). There is currently a wide range of accelerometers available, for use in both research (Henriksen, Haugen Mikalsen, et al., 2018) and the consumer market (Butte et al., 2012). The latter are becoming increasingly popular due to their potential to encourage increased physical activity. While some accelerometers provide very precise measurements, the accuracy of newer accelerometers remains largely unexplored and should therefore be compared with established research instruments (Henriksen, Svartdal, et al., 2022). Furthermore, the agreement between these wearable devices may be influenced by the characteristics of the applicants or measurement settings. However, the well-established approach to measure agreement, the Bland–Altman plot, does not capture the potential dependence of agreement on applicant characteristics. The new concept of conditional method agreement introduced here allows to assess covariate-dependent agreement, and is applied here using accelerometer data.

The accelerometer data consists of 24-hr accelerometer measurements and socio-demographic information from $n = 50$ participants of the original study (Henriksen, Grimsgaard, et al., 2019). Figure 1a shows a respective Bland–Altman plot of the agreement of activity energy expenditure (AEE) (in kilocalories) measured by two investigated accelerometers, namely ActiGraph and Actiheart. More details are given in Section 3. Using conditional method agreement trees (COAT) by MOB, it can be shown that this agreement is related to the age of the participants. There are two subgroups with statistically significantly different agreement ($p = .023$), especially in terms of bias, which is divided from −385 in the whole sample into −536 and −207 in the subgroups defined by a split point of 41 years (cf. Figure 1b). Also, the LoA within the defined subgroups are less wide than for the whole sample. Comparing the subgroups, the LoA are wider within subjects of increased age of >41 years. This result is of interest to scientists, health professionals, users, and manufacturers of accelerometers who develop the wearable devices or rely on their functionality and who may want to discuss the reasons for this difference in agreement and possible solutions or implications for proper use.

2 Modelling approaches

The following subsections outline the concept of conditional method agreement, corresponding modelling through recursive partitioning and a two-sample test for hypothesis testing of group

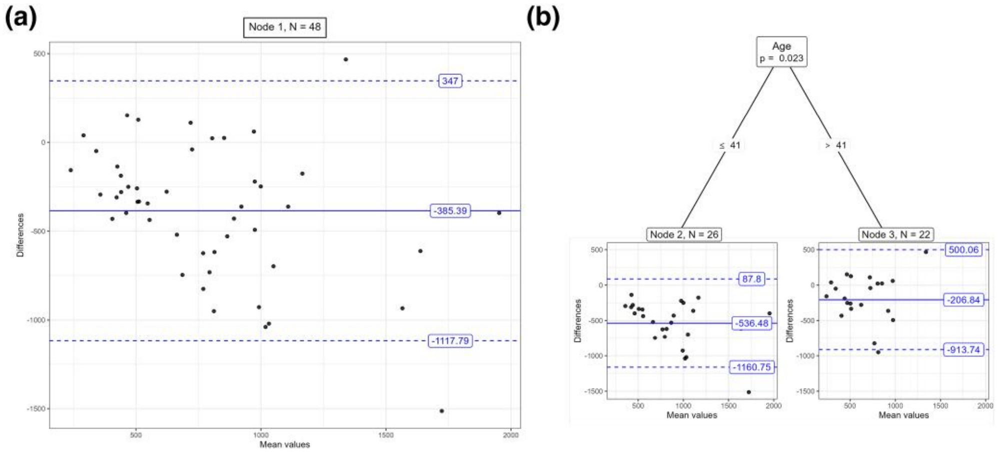


Figure 1. Agreement (a) Bland–Altman plot and conditional agreement (b) Conditional method agreement trees plot of activity energy expenditure (kilocalories) measured by two different accelerometers. ActiGraph based on uniaxial activity counts and Actiheart in lower position are compared. See Section 3 for details.

Downloaded from https://academic.oup.com/jrsssc/advance-article/doi/10.1093/jrsssc/qlae077/7942358 by UiT The Arctic University of Norway user on 06 January 2025

differences in agreement. The models used for analysis are called conditional method agreement trees (COAT).

2.1 Conditional method agreement

As shown and discussed in Section 1, in a Bland–Altman analysis, we are essentially interested in the first and second moments of the marginal density function $f_Y(y)$. Here, $Y = M_1 - M_2$ is a random variable of independent differences between two methods' paired measurements (M_1, M_2). The moments of $f_Y(y)$ are the expectation $\mathbb{E}(Y)$ and the variance $\mathrm{Var}(Y)$ with corresponding estimates given by the mean $\bar{y} = \sum_{i=1}^{n} y_i$ and the empirical variance $s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2$ of the observed differences y_i, $i \in \{1, \ldots, n\}$, of n subjects or observational units. Thereby, $\bar{y}$ describes the overall deviation between methods, which is often referred to as the 'bias' (Hanneman, 2008). However, as discussed in Section 1, the bias is of limited use because it does not provide information about the individual agreement of measurements made for the same subject or observational unit. Interpretation of agreement in a Bland–Altman plot therefore relies mainly on 95% prediction intervals, that is the LoA, which are calculated from the normal distribution using $\bar{y}$ and s, if the differences are normally distributed.

Another assumption of a Bland–Altman analysis is that $\mathbb{E}(Y)$ and $\mathrm{Var}(Y)$ are independent of the magnitude of measurements, implying that the differences are independent and identically distributed (iid). However, if the observed distribution of data suggests that such an association has to be assumed, Bland and Altman propose either to remove this relationship by transforming the differences, for example to establish homoscedasticity by using a log-transformation, or to use a regression model considering the differences Y as the outcome and the mean measurements $M = \frac{1}{2}(M_1 + M_2)$ as an explanatory variable (Bland & Altman, 1999). We generalize this approach to define conditional method agreement as follows.

Given a random variable $Y = M_1 - M_2$ of differences between two methods' measurements M_1 and M_2 and random variables of any scale serving as covariates X_j, $j = 1, \ldots, m$, conditional method agreement is based on the following assumption:

$$f_Y(y \mid x_j) \neq f_Y(y).$$

Here $f_Y(y \mid x_j)$ is the conditional density function of Y given $X_j = x_j$. The realizations y are the observed differences and x are the measured covariate values which can also include mean values $m = \frac{1}{2}(m_1 + m_2)$ of paired measurements. In the present work, we use COAT to obtain estimates of the conditional expectation $\mathbb{E}(Y \mid X_j)$ and the conditional variance $\mathrm{Var}(Y \mid X_j)$ to assess conditional method agreement, with and without using distributional assumptions about $f_Y(y \mid x_j)$ as detailed in Table 1. Respective null-hypotheses

$$H_0 : \mathbb{E}(Y \mid X_j) = \mathbb{E}(Y) \cap \mathrm{Var}(Y \mid X_j) = \mathrm{Var}(Y) \tag{1}$$

are tested by COAT to determine the statistical significance of the association of the agreement and covariates in terms of expectation and variance. The covariates can be of any scale. With a binary covariate, the procedure can also be used to perform a two-sample test to compare agreement between predefined subgroups as outlined in detail in Section 3.3. With a continuous covariate or a multicategorical covariate, subsetting of the data is determined after a significant association with agreement has been detected as described in the following section.

2.2 Recursive partitioning of method agreement

The general idea of recursive partitioning is to assess sequentially whether an investigated outcome variable (or model) is homogeneous across all available covariates and, if this is not the case, to capture the differences by splits into more homogeneous subsets of the data (Breiman et al., 1984). The procedure continues recursively until some kind of stopping criterion is reached. The resulting model is often referred to as a tree because of its structure. The subsets considered for splitting or emerging from splitting are termed parent nodes or daughter/child nodes,

Downloaded from https://academic.oup.com/jrsssc/advance-article/doi/10.1093/jrsssc/qlae077/7942358 by UiT The Arctic University of Norway user on 06 January 2025

Table 1. Combinations of fitted model type, test type, test statistics, and transformation function considered in COAT models

	Fit	Test	Statistic	Transformation	Distribution
CTreeTrafo	nonparametric	permutation	quadratic	$(y_i, (y_i - \bar{y}_\omega)^2)$	non
DistTree	parametric	permutation	quadratic	$s(\hat{\boldsymbol{\theta}}, y_i)$	normal
MOB	parametric	fluctuation	quadratic	$s(\hat{\boldsymbol{\theta}}, y_i)$	normal

respectively. A so called stump is obtained if a single split is performed. The definition of the splits performed in the covariates provides decision rules that specify the subsets.

To define heterogeneous subsets in terms of $\mathbb{E}(Y \mid X_j)$ and $\text{Var}(Y \mid X_j)$, referring to the mean (bias) and standard deviation of the differences y, we consider the following tree-based algorithms: conditional inference tree with an appropriate transformation of the outcome (CTreeTrafo) (Hothorn, Hornik, et al., 2006), distributional tree (DistTree) (Schlosser et al., 2019), and model-based recursive partitioning (MOB) (Zeileis, Hothorn, et al., 2008). All of these modelling approaches are based on the same basic steps:

1. In DistTree and MOB, a model is fit to the data by optimizing some objective function. In CTreeTrafo, a transformation function is applied to the data.
2. A split variable is selected based on the association of some goodness-of-fit measure or the transformed data with each possible variable. The variable with the highest significant association is selected.
3. A split point of the selected variable is chosen so the goodness-of-fit is maximized in the resulting subsets.
4. Steps (1)–(3) are repeated in the subsets until no more significant associations are found or the subsets become too small for further splits.

The basic algorithm of the three models considered is thus similar. However, they differ in the implementation of the individual steps, as explained in more detail in the following. Default features of all of the aforementioned models are summarized in Table 1.

2.2.1 *Conditional inference tree*

The algorithm uses the asymptotic distribution of permutation statistics (Hothorn, Hornik, et al., 2006; Strasser & Weber, 1999), to explore whether there is a statistically significant dependence of the outcome on a covariate. Therefore, j partial hypotheses of independence $H_0^j : f_Y(y \mid x_j) = f_Y(y)$ are defined for $j = 1, \ldots, J$ covariates. The respective linear test statistics are

$$t_j = \text{vec}\left(\sum_{i=1}^{n} \omega_i g_j(x_{ji}) h(y_i, (y_1, \ldots, y_n))^\top\right) \in \mathbb{R}^{pq},$$

where ω_i is a case weight of zero or one, indicating the correspondence of an observation to the node or subset in which the test is performed. $g_j(\cdot)$ and $h(\cdot)$ represent nonrandom transformation functions. The choice of $g_j(\cdot)$ depends on the type of the jth covariate. The identity function, $g_j(x_{ji}) = x_{ji}$, is a natural choice for a continuous variable, while the indicator function $g_j(x_{ji}) = (I(x_{ji} = 1), \ldots, I(x_{ji} = K))$ is more appropriate for a categorical variable with K levels. With the vec$(\cdot)$ operator, the test statistic becomes a pq column vector, where $p = K$ for categorical covariates and $p = 1$ for continuous covariates with identity transformation. q depends on the choice of $h(\cdot)$ and takes a value of 2 in our case, as outlined below.

In the present setting, that is to model method agreement through the estimation of $\mathbb{E}(Y \mid X)$ and $\text{Var}(Y \mid X)$, we define $h(y_i) = (y_i, (y_i - \bar{y}_\omega)^2)$, which corresponds to the first step in the basic

Downloaded from https://academic.oup.com/jrsssc/advance-article/doi/10.1093/jrsssc/qlae077/7942358 by UiT The Arctic University of Norway user on 06 January 2025

algorithm. The multivariate linear test statistic t_j is then defined as

$$t_j = \mathrm{vec}\left(\sum_{i=1}^{n} \omega_i g_j(x_{ji})\left(y_i, \left(y_i - \bar{y}_\omega\right)^2\right)^\top\right) \in \mathbb{R}^{p2},$$

where $\bar{y}_\omega = \sum_{i=1}^{n} \omega_i y_i / \sum_{i=1}^{n} \omega_i$ is the mean outcome in the node or subset in which the test is performed. Under the null hypothesis H_0^j, the expectation μ_j and covariance matrix Σ_j has been derived by Strasser and Weber (1999) who also show that the asymptotic conditional distribution is normal. This result can be leveraged to obtain critical values or p-values relatively easily for two types of univariate test statistics based on t_j. The first is a maximum standardized test statistic:

$$c_{\max}\left(t_j, \mu_j, \Sigma_j\right) = \max_{z=1,\ldots,p2} \left| \frac{\left(t_j - \mu_j\right)_z}{\sqrt{\left(\Sigma_j\right)_{zz}}} \right|.$$

The second one is a quadratic form

$$c_{\mathrm{quad}}\left(t_j, \mu_j, \Sigma_j\right) = \left(t_j - \mu_j\right)\Sigma_j^{+}\left(t_j - \mu_j\right)^\top,$$

where the asymptotic conditional distribution is χ^2 with degrees of freedom given by the rank of Σ_j. Σ_j^{+} is the Moore-Penrose inverse of Σ_j. Thus, both of these test statistics enable the computation of a p-value, where H_0^j can be rejected if this value falls below a specified significance level. The jth covariate with the minimum and statistically significant p-value is selected for splitting, corresponding to the second step in the basic algorithm. Note, that the multiple testing problem is present, as hypotheses for several covariates are checked. Therefore, the CTree algorithm uses Bonferroni-adjusted p-values by default. All theoretical details for the test statistics are derived by Strasser and Weber (1999) and discussed from a practical perspective in Hothorn, Hornik, et al. (2006).

After selecting the split variable j^*, the subsequent and third step of the basic algorithm is to find the optimal split point in a continuous variable or dichotomization of the K categories of a categorical variable for binary splitting, which is again determined through a test statistic

$$t_{j^*}^{A} = \mathrm{vec}\left(\sum_{i=1}^{n} \omega_i I\left(x_{j^*i} \in A\right)\left(y_i, \left(y_i - \bar{y}_\omega\right)^2\right)^\top\right) \in \mathbb{R}^2.$$

Here, $t_{j^*}^{A}$ implicitly measures the discrepancy between the subsets $\{y_i \mid \omega_i = 1 \text{ and } x_{j^*i} \in A;\ i = 1, \ldots, n\}$ and $\{y_i \mid \omega_i = 1 \text{ and } x_{j*i} \notin A;\ i = 1, \ldots, n\}$ in terms of a metric defined by $h(\cdot)$, where A represents all possible subsets. The best split point is found by maximizing $c(t_{j^*}^{A}, \mu_{j^*}^{A}, \Sigma_{j^*}^{A})$ over all possible subsets A using the conditional expectation $\mu_{j^*}^{A}$ and covariance $\Sigma_{j^*}^{A}$ of $t_{j^*}^{A}$. This procedure is recursively repeated until no further statistically significant associations are found or subsets become too small for further splitting (which is the fourth step in the basic algorithm).

It is important to note that by choosing the transformation functions $h(\cdot)$ and $g_j(\cdot)$ a wide range of classical tests are special cases of this conditional inference framework. This includes rank-based procedures such as the Wilcoxon–Mann–Whitney or the Spearman test but also ANOVA statistics and Pearson correlation tests (see Hothorn, Hornik, et al., 2006, for details). Using this flexible inference framework as the basis for the CTree algorithm is attractive for two reasons: First, by choosing the bivariate transformation $h(y_i) = (y_i, (y_i - \bar{y}_w)^2)$ we readily obtain tests that simultaneously assess both parts (expectation and variance) of the null hypothesis H_0 from (1). Second, the tests can be re-used both for split variable and split point selection in the CTree algorithm. Both properties are provided in a unifying framework. In the following simulation study

Downloaded from https://academic.oup.com/jrsssc/advance-article/doi/10.1093/jrsssc/qlae077/7942358 by UiT The Arctic University of Norway user on 06 January 2025

(see Section 4), we will revisit this aspect as we compare the performance of COAT against the benchmark of a CTree that uses the default settings of the algorithm, that is $h(y_i) = y_i$, which implies that the correlation between the continuous outcome and the continuous covariates is tested. We also discuss in the next section that using transformation of the outcome $h(y_i) = (y_i, (y_i - \bar{y}_\omega)^2)$ makes the association test of CTree test the null-hypothesis (1).

2.2.2 Distributional tree

DistTree is similar to CTreeTrafo while a parametric model is fit to the data and the transformation function is replaced with the resulting score function. In particular, DistTree models all parameters of a given distribution (Schlosser et al., 2019). In the present setting of a Bland–Altman analysis, a normal distribution with the location and scale parameters μ and σ^2 for the differences Y is assumed (Bland & Altman, 1986). This allows the specification of the corresponding log-likelihood

$$l(\boldsymbol{\theta}; Y) = \log\left\{\frac{1}{\sigma\sqrt{2\pi}}\phi\left(\frac{Y-\mu}{\sigma}\right)\right\}; \quad \boldsymbol{\theta} = (\mu, \sigma),$$

and its score function $s(\boldsymbol{\theta}, Y) = \partial l(\boldsymbol{\theta}; Y)/\partial\boldsymbol{\theta}$ as a measure of goodness-of-fit. $\phi(\cdot)$ is the density function of a standard normal distribution. A maximum-likelihood (ML) estimate of $\boldsymbol{\theta}$ is $\hat{\boldsymbol{\theta}} = \arg\max \sum_{i=1}^{n} l(\boldsymbol{\theta}; y_i)$. This corresponds to the first step in the basic algorithm.

When it is assumed that the differences y are not iid, DistTree can be used to model the conditional expectation $\mathbb{E}(Y \mid X)$ and variance $\text{Var}(Y \mid X)$. To do so, a possible association of $\boldsymbol{\theta}$ and a covariate X_j is tested in terms of the null-hypothesis $H_0^j : s(\boldsymbol{\theta}, Y) \perp X_j$, based on the multivariate linear test statistic

$$t_j = \text{vec}\left(\sum_{i=1}^{n} g_j(x_{ji})\, s\left(\hat{\boldsymbol{\theta}}, y_i\right)\right).$$

The asymptotic conditional distribution of the linear test statistic t_j has been shown to be multivariate normal with parameters μ_j and Σ_j (Strasser & Weber, 1999). Here, $\hat{\boldsymbol{\theta}}$ is substituted into the score function to obtain $s(\hat{\boldsymbol{\theta}}, y_i)$ as a measure of goodness-of-fit for each of the observations y_i. The transformation function g_j, as well as the standardized test statistics $c_{\text{quad}}(t_j, \mu_j, \Sigma_j)$ and $c_{\max}(t_j, \mu_j, \Sigma_j)$ are defined as outlined in Section 2.2.1. The split variable X_{j^*} is determined by the lowest and statistically significant p-value, which is by default corrected for multiple testing (equals step 2 of the basic algorithm). In the third step the split point is chosen so that it leads to the largest discrepancy in the sum of scores between the resulting subsets. This procedure is repeated recursively in each subset until no further significant associations are found or the resulting subsets become too small for further splitting.

It is important at this point to draw attention to the similarity of the statistics t_j of CTreeTrafo and DistTree, with CTreeTrafo using a transformation function $h(\cdot)$ instead of the score function $s(\cdot)$ in the calculation. We show the equality of the resulting quadratic test statistics $c_{\text{quad}}(\cdot)$ of CTreeTrafo (with the transformation function $h(\cdot)$ defined as given in the previous Section 2.2.1) and DistTree analytically for the case of a continuous predictor in Appendix A.

2.2.3 Model-based recursive partitioning

MOB is similar to DistTree, but uses a different underlying model and hypothesis test. MOB uses fluctuation tests for parameter instability in regression model fits to build a tree model (Zeileis & Hornik, 2007). In the first step of MOB, a parametric model is fit to the data by maximum-likelihood estimation. In the present case, we consider an intercept-only linear regression model $y_i = \beta_0 + \epsilon_i,\ \epsilon_i \sim \mathcal{N}(0, \sigma)$, to obtain estimates of the expectation $\mathbb{E}(Y) = \beta_0$ and variance $\text{Var}(Y) = \sigma^2$. The second step is to assess parameter instability of the estimated model parameters $\hat{\theta} = (\widehat{\beta}_0, \widehat{\sigma})$ across the values x_j of a potential split variable X_j. Instability is concluded when the scores $s(\hat{\boldsymbol{\theta}}, y_i)$ do not fluctuate randomly along the ordered values x_j (see Zeileis, Hothorn,

Downloaded from https://academic.oup.com/jrsssc/advance-article/doi/10.1093/jrsssc/qlae077/7942358 by UiT The Arctic University of Norway user on 06 January 2025

et al., 2008, for details). The split variable X_j^* is selected as it provides the minimal and statistically significant p-value, which is by default corrected for multiple testing. The split in x_j^* is determined so it maximizes the sum of the log-likelihoods of models that are refit to the resulting subsets, corresponding to the third step in the basic algorithm. As with CTreeTrafo and DistTree, the procedure is repeated recursively in each subset until no further significant associations are found or the resulting subsets become too small for further splitting.

3 Application to accelerometer data

As indicated in Section 1.2, it is of importance to explore the agreement between different wearable devices for measuring physical activity. To detect whether this depends on the characteristics of the participants—and if so, to account for this dependency in the model—we employ the COAT approach proposed in the previous section.

3.1 Data

The data consist of 50 study participants who wore different accelerometers, namely one ActiGraph and two Actiheart devices, simultaneously for 24 hr (Henriksen, Grimsgaard, et al., 2019). The ActiGraph was placed on their right hip, one Actiheart was placed in the upper position of chest, and the second Actiheart in the lower position. Both accelerometers are considered valid for estimating activity energy expenditure (AEE). The difference is that the Actiheart directly reports AEE, using an internal branching model (Brage et al., 2004) where heart rate and acceleration is used together to estimate energy expenditure, while the ActiGraph uses both uniaxial and triaxial activity counts for its calculation.

Using the proprietary software of ActiGraph and the Actiheart, activity counts are exported and subsequently transformed into 60-s epochs. Counts per minute (CPM) are then employed to calculate minutes in the various physical activity (PA) intensity zones—namely, sedentary, light, moderate, vigorous, and very vigorous—using cut-offs defined by several algorithms. Minutes spent in vigorous and very vigorous intensity are combined into one variable. For a more in-depth explanation of variable generation process, refer to (Henriksen, Grimsgaard, et al., 2019).

In the present application, the agreement of daily measurements of PA (in minutes) and AEE (in kilocalories) are compared between different pairs of two accelerometers each, conditional on the participants' age, sex, height, and weight. As described in Section 2.1, we also include the mean PA and AEE measurements along with the other covariates as a potential explanatory variable. Two cases with missing values were removed from the data. Characteristics of the participants are presented in Table 2.

3.2 Fitting the model

3.2.1 Physical activity

When COAT is applied to physical activity measurements derived from different accelerometers, it becomes clear that the magnitude of the measurements can play a role in determining the agreement. Regarding light physical activity, no differences are detected between the agreement in terms of potential split variables and the size of the measurements, resulting in the classical Bland–Altman plot (cf. Figure 2a). However, COAT by CTreeTrafo reveals that the agreement in moderate-to-vigorous physical activity (MVPA) measurements may depend on the magnitude

Table 2. Participant characteristics of the accelerometer study ($n = 48$)

Variables	n(%); **Median (IQR)**
Female	24(50%)
Age (years)	40(35, 57)
Height (cm)	174(166, 182)
Weight (kg)	75(63, 86)

Downloaded from https://academic.oup.com/jrsssc/advance-article/doi/10.1093/jrsssc/qlae077/7942358 by UiT The Arctic University of Norway user on 06 January 2025

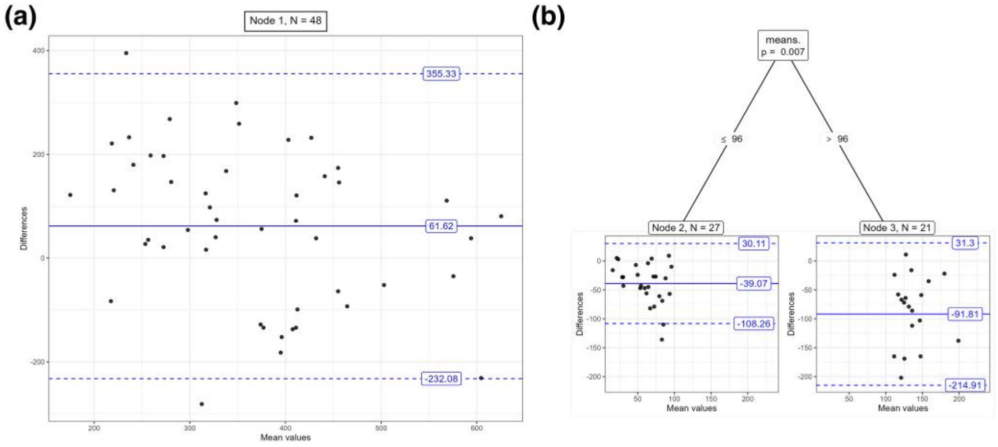


Figure 2. Conditional method agreement trees by CTreeTrafo for conditional agreement of light physical activity (a) and moderate-to-vigorous physical activity (b) measurements of two accelerometers. ActiGraph based on triaxial activity counts and Actiheart in upper position are compared. The ActiGraph was placed on the right hip, the Actiheart was placed in the upper position of chest.

of the measurements. Heterogeneous subgroups, defined by a split point of 96 mean minutes in MVPA, show significant differences in bias and width of LoA (cf. Figure 2b).

This observation also shows that transforming raw data, as proposed by Bland and Altman, in order to remove dependencies between mean measurements and differences, may become redundant in specific cases as the new approach is able to explain this dependency. In other cases there might still be the need for additional transformation of data, for example when the dependency cannot be well explained solely by the definition of subgroups. The application of COAT to various physical activity measurements, computed using different combinations of accelerometers, demonstrates its ability to identify dependencies in agreement based on covariates or the magnitude of measurements. In particular, we have found that the ActiGraph corresponds better with the Actiheart among younger participants or with lower mean measurements.

3.2.2 *Activity energy expenditure*

Figure 3a shows that for one pair of compared accelerometers, COAT by MOB is able to identify subgroups of participants, which are heterogeneous regarding the bias and width of LoA depending on age ($p = .034$). Better agreement, in terms of bias decreasing from about 424 to 100 kcal, is obtained for patients older than 41 years. With two other accelerometers, performing COAT by CTreeTrafo shows that agreement may be conditional on the magnitude of measurements (Figure 3b). With an average AEE >1040 kcal, the bias in agreement increases from about 189 to 607 kcal and the width of the LoA increases from about 487 to 1,198 kcal ($p = .006$).

3.3 A two-sample test of differences in method agreement

It has been proposed in Section 2.1 to apply COAT to perform a two-sample test of the null-hypotheses (1) for comparison of agreement between (pre)defined subgroups. For example, in the application of the previous Section 3.2, a researcher may be interested in a potential difference of agreement between the sexes. Figure 4 shows the result of COAT by CTree, when a stump tree is generated for sex as the only covariate. In this implementation of COAT, the χ^2 test statistic c_{quad}, the degrees of freedom and the respective p-value (cf. Section 2.2.1) are presented for testing the null-hypothesis (1) concerning differences in bias and width of LoA between the considered subgroups. Corresponding estimates of $\mathbb{E}(Y \mid X)$ and $\text{Var}(Y \mid X)$ are provided for each subgroup, too. The choice of COAT by CTree is motivated by the fact that we can additionally test associations with respect to each of $\mathbb{E}(Y \mid X)$ and $\text{Var}(Y \mid X)$ separately, using $h(y_i) = y_i$ or $h(y_i) = (y_i - \bar{y}_\omega)^2$ in the respective test statistic (see Section 2.2.1 for details on $h(\cdot)$ and the test statistic). In the present case, no statistically significant association with sex was found in terms of bias ($p = .619$), width of

Downloaded from https://academic.oup.com/jrsssc/advance-article/doi/10.1093/jrsssc/qlae077/7942358 by UiT The Arctic University of Norway user on 06 January 2025

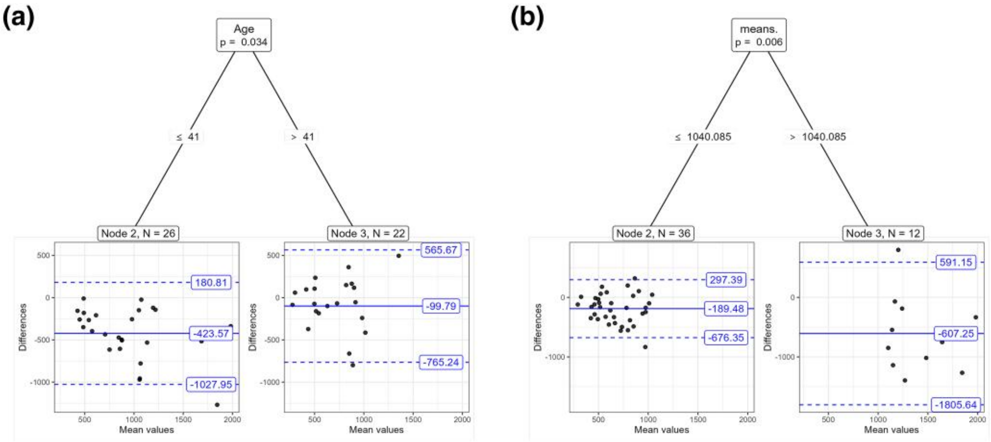
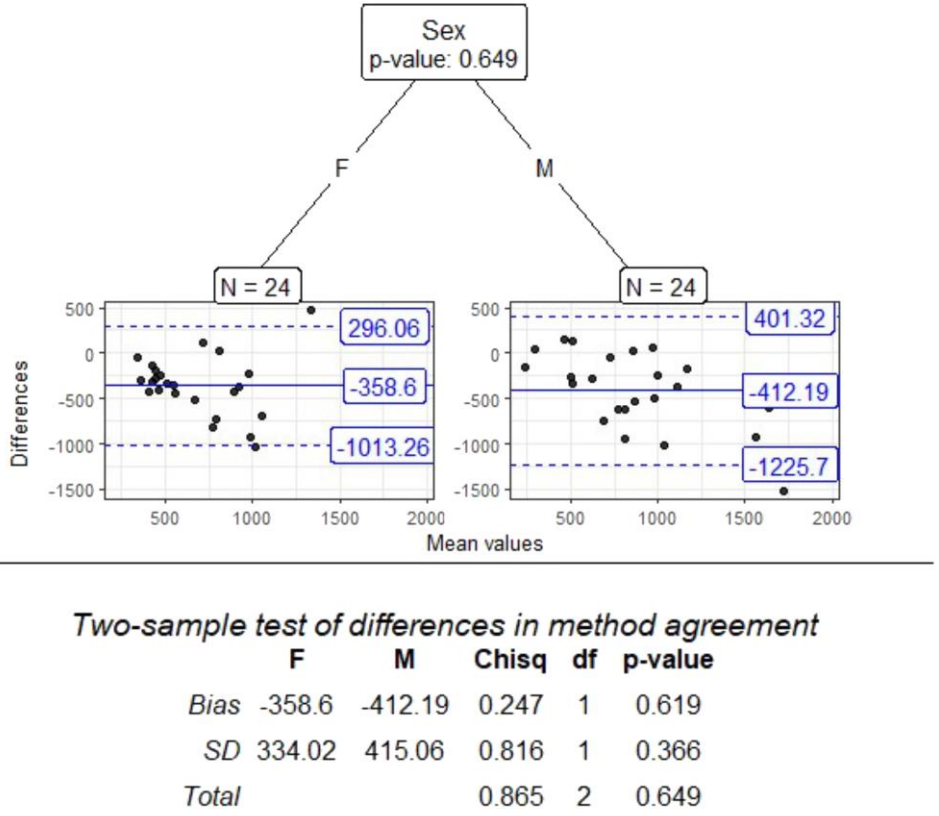


Figure 3. Conditional method agreement trees (COAT) for conditional agreement of activity energy expenditure measurements of two accelerometers. Note that different pairs of accelerometers are compared in (a) COAT by model-based trees, ActiGraph based on triaxial activity counts and Actiheart in upper position and (b) COAT by CTreeTrafo, ActiGraph based on triaxial activity counts, and Actiheart in lower position. See Section 3 for details on accelerometers.



Two-sample test of differences in method agreement

	F	M	Chisq	df	p-value
Bias	-358.6	-412.19	0.247	1	0.619
SD	334.02	415.06	0.816	1	0.366
Total			0.865	2	0.649

Figure 4. Two-sample test of differences in method agreement of activity energy expenditure measurements between female (F) and male (M) participants in the application study. ActiGraph based on uniaxial activity counts and Actiheart in the upper position are compared.

Downloaded from https://academic.oup.com/jrsssc/advance-article/doi/10.1093/jrsssc/qlae077/7942358 by UiT The Arctic University of Norway user on 06 January 2025

Downloaded from https://academic.oup.com/jrsssc/advance-article/doi/10.1093/jrsssc/qlae077/7942358 by UiT The Arctic University of Norway user on 06 January 2025

the LoA ($p = .366$) and both of these quantities ($p = .649$). Please note that these three p-values are not adjusted for the multiple testing problem, but are easily suitable for conducting a sequential test procedure (starting with the test of both quantities, followed by the test of the individual quantities). Other corrections, such as the Bonferroni correction, are of course also possible. Similarly to the proposed approach, a single predictor variable of any scale could be used to simultaneously derive and test two subgroups. The type-I error is still controlled in this case as the tree algorithms used detach the test of association from the search of an optimal split point or definition of subsets (Hothorn, Hornik, et al., 2006, see also Section 2.2). However, this case is already covered by COAT, as described above.

4 Simulation studies

To further investigate the performance of COAT, several simulation studies are carried out in this section, assessing the properties of the model in different controlled settings. In particular, the assessment of performance is based on the type-I error and the power to reject H_0 as defined in (1), and the Adjusted Rand Index (ARI). The latter is a measure of concordance of two classifications (Hubert & Arabie, 1985), as it quantifies the proportion of paired observations that belong to the same or different class levels in either classification among the total number of paired observations (Rand, 1971). In the case of independent or random classifications, the ARI takes a value of 0. Higher values indicate a higher concordance, with 1 indicating perfect agreement. Here, the ARI is used to assess the concordance between the given subgroups and the subgroups defined by COAT.

4.1 Design

We run 10,000 simulations, and consider sample sizes $n \in \{50, 100, 150, \ldots, 1, 000\}$. A CTree with the default transformation function $h(y_i, (y_i, \ldots, y_n)) = y_i$, as implemented through the function `ctree()` of the R package `partykit` (Hothorn & Zeileis, 2015) serves as a benchmark. It implements a classical test of correlation between the continuous outcome and the continuous covariates, and may therefore be suitable to detect heterogeneity in terms of bias but not in terms of variance/LoA. By contrast, it has been shown for each implementation of COAT (see Section 2), that it tests the null-hypothesis (1) and should therefore be able to detect heterogeneity with respect to both, bias and LoA.

Due to the equivalence of the statistics t_j of CTreeTrafo and DistTree, they are also referred to jointly as CTreeTrafo/DistTree in the following.

Three different simulation scenarios are considered as follows. In the Null Case, the method agreement does not depend on any covariates. The simulated data consist of six independent, standard-normally distributed variables including the outcome Y, which is the simulated differences between the methods, and five uninformative covariates X. The Null Case allows the exploration of the type I error as we look for statistically significant p-values in the root nodes of COAT models that were fit to the simulated data. The nominal significance level is set to $\alpha = 0.05$.

The Stump Case covers three different scenarios. In each of them there are five standard-normally distributed covariates X, where method agreement depends on the informative covariate X_1 such that $Y \sim \mathcal{N}(\mu_k, \sigma_k)$, $k \in \{1, 2, 3\}$, where

$$(\mu_k, \sigma_k) = \begin{cases} (\mu_1 = 0.3 \cdot I(X_1 > Q_{0.25}), \sigma_1 = 1) & \text{if } k = 1, \\ (\mu_2 = 0, \sigma_2 = 1 + I(X_1 > Q_{0.25})) & \text{if } k = 2, \\ (\mu_3 = 0.4 \cdot I(X_1 > Q_{0.25}), \sigma_3 = 1 + I(X_1 > Q_{0.25})) & \text{if } k = 3 \end{cases}$$

Here, $Q_{0.25}$ is the 25th percentile of the standard normal distribution and has been chosen as a split point in X_1 to create subgroups that approximately comprise 25% and 75% of the observations. The subgroups consequently differ only in $\mu_k = \mathbb{E}(Y \mid X)$, that is in the bias of method agreement in the scenario $k = 1$, they differ in $\sigma_k = \text{Var}(Y \mid X)$, that is in the width of the LoA in the scenario $k = 2$, and they differ in both quantities in the scenario $k = 3$. See also Figure 5a for a respective illustration. The performance of COAT is assessed in terms of its power to reject the null-hypothesis (1) for the informative covariate X_1, and to uncover the correct subgroups as measured by the ARI. In this respect, the values of μ_k and σ_k have been chosen in such a way that the power of

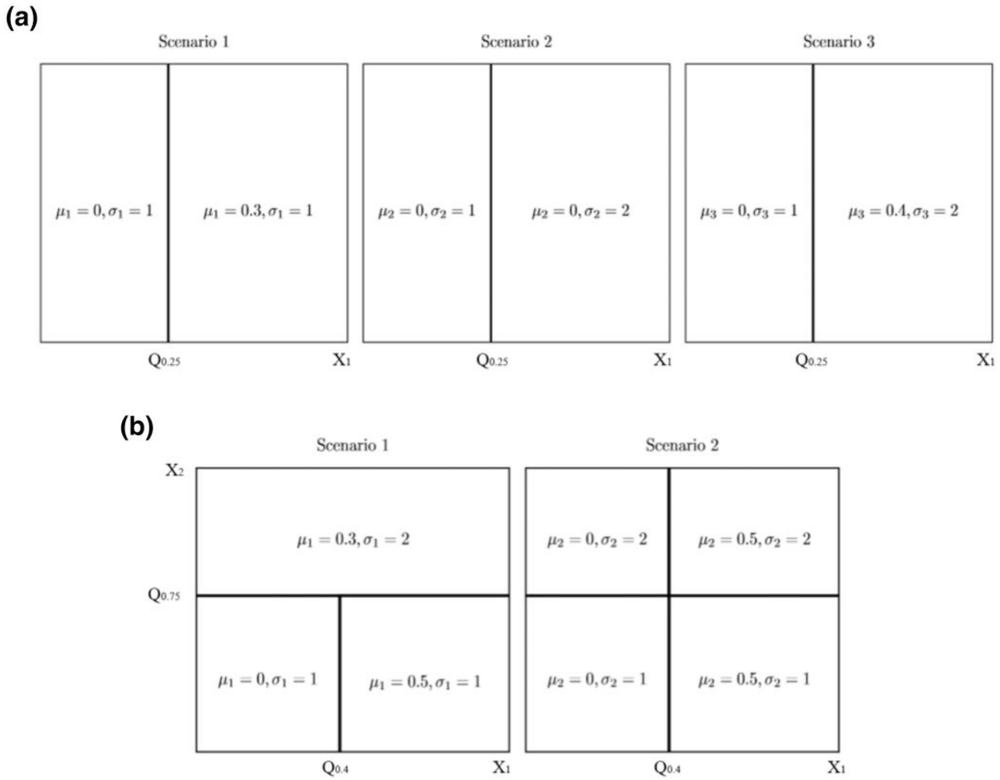

(a)
Scenario 1
$\mu_1 = 0, \sigma_1 = 1$
$\mu_1 = 0.3, \sigma_1 = 1$
Q0.25
X1
Scenario 2
$\mu_2 = 0, \sigma_2 = 1$
$\mu_2 = 0, \sigma_2 = 2$
Q0.25
X1
Scenario 3
$\mu_3 = 0, \sigma_3 = 1$
$\mu_3 = 0.4, \sigma_3 = 2$
Q0.25
X1
(b)
Scenario 1
X2
$\mu_1 = 0.3, \sigma_1 = 2$
Q0.75
$\mu_1 = 0, \sigma_1 = 1$
$\mu_1 = 0.5, \sigma_1 = 1$
Q0.4
X1
Scenario 2
$\mu_2 = 0, \sigma_2 = 2$
$\mu_2 = 0.5, \sigma_2 = 2$
$\mu_2 = 0, \sigma_2 = 1$
$\mu_2 = 0.5, \sigma_2 = 1$
Q0.4
X1

Figure 5. Partitions of X used to define the subgroups in the simulation studies. (a) Illustration of the Stump Case with three scenarios, (b) Illustration of the Tree Case with two scenarios. Detailed explanation of the scenarios can be found in Section 4.1.

a respective two-sample t test would range between 0.372 and 0.995 for the given sample sizes (Chow et al., 2017).

Finally, in the Tree Case, we again consider an outcome $Y \sim \mathcal{N}(\mu_k, \sigma_k)$ with $k \in \{1, 2\}$ and two informative, X_1 and X_2, and three uninformative, X_3, X_4 and X_5, standard-normally distributed covariates, resulting in three or four subgroups (see Figure 5b), according to

$$(\mu_k, \sigma_k) = \begin{cases} \begin{array}{l} \big(\mu_1 = 0.3 \cdot I\big(X_2 \geq Q_{0.75}\big) + 0.5 \cdot I\big(X_2 < Q_{0.75}\big). \\ \quad I\big(X_1 \geq Q_{0.4}\big), \sigma_1 = 1 + I\big(X_2 \geq Q_{0.75}\big)\big) \end{array} & \text{if } k = 1, \\ \big(\mu_2 = 0.5 \cdot I\big(X_1 \geq Q_{0.4}\big), \sigma_2 = 1 + I\big(X_2 \geq Q_{0.6}\big)\big) & \text{if } k = 2. \end{cases}$$

The values of μ_k and σ_k in scenario $k = 1$ have been chosen such that it offers a first split with respect to $\sigma_1^2 = \mathrm{Var}(Y \mid X)$, which deviates between the subgroups defined by the split point $Q_{0.75}$ in X_2, while μ_1 takes the same value $0.4 \cdot 0 + 0.6 \cdot 0.5 = 0.3$ on both sides of this split point. Subsequently, a second split could be performed with respect to $\mu_1 = \mathbb{E}(Y \mid X)$ as it differs between the subgroups defined by the split point $Q_{0.4}$ in X_1 where $X_2 < Q_{0.75}$. In the second scenario, the split point $Q_{0.6}$ in X_2 defines a split with respect to $\sigma_2^2 = \mathrm{Var}(\mathrm{Y} \mid \mathrm{X})$, and the split point $Q_{0.4}$ in X_1 defines a split with respect to $\mu_2 = \mathbb{E}(Y \mid X)$, resulting in four subgroups (see Figure 5b).

4.2 Results

We first investigate the estimated type-I error probabilities of COAT in dependence of sample size in the Null Case. CTree and COAT by CTreeTrafo/DistTree show similar performance with relative rejection frequencies of the null-hypothesis (1) reaching from 3.8% to 5.5%, which are close to the nominal significance level of 0.05 and appear to be independent of sample size (Figure 6). On

Downloaded from https://academic.oup.com/jrsssc/advance-article/doi/10.1093/jrsssc/qlae077/7942358 by UiT The Arctic University of Norway user on 06 January 2025

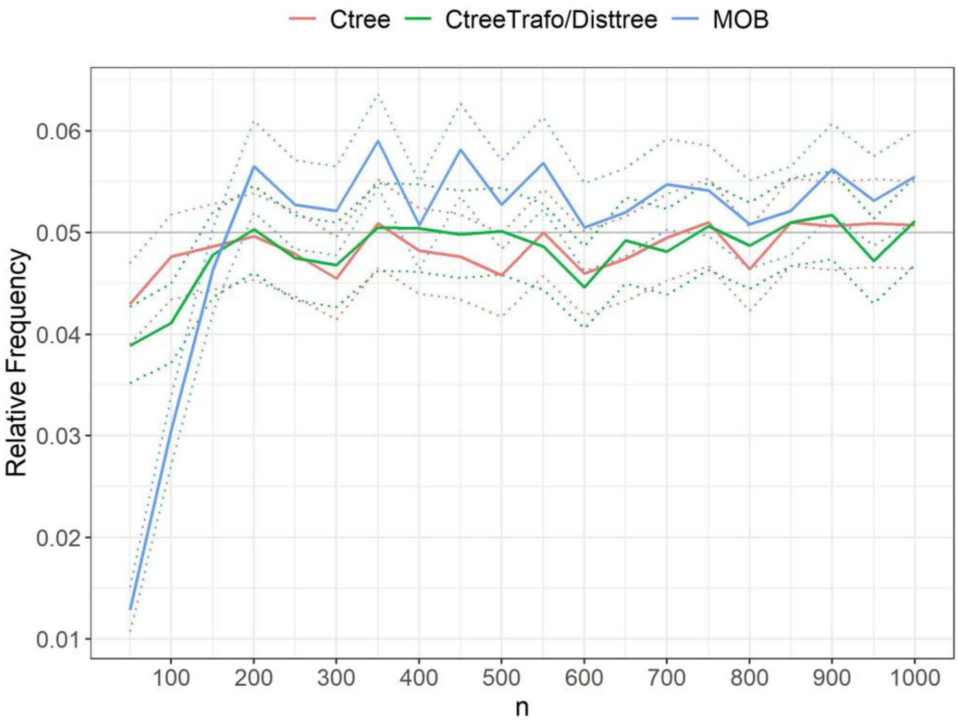


Figure 6. Relative frequency of statistically significant *p*-values observed in the root nodes of conditional method agreement trees models fit to data of increasing sample size in the Null Case with 10,000 replications. These estimates of the type-I error probability are presented with pointwise 95% confidence intervals (dashed lines).

the contrary, COAT implementation by MOB does not seem to achieve the nominal significance level of 0.05 well for smaller sample sizes as it rejects the null-hypothesis in only 1.3% and 3.3% of the simulated cases with $n \leq 100$. With larger sample sizes of $n \geq 200$, it showed relative frequencies for the type-I error between 5.1% and 5.8%, which are slightly but clearly increased beyond the nominal significance level of 0.05.

The performance of COAT in the Stump Case in terms of the power to reject the null-hypothesis (1) for the informative covariate X_1 is estimated by the respective relative frequencies of the association between X_1 and the outcome being significant at the 5% level in the root node of the tree models (Figure 7). When only the expectation μ_1 but not the variance σ_1^2 varies between the defined subgroups (i.e. scenario $k = 1$), CTree and MOB perform best. However, for the case where only the variance σ_1^2 varies (i.e. scenario $k = 2$), the performance of CTree decreases as it has not been enabled through a respective definition of the transformation function $h(\cdot)$ to detect such variation. Again, the MOB tree performs best, closely followed by CTreeTrafo and DistTree.

However, power estimates do not indicate whether the true subgroups are correctly specified. Therefore, the ARI has also been investigated for the Stump Case and the Tree Case. The average ARI is plotted against increasing sample size in Figure 7 and in Figure 8, respectively. As expected, the ARI increases as the sample size increases in both cases. However, Ctree can only keep up with COAT implementations when there is only variation in the expectation μ_k and not in the variance σ_k. COAT seems to be able to cope even with the more complex setting when there are more than two true subgroups. Overall, the results for estimated power and ARI are largely comparable and lead to identical conclusions regarding the performance of the modelling approaches. An example of a tree case is given in Figure 9. It can be seen that after splitting in x_1 COAT splits in x_2, which represents a possible association of x_2 to agreement conditional on x_1. This analysis is exploratory per se, but it is also possible to perform confirmatory tests by predefining subgroups and performing a two-sample test of differences in method agreement.

Downloaded from https://academic.oup.com/jrsssc/advance-article/doi/10.1093/jrsssc/qlae077/7942358 by UiT The Arctic University of Norway user on 06 January 2025

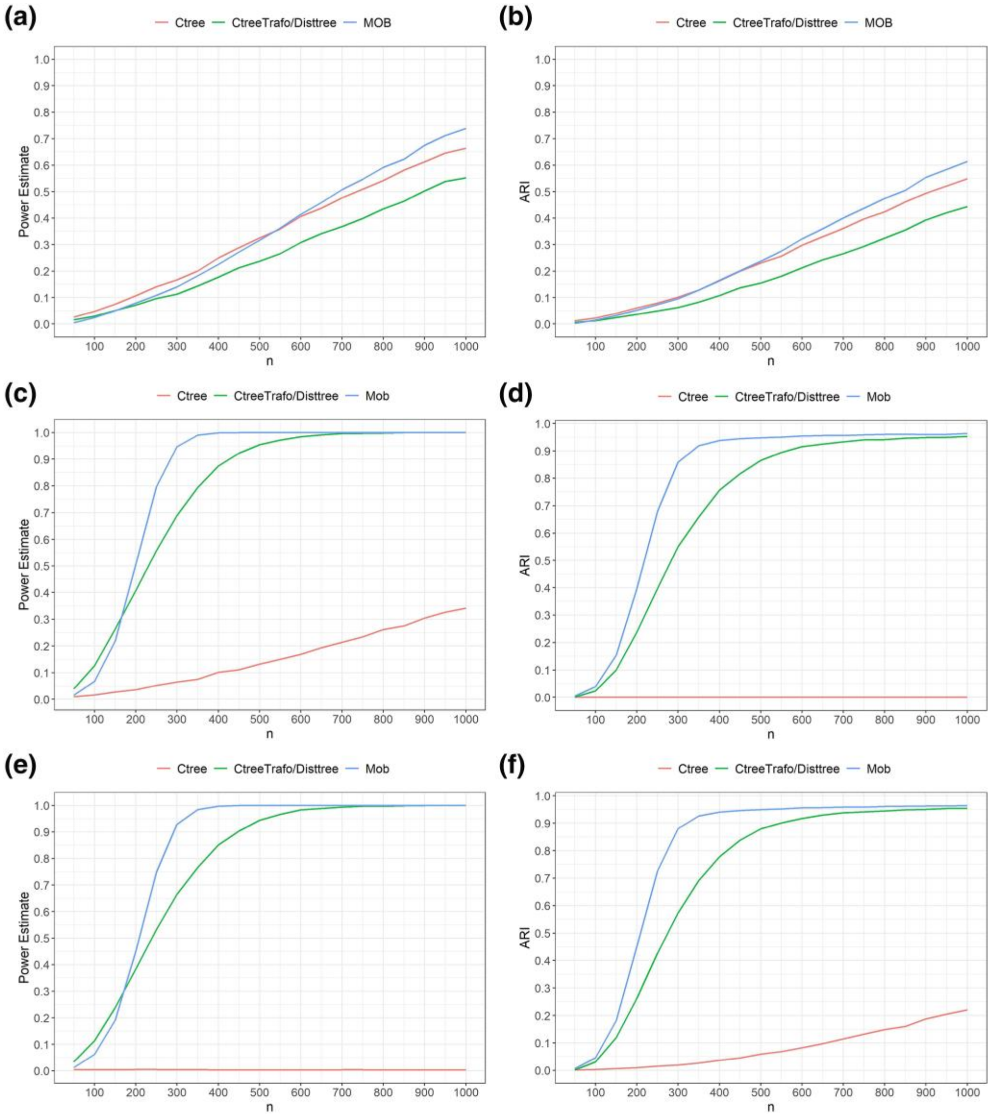


Figure 7. Power estimates (a), (c), (e) and Adjusted Rand Index (ARI) (b), (d), (f) for CTree, CTreeTrafo, DistTree, and model-based trees in the three Stump Case scenarios $k \in \{1, 2, 3\}$, for increasing sample size. The ARI measures the concordance of the subgroups detected by conditional method agreement trees and the true underlying subgroups on a range from 0 (= random concordance) and 1 (= perfect concordance). In scenario $k = 1$, there are two subgroups with different bias, in scenario $k = 2$, there are two subgroups with different variance, and i scenario $k = 3$, there are two subgroups with different bias and variance as detailed in Section 4.1. The maximum width of the pointwise 95% confidence intervals was only 1.97%, which is why they are not presented in the plots.

5 Discussion

The contribution of the present work to the field of method comparison studies is fourfold. First, the concept of conditional method agreement is introduced and formalized. Second, respective statistical modelling by recursive partitioning is proposed introducing COAT. Third, a respective two-sample test is suggested to test for differences in agreement, with respect to the bias and width of LoA, between (pre)defined subgroups. Fourth, COAT is made publicly available through the R package `coat`. Additionally, this work provides an empirical contribution by showing that the agreement of activity measurements depends on the age of the participants.

Downloaded from https://academic.oup.com/jrsssc/advance-article/doi/10.1093/jrsssc/qlae077/7942358 by UiT The Arctic University of Norway user on 06 January 2025

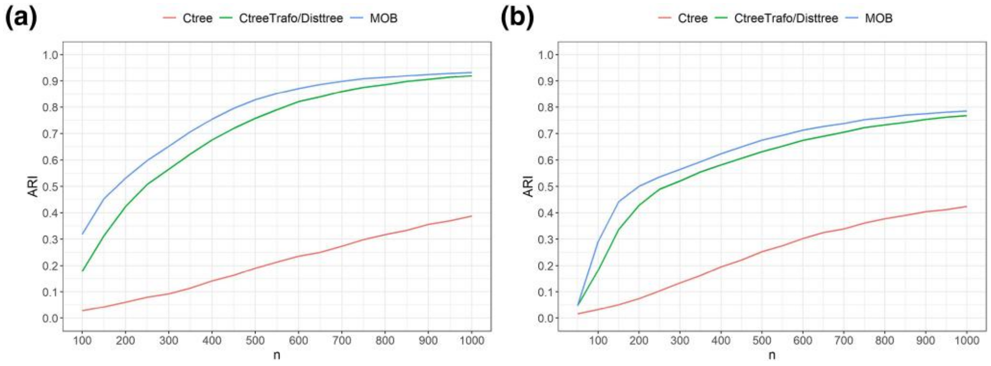


Figure 8. Adjusted Rand Index (ARI) of CTree, CTreeTrafo, DistTree, and model-based trees in the two Tree Case scenarios $k \in \{1, 2\}$, for increasing sample size. The ARI measures the concordance of the subgroups detected by conditional method agreement trees and the true underlying subgroups on a range from 0 (= random concordance) and 1 (= perfect concordance). In scenario $k = 1$ (a), there are three subgroups with different bias and variance, and in scenario $k = 2$ (b), there are four different subgroups as detailed in Section 4.1. The maximum width of the pointwise 95% confidence intervals was only 0.96%, which is why they are not presented in the plots.

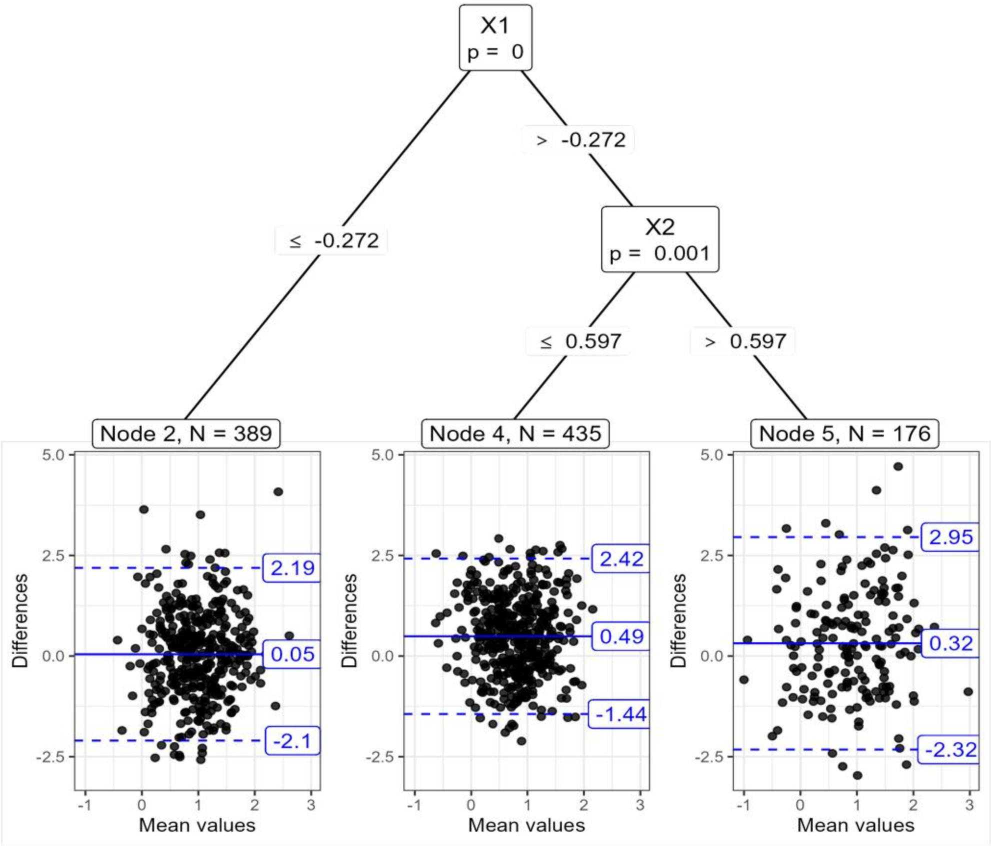


Figure 9. Conditional method agreement trees by CTreeTrafo for conditional agreement of simulated data in the Tree Case scenario $k = 1$. Three subgroups with heterogeneous agreement are defined.

Downloaded from https://academic.oup.com/jrsssc/advance-article/doi/10.1093/jrsssc/qlae077/7942358 by UiT The Arctic University of Norway user on 06 January 2025

Downloaded from https://academic.oup.com/jrsssc/advance-article/doi/10.1093/jrsssc/qlae077/7942358 by UiT The Arctic University of Norway user on 06 January 2025

When examining accelerometer data by applying COAT, the potential influence of covariates and mean measurements on the agreement between physical activity measurements from different accelerometers becomes apparent. In particular, better agreement in terms of bias was observed in younger participants or those with lower mean measurements on most physical activity measures. Consequently, measurements of larger values may be less reliable. In general, this application demonstrates the ability of COAT to provide a solution to simultaneously address the research questions of method agreement and potential dependence on covariates in a unifying framework. It therefore exploits the fact that conditional method agreement can be parameterized through the expectation $\mathbb{E}(Y \mid X)$ and variance $\mathrm{Var}(Y \mid X)$ of paired differences between two methods' measurements. Correctly specified tree-based models are used for estimation of these conditional parameters and enable the definition of subgroups with different agreement.

The application study shows the potential of COAT for epidemiological research. Therefore, subgroups with heterogeneous method agreement in activity energy expenditure (AEE) measurements could be identified in terms of bias and width of LoA depending on covariates and the size of the AEE measurements. Therefore, we recommend that the inclusion of covariates should already be considered in the planning phase of method agreement studies. From the perspective of the applicant, which could be a manufacturer of accelerometers, a researcher, investigator or treating physician, one can then recommend which accelerometer to use or how to improve measurements in a particular setting for a particular person.

Results of the simulation study indicate that the implementations of COAT by CTree (i.e. CTreeTrafo) and DistTree are able to control the type-I error probability at the nominal significance level, independent of sample size. By contrast, the implementation by MOB showed a decisively decreased error rate with small sample sizes and a slightly increased error rate with larger sample sizes. Therefore, it cannot be recommended for COAT in its present form, and further research could be directed towards robust variance estimation and improvements in distributional approximations for possible correction. All implementations of COAT performed well in detecting existent subgroups with increasing sample size. The comparison to the default specification of the CTree algorithm shows the disadvantage of implementing classical tests only, that is that CTree without the proposed transformation only captures differences in the bias, that is in the conditional expectation $\mathbb{E}(Y \mid X)$, but cannot uncover differences in the width of LoA, that is in the variance $\mathrm{Var}(Y \mid X)$. The present simulation studies are based on normally distributed outcomes and covariates. The performance of the approaches studied might have been different if other scalings had been used. Such cases will be addressed in future studies, but are beyond the scope of this introductory work.

Observed differences between the implementations of COAT arise from the testing strategy. Both CTreeTrafo and DistTree compute quadratic test statistics which are equivalent, as has been analytically shown in Appendix A. In this respect, DistTree can be considered a special case of CTree with the appropriate transformation function $h(\cdot)$ as defined in Section 2.2.1. By contrast, MOB is based on fluctuation tests for parameter instability in regression model fits.

Based on our findings, COAT by CTreeTrafo or DistTree is the better choice to strictly control type-I error. It should be noted that the results of COAT are exploratory, unless it is used to conduct a two-sample test of different agreement between (pre)defined subgroups. In the latter case, it can be used for confirmatory hypothesis testing. In this context, it should also be mentioned that CTree, DistTree and MOB by default apply a Bonferroni correction to the multiple testing problem that occurs when a test-based splitting is performed based on multiple covariates. At present, COAT is limited to the case of single measurements per observational unit or subject. A modification for repeated measurements is currently being developed.

6 Conclusion

COAT enables the uni- and multivariable analysis of method agreement in dependence of covariates and mean measurements by conditional modelling and exploratory or confirmatory hypothesis testing. It is made publicly available through the R package `coat`.

Acknowledgments

We thank Alexander Horsch for fruitful discussions and recommendations.

Author contributions

S.K. and A.H. drafted the manuscript, performed the statistical analyses and interpreted the results. André H. extracted and prepared the data used in the application study. All authors revised the manuscript for its content and approved the submission of the final manuscript.

Conflict of interests: None to declare.

Funding

This study was funded by the Deutsche Forschnugsgemeinschaft (DFG, German Research Foundation)—Projektnummer (grant number) 447467169.

Data availability

Data underlying the application study may be obtained from the authors of the original study upon reasonable request (Henriksen, Grimsgaard, et al., 2019). The code of the simulation studies is provided as supplementary material. COAT is made publicly available through the associated R package `coat` on the Comprehensive R Archive Network (CRAN). R code of the performed simulation and application studies is provided as supplementary material. The associated R package `coat` is available from GitHub.

Supplementary material

Supplementary material is available online at *Journal of the Royal Statistical Society: Series C*.

Appendix. Equality of test statistics of CTreeTrafo and DistTree

In this section, we analytically show that the test statistics $c_{\text{quad}}(\cdot)$ of CTreeTrafo and DistTree are equivalent for the case of numeric split variables. For a clearer and more comprehensible presentation of the already very extensive proof, the slightly more complex case of categorical variables has been omitted. However, it can be shown analogously. Recall the test statistic used for CTreeTrafo and DistTree:

$$c_{\text{quad}}(t_j, \mu_j, \Sigma_j) = (t_j - \mu_j)\Sigma_j^{+}(t_j - \mu_j)^{\top}. \tag{A1}$$

In the following, we define each element t_j, μ_j and Σ_j in (A1- based on the formulas from the original publication (Strasser & Weber, 1999) and as outlined in Sections 2.2.1 and 2.2.2. To simplify notation, we omit the index j, which specifies a particular split variable. The weights ω_i are chosen to be 1 focusing on the observations of a given node in a tree. Σ^{+} is in our case equivalent to Σ^{-1}.

A.1 CTreeTrafo

In CTreeTrafo the statistic

$$t = \operatorname{vec}\left(\sum_{i=1}^{n} \underbrace{\omega_i}_{=1} \underbrace{g(x_i)}_{=x_i} (y_i, (y_i - \bar{y})^2)^{\top}\right) = \sum_{i=1}^{n} x_i (y_i, (y_i - \bar{y})^2)^{\top}$$

$$= \left(\sum_{i=1}^{n} x_i y_i, \sum_{i=1}^{n} x_i \underbrace{(y_i - \bar{y})^2}_{=s_i}\right)^{\top} = \left(\sum_{i=1}^{n} x_i y_i, \sum_{i=1}^{n} x_i s_i\right)^{\top}$$

Downloaded from https://academic.oup.com/jrsssc/advance-article/doi/10.1093/jrsssc/qlae077/7942358 by UiT The Arctic University of Norway user on 06 January 2025

has the expectation

$$
\begin{aligned}
\mu &= \operatorname{vec}\left(\left(\sum_{i=1}^{n}\omega_i g(x_i)\right)\frac{1}{n}\sum_{i=1}^{n}\omega_i(y_i,\,(y_i-\overline{y})^2)^\top\right) = \left(\sum_{i=1}^{n}x_i\right)\frac{1}{n}\sum_{i=1}^{n}(y_i,\,(y_i-\overline{y})^2)^\top \\
&= \left(\frac{1}{n}\sum_{i=1}^{n}x_i\sum_{i=1}^{n}y_i,\ \frac{1}{n}\sum_{i=1}^{n}x_i\sum_{i=1}^{n}(y_i-\overline{y})^2\right)^T = \left(n\overline{x}\overline{y},\ \overline{x}\sum_{i=1}^{n}(y_i-\overline{y})^2\right)^T \\
&= \left(n\overline{x}\overline{y},\ n\overline{x}\overline{s}\right)^T
\end{aligned}
$$

and covariance

$$
\begin{aligned}
\Sigma &= \frac{n}{n-1}V\otimes\left(\sum_{i=1}^{n}\omega_i g(x_i)\otimes\omega_i g(x_i)^\top\right) - \frac{1}{n-1}V\otimes\left(\sum_{i=1}^{n}\omega_i g(x_i)\right)\left(\sum_{i=1}^{n}\omega_i g(x_i)\right)^\top \\
&= \frac{n}{n-1}V\left(\sum_{i=1}^{n}x_i x_i\right) - \frac{1}{n-1}V\left(\sum_{i=1}^{n}x_i\right)\left(\sum_{i=1}^{n}x_i\right) \\
&= \frac{n}{n-1}V\left(\sum_{i=1}^{n}x_i^2\right) - \frac{1}{n-1}V\left(\sum_{i=1}^{n}x_i\right)^2 = \frac{n}{n-1}V\left(\sum_{i=1}^{n}x_i^2 - \frac{1}{n}\left(\sum_{i=1}^{n}x_i\right)^2\right) \\
&= \frac{n}{n-1}V\left(\sum_{i=1}^{n}x_i^2 - n\overline{x}^2\right),
\end{aligned}
\tag{A2}
$$

where ⊗ is the Kronecker product and V is defined as follows:

$$
\begin{aligned}
V &= \frac{1}{n}\sum_{i=1}^{n}\omega_i\left((y_i,\,(y_i-\overline{y})^2) - \frac{1}{n}\sum_{i=1}^{n}\omega_i(y_i,\,(y_i-\overline{y})^2)\right)\left((y_i,\,(y_i-\overline{y})^2) - \frac{1}{n}\sum_{i=1}^{n}\omega_i(y_i,\,(y_i-\overline{y})^2)\right)^\top \\
&= \frac{1}{n}\sum_{i=1}^{n}\left(y_i - \frac{1}{n}\sum_{i=1}^{n}y_i,\ (y_i-\overline{y})^2 - \frac{1}{n}\sum_{i=1}^{n}(y_i-\overline{y})^2\right)\left(y_i - \frac{1}{n}\sum_{i=1}^{n}y_i,\ (y_i-\overline{y})^2 - \frac{1}{n}\sum_{i=1}^{n}(y_i-\overline{y})^2\right)^\top \\
&= \frac{1}{n}\sum_{i=1}^{n}\left(y_i-\overline{y},\ (y_i-\overline{y})^2 - \frac{1}{n}\sum_{i=1}^{n}(y_i-\overline{y})^2\right)\left(y_i-\overline{y},\ (y_i-\overline{y})^2 - \frac{1}{n}\sum_{i=1}^{n}(y_i-\overline{y})^2\right)^\top \\
&= \frac{1}{n}\begin{pmatrix}
\sum_{i=1}^{n}(y_i-\overline{y})^2 & \sum_{i=1}^{n}(y_i-\overline{y})((y_i-\overline{y})^2 - \frac{1}{n}\sum_{i=1}^{n}(y_i-\overline{y})^2) \\
\sum_{i=1}^{n}(y_i-\overline{y})((y_i-\overline{y})^2 - \frac{1}{n}\sum_{i=1}^{n}(y_i-\overline{y})^2) & \sum_{i=1}^{n}((y_i-\overline{y})^2 - \frac{1}{n}\sum_{i=1}^{n}(y_i-\overline{y})^2)^2
\end{pmatrix} \\
&= \frac{1}{n}\begin{pmatrix}
\sum_{i=1}^{n}s_i & \sum_{i=1}^{n}\sqrt{s_i}(s_i-\overline{s}) \\
\sum_{i=1}^{n}\sqrt{s_i}(s_i-\overline{s}) & \sum_{i=1}^{n}(s_i-\overline{s})^2
\end{pmatrix}
\end{aligned}
$$

Downloaded from https://academic.oup.com/jrsssc/advance-article/doi/10.1093/jrsssc/qlae077/7942358 by UiT The Arctic University of Norway user on 06 January 2025

Taken together we now obtain

$$
\begin{aligned}
c_{\text{quad}}(t,\mu,\Sigma) &= (t-\mu)\Sigma^{-1}(t-\mu)^{\top}\\
&= \left(\sum_{i=1}^{n} x_i y_i - n\overline{xy},\ \sum_{i=1}^{n} x_i s_i - n\overline{xs}\right)\Sigma^{-1}\left(\sum_{i=1}^{n} x_i y_i - n\overline{xy},\ \sum_{i=1}^{n} x_i s_i - n\overline{xs}\right)^{\top}\\
&= \left(\sum_{i=1}^{n} x_i y_i - n\overline{xy},\ \sum_{i=1}^{n} x_i s_i - n\overline{xs}\right)\left(\frac{n}{n-1}V\left(\sum_{i=1}^{n} x_i^2 - n\bar{x}^2\right)\right)^{-1}\\
&\quad\times\left(\sum_{i=1}^{n} x_i y_i - n\overline{xy},\ \sum_{i=1}^{n} x_i s_i - n\overline{xs}\right)^{\top}\\
&= \left(\sum_{i=1}^{n} x_i y_i - n\overline{xy},\ \sum_{i=1}^{n} x_i s_i - n\overline{xs}\right)\underbrace{\left(\frac{n}{n-1}\left(\sum_{i=1}^{n} x_i^2 - n\bar{x}^2\right)\right)^{-1}}_{:=a}V^{-1}\\
&\quad\times\left(\sum_{i=1}^{n} x_i y_i - n\overline{xy},\ \sum_{i=1}^{n} x_i s_i - n\overline{xs}\right)^{\top}\\
&= \left(\sum_{i=1}^{n} x_i y_i - n\overline{xy},\ \sum_{i=1}^{n} x_i s_i - n\overline{xs}\right)a\underbrace{\frac{1}{\sum_{i=1}^{n} s_i\sum_{i=1}^{n} x_i^2(s_i-\bar{s})^2 - \left(\sum_{i=1}^{n} x_i^2\sqrt{s_i}(s_i-\bar{s})\right)^2}}_{:=b}\\
&\quad\times n\begin{pmatrix} \sum_{i=1}^{n}(s_i-\bar{s})^2 & -\sum_{i=1}^{n}\sqrt{s_i}(s_i-\bar{s})\\ -\sum_{i=1}^{n}\sqrt{s_i}(s_i-\bar{s}) & \sum_{i=1}^{n} s_i\end{pmatrix}\left(\sum_{i=1}^{n} x_i y_i - n\overline{xy},\ \sum_{i=1}^{n} x_i s_i - n\overline{xs}\right)^{\top}\\
&= a\cdot b\cdot n\left(\sum_{i=1}^{n} x_i y_i - n\overline{xy},\ \sum_{i=1}^{n} x_i s_i - n\overline{xs}\right)\begin{pmatrix} \sum_{i=1}^{n}(s_i-\bar{s})^2 & -\sum_{i=1}^{n}\sqrt{s_i}(s_i-\bar{s})\\ -\sum_{i=1}^{n}\sqrt{s_i}(s_i-\bar{s}) & \sum_{i=1}^{n} s_i\end{pmatrix}\\
&\quad\times\left(\sum_{i=1}^{n} x_i y_i - n\overline{xy},\ \sum_{i=1}^{n} x_i s_i - n\overline{xs}\right)^{\top}\\
&= a\cdot b\cdot n\begin{pmatrix} \left(\sum_{i=1}^{n} x_i y_i - n\overline{xy}\right)\sum_{i=1}^{n}(s_i-\bar{s})^2 - \left(\sum_{i=1}^{n} x_i s_i - n\overline{xs}\right)\sum_{i=1}^{n}\sqrt{s_i}(s_i-\bar{s}),\\ -\left(\sum_{i=1}^{n} x_i y_i - n\overline{xy}\right)\sum_{i=1}^{n}\sqrt{s_i}(s_i-\bar{s}) + \left(\sum_{i=1}^{n} x_i s_i - n\overline{xs}\right)\sum_{i=1}^{n} s_i\end{pmatrix}\\
&\quad\times\left(\sum_{i=1}^{n} x_i y_i - n\overline{xy},\ \sum_{i=1}^{n} x_i s_i - n\overline{xs}\right)^{\top}\\
&= a\cdot b\cdot n\\
&\quad\times\begin{pmatrix} \underbrace{\sum_{i=1}^{n} x_i y_i\sum_{i=1}^{n}(s_i-\bar{s})^2 - n\overline{xy}\sum_{i=1}^{n}(s_i-\bar{s})^2 - \sum_{i=1}^{n} x_i s_i\sum_{i=1}^{n}\sqrt{s_i}(s_i-\bar{s}) - n\overline{xs}\sum_{i=1}^{n}\sqrt{s_i}(s_i-\bar{s}),}_{:=k}\\ \underbrace{-\sum_{i=1}^{n} x_i y_i\sum_{i=1}^{n}\sqrt{s_i}(s_i-\bar{s}) + n\overline{xy}\sum_{i=1}^{n}\sqrt{s_i}(s_i-\bar{s}) + \sum_{i=1}^{n} x_i s_i\sum_{i=1}^{n} s_i - n\overline{xs}\sum_{i=1}^{n} s_i}_{:=m}\end{pmatrix}\\
&\quad\times\left(\sum_{i=1}^{n} x_i y_i - n\overline{xy},\ \sum_{i=1}^{n} x_i s_i - n\overline{xs}\right)^{\top}\\
&= a\cdot b\cdot n\cdot(k,m)\cdot\left(\sum_{i=1}^{n} x_i y_i - n\overline{xy},\ \sum_{i=1}^{n} x_i s_i - n\overline{xs}\right)^{\top}\\
&= a\cdot b\cdot n\cdot\left(k\left(\sum_{i=1}^{n} x_i y_i - n\overline{xy}\right) + m\left(\sum_{i=1}^{n} x_i s_i - n\overline{xs}\right)\right)
\end{aligned}
$$

Downloaded from https://academic.oup.com/jrsssc/advance-article/doi/10.1093/jrsssc/qlae077/7942358 by UiT The Arctic University of Norway user on 06 January 2025

A.1.1 DistTree

Since score functions are used in the test statistic for DistTree, we will define them first (cf. Fahrmeir et al., 2016):

$$s(\hat{\boldsymbol{\mu}}, y_i) = \frac{y_i - \hat{\mu}}{\hat{\sigma}^2}; \quad s(\hat{\boldsymbol{\sigma}}, y_i) = -\frac{1}{\hat{\sigma}} + \frac{(y_i - \hat{\mu})^2}{\hat{\sigma}^3}.$$

From the maximum-likelihood estimation, it follows that

$$\hat{\mu} = \bar{y}; \quad \hat{\sigma} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})^2} = \sqrt{\bar{s}}.$$

Therefore, we can express the score functions as follows:

$$s(\hat{\boldsymbol{\mu}}, y_i) = \frac{y_i - \bar{y}}{\bar{s}} = \frac{\sqrt{s_i}}{\bar{s}}; \quad s(\hat{\boldsymbol{\sigma}}, y_i) = -\frac{1}{\sqrt{\bar{s}}} + \frac{\overbrace{(y_i - \bar{y})^2}^{=s_i}}{\bar{s}\sqrt{\bar{s}}} = \frac{s_i - \bar{s}}{\bar{s}\sqrt{\bar{s}}}.$$

In DistTree, the statistic

$$\begin{aligned} t &= \operatorname{vec}\left(\sum_{i=1}^{n} g(x_i) s(\hat{\boldsymbol{\theta}}, y_i)\right) = \sum_{i=1}^{n} x_i (s(\hat{\boldsymbol{\mu}}, y_i), s(\hat{\boldsymbol{\sigma}}, y_i)) \\ &= \left(\sum_{i=1}^{n} x_i s(\hat{\boldsymbol{\mu}}, y_i), \sum_{i=1}^{n} x_i s(\hat{\boldsymbol{\sigma}}, y_i)\right) = \left(\sum_{i=1}^{n} x_i \frac{\sqrt{s_i}}{\bar{s}}, \sum_{i=1}^{n} x_i \frac{s_i - \bar{s}}{\bar{s}\sqrt{\bar{s}}}\right) \end{aligned}$$

has expectation

$$\begin{aligned} \mu &= \operatorname{vec}\left(\sum_{i=1}^{n} g(x_i) \frac{1}{n} \sum_{i=1}^{n} (s(\hat{\boldsymbol{\mu}}, y_i), s(\hat{\boldsymbol{\sigma}}, y_i))\right) = \left(\frac{1}{n}\sum_{i=1}^{n} x_i \sum_{i=1}^{n} \frac{\sqrt{s_i}}{\bar{s}}, \frac{1}{n}\sum_{i=1}^{n} x_i \sum_{i=1}^{n} \frac{s_i - \bar{s}}{\bar{s}\sqrt{\bar{s}}}\right) \\ &= \left(\bar{x}\sum_{i=1}^{n} \frac{\sqrt{s_i}}{\bar{s}}, \bar{x}\sum_{i=1}^{n} \frac{s_i - \bar{s}}{\bar{s}\sqrt{\bar{s}}}\right). \end{aligned}$$

Downloaded from https://academic.oup.com/jrsssc/advance-article/doi/10.1093/jrsssc/qlae077/7942358 by UiT The Arctic University of Norway user on 06 January 2025

The covariance Σ is defined similarly to CTreeTrafo (see Equation (A2)), where V is defined as follows:

$$
\begin{aligned}
V &= \frac{1}{n}\sum_{i=1}^{n}\left(\left(\frac{y_i-\overline{y}}{\overline{s}},\ \frac{s_i-\overline{s}}{\overline{s}\sqrt{\overline{s}}}\right)-\frac{1}{n}\sum_{i=1}^{n}\left(\frac{y_i-\overline{y}}{\overline{s}},\ \frac{s_i-\overline{s}}{\overline{s}\sqrt{\overline{s}}}\right)\right)\\
&\quad\times\left(\left(\frac{y_i-\overline{y}}{\overline{s}},\ \frac{s_i-\overline{s}}{\overline{s}\sqrt{\overline{s}}}\right)-\frac{1}{n}\sum_{i=1}^{n}\left(\frac{y_i-\overline{y}}{\overline{s}},\ \frac{s_i-\overline{s}}{\overline{s}\sqrt{\overline{s}}}\right)\right)^{\top}\\
&= \frac{1}{n}\sum_{i=1}^{n}\left(\frac{1}{\overline{s}}\left(y_i-\overline{y}-\frac{1}{n}\sum_{i=1}^{n}y_i+\frac{1}{n}\sum_{i=1}^{n}\overline{y}\right),\ \frac{1}{\overline{s}\sqrt{\overline{s}}}\left(s_i-\overline{s}-\frac{1}{n}\sum_{i=1}^{n}s_i+\frac{1}{n}\sum_{i=1}^{n}\overline{s}\right)\right)\\
&\quad\times\left(\frac{1}{\overline{s}}\left(y_i-\overline{y}-\frac{1}{n}\sum_{i=1}^{n}y_i+\frac{1}{n}\sum_{i=1}^{n}\overline{y}\right),\ \frac{1}{\overline{s}\sqrt{\overline{s}}}\left(s_i-\overline{s}-\frac{1}{n}\sum_{i=1}^{n}s_i+\frac{1}{n}\sum_{i=1}^{n}\overline{s}\right)\right)^{\top}\\
&= \frac{1}{n}\sum_{i=1}^{n}\left(\frac{1}{\overline{s}}(y_i-\overline{y}-\overline{y}+\overline{y}),\ \frac{1}{\overline{s}\sqrt{\overline{s}}}(s_i-\overline{s}-\overline{s}+\overline{s})\right)\\
&\quad\times\left(\frac{1}{\overline{s}}(y_i-\overline{y}-\overline{y}+\overline{y}),\ \frac{1}{\overline{s}\sqrt{\overline{s}}}(s_i-\overline{s}-\overline{s}+\overline{s})\right)^{\top}\\
&= \frac{1}{n}\sum_{i=1}^{n}\left(\frac{1}{\overline{s}}(y_i-\overline{y}),\ \frac{1}{\overline{s}\sqrt{\overline{s}}}(s_i-\overline{s})\right)\left(\frac{1}{\overline{s}}(y_i-\overline{y}),\ \frac{1}{\overline{s}\sqrt{\overline{s}}}(s_i-\overline{s})\right)^{\top}\\
&= \frac{1}{n}\begin{pmatrix}\frac{1}{\overline{s}^2}\sum_{i=1}^{n}(y_i-\overline{y})^2 & \frac{1}{\overline{s}^2\sqrt{\overline{s}}}\sum_{i=1}^{n}(y_i-\overline{y})(s_i-\overline{s})\\ \frac{1}{\overline{s}^2\sqrt{\overline{s}}}\sum_{i=1}^{n}(y_i-\overline{y})(s_i-\overline{s}) & \frac{1}{\overline{s}^2\overline{s}}\sum_{i=1}^{n}(s_i-\overline{s})^2\end{pmatrix}\\
&= \frac{1}{n}\frac{1}{\overline{s}^2}\begin{pmatrix}\sum_{i=1}^{n}s_i & \frac{1}{\sqrt{\overline{s}}}\sum_{i=1}^{n}\sqrt{s_i}(s_i-\overline{s})\\ \frac{1}{\sqrt{\overline{s}}}\sum_{i=1}^{n}\sqrt{s_i}(s_i-\overline{s}) & \frac{1}{\overline{s}}\sum_{i=1}^{n}(s_i-\overline{s})^2\end{pmatrix}
\end{aligned}
$$

Downloaded from https://academic.oup.com/jrsssc/advance-article/doi/10.1093/jrsssc/qlae077/7942358 by UiT The Arctic University of Norway user on 06 January 2025

Combined we get

$$
\begin{aligned}
c_{\text{quad}}(t,\mu,\Sigma) &= (t-\mu)\Sigma^{-1}(t-\mu)^{\top} \\
&= \left(\frac{1}{\bar{s}}\left(\sum_{i=1}^{n} x_i y_i - n\overline{xy}\right), \frac{1}{\bar{s}\sqrt{\bar{s}}}\left(\sum_{i=1}^{n} x_i s_i - n\overline{xs}\right)\right)\Sigma^{-1} \\
&\quad \times \left(\frac{1}{\bar{s}}\left(\sum_{i=1}^{n} x_i y_i - n\overline{xy}\right), \frac{1}{\bar{s}\sqrt{\bar{s}}}\left(\sum_{i=1}^{n} x_i s_i - n\overline{xs}\right)\right)^{\top} \\
&= \left(\frac{1}{\bar{s}}\left(\sum_{i=1}^{n} x_i y_i - n\overline{xy}\right), \frac{1}{\bar{s}\sqrt{\bar{s}}}\left(\sum_{i=1}^{n} x_i s_i - n\overline{xs}\right)\right)\left(\frac{n}{n-1}V\left(\sum_{i=1}^{n} -n\bar{x}^2\right)\right)^{-1} \\
&\quad \times \left(\frac{1}{\bar{s}}\left(\sum_{i=1}^{n} x_i y_i - n\overline{xy}\right), \frac{1}{\bar{s}\sqrt{\bar{s}}}\left(\sum_{i=1}^{n} x_i s_i - n\overline{xs}\right)\right)^{\top} \\
&= \left(\frac{1}{\bar{s}}\left(\sum_{i=1}^{n} x_i y_i - n\overline{xy}\right), \frac{1}{\bar{s}\sqrt{\bar{s}}}\left(\sum_{i=1}^{n} x_i s_i - n\overline{xs}\right)\right)\underbrace{\left(\frac{n}{n-1}\left(\sum_{i=1}^{n} -n\bar{x}^2\right)\right)^{-1}}_{:=a} V^{-1} \\
&\quad \times \left(\frac{1}{\bar{s}}\left(\sum_{i=1}^{n} x_i y_i - n\overline{xy}\right), \frac{1}{\bar{s}\sqrt{\bar{s}}}\left(\sum_{i=1}^{n} x_i s_i - n\overline{xs}\right)\right)^{\top} \\
&= \left(\frac{1}{\bar{s}}\left(\sum_{i=1}^{n} x_i y_i - n\overline{xy}\right), \frac{1}{\bar{s}\sqrt{\bar{s}}}\left(\sum_{i=1}^{n} x_i s_i - n\overline{xs}\right)\right)a \\
&\quad \times \frac{1}{\underbrace{\sum_{i=1}^{n} s_i \frac{1}{\bar{s}} \sum_{i=1}^{n} (s_i-\bar{s})^2 - \left(\frac{1}{\sqrt{\bar{s}}}\sum_{i=1}^{n}\sqrt{s_i}(s_i-\bar{s})\right)^2}_{(\star_1)}} n\cdot\bar{s}^2 \\
&\quad \times \begin{pmatrix} \frac{1}{\bar{s}}\sum_{i=1}^{n}(s_i-\bar{s})^2 & \frac{1}{\sqrt{\bar{s}}}-\sum_{i=1}^{n}\sqrt{s_i}(s_i-\bar{s}) \\ \frac{1}{\sqrt{\bar{s}}}-\sum_{i=1}^{n}\sqrt{s_i}(s_i-\bar{s}) & \sum_{i=1}^{n} s_i \end{pmatrix}\left(\frac{1}{\bar{s}}\left(\sum_{i=1}^{n} x_i y_i - n\overline{xy}\right), \frac{1}{\bar{s}\sqrt{\bar{s}}}\left(\sum_{i=1}^{n} x_i s_i - n\overline{xs}\right)\right)^{\top} \\
&= a\cdot b\cdot n\cdot\bar{s}^3\left(\frac{1}{\bar{s}}\left(\sum_{i=1}^{n} x_i y_i - n\overline{xy}\right), \frac{1}{\bar{s}\sqrt{\bar{s}}}\left(\sum_{i=1}^{n} x_i s_i - n\overline{xs}\right)\right) \\
&\quad \times \begin{pmatrix} \frac{1}{\bar{s}}\sum_{i=1}^{n}(s_i-\bar{s})^2 & \frac{1}{\sqrt{\bar{s}}}-\sum_{i=1}^{n}\sqrt{s_i}(s_i-\bar{s}) \\ \frac{1}{\sqrt{\bar{s}}}-\sum_{i=1}^{n}\sqrt{s_i}(s_i-\bar{s}) & \sum_{i=1}^{n} s_i \end{pmatrix}\left(\frac{1}{\bar{s}}\left(\sum_{i=1}^{n} x_i y_i - n\overline{xy}\right), \frac{1}{\bar{s}\sqrt{\bar{s}}}\left(\sum_{i=1}^{n} x_i s_i - n\overline{xs}\right)\right)^{\top} \\
&= a\cdot b\cdot n\cdot\bar{s}^3 \underbrace{\begin{pmatrix} \frac{1}{\bar{s}}\left(\sum_{i=1}^{n} x_i y_i - n\overline{xy}\right)\frac{1}{\bar{s}}\sum_{i=1}^{n}(s_i-\bar{s})^2 - \frac{1}{\bar{s}\sqrt{\bar{s}}}\left(\sum_{i=1}^{n} x_i s_i - n\overline{xs}\right)\frac{1}{\sqrt{\bar{s}}}\sum_{i=1}^{n}\sqrt{s_i}(s_i-\bar{s}), \\ -\frac{1}{\bar{s}}\left(\sum_{i=1}^{n} x_i y_i - n\overline{xy}\right)\frac{1}{\sqrt{\bar{s}}}\sum_{i=1}^{n}\sqrt{s_i}(s_i-\bar{s}) + \frac{1}{\bar{s}\sqrt{\bar{s}}}\left(\sum_{i=1}^{n} x_i s_i - n\overline{xs}\right)\sum_{i=1}^{n} s_i \end{pmatrix}}_{(\star_2)} \\
&\quad \times \left(\frac{1}{\bar{s}}\left(\sum_{i=1}^{n} x_i y_i - n\overline{xy}\right), \frac{1}{\bar{s}\sqrt{\bar{s}}}\left(\sum_{i=1}^{n} x_i s_i - n\overline{xs}\right)\right)^{\top} \\
&= a\cdot b\cdot n\cdot\bar{s}^3\cdot\left(\frac{1}{\bar{s}^2}k, \frac{1}{\bar{s}\sqrt{\bar{s}}}m\right)\cdot\left(\frac{1}{\bar{s}}\left(\sum_{i=1}^{n} x_i y_i - n\overline{xy}\right), \frac{1}{\bar{s}\sqrt{\bar{s}}}\left(\sum_{i=1}^{n} x_i s_i - n\overline{xs}\right)\right)^{\top} \\
&= a\cdot b\cdot n\cdot\bar{s}^3\cdot\left(\frac{1}{\bar{s}^2}k\frac{1}{\bar{s}}\left(\sum_{i=1}^{n} x_i y_i - n\overline{xy}\right) + \frac{1}{\bar{s}\sqrt{\bar{s}}}m\frac{1}{\bar{s}\sqrt{\bar{s}}}\left(\sum_{i=1}^{n} x_i s_i - n\overline{xs}\right)\right) \\
&= a\cdot b\cdot n\cdot\bar{s}^3\cdot\left(\frac{1}{\bar{s}^3}k\left(\sum_{i=1}^{n} x_i y_i - n\overline{xy}\right) + \frac{1}{\bar{s}^3}m\left(\sum_{i=1}^{n} x_i s_i - n\overline{xs}\right)\right) \\
&= a\cdot b\cdot n\cdot\bar{s}^3\cdot\frac{1}{\bar{s}^3}\cdot\left(k\left(\sum_{i=1}^{n} x_i y_i - n\overline{xy}\right) + m\left(\sum_{i=1}^{n} x_i s_i - n\overline{xs}\right)\right) \\
&= a\cdot b\cdot n\cdot\left(k\left(\sum_{i=1}^{n} x_i y_i - n\overline{xy}\right) + m\left(\sum_{i=1}^{n} x_i s_i - n\overline{xs}\right)\right)
\end{aligned}
$$

Downloaded from https://academic.oup.com/jrsssc/advance-article/doi/10.1093/jrsssc/qlae077/7942358 by UiT The Arctic University of Norway user on 06 January 2025

In the following we resolve the individual components ($\star_1$, $\star_2$) of c_{quad}.

$$\begin{aligned}
\star_1 &= \frac{1}{\frac{1}{\bar{s}}\sum_{i=1}^{n} s_i \sum_{i=1}^{n} (s_i - \bar{s})^2 - \frac{1}{\bar{s}}\left(\sum_{i=1}^{n} \sqrt{s_i}(s_i - \bar{s})\right)^2} \\
&= \frac{1}{\frac{1}{\bar{s}}} \cdot \frac{1}{\underbrace{\sum_{i=1}^{n} s_i \sum_{i=1}^{n} (s_i - \bar{s})^2 - \left(\sum_{i=1}^{n} \sqrt{s_i}(s_i - \bar{s})\right)^2}_{:=b}} \\
&= \bar{s} \cdot b
\end{aligned}$$

$$\begin{aligned}
\star_2 &= \begin{pmatrix} \frac{1}{\bar{s}}\sum_{i=1}^{n} x_i y_i \frac{1}{\bar{s}} \sum_{i=1}^{n} (s_i - \bar{s})^2 - \frac{1}{\bar{s}} n\overline{xy} \frac{1}{\bar{s}} \sum_{i=1}^{n} (s_i - \bar{s})^2 - \frac{1}{\bar{s}\sqrt{\bar{s}}} \sum_{i=1}^{n} x_i s_i \frac{1}{\sqrt{\bar{s}}} \sum_{i=1}^{n} \sqrt{s_i}(s_i - \bar{s}) \\
+ \frac{1}{\bar{s}\sqrt{\bar{s}}} n\overline{xs} \frac{1}{\sqrt{\bar{s}}} \sum_{i=1}^{n} \sqrt{s_i}(s_i - \bar{s}), \\
- \frac{1}{\bar{s}} \sum_{i=1}^{n} x_i y_i \frac{1}{\sqrt{\bar{s}}} \sum_{i=1}^{n} \sqrt{s_i}(s_i - \bar{s}) + \frac{1}{\bar{s}} n\overline{xy} \frac{1}{\sqrt{\bar{s}}} \sum_{i=1}^{n} \sqrt{s_i}(s_i - \bar{s}) + \frac{1}{\bar{s}\sqrt{\bar{s}}} \sum_{i=1}^{n} x_i s_i \sum_{i=1}^{n} s_i - \frac{1}{\bar{s}\sqrt{\bar{s}}} n\overline{xs} \sum_{i=1}^{n} s_i \end{pmatrix} \\
&= \begin{pmatrix} \frac{1}{\bar{s}^2} \sum_{i=1}^{n} x_i y_i \sum_{i=1}^{n} (s_i - \bar{s})^2 - \frac{1}{\bar{s}^2} n\overline{xy} \sum_{i=1}^{n} (s_i - \bar{s})^2 - \frac{1}{\bar{s}^2} \sum_{i=1}^{n} x_i s_i \sum_{i=1}^{n} \sqrt{s_i}(s_i - \bar{s}) + \frac{1}{\bar{s}^2} n\overline{xs} \sum_{i=1}^{n} \sqrt{s_i}(s_i - \bar{s}), \\
- \frac{1}{\bar{s}\sqrt{\bar{s}}} \sum_{i=1}^{n} x_i y_i \sum_{i=1}^{n} \sqrt{s_i}(s_i - \bar{s}) + \frac{1}{\bar{s}\sqrt{\bar{s}}} n\overline{xy} \sum_{i=1}^{n} \sqrt{s_i}(s_i - \bar{s}) + \frac{1}{\bar{s}\sqrt{\bar{s}}} \sum_{i=1}^{n} x_i s_i \sum_{i=1}^{n} s_i - \frac{1}{\bar{s}\sqrt{\bar{s}}} n\overline{xs} \sum_{i=1}^{n} s_i \end{pmatrix} \\
&= \begin{pmatrix} \frac{1}{\bar{s}^2} \left(\underbrace{\sum_{i=1}^{n} x_i y_i \sum_{i=1}^{n} (s_i - \bar{s})^2 - n\overline{xy} \sum_{i=1}^{n} (s_i - \bar{s})^2 - \sum_{i=1}^{n} x_i s_i \sum_{i=1}^{n} \sqrt{s_i}(s_i - \bar{s}) + n\overline{xs} \sum_{i=1}^{n} \sqrt{s_i}(s_i - \bar{s})}_{:=k} \right), \\
\frac{1}{\bar{s}\sqrt{\bar{s}}} \left(\underbrace{- \sum_{i=1}^{n} x_i y_i \sum_{i=1}^{n} \sqrt{s_i}(s_i - \bar{s}) + n\overline{xy} \sum_{i=1}^{n} \sqrt{s_i}(s_i - \bar{s}) + \sum_{i=1}^{n} x_i s_i \sum_{i=1}^{n} s_i - n\overline{xs} \sum_{i=1}^{n} s_i}_{:=m} \right) \end{pmatrix} \\
&= \left(\frac{1}{\bar{s}^2} k, \frac{1}{\bar{s}\sqrt{\bar{s}}} m \right)
\end{aligned}$$

Substituting the above components into the statistics c_{quad} of CTreeTrafo and DistTree, we find that both statistics are

$$a \cdot b \cdot n \cdot \left(k \left(\sum_{i=1}^{n} x_i y_i - n\overline{xy} \right) + m (\sum_{i=1}^{n} x_i s_i - n\overline{xs}) \right).$$

References

Abu-Arafeh A., Jordan H., & Drummond G. (2016). Reporting of method comparison studies: A review of advice, an assessment of current practice, and specific suggestions for future reports. *British Journal of Anaesthesia*, *117*(5), 569–575. https://doi.org/10.1093/bja/aew320

Altman D. G., & Bland J. M. (1983). Measurement in medicine: The analysis of method comparison studies. *Journal of the Royal Statistical Society D*, *32*(3), 307–317. https://doi.org/10.2307/2987937

Bland J. M., & Altman D. G. (1986). Statistical methods for assessing agreement between two methods of clinical measurement. *The Lancet*, *327*(8476), 307–310. https://doi.org/10.1016/S0140-6736(86)90837-8

Bland J. M., & Altman D. G. (1999). Measuring agreement in method comparison studies. *Statistical Methods in Medical Research*, *8*(2), 135–160. https://doi.org/10.1177/096228029900800204

Brage S., Brage N., Franks P. W., Ekelund U., Wong M.-Y., Andersen L. B., Froberg K., & Wareham N. J. (2004). Branched equation modeling of simultaneous accelerometry and heart rate monitoring improves estimate of directly measured physical activity energy expenditure. *Journal of Applied Physiology*, *96*(1), 343–351. https://doi.org/10.1152/japplphysiol.00703.2003

Downloaded from https://academic.oup.com/jrsssc/advance-article/doi/10.1093/jrsssc/qlae077/7942358 by UiT The Arctic University of Norway user on 06 January 2025

Breiman L., Friedman J. H., Olshen R. A., & Stone C. J. (1984). *Classification and regression trees*. Routledge.

Bunce C. (2009). Correlation, agreement, and Bland-Altman analysis: Statistical analysis of method comparison studies. *American Journal of Ophthalmology*, *148*(1), 4–6. https://doi.org/10.1016/j.ajo.2008.09.032

Butte N. F., Ekelund U., & Westerterp K. R. (2012). Assessing physical activity using wearable monitors: Measures of physical activity. *Medicine and Science in Sports and Exercise*, *44*(Suppl 1), S5–S12. https://doi.org/10.1249/MSS.0b013e3182399c0e

Carstensen B. (2010). Comparing methods of measurement: Extending the LoA by regression. *Statistics in Medicine*, *29*(3), 401–410. https://doi.org/10.1002/sim.v29:3

Carstensen B. (2011). *Comparing clinical measurement methods: A practical guide*. (*Vol. 108*). John Wiley & Sons.

Chhapola V., Kanwal S. K., & Brar R. (2015). Reporting standards for Bland-Altman agreement analysis in laboratory research: A cross-sectional survey of current practice. *Annals of Clinical Biochemistry*, *52*(3), 382–386. https://doi.org/10.1177/0004563214553438

Chow S.-C., Shao J., Wang H., & Lokhnygina Y. (2017). *Sample size calculations in clinical research*. Chapman & Hall/CRC.

Fahrmeir L., Heumann C., Künstler R., Pigeot I., & Tutz G. (2016). *Statistik: Der weg zur datenanalyse*. Springer-Verlag.

Francq B. G., & Govaerts B. (2016). How to regress and predict in a Bland–Altman plot? Review and contribution based on tolerance intervals and correlated-errors-in-variables models. *Statistics in Medicine*, *35*(14), 2328–2358. https://doi.org/10.1002/sim.v35.14

Gerke O. (2020). Reporting standards for a Bland-Altman agreement analysis: A review of methodological reviews. *Diagnostics*, *10*(5), 334. https://doi.org/10.3390/diagnostics10050334

Giavarina D. (2015). Understanding Bland Altman analysis. *Biochemia Medica*, *25*(2), 141–151. https://doi.org/10.11613/issn.1846-7482

Haghayegh S., Kang H.-A., Khoshnevis S., Smolensky M. H., & Diller K. R. (2020). A comprehensive guideline for Bland-Altman and intra class correlation calculations to properly compare two methods of measurement and interpret findings. *Physiological Measurement*, *41*(5), 055012. https://doi.org/10.1088/1361-6579/ab86d6

Hanneman S. K. (2008). Design, analysis, and interpretation of method-comparison studies. *AACN Advanced Critical Care*, *19*(2), 223–234. https://doi.org/10.1097/01.AACN.0000318125.41512.a3

Hapfelmeier A., Cecconi M., & Saugel B. (2016). Cardiac output method comparison studies: The relation of the precision of agreement and the precision of method. *Journal of Clinical Monitoring and Computing*, *30*(2), 149–155. https://doi.org/10.1007/s10877-015-9711-x

Henriksen A., Grimsgaard S., Horsch A., Hartvigsen G., & Hopstock L. (2019). Validity of the polar M430 activity monitor in free-living conditions: Validation study. *JMIR Formative Research*, *3*(3), e14438. https://doi.org/10.2196/14438

Henriksen A., Haugen Mikalsen M., Woldaregay A. Z., Muzny M., Hartvigsen G., Hopstock L. A., & Grimsgaard S. (2018). Using fitness trackers and smartwatches to measure physical activity in research: Analysis of consumer wrist-worn wearables. *Journal of Medical Internet Research*, *20*(3), e110. https://doi.org/10.2196/jmir.9157

Henriksen A., Svartdal F., Grimsgaard S., Hartvigsen G., & Hopstock L. A. (2022). Polar vantage and oura physical activity and sleep trackers: Validation and comparison study. *JMIR Formative Research*, *6*(5), e27248. https://doi.org/10.2196/27248

Hothorn T., Hornik K., & Zeileis A. (2006). Unbiased recursive partitioning: A conditional inference framework. *Journal of Computational and Graphical Statistics*, *15*(3), 651–674. https://doi.org/10.1198/106186006X133933

Hothorn T., & Zeileis A. (2015). Partykit: A modular toolkit for recursive partytioning in R. *Journal of Machine Learning Research*, *16*(1), 3905–3909.

Huber W., Kraski T., Haller B., Mair S., Saugel B., Beitz A., Schmid R. M., & Malbrain M. L. N. G. (2014). Room-temperature vs iced saline indicator injection for transpulmonary thermodilution. *Journal of Critical Care*, *29*(6), 1133.e7–1133.e14. https://doi.org/10.1016/j.jcrc.2014.08.005

Hubert L., & Arabie P. (1985). Comparing partitions. *Journal of Classification*, *2*(1), 193–218. https://doi.org/10.1007/BF01908075

Jensen A. L., & Kjelgaard-Hansen M. (2006). Method comparison in the clinical laboratory. *Veterinary Clinical Pathology*, *35*(3), 276–286. https://doi.org/10.1111/vcp.2006.35.issue-3

Möller S., Debrabant B., Halekoh U., Petersen A. K., & Gerke O. (2021). An extension of the Bland-Altman plot for analyzing the agreement of more than two raters. *Diagnostics*, *11*(1), 54. https://doi.org/10.3390/diagnostics11010054

Nawarathna L. S., & Choudhary P. K. (2013). Measuring agreement in method comparison studies with heteroscedastic measurements. *Statistics in Medicine*, *32*(29), 5156–5171. https://doi.org/10.1002/sim.v32.29

Downloaded from https://academic.oup.com/jrsssc/advance-article/doi/10.1093/jrsssc/qlae077/7942358 by UiT The Arctic University of Norway user on 06 January 2025

Nawarathna L. S., & Choudhary P. K. (2015). A heteroscedastic measurement error model for method comparison data with replicate measurements. *Statistics in Medicine*, *34*(7), 1242–1258. https://doi.org/10.1002/sim.v34.7

Rand W. M. (1971). Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical Association*, *66*(336), 846–850. https://doi.org/10.1080/01621459.1971.10482356

Schlosser L., Hothorn T., Stauffer R., & Zeileis A. (2019). Distributional regression forests for probabilistic precipitation forecasting in complex terrain. *The Annals of Applied Statistics*, *13*(3), 1564–1589. https://doi.org/10.1214/19-AOAS1247

Stöckl D., Rodríguez Cabaleiro D., Van Uytfanghe K., & Thienpont L. M. (2004). Interpreting method comparison studies by use of the Bland-Altman plot: Reflecting the importance of sample size by incorporating confidence limits and predefined error limits in the graphic. *Clinical Chemistry*, *50*(11), 2216–2218. https://doi.org/10.1373/clinchem.2004.036095

Strasser H., & Weber C. (1999). On the asymptotic theory of permutation statistics. *Mathematical Methods of Statistics*, *8*, 220–250. https://doi.org/10.57938/ff565ba0-aa64-4fe0-a158-86fd331bee78

Taffé P. (2018). Effective plots to assess bias and precision in method comparison. *Statistical Methods in Medical Research*, *27*(6), 1650–1660. https://doi.org/10.1177/0962280216666667

Taffé P. (2020). Assessing bias, precision, and agreement in method comparison studies. *Statistical Methods in Medical Research*, *29*(3), 778–796. https://doi.org/10.1177/0962280219844535

WHO (2022). Fact sheet: Physical activity. https://www.who.int/news-room/fact-sheets/detail/physical-activity. [Accessed 2023-12-22].

Zeileis A., & Hornik K. (2007). Generalized M-fluctuation tests for parameter instability. *Statistica Neerlandica*, *61*(4), 488–508. https://doi.org/10.1111/stan.2007.61.issue-4

Zeileis A., Hothorn T., & Hornik K. (2008). Model-based recursive partitioning. *Journal of Computational and Graphical Statistics*, *17*(2), 492–514. https://doi.org/10.1198/106186008X319331

Downloaded from https://academic.oup.com/jrsssc/advance-article/doi/10.1093/jrsssc/qlae077/7942358 by UiT The Arctic University of Norway user on 06 January 2025