G-TWYST
GM PLANTS TWO YEAR SAFETY TESTING

# G-TwYST

# Harmonisation of statistical methods for use of omics data in food safety assessment

Jasper Engel & Hilko van der Voet

G-TwYST (EU grant agreement no: 632165)

Biometris report 41.05.18 May 2018
https://doi.org/10.18174/455159

WAGENINGEN
UNIVERSITY & RESEARCH

# G-TwYST

# Harmonisation of statistical methods for use of omics data in food safety assessment

Jasper Engel & Hilko van der Voet

Biometris, one of the largest groups of quantitative scientists in North-Western Europe, develops statistical and mathematical methods for the quantification of biological processes and processes in our living environment. These methods are applied and validated in practice and are often available as software packages. In addition, we provide education at the undergraduate, Master, PhD, and PostDoc levels, as well as training and consultancy for industry. We cover a wide range of application areas, from gene to ecosystem and from product to production chain. Our goal is to bring quantitative methods to life.

Biometris is the full integration of the chair groups Applied Mathematics (Molenaar) and Applied Statistics (Van Eeuwijk) with the Wageningen Plant Research business unit Biometris (Wehrens).

For more information please visit the website www.biometris.nl or contact:

Biometris, Wageningen University & Research
P.O. Box 16
6700 AA Wageningen, The Netherlands

Visiting address:
Buildingnumber 107
Droevendaalsesteeg 1, 6708 PD Wageningen, The Netherlands

Phone: +31 317 480798 or +31 317 486001
E-mail: biometris@wur.nl

May 2018

# Abstract

The G-TwYST project mainly focused on the statistical analysis of results from animal studies where a limited number of variables based on OECD guidance was measured. In the statistical analysis of these measurements, emphasis was placed on the possibilities to test the equivalence of the GM groups relative to historical non-GM groups from the GRACE study. A new univariate statistical approach was proposed for this purpose.

In G-TwYST, only limited attention was paid to food safety assessment using the plant material that was used as feed in the animal studies, even though such material can be easily obtained. Especially high-throughput untargeted omics measurements (e.g. metabolomics or transcriptomics) of this material could be an excellent approach for checks on unintended effects due to the larger number of compounds (e.g. metabolites, transcripts) that are simultaneously measured. In G-TwYST a limited number of samples from the maize harvests has been analysed by transcriptomics and metabolomics. Subsequently, an already available multivariate one-class model was used to compare the G-TwYST samples to a reference set of measurements, as had already been done before in the GRACE project.

The pilot studies in GRACE and G-TwYST highlight the potential of omics-based for food safety assessment. However, there is a need to study the statistical properties of the applied multivariate one-class model and among a range of alternative models it is unclear whether it is the most appropriate method for omics-based food safety assessment. Additionally, the applied approach is distinctly different from the univariate methodology that was developed in G-TwYST for animal studies. There is no reason why statistical methodology should be different for similar data from plant or animal studies. From a regulator’s point of view the statistical criteria for evaluating food safety data should as much as possible be the same. However, statistical methods will need to be different for highly multivariate omics data and for the more traditional univariate measurements.

The main purpose of this short report is therefore to identify the differences between statistical approaches for food safety assessment as were applied in the context of the G-TwYST project, and to suggest directions for future research to improve the harmonisation of statistical methodology for analysing omics data.

# 1 Introduction

Food safety assessment requires the specification of a set of measured variables or for derived statistics representing adverse outcomes. In addition, limit values are needed to establish minimal regions of safety. Ideally, limit values discriminating safe from (toxicologically) unsafe values would be based on thorough toxicological experiments. However, such values are not commonly available. A more feasible approach is the specification of equivalence regions, where values inside the limits are interpreted as not giving any concern, but without giving any verdict about values outside the limits. Typically, one would expect the true, but unknown, toxicity limits to be outside the specified equivalence limits. However, experts still find it difficult to establish such equivalence limits. For animal studies, a recent paper proposed 'targeted effect sizes' for 9 variables, although the total set of variables was around 140 (Hong et al. 2017). For plant omics data, which will be discussed below, we are not aware of established equivalence limits for the omics variables based on expert opinions.

It could be thought that the need to establish equivalence limits for food safety assessment is far more complex than the classical approach based on difference testing. It should be noted, however, that classical difference tests are always required to be used after an a priori power analysis to establish that relevant targeted effect sizes can be determined with sufficient statistical power. Therefore, the need to set limit values is not different for classical difference and equivalence tests, and the two approaches may have more in common than is usually thought.

In the absence of (sufficient) external equivalence limits, the measurements on the GMOs have to be compared to historical data. Such data are typically available both for animal and plant studies, although the organisation of such data is almost always sub-optimal for direct use. In G-TwYST a univariate equivalence testing method was proposed for animal feeding studies, which allows to use historical reference data to replace externally set equivalence limits (van der Voet et al. 2017). The method is based on well-developed principles for univariate equivalence testing in the field of medicinal equivalence (FDA, EMA). Plant studies typically allow to include a higher number of experimental units than animal studies. Consequently, it may be possible to include moderate amount of reference groups in the study. Univariate statistical tests that use these internal references (rather than historical data) to set equivalence limits have been proposed as well (van der Voet et al. 2011, Kang and Vahl 2014). The use of multiple internal references is not feasible in animal experiments.

So far, the univariate equivalence testing procedures proposed in G-TwYST have been applied to classical measurements from animal studies (e.g. based on OECD guidance) (Goedhart & van der Voet 2017, 2018abcdef). Typically, as shown in Table 1, the dimensionality of the acquired data is moderate to large (e.g. 50 measurements). Nowadays, using omics technologies it is possible to characterise a sample in great molecular detail. In other words, using e.g. transcriptomics or metabolomics it is possible to simultaneously measure the expression of thousands to tens of thousands of genes or the abundances of hundreds to thousands of metabolites in a single sample. Therefore, the use of omics holds great promise for identification of unintended effects in comparison to classical measurements. In the pilot study carried out in the GRACE project it was shown how an already available one-class multivariate statistical model (van Dijk et al. 2014) could be used to compare the GM samples to a set of references on the basis of the acquired omics data (Kok et al., submitted; Corujo et al., submitted). However, the statistical properties of the multivariate method applied have not been studied closely and it is currently unclear whether it is, in comparison to alternative methods, the most appropriate procedure for food safety assessment. Additionally, the method is distinctly different from the univariate method developed so far for food safety assessment, while, from a regulator’s perspective, harmonization of statistical methodology for food safety is to be preferred.

**Table 1. Comparison of statistical characteristics applied in GRACE and/or G-TwYST for animal feeding tests and plant omics tests**

| G-TwYST approach | Equivalence testing | One-class model |
|---|---|---|
| Description | univariate (variable-wise), equivalence tests | multivariate, PCA-based one-class classification model |
| References | van der Voet et al. (2017),<br>Goedhart & van der Voet (2017, 2018a-f) | van Dijk et al. (2014),<br>Kok et al. (subm.),<br>Corujo et al. (subm.) |
| Type of study | Animal | Plant |
| Type of measurements | classical (body and organ weights, clinical chemistry, haematology, etc.) | omics (transcriptomics NGS, metabolomics LCMS, etc.) |
| Dimensionality | moderate to large (e.g. 50) | very large (e.g. 10,000) |
| Biological interpretation of variables | yes | partly |
| Number of samples GM | small (e.g. 8) | small (e.g. 8) |
| Within-study reference groups | none / very few (e.g. 2) | none / very few / some (e.g. 10) |
| External effect sizes available | no / hardly | no |
| Historical database | available / possible | available / possible |
| Correction for between-study variation | yes | no |
| Additional statistical methods | univariate, classical difference tests | - |
| Correction for multiplicity | no | not needed |
| Power defined for conclusion ‘Food is safe’ | yes | no |

Table 1 summarizes the main characteristics of data obtained from classical animal studies or modern omics data of plant material. As already mentioned above, the most striking difference is a much higher dimensionality of a typical omics data set compared to the data from animal studies. However, in principle, the data can be analysed by applying the same univariate methodology that is also used in animal studies to each variable. Crucial in this respect is that the multiplicity of the test results needs to be properly addressed. This is already somewhat the case in animal studies where, as shown in Table 1, tests are also applied to multiple variables. For univariate tests in G-TwYST, multiplicity was only addressed by tabulating the proportions of significant test results and comparing these to the nominal test level (usually $\alpha = 0.05$). The argument is that rates similar to the nominal level can be expected even under the null hypothesis, and therefore are not a reason for concern themselves. However, a univariate approach requires univariate limits, and therefore might not directly be applicable to high-dimensional omics data. Typically, not all variables in an omics data set have a direct biological interpretation. For example, in a typical metabolomics data set only a small percentage of the measured variables (20 – 30%) can be putatively annotated. Therefore, setting univariate equivalence limits for each single variable might also not be preferred.

Alternative to univariate analysis is multivariate analysis, where all variables in the data set are taken into account together. An obvious advantage is that the need to address the multiplicity of (univariate) test results is avoided this way. Additionally, more subtle unintended effects that don't manifest themselves in a single variable may potentially be detected. However, multivariate analyses are also more likely to over-fit the data, i.e. identify a spurious pattern in the data (noise) as an unintended effect. Therefore, a larger number of experimental units may be required for multivariate analysis. Also, in the multivariate case the equivalence limits can be specified in different ways (see section 2), further complicating analysis. Additionally, given the low number of experimental units in food safety assessment (see Table 1) some form of regularization (e.g. dimension reduction or shrinkage) needs to be applied to handle the curse of dimensionality (see section 3). Finally, whereas univariate methodology for equivalence testing in food safety assessment has been developed from rigorous statistical principles regarding hypothesis testing, the multivariate one-class model strategy has been developed as a classification tool. For example, there is no defined null hypothesis, but only a decision rule, where based on a threshold GM samples are classified as inside or outside the 'safe class'. Because of this, it is not straightforward to make statements about the statistical power of a proof-of-safety test.

The main purpose of this short report is to identify the differences between the currently available statistical approaches, and to suggest directions for future research to improve the harmonisation of statistical methodology. Related models that have been introduced in other fields such as industrial process monitoring are selectively reviewed as well. Section 2 discusses variable-wise and multivariate statistical models for multivariate data sets where the number of samples (reference genotypes, $n$) is still larger than the number of variables ($m$). In section 3 models are discussed for high-dimensional data, where $n < m$. In both sections simulations are carried out to assess and compare the statistical power of various methods. In both sections we focus on difference tests for the very pragmatic reason that many multivariate models have not yet been fully elaborated in the context of equivalence testing. Nevertheless, there is a strong link between setting equivalence limits in equivalence tests and the need to evaluate the power of difference tests at pre-set values. Finally, section 4 discusses the differences between the univariate and multivariate statistical tests used in GRACE and G-TwYST for food safety assessment and proposes directions for future research.

# 2 Variable-wise and multivariate difference and equivalence tests for use in food safety assessment

In this section, we selectively review variable-wise and multivariate difference and equivalence tests, where the main purpose is to present different categories of approaches. We focus on tests that can be seen as a multivariate extension of the *t*-test and TOST presented shortly in section 2.1. Applications of the tests to simulated data sets are used to highlight some of their main similarities and differences. In section 2.2, we consider variable-wise (endpoint-wise) application of the *t*-test or TOST. Section 2.3, covers tests based on the sum of test statistics over all endpoints. In section 2.4, fully multivariate approaches taking into account the correlations between the different measurements / endpoints are discussed. The different methods are compared in section 2.5. Although sections 2.2 – 2.5 highlight the potential of multivariate statistical testing in risk assessment for food safety, many issues also remain. For example, it is currently unclear how the equivalence limit should be extended to the multivariate case and several options are available (as highlighted in sections 2.2 – 2.5). Also, some of the methods discussed are not applicable to high-dimensional data where the number of endpoints measured is much larger than the sample size (e.g. omics data), or they perform poorly. These issues amongst others are discussed in section 2.6 and some suggestions for future research are made.

## 2.1 Introduction

In risk assessment studies of food safety the primary objective is to demonstrate equivalence between a test food (e.g. a GMO) and a set of reference varieties that have a history of safe use as foods. Traditionally, equivalence in such comparative studies for some relevant endpoint was established by testing statistically whether the means of the reference ($\mu_{ref}$) and GMO-group ($\mu_{GMO}$) can be assumed to be the same or that they are found to be significantly different [van der Voet et al 2011; Vahl and Kang 2016]. The tested hypotheses for this difference test are:

$$H_0: \mu_{ref} = \mu_{GMO} \text{ vs. } H_A: \mu_{ref} \neq \mu_{GMO} \qquad (1)$$

The two-sample *t*-test is a suitable statistical test for this problem, when the GMO and reference data are normally distributed [Limentani et al 2005]. Often, homogenous variances of the GMO and reference populations are also assumed. Note that this assumption has to be made when only a single GMO is measured. The finding of a statistically significant difference by the *t*-test can be used to argue that more experiments are required to assess the safety of the GMO. Traditionally, non-rejection of the null-hypothesis of ‘no difference’ is taken to mean that there is no relevant difference between the GMO and references for that endpoint and that this endpoint on its own provides no trigger for a further safety evaluation [van der Voet 2011 et al; Vahl and Kang 2016].

Although the *t*-test is an appropriate test for demonstrating a difference, problems arise when it is used to show equivalence [Wellek 2010; Limentani et al 2005]. Most notably, non-significance is not sufficient evidence for the null hypothesis (absence of evidence is not evidence of absence) [Altman and Bland 1995]. In other words, the approach of declaring equivalence between the GMO and references when the null-hypothesis is not rejected is flawed [Van der Voet et al 2011; Vahl and Kang 2016]. Additionally, the *t*-test may declare that a significant difference exists even though its magnitude is of no practical relevance [Limentani et al 2005]. Finally, the *t*-test rewards poor experimental precision and low sample sizes in the sense that it becomes more difficult to detect differences [Limentani et al 2005]. Simply switching the roles of the null and alternative hypotheses in (1) to resolve these issues it not an option, since null hypotheses have to be specified by an equality rather than an inequality [Wellek 2010; Hoffelder et al 2015]. The appropriate statistical approach in this setting is an equivalence test, where the aim is not to show exact equality of the GMO and reference means but rather to show that they are equivalent up to a practical difference [Wellek 2010; Hoffelder et al 2015; van der Voet et al 2011]. This is, for example, the recommended approach for GMO safety evaluation of the European Food Safety Authority (EFSA) [Van der Voet et al 2011; Vahl and Kang 2016]. The equivalence hypotheses are given as:

$$H_0\colon \left|\mu_{ref} - \mu_{GMO}\right| \geq \Delta \text{ vs. } H_A\colon \left|\mu_{ref} - \mu_{GMO}\right| < \Delta \tag{2}$$

where Δ corresponds to an a priori specified equivalence limit. It defines the range of practical differences between the GMO and references for which equivalence is assumed, as specified by a regulatory authority. A univariate statistical test for (2), when the GMO and references are normally distributed, is the *two one-sided t-tests* (*TOST*) [Limentani et al 2005]. Similar to the *t*-test, often equal variances of the GMO and reference distributions are assumed. Significance of the TOST does indeed provide evidence of the similarity of the GMO and references for the tested endpoint [Van der Voet et al 2011; Vahl and Kang 2016]. Currently, experts find it difficult to specify the equivalence limit Δ. In the absence of equivalence limits specified by experts, they have to be determined experimentally. In G-TwYST a method was proposed that allows to use data from previous (historical) studies to replace externally set equivalence limits (expert knowledge) [van der Voet et al 2017]. Tests that use internal references (rather than historical data) to set equivalence limits have been proposed as well [van der Voet et al 2011, Kang and Vahl 2014].

Figure 1 shows the performance of difference and equivalence tests for simulated data, for an example equivalence limit of 3. Here, $\mu_{GMO}$ was varied systematically and its difference to $\mu_{ref}$ is specified along the x-axis of the graph. For the difference test (*t*-test, green curve) it can be seen that its power quickly increases to effectively 100% at $\mu_{GMO} - \mu_{ref} = 5$ . In contrast, the power of the equivalence test (TOST, red curve) increases when the difference between GMO and references becomes smaller. Its power is 82% at an effect size of zero, which is close to the theoretically expected power of 80%. Note that (at the equivalence limit 3) the attained significance level of the equivalence test is close to the nominal level of 5%. This shows that the equivalence test controls the consumer's risk, i.e. the probability of accepting equivalence when the GMO mean relevantly differs from the reference mean (at or beyond the equivalence limit) [Vahl et al 2016]. The nominal significance level of the difference (t-test) of 5% is reproduced when the effect size is zero. The difference testing approach controls the producer's risk of concluding that the GMO and reference means are unequal when they are actually equal [Vahl et al 2016]. This shows that equivalence and difference tests have their own role in risk assessment for food safety.

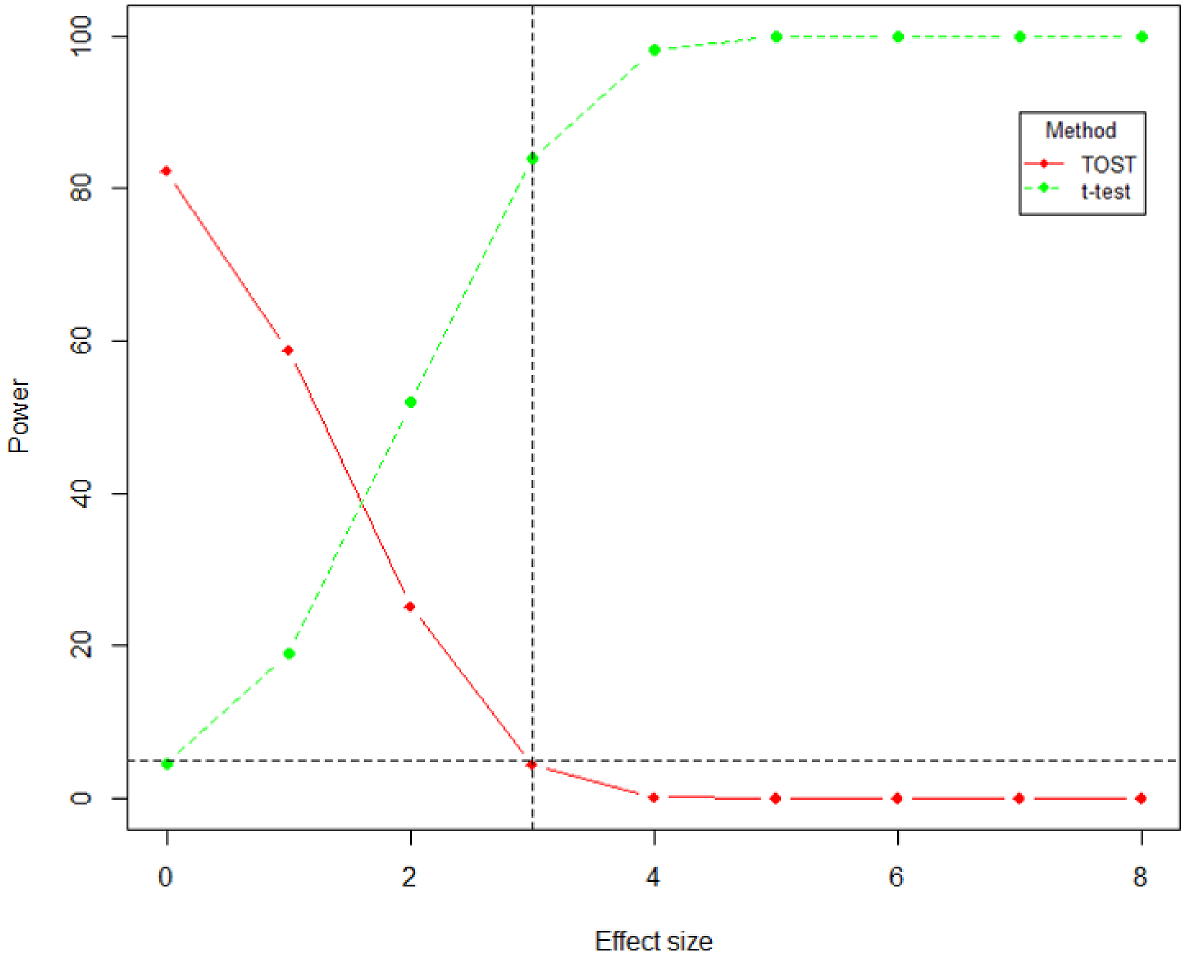

*Figure 1: power versus effect size of TOST (red curve) and t-test (green curve) for simulated data. The vertical dotted line indicates the equivalence limit used by TOST. The power is defined as the percentage of significant outcomes over 1000 simulations. For each simulation 40 reference observations were drawn from a standard normal distribution. A single GMO sample was drawn from the same distribution, and subsequently its mean was shifted according to the value specified on the x-axis in the figure.*

The *t*-test and TOST are one-dimensional, i.e. they can be applied to measurements for one specific endpoint. Typically, however, multiple characteristics are measured to compare the GMO and references. It is of interest to consider multivariate tests that simultaneously take into account all of these characteristics [Wellek 2010; Hoffelder et al 2015]. An obvious advantage is that this allows one to make statements on differences or equivalence between the GMO and the references on the basis of all endpoints simultaneously rather than on an endpoint-by-endpoint basis. Additionally, some of these multivariate tests have the potential to detect subtle unintended effects that don’t manifest themselves in a single variable (and are therefore not considered by the *t*-test) [Saccenti et al 2014]. A wealth of literature on multivariate data analysis is available, especially in the context of difference testing [Kent et al 2006; Friedman et al 2001; James et al 2013; ].

### 2.2 Variable-wise Union-Intersection and Intersection-Union tests

Before we consider analysis of multivariate data, we briefly revisit the univariate case. For the discussion below we remind the reader that rejection of the null hypothesis of no difference at an $\alpha$ significance level by the two-sample *t*-test is the same as observing that the $(100-\alpha)$% confidence interval (CI) of $diff = \mu_{ref} - \mu_{GMO}$ does not contain the value zero [Litamani et al 2005]. Using TOST, equivalence is established at the 5% level if the $(100-2\alpha)$% CI of *diff* is contained within the equivalence interval $(-\Delta, \Delta)$ [Litamani et al 2005; Lakens 2017]. In other words, the difference or equivalence testing can be carried out by visual inspection of appropriate confidence intervals. A few examples are presented in figure 2.

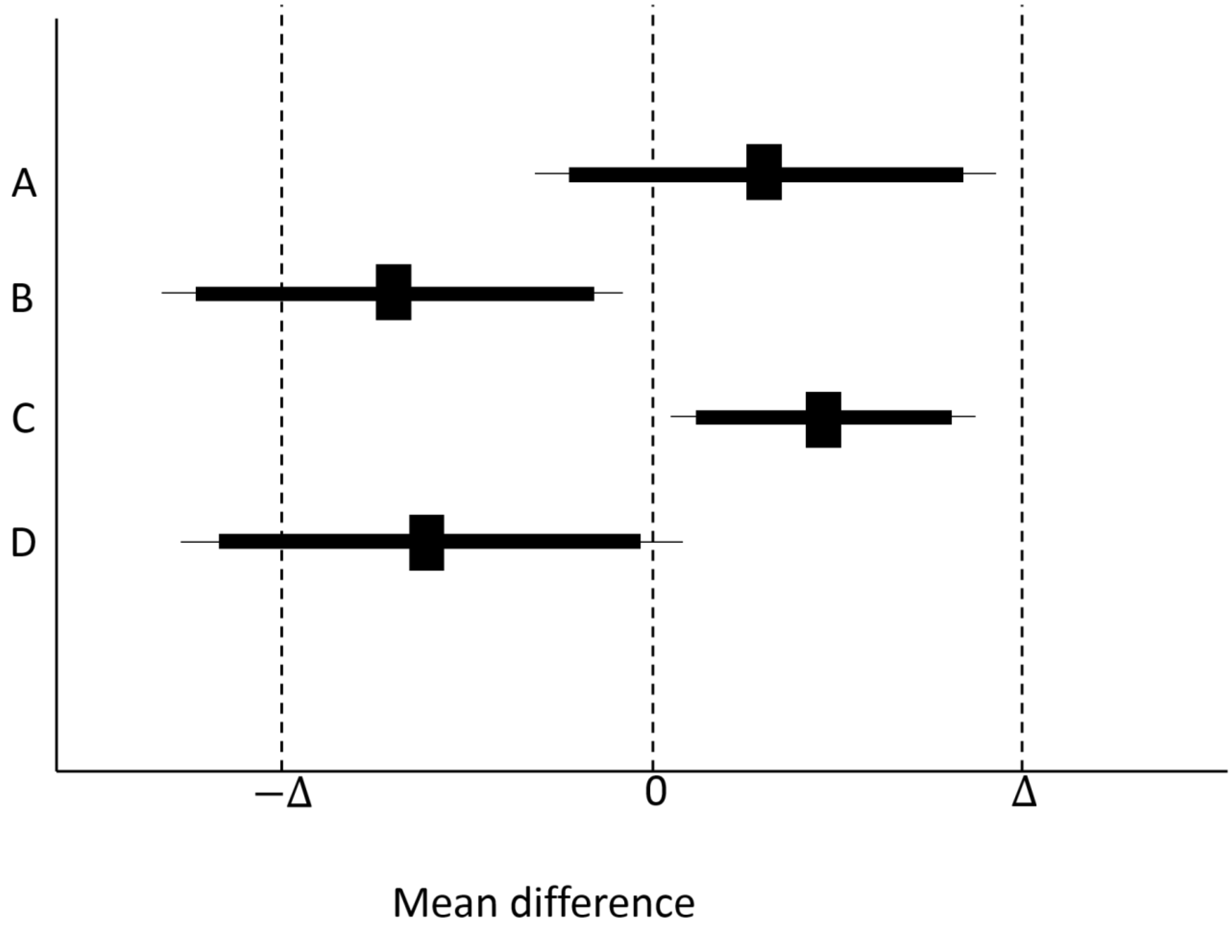

*Figure 2. Mean differences (black squares), 90% confidence intervals (thick horizontal lines) and 95% confidence intervals (thin horizontal lines) for four combinations of test results. The vertical lines indicate the equivalence limits used by TOST. The conclusions would be for case*

*(A) no proof of difference and equivalent, (B) different and no proof of equivalence, (C) different, but equivalent, and (D) no proof of difference, no proof of equivalence. Based on Lakens 2017.*

The tests discussed below essentially extend the confidence interval approach to the multivariate case, with *m* variables. Here, for the *m*–dimensional vector $\boldsymbol{\mu}_{GMO} - \boldsymbol{\mu}_{ref}$ a confidence region (CR) rather than a one-dimensional CI is established (see figure 4 for an example), which is subsequently compared to an equivalence region (ER) for equivalence testing, or the vector **0** for difference testing [Berger 1996; Hoffelder 2015; Munk and Pfluger 1999; Pallmann and Jaki 2017]. The difference and equivalence tests considered in this subsection consist of a combination of the univariate tests applied to each variable. First, we consider difference testing where the two-sample *t*-test is applied to each variable (symbol *i*) to test the following hypothesis:

$$H_0: \quad \mu_{ref,i} = \mu_{GMO,i} \quad \forall i \qquad \text{vs.} \qquad H_A: \quad \exists i: \quad \mu_{ref,i} \neq \mu_{GMO,i} \tag{3}$$

The null hypothesis is rejected if a significant difference is detected by *at least one* of the *t*-tests. In other words, the global null hypothesis of no effect is rejected if there is evidence of a possible difference in at least one of the measured variables / endpoints. Note that null hypothesis (3) is the intersection of the variable-wise null hypotheses (1), while the alternative hypothesis is the union of all variable-wise alternatives. This type of test is known as an Union-Intersection (UI) test. It is well known for UI tests that carrying out each individual test (two-sample *t*-test) at a significance level of $\alpha$ leads to an inflated probability of rejecting the null hypothesis, i.e. the type I error rate for problem (3) is not controlled [Goeman and Solari 2014]. An example of this phenomenon is shown in figure 3. By comparing figures 1 and 3a it can be seen that the multivariate difference test (3) returns the sample results at the two-sample *t*-test when applied to one variable. In this case the size of the test is 5%. By comparing the green curve in panels 3a – d at effect size zero it can be seen that the type I error rate increases with increasing numbers of variables. To alleviate this problem some form of multiplicity correction needs to be applied. Here, we combine variable-wise *t*-test with multiplicity adjustment by the Bonferroni procedure [Goeman and Solari 2014]. As shown in figure 3 (blue curve), now the expected size of the test of 5% is indeed reproduced when the effect size is zero. At the same time, however, the power to detect differences is greatly reduced compared to the univariate case in panel a, and this effect becomes larger with increasing dimension of the data. This can be understood by considering the CR used by this test. As shown in the bivariate example in figure 4a, the CR of the uncorrected test is of rectangular shape and spans all values that fall inside the 95 CIs of *diff* for each variable. Due to the Bonferroni adjustment the width of these 95% CIs is increased resulting in a larger rectangular area spanned by the CR. Because of this, the null hypothesis of no difference is less likely to be rejected when there is a deviation in a single variable in comparison to a univariate test for that variable alone. Another way to see that null hypothesis (3) is less often rejected is noting that a larger square is more likely to contain the point **0**. Note that in some cases the power of the test may seemingly be improved by using other multiplicity correction procedures such as the Benjamini-Hochberg False Discovery Rate (FDR) or Meinshausen's hierarchical hypothesis testing procedure [Goeman and Solari 2014; Meinshausen 2008]. However, the use of FDR in the context of safety assessment has been criticised as being inappropriate because it protects against the wrong type of error [van der Voet 2018].

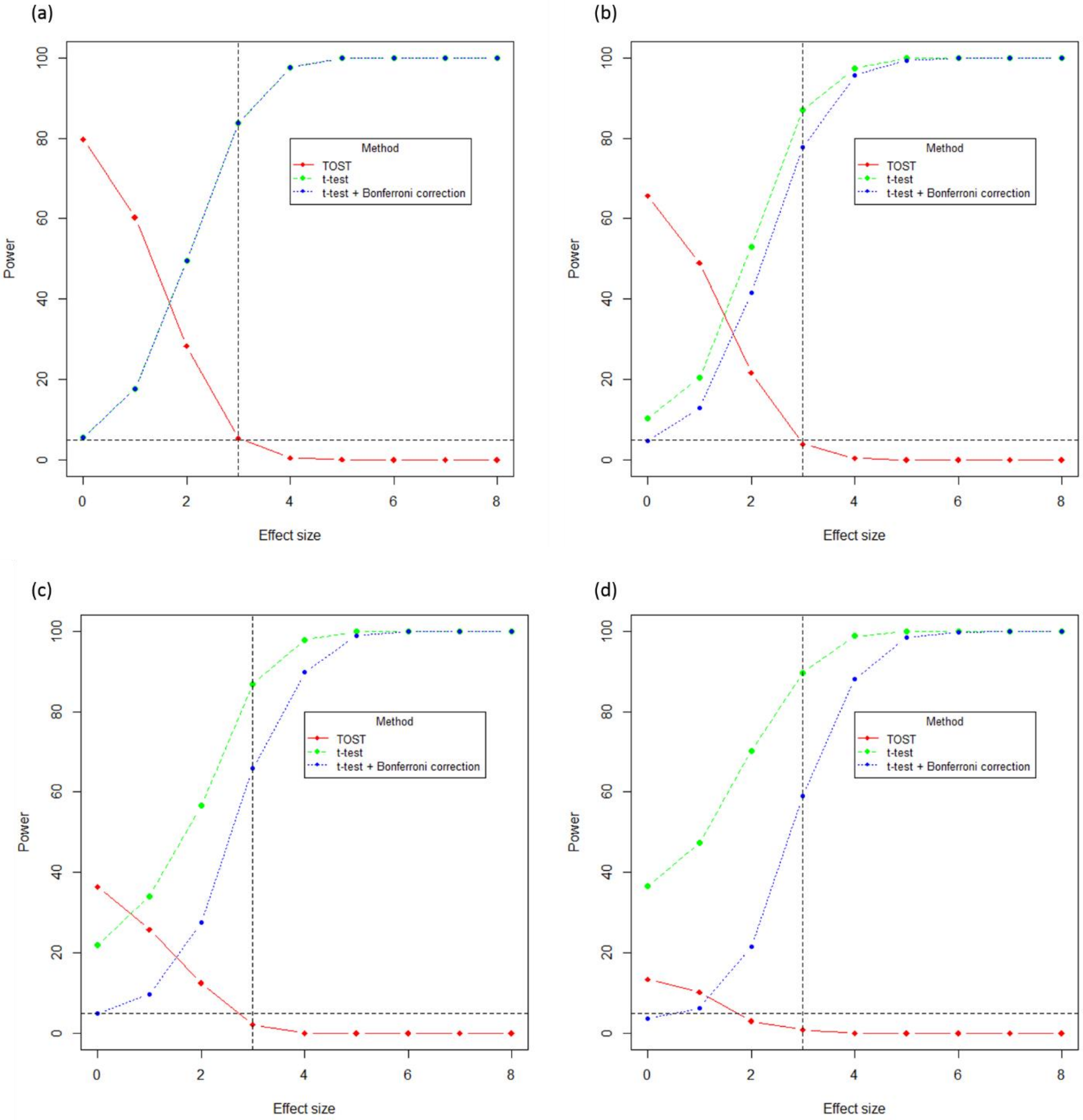

*Figure 3: power versus effect size of the variable-wise TOST (red curve), variable-wise t-test (green curve) and variable-wise t-test with Bonferroni correction (blue curve) for simulated multivariate data with (a) 1, (b) 2, (c) 5 and (d) 10 variables. The vertical dotted line indicates the variable-wise equivalence limit. The power is defined as the percentage of significant outcomes over 1000 simulations. For each simulation 40 reference observations were drawn from a standard multivariate normal distribution. A single GMO sample was drawn from the same distribution where the mean of the first variable was shifted according to the effect size specified on the x-axis in the figure.*

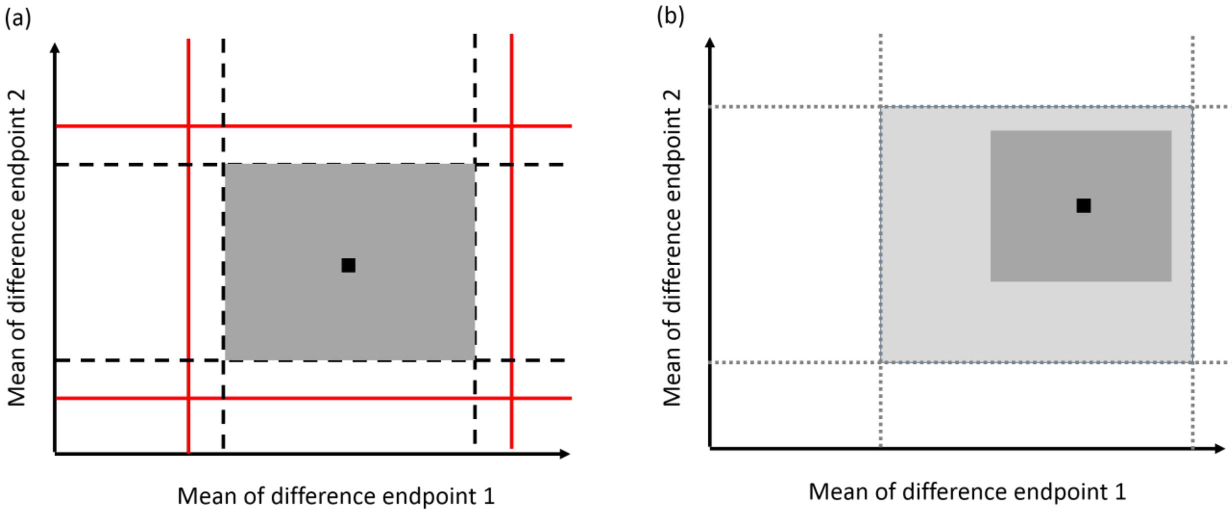
(a)
Mean of difference endpoint 2
Mean of difference endpoint 1
(b)
Mean of difference endpoint 2
Mean of difference endpoint 1

*Figure 4: schematic overview of multivariate confidence regions of rectangular shape for (a) the variable-wise t-test procedure (dark grey area within the black dotted lines lines) and (b) the variable-wise TOST (dark grey area). The dotted lines in panel (a) indicate variable-wise 95% CIs while the red lines correspond to the Bonferroni-adjusted intervals. Note that the Bonferroni-adjusted CR is the rectangle within the red lines. The dotted lines in panel (b) correspond to equivalence limits, which together define an equivalence region of rectangular shape (light grey area).*

For similar reasons as the univariate case, non-rejection of hypothesis (3) is not proof of equivalence between the GMO and reference samples. In this case variable-wise application of TOST seems more appropriate to test the hypothesis:

$$H_0: \quad \exists i: \quad |\mu_{ref,i} - \mu_{GMO,i}| \geq \Delta \quad \text{vs.} \quad H_A: \quad |\mu_{ref,i} - \mu_{GMO,i}| < \Delta \quad \forall i \qquad (4)$$

Note that a different equivalence limit can be specified for each variable, in which case Δ in (4) should be replaced by ($\Delta_i$) [Wellek 2010; Hoffelder et al 2015]. The null hypothesis of the variable-wise TOST is the union of the variable-wise null-hypotheses (2), and the alternative hypothesis is the intersection of the corresponding variable-wise alternatives. This type of test is known as an Intersection-Union (IU) test. The alternative hypothesis defines an equivalence region of rectangular shape (see figure 4b). The null hypothesis of the variable-wise TOST is rejected when for each variable the null-hypothesis (2) is rejected [Pallmann and Jaki 2017]. In other words this global test declares the GMO to be equivalent to the references if equivalence is established in each of the variables / endpoints. As shown in figure 4b, this happens when the rectangular CR (defined by the variable-wise 90% CI's) falls completely inside the equivalence region of rectangular shape [Pallmann and Jaki 2017]. As shown in figure 3, also for the variable-wise TOST and a fixed rectangular equivalence region the power to detect equivalence goes down with increasing dimension of the data. This means that if a certain power, say 80%, is desired for large-dimensional data much higher equivalence limits need to be specified for each endpoint compared to the univariate case shown in figure 1. In contrast to the difference test, it is well known that the attained significance level of the variable-wise TOST never exceeds the level at which the univariate TOSTs are carried out (typically 5%) [Berger and Hsu 1996]. Because of this, multiplicity correction is less of a concern compared to difference testing. However, by looking at the observed power of the test at effect size 3 in figures 3a-d it can be seen that the attained significance level of the test can be much smaller than 5%. In other words, the test is conservative and therefore has less power to declare equivalence. The magnitude of this effect depends on the number of end points that are tested (as seen in figure 3), their pairwise correlation, and the type of unintended effect (e.g. in which variables does the effect occur).

For all simulation results shown so far the observations were drawn from a standard multivariate normal distribution where within-reference correlation between the variables is zero. Figure 5 displays the size of the test when the observations are drawn from a bivariate normal distribution with correlated variables. The correlation between the variables was systematically varied from -1 to 1. For this example it can be seen that the attained significance level was close to 5% when the correlation was close to zero, but approached zero when the variables were almost perfectly (anti)correlated. This example suggests that the size of the variable-wise TOST can possibly be better controlled when the correlation structure of the data is specifically taken into account during modelling. Wellek et al proposed a maximum likelihood estimator of the rectangular CR when the distribution of the references is assumed to be multivariate normal. Essentially, this approaches decreases the size of the rectangular CR such that its coverage probability, given the underlying distribution of the data, is 90% (when an equivalence test at the 5% level is desired) [Wellek 2011]. This might offer an interesting improvement upon the variable-wise TOST. Currently, however, because of numerical reasons this approach is not applicable when the dimension of the data is greater than about 25.

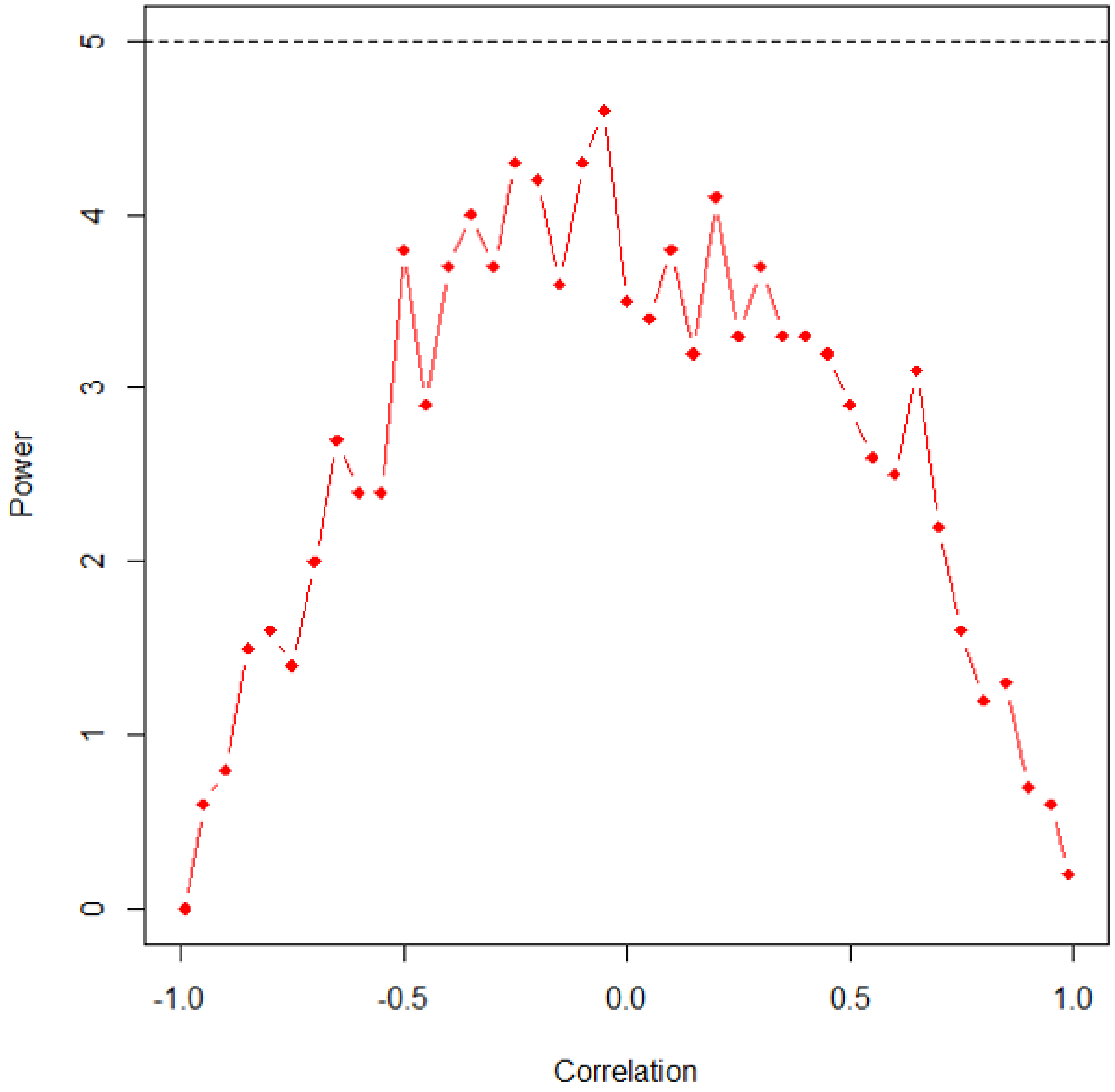

*Figure 5: Test size (probability of finding equivalence when in fact the first endpoint had a difference of 3) versus correlation of the bivariate variable-wise TOST for simulated data. The test size is defined as the percentage of significant outcomes over 1000 simulations. For each simulation 40 reference observations were drawn from a bivariate normal distribution with unit variances. The correlation between the variables was systematically varied from -1 to 1.  A single GMO sample was drawn from the same distribution, but where the mean of the first variable was shifted according to the effect size of 3. For each variable, the equivalence limit was also set to 3.*

### 2.3 Multivariate tests based on Euclidean distance

Although the variable-wise tests discussed above offer a global test for difference and equivalence testing, they only take into account the characteristics of each variable one at a time, resulting in tests based on confidence regions of rectangular shape. Below, we show that other multivariate measures to quantify the discrepancy between the GMO and references are also available, which result in circular or ellipsoidal confidence and equivalence regions [Hoffelder et al 2015; Munck and Pfluger 1999]. The main difference compared to the variable-wise approach is that now a possible difference (or equivalence) between the GMO and references is seen as a combined measure of several variables, i.e. the deviation from equality (the point zero) along each endpoint is taken into account simultaneously.  Because of this, more subtle differences between the references and the GMO may be detected. An example is shown in figure 6 in the context of equivalence testing. The hyper-rectangle indicates the equivalence region of the variable-wise TOST, while the circle and ellipsoid corresponds to equivalence regions of the tests discussed below. Observe that for the rectangle the points 1 - 3 can be considered to be "equally" equivalent. On the other hand, with respect to the circle and ellipsoidal region point 3 is outside the equivalence area.

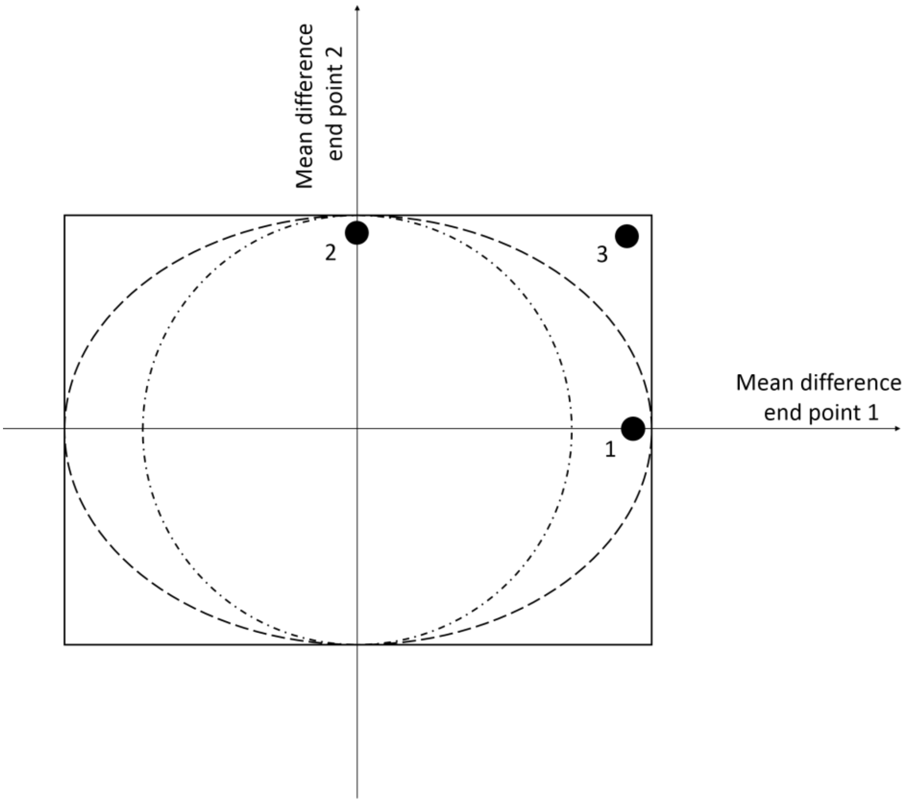

*Figure 6: Bivariate rectangular (solid line), circular (dashed – dotted line) and ellipsoidal (dashed line) regions for bivariate data. Note that the point 2 falls inside all equivalence regions, point 1 falls inside the elliptical and rectangular equivalence region, while point 3 is only marked as equivalence when a rectangular equivalence region is considered. Based on Munk and Pfluger 1999.*

Testing the difference between the multivariate vector of the mean of the references ($\boldsymbol{\mu}_{ref}$) and the mean of the GMO ($\boldsymbol{\mu}_{GMO}$) corresponds to the following hypothesis:

$$H_0\colon \boldsymbol{\mu}_{ref} = \boldsymbol{\mu}_{GMO} \text{ vs. } H_A\colon \boldsymbol{\mu}_{ref} \neq \boldsymbol{\mu}_{GMO} \tag{5}$$

A difference test based on a circular confidence region is obtained when the test-statistic is based on the Euclidean distance between $\boldsymbol{\mu}_{ref}$ and $\boldsymbol{\mu}_{GMO}$, i.e. the sum of all variable-wise differences between the means. Dempster proposed such a test when the distribution of the reference and GMO populations is multivariate normal, which was further refined by Bai and Saranadasa (1996) and Chen and Qin (2010) [Dempster 1958; Dempster 1960; Bai and Saranadasa 1996; Chen and Qin 2010]. Note that these tests lack desirable invariance properties under rescaling of the data (Zhang 2016) [Zhang and Pan 2016]. They do not take into account that the within-reference variance can be different for each endpoint. Therefore, Srivastava and Du proposed a test based on the Euclidean distance between $\boldsymbol{\mu}^*_{ref}$ and $\boldsymbol{\mu}^*_{GMO}$, i.e. the means of the data after scaling by the pooled within-group standard deviations (in this case the between-genotype variation) [Srivastava and Du 2008]. The resulting test statistic is proportional to the sum of the variable-wise two-sample *t*-test statistics discussed above. The resulting confidence region has ellipsoidal shape instead of circular shape, where the major axes of the ellipse correspond to endpoints with large within-reference variance. We will refer to this test at the Srivastava test.

Figure 6 compares the power of the *t*-test, variable-wise *t*-test and the Srivastava test. It can be seen that the power versus effect size curves are the same for all tests when only a single end point is considered. In other words, all tests are equal to the two-sample *t*-test when applied to a single variable. With increasing numbers of endpoints it can be seen that the attained significance level of the *t*-test is much higher than the desired significance level and some form of multiplicity correction is required. In contrast, the attained significance level of the Srivastava test was close to the nominal level in all cases. In this particular simulation, all tests had similar power, which decreased with increasing number of endpoints considered. This is not unexpected since the unintended effect considered occurred in only a single variable, i.e. we consider a case for which hypothesis (3) seems optimal. Figure 9c suggests that the Srivastava test might have higher power than the variable-wise *t*-test for detection of unintended effects that express themselves in multiple endpoints simultaneously. More examples are presented in section 3.

Figure 8 compares the power of the variable-wise *t*-test and the Srivastava test as a function of the correlation between the endpoints. Again, the attained significance level of the variable-wise *t*-test is much higher than the nominal level and multiplicity correction is required. Note, however, that the significance level is much closer to 5% when all variables are highly correlated. The power of the variable-wise *t*-test to detect unintended effects does not depend on the correlations between the variables. In contrast, in this particular simulation the power of the Srivastava test was greatly reduced when the correlation between the variables increased. This can be understood by the fact that the test does not take into account the correlation between the variables. Often, great improvements can be obtained using oriented confidence ellipses where the direction of the semi-major axes of the ellipse depends on the covariance structure of the data. These tests are discussed in the next section.

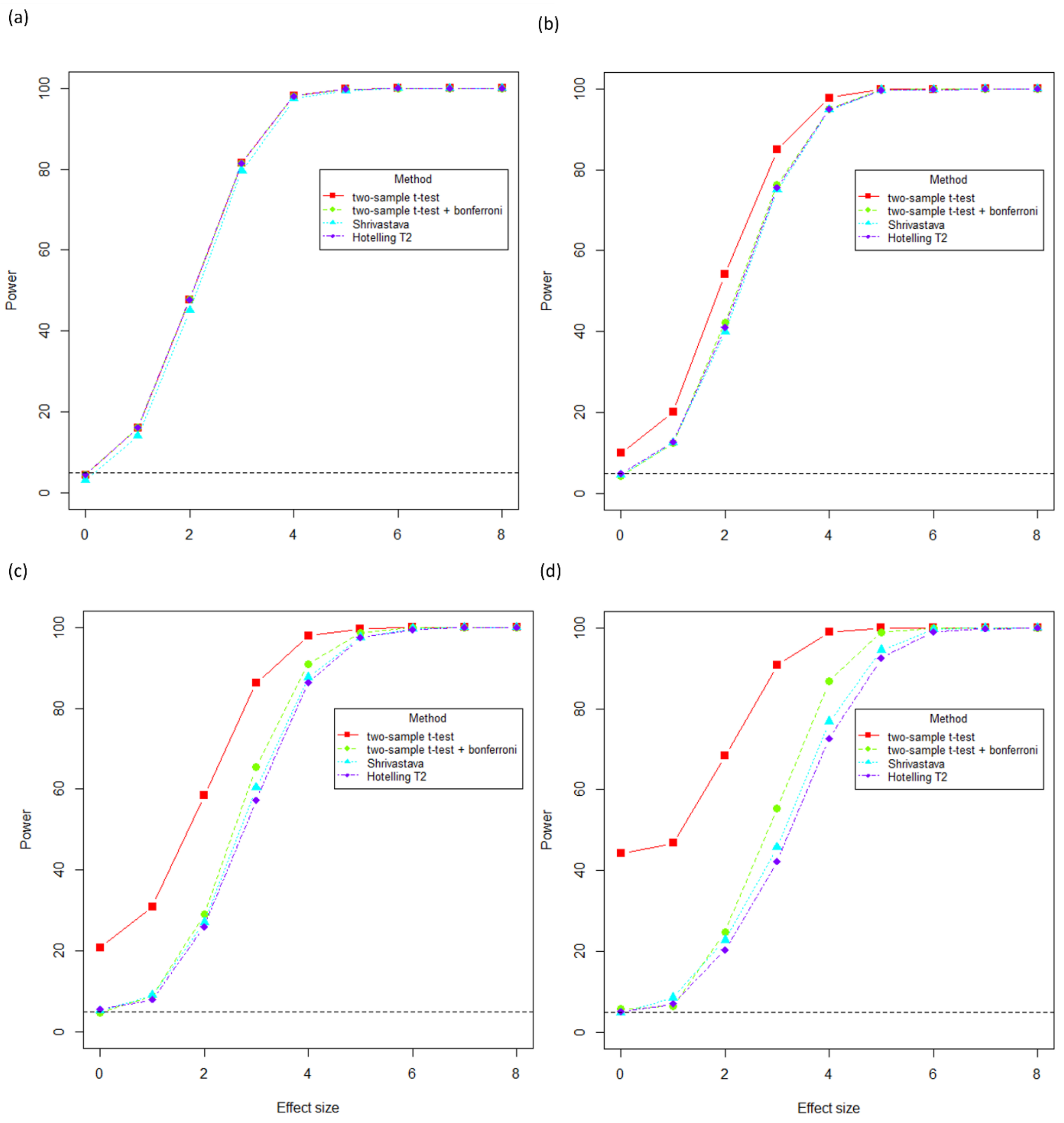

*Figure 7: power versus effect size of variable-wise t-test (red curve), variable-wise t-test with Bonferroni correction (green curve), Srivastava test (cyan curve) and Hotelling $T^2$ test (purple curve) for simulated multivariate data with (a) 1, (b) 2, (c) 5 and (d) 10 variables. The power is defined as the percentage of significant outcomes over 1000 simulations. For each simulation 40 reference observations were drawn from a standard multivariate normal distribution. A single GMO sample was drawn from the same distribution where the mean of the first variable was shifted according to the effect size specified on the x-axis in the figure.*

An equivalence test based on Euclidean distances tests the following hypothesis:

$$H_0: D(R,G)^2 \geq \Delta^2 \text{ vs. } H_A: D(R,G)^2 < \Delta^2 \qquad (6)$$

where $D(R,G)^2$ corresponds to the squared Euclidean distance between the mean of the references and the GMO [Munk and Pfluger 1999; Hoffelder et al 2015]. Note that (6) defines an equivalence region of circular shape (see figure 6). The null hypothesis of the equivalence test is rejected when the 90% circular CR completely lies inside the equivalence region of circular shape [Munk and Pfluger 1999]. Similar to the difference test, often one wants to study the difference in means relative to the within-reference variance for each endpoint. In that case, $D(R,G)^2$ in (6) is replaced by the sum of the variable-wise two-sample $t$-test statistics, i.e. the Euclidean distance is used to assess the discrepancy between the mean of the references and GMO relative to the between-genotype variance. The resulting equivalence region is of ellipsoidal shape (see figure 6). Note that point 1 in figure 6 is inside the ellipsoidal equivalence region (which takes into account between-genotype variance), but outside the circular region. In other words, this point does not seem to be abnormal if the within-reference variances are taken into account. Hoffelder et al. propose a circular and ellipsoidal equivalence test [Hoffelder et al 2015]. The test proposed by Wellek has an equivalence region of ellipsoidal shape [Wellek 2010]. It seems that many of the test-statistics of the difference tests (5) discussed above can also be extended to the equivalence-testing framework. To the best of our knowledge this has not been discussed in the literature.

In contrast to the difference testing, more work is needed to specify how a comparison between different equivalence testing approaches with respect to their power should be carried out. This is largely due to the fact that it is not directly evident how the equivalence limits for each test should be specified such that their equivalence regions are comparable for different types of unintended effects. **Therefore, we leave such a comparison for future research.**

Hoffelder et al, resolve this issue by only considering the power of various equivalence tests for a very specific unintended effect [Hoffelder et al 2015]. Simulated data was used to set the limits for each test such that they all attained the same size (5%) for differences of the unintended effect that were equal to the desired limit. Using these predefined equivalence limits the power of the various methods at an effect size of 0 was compared. For the unintended effect considered, Hoffelder et al showed that Euclidean-distance based tests performed quite favourably compared to the variable-wise approaches, especially for small sample sizes [Hoffelder et al 2015]. **We do not present an analogous power comparison in this report since first further research is required with respect to how equivalence limits should be defined based on multivariate data of reference genotypes and how such limits relate to possible unintended effects. Currently, the optimality of a specific equivalence test in a simulation study might be mistakenly generalized by the reader beyond the pre-specified equivalence limit.**

### 2.4 Multivariate tests based on Mahalanobis distance

The distance measure used in subsection 2.3 might be criticized for leading to confidence and equivalence regions whose geometrical shape does not depend on underlying correlation structure of the endpoints that have been measured. As shown in figure 8, for difference testing this might result in a loss of power when the test is applied to highly correlated data, although this also depends on the type of unintended effect and the number of end points considered. This subsection discusses difference and equivalence testing procedures where the distribution of the references and the GMO is multivariate normal. In the univariate case, the two-sample t-test and TOST can be used for difference and equivalence testing, respectively (see subsection 2.1). The multivariate analogue for the situation where the GMO and the reference-group have equal covariance matrices is the Hotelling T$^2$ test for differences or equivalence [Kent et al 2006; Wellek 2010]. In the T$^2$-test, the Mahalanobis distance (MD) is chosen as a measure of discrepancy between the GMO and references. The same hypothesis as in 2.3 are tested, but with the squared Euclidean distance ($D(R,G)^2$) in (6) replaced by the squared Mahalanobis distance ($MD(R,G)^2$). The test can be seen as an extension of the ellipsoidal tests shown in figure 6 towards oriented ellipsoids where the direction of the axes of the ellipse depend on the covariance structure of the data (i.e. correlations between end points) as shown in figure 7.

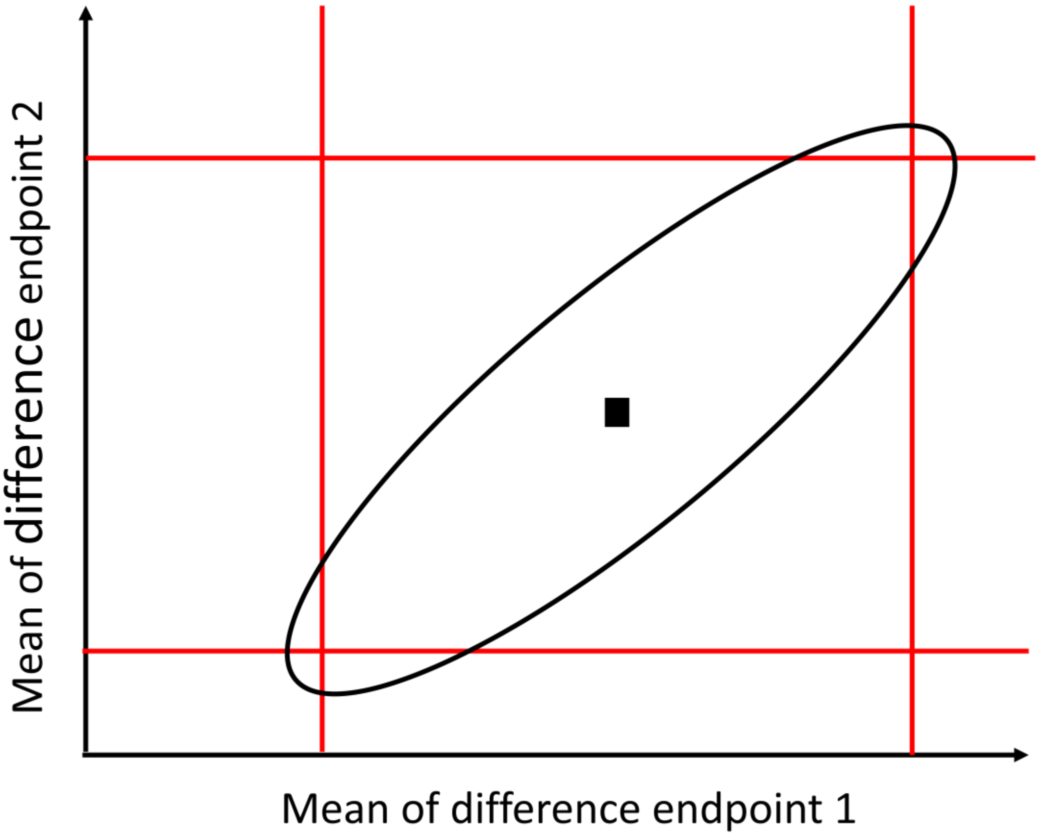

*Figure 8: Bivariate Hotelling's T$^2$ confidence ellipse. Note that the confidence ellipse uses the covariance to create an oriented confidence region. In contrast, the rectangular confidence region defined by the Bonferroni 95% CIs (red lines) does not depend on the covariance structure of the data.*

Application of Hotelling's $T^2$ for testing of hypothesis (5) is discussed in many texts such as [Kent 2006]. As shown in figure 7, in our simulation study the power of the Hotelling's $T^2$ test was similar to that of the Srivastava test when the endpoints where uncorrelated. This is not unexpected since the Srivastava test can be seen as a special case of the Hotelling's $T^2$ test when it is assumed that all endpoints are uncorrelated. As shown in figure 9, this assumption may lead to a decrease in power when the endpoints are in fact correlated. In contrast, since the Hotelling $T^2$ test does take these correlations into account it often has much more power to detect subtle unintended effects as shown in figure 9. Equivalence testing based on Hotellings $T^2$ is discussed by Wellek [Wellek 2010] and Hoffelder [Hoffelder et al 2015]. Hoffelder et al compare the power of the Hotelling $T^2$-based equivalence test to Euclidean distance-based tests as well as the variable-wise TOST [Hoffelder et al 2015]. A specific unintended effect was considered, namely a difference between the two groups of interest, which in our case are the GMO and references, in each of the endpoints. In this particular simulation the distribution of the data was multivariate normal and three endpoints were considered. Similarly to results of the difference test discussed above, it was observed that the power of the tests depended on the covariance structure in the data. In general, they observed the highest power for Hotelling's $T^2$ followed by the Euclidean-distance based equivalence test. The attained significance level of Hotelling's $T^2$ was close to the nominal level of 5%, although it is well known from Wellek [Wellek 2010] that the test can be conservative for non-normal data. The Euclidean-distance based equivalence test appeared to be anti-conservative, while the variable-wise approaches from section 2.2 were quite conservative. These variable-wise approaches showed little power for small sample sizes.

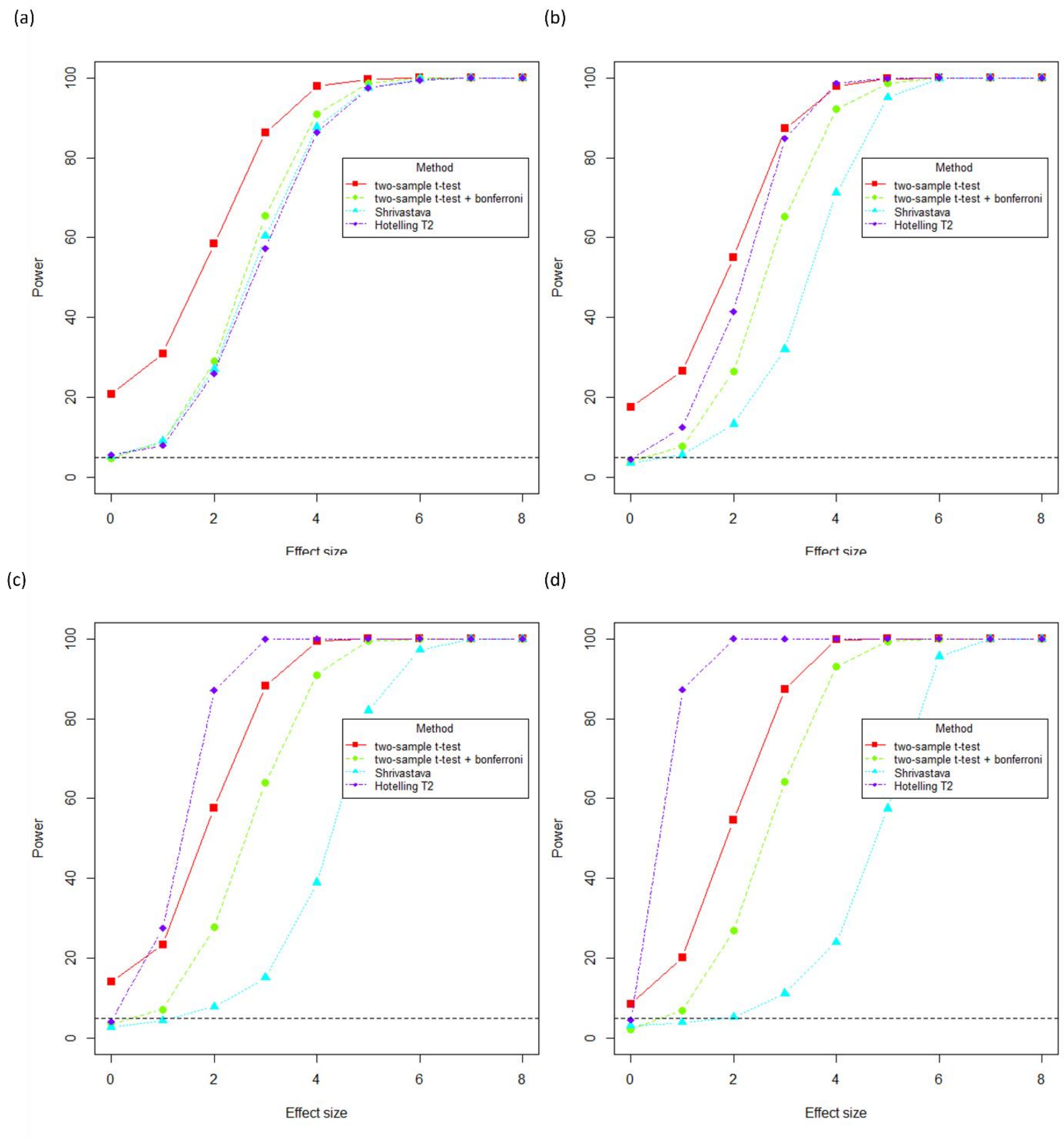

*Figure 9: power versus effect size of variable-wise t-test (red curve), variable-wise t-test with Bonferroni correction (green curve), Srivastava test (cyan curve) and Hotelling T2 test (purple curve) for simulated bivariate data with correlations of (a) 0, (b) 0.5, (c) 0.8 and (d) 0.95 between the two variables. The power is defined as the percentage of significant outcomes over 1000 simulations. For each simulation 40 reference observations were drawn from a bivariate normal with the specified correlation and unit variances. A single GMO sample was drawn from the same distribution where the mean of the first variable was shifted according to the effect size specified on the x-axis in the figure.*

## 2.5 Comparison of methods

In sections 2.2 – 2.4 various methods for difference and equivalence testing between the multivariate vectors of the mean of the references and the GMO were presented. It was shown, that the power of the methods depends on the number of variables in the data set as well as the correlation between them. However, the unintended effect (a difference in the first variable) was kept the same for all simulations. In this subsection we consider different types of unintended effects and show that the methods presented in sections 2.2 – 2.4 are biased towards different types of effects with respect to their direction in multivariate space. Figure 10a considers an unintended effect in one (out of 5) endpoint. In this case, Hotelling $T^2$ clearly outperforms the other methods since the 5 endpoints were highly correlated. When an unintended effect is simultaneously present in 3 out of 5 endpoints the methods perform much more similarly, although Hotelling $T^2$ still has higher power (figure 10b). In figure 10c, the unintended effect lies in the direction of the major-axis of the confidence ellipse in figure 8. It is well known in the literature that statistical tests that do not take into account the correlations between variables are biased towards such differences, and, indeed, it is seen that the Srivastava test outperforms Hotelling's $T^2$. The Srivastava test and two-sample *t*-test with Bonferroni correction perform similarly. In figure 10d, the unintended effect is along the minor axis of the confidence ellipse. In other words, the GMO differs from the references along a direction in multivariate space where the reference genotypes are quite similar. Such unintended effects cannot be identified well if the correlations between the end points are not taken into account. Hotelling's $T^2$ clearly outperforms the other methods in this case.

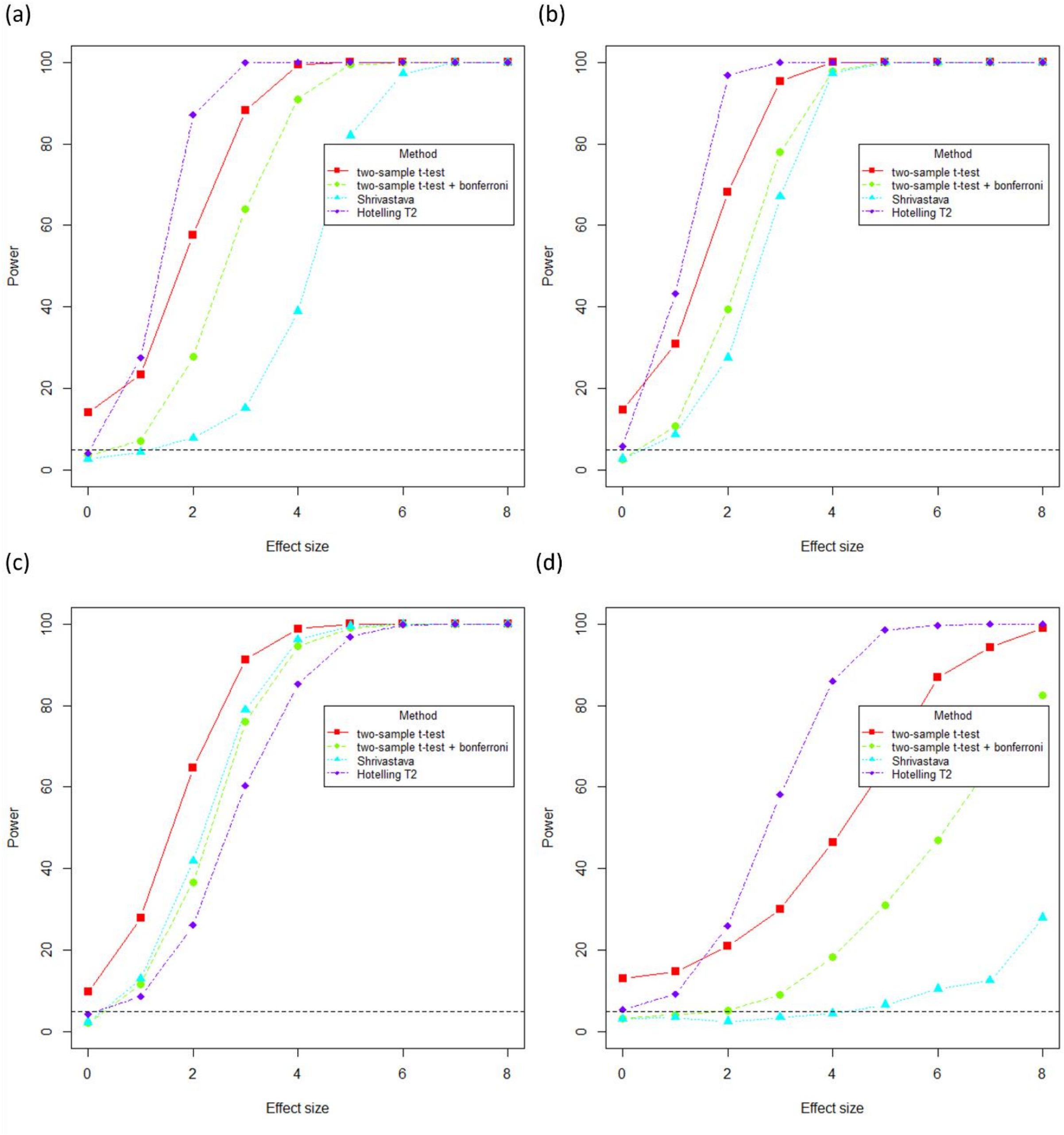

*Figure 10: power versus effect size of variable-wise t-test (red curve), variable-wise t-test with Bonferroni correction (green curve), Srivastava test (cyan curve) and Hotelling T2 test (purple curve) for simulated multivariate data with an unintended effect along (a) one variable, (b) three variables, (c) along major axis of the "oriented" ellipsoidal CR and (d) along the minor axis of the "oriented" ellipsoidal CR . The power is defined as the percentage of significant outcomes over 1000 simulations. For each simulation 40 reference observations were drawn from a 5 dimensional normal distribution with common correlation structure, i.e. all variables had unit variance and pairwise correlations of 0.8 to all other variables. A single GMO sample was drawn from the same distribution where the mean of the first variable was shifted as indicated above.*

## 2.6 Discussion and future prospects

This section discussed multivariate extensions of the two-sample *t*-test and TOST for difference and equivalence testing, respectively. It was shown that this extension requires one to define multivariate confidence and equivalence regions as opposed to univariate confidence and equivalence intervals. This can be done in different ways resulting in confidence regions of rectangular, circular, along-axes ellipsoidal, or oriented ellipsoidal shape. Based on simulated data examples it was shown that:

1. The size of the confidence and equivalence regions is typically larger than their univariate counterparts. This means, for example, that compared to the univariate case higher equivalence limits need to be specified for each endpoint to attain the same power in a proof of safety test.
2. In general, the power for detecting equivalence or differences depends on the number of endpoints considered, the correlation between the end points, and the type of unintended effect. There seems to be no method that always outperforms all other methods.

A major challenge to be addressed in applying multivariate difference or equivalence assessment procedures of any kind is to reach a consensus about the question of what kind of unintended effects are of interest and how to specify multivariate equivalence regions for these effects. For example, equivalence tests based on rectangular equivalence regions require equivalence in each endpoint for the GMO to be marked as equivalent to the references. In contrast, the circular and ellipsoidal approaches define absence of relevant differences (in all endpoints simultaneously) as overall equivalence [Munk and Pfluger 1999]. This situation is similar to the challenges encountered in, for example, industrial process monitoring or food authenticity studies. For a useful and effective application of the difference and equivalence testing approaches discussed the expected unintended effect should be also be considered with respect to the type of endpoint considered. For example, for omics data most likely only a few variables (endpoints) are expected to be changed with respect to the unintended effect. After the type of change from unintended effects has been chosen, simulation studies similar to those shown in this section seem to be useful to better understand the properties of the methods and to determine the most promising difference and equivalence testing procedure. Practical application by non-experts of the methodology should also be taken into account here. In this sense the variable-wise procedures resulting in a rectangular confidence and equivalence region might be preferred since they are much more interpretable compared to ellipsoidal regions whose shape (and direction) depends on the covariance structure of the end points.

An additional advantage of the variable-wise testing procedures is that they are based on well-developed univariate statistical tests. For example, for more complicated experimental designs such as randomised block designs with both fixed and random factors, mixed model methodology is available for univariate difference and equivalence testing, which can be easily extended to variable-wise approaches. In contrast, for the ellipsoidal multivariate tests it is not directly evident how the experimental design should be taken into account. Additionally, with respect to specification of the equivalence region currently no approaches exist to use internal references or historical data for this purpose.  In contrast, such approaches have been developed for the univariate case, as mentioned in section 1. Their extension is more straightforward for the variable-wise approaches compared to the ellipsoidal methods such as Hotelling $T^2$.

Finally, we would like to remark that not all methods are directly applicable to high-dimensional data where the number of endpoints measured is much larger than the sample size (e.g. omics data), or they perform poorly. Methods based on Hotelling's $T^2$ are not applicable in this case due to the "curse of dimensionality", and most likely some form of regularization such as dimensional reduction or variable selection is required. This issue is also encountered in Industrial process monitoring and more difference regularization approaches have been proposed in the context of difference testing. Some of these are presented and compared in section 3. In general, extensive simulations studies are required to better understand the potential of difference and equivalence testing using high-dimensional omics data for food safety assessment.

# 3 Multiple and multivariate difference tests of high dimensional means for use in food safety assessment

In this section several multivariate difference tests of high-dimensional means are presented. We specifically focus on typical omics data sets where the number of references is (much) lower than the number of endpoints measured. Such data poses specific challenges for data analysis due to the curse of dimensionality. Most notably, the traditional Hotelling's $T^2$ test is not applicable to such data. Several regularized Hotelling's $T^2$ tests are presented, including the one-class model of van Dijk et al [van Dijk et al 2014]. Examples using simulated data are used to highlight some of the similarities and differences between the approaches. The section is structured as follows. Subsection 3.2 discusses testing for differences between GMs and references when the data has a latent low-dimensional structure. More specifically, the one-class model of van Dijk [van Dijk et al 2014] is placed in the context of the difference tests discussed in section 2. Borrowing on ideas from the field of Multivariate Statistical Process Control (MSPC) some possible improvements of the one-class model are discussed as well. Section 3.3 shows how a modern shrinkage estimator of the covariance matrix can be used to obtain a Hoteling's $T^2$-like difference test for high-dimensional data. In subsection 3.4 the dimension reduction and shrinkage-based tests are compared to each other. In subsection 3.5, equivalence testing in high-dimensional data is discussed. Finally, a summary and some suggestions for further research are presented in subsection 3.6 focusing on the application of these difference tests to high-dimensional omics data.

## 3.1 Introduction

In classical experiments for food safety assessment a relatively small number of endpoints is measured. Due to the omics revolution it is now possible to characterise samples in much greater molecular detail. Using technologies such as transcriptomics or metabolomics it is possible to simultaneously measure the expression of thousands to tens of thousands of genes or the abundances of hundreds to thousands of metabolites in a single sample. In comparison to classical experiments, the use of omics therefore holds great promise for identification of unintended effects in a GM.

The high-dimensional nature of omics data poses specific challenges for data analysis, which together are often referred to as 'the curse of dimensionality' and include amongst others inaccurate parameter estimation, noise accumulation, and spurious correlation [Zimek et al 2012]. Additionally, many traditional statistical methods were originally developed for data from classical experiments where the number of reference varieties measured is (much) larger than the number of endpoints considered. With omics data the opposite is often encountered (many more end points than reference varieties), and many multivariate statistical techniques break down or their estimates are unreliable. This is, for example, the case for the difference and equivalence tests based on Hoteling's $T^2$-statistic that were discussed in section 2.4 [Engel et al 2017]. As shown in 2.4, these tests take into account the covariance structure in the data resulting in oriented confidence regions of ellipsoidal shape. Estimation of the covariance structure (covariance matrix) requires the determination of $(m^2 + m)/2$ parameters ($m$ is the number of variables or end points in the data). Clearly, this number becomes quickly too large when the dimension of the data (*m*) increases with respect to the number of reference varieties that is measured [Engel et al 2017; Ledoit and Wolf 2004]. The effect of high-dimensional data can for example be clearly seen when considering the variances of the reference varieties along the major and minor axes of the Hoteling's $T^2$ confidence ellipse. It is well known that for high-dimensional data the between-reference variance along the major axes of the ellipse are overestimated, while those along the minor axis are underestimated [Schafer and Strimmer 2005]. Because of this, a small observed difference between the GM and references along one of the minor axis of the confidence ellipse may be incorrectly marked as significant, while a large difference along the major axes may be incorrectly marked as non-significant. An example is shown in figure 1, which is more thoroughly discussed in section 3.3. When the number of endpoints measured is (much) larger than the number of references, some (many) of the estimates of the variance along the axes of the ellipse will be zero in which case Hoteling's $T^2$ is clearly not applicable [Engel et al 2017; Ledoit and Wolf 2004].

A (naive) strategy to obtain a more efficient estimator can be to consider an estimator with a lot of structure imposed. This way fewer parameters need to be estimated. For example, one could assume that all variables (endpoints) are uncorrelated, giving rise to statistical tests similar to those discussed in sections 2.2 and 2.3. As will be shown below, these tests are indeed applicable to high-dimensional data, but clearly relevant information in the data might be discarded when correlations are ignored resulting in a loss of power. Another (perhaps more often used) strategy in many fields is to assume that the data has a latent low-dimensional structure. In such cases the dimension of the data can be reduced by e.g. principal component analysis, where a few principal components explain a large percentage of variance, and the analysis is restricted to this low-dimensional space and its orthogonal complement [Qin 2003]. Note that the one-class model of van Dijk [van Dijk et al 2014], which was used in G-TwYST to compare GM samples to a set of references on the basis of acquired omics data, is of this type.

Larger and noisier data, such as encountered in omics, are less likely to completely respect the assumption of low-dimensionality. In recent years, much effort in fields such as statistics, machine learning, and econometrics has focused on obtaining improved (regularized) estimators of the covariance matrix [Engel et al 2017]. Many of these methods are computationally inexpensive and can be easily combined with well-known multivariate statistical techniques in high-dimensional problems to potentially obtain improved (in some relevant sense) models. These improved estimators of the covariance matrix can be roughly divided into two categories, depending on whether they do or do not make specific assumptions about the structure of the covariance matrix. In this section, we focus on estimators that do not make specific assumptions with respect to the structure of the covariance matrix. These methods generally aim to obtain better estimates of the variance along the axes of the confidence ellipse in figure 1 without adjusting the direction of the axes. We will refer to these estimators as shrinkage estimators. It is well known that for high-dimensional data also the direction of the axes of the confidence ellipse are estimated with  (possibly large) error. Better estimates of these directions, and the variance along them, can sometimes be obtained by assuming a special structure for the covariance matrix. A common assumption in omics, for example, is that the covariance matrix is sparse, i.e. a large number of endpoints are uncorrelated or nearly uncorrelated to each other. However, whether such a structure does indeed exist cannot be verified from the data. Therefore, in some applications, a structure free approach as used here can be preferred.

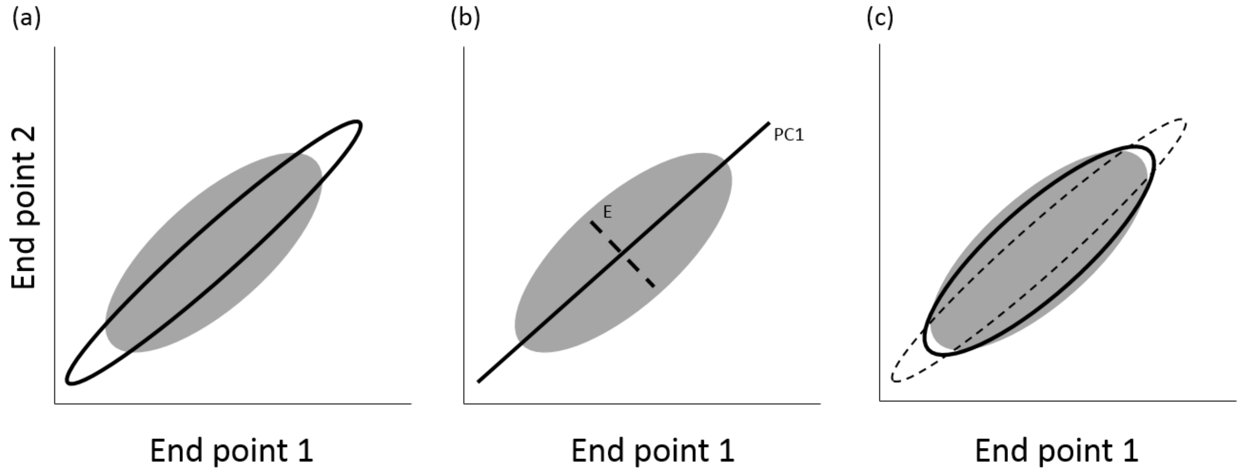

*Figure 1: Schematic representation illustrating in two dimensions the confidence region of various multivariate difference tests. In each panel the grey ellipse corresponds to the confidence region of Hoteling's $T^2$ based on the population covariance structure, i.e. the "real" confidence region. The ellipse indicated by the black line in panel (a) shows the confidence region that is obtained using the traditional sample estimator of the covariance matrix. Panel (b) corresponds to a difference test using principal component analysis, where one tests for differences along the selected principal component(s) (PC1) and the orthogonal directions (E) separately. In panel (c) the confidence region of Hoteling's $T^2$ (black line) is based on a shrinkage estimator of the covariance matrix. For clarity, the sample covariance ellipse is indicated by the dotted line in panel c.*

### 3.2 Methods based on dimension reduction

A common strategy to deal with the curse of dimensionality is to first reduce the dimension of the data using principal component analysis (PCA) [Qin 2003]. Here, it is assumed that the data of the reference varieties has a latent low-dimensional structure. In other words, it is assumed that the majority of the differences between e.g. the metabolic profiles of the reference varieties can be captured by a limited number of principal components (PCs). Hoteling's $T^2$ can be used to detect differences between a GM and the reference varieties in the low-dimensional space. Often, the orthogonal complement to the PC-space is assumed to contain only (uncorrelated) noise and a test based on the Euclidean distance (see section 2.3) is used to compare the GM to the references in this space. Although both subspaces can be useful for detecting unintended effects in a GM, we remark that they measure different types of effects and therefore have distinct roles in difference testing. As shown in figure 1b, a difference in the PC-space occurs when e.g. several metabolites shift in abundance in accordance to the between-metabolite correlations in the reference varieties (i.e. a shift along one or multiple of the major axes of the confidence ellipse). A difference in the orthogonal complement (marked E in figure 1b) suggests that e.g. the metabolic profile of the GM either contains metabolites that are not present in the reference varieties and / or patterns of metabolite abundances that break the between-metabolite correlation structure. Unintended effects in both subspaces may be observed as well.

Van Dijk et al. [van Dijk et al 2014] propose a PCA-based approach to compare GMs to reference varieties. They refer to their model as soft independent modelling of class analogy (SIMCA), which we will also do in this section. Their model is a PCA-based one-class model such as shown in figure 1b [Qin 2003]. These models have been used with great success in many applications such as industrial process monitoring, food authentication and health monitoring. Many variants of PCA-based one-class models have been proposed in the literature. They mainly differ with respect to the exact test statistics used and how the number of selected principal components is chosen. Van Dijk et al (2014) propose the following procedure:

1. Apply PCA to the data of the reference varieties
2. Select *k* principal components
3. Estimate the dissimilarity (in SIMCA customarily called the class distance $\hat{s}_i$) between a GM and the references in the subspace orthogonal to the *k* selected components as
$$\hat{s}_i = \left(\frac{\hat{e}_i\hat{e}_i^T}{m-k}\right)^{\frac{1}{2}}$$
where $\hat{e}_i\hat{e}_i$ corresponds to the squared Euclidean distance between observation *i* and the mean of the references in the subspace orthogonal to the selected *k* components.
4. Determine the cut-off-value $s_{crit}$ for the class distance as
$$s_{crit} = s_0(F_{crit})^{\frac{1}{2}}$$
where $s_0^2$ is the average of the $\hat{s}_i^2$ , and where $F_{crit}$ is a 95% point of an F distribution with 1 and $n-k-1$ degrees of freedom.
5. Mark observations with a class distance smaller than the cut-off as safe and observations with a larger distance as potentially unsafe.

A double cross-validation procedure is used to select the optimal number of PCs (*k*) in step 2. More specifically, the optimal number is defined as the highest number still causing all validation samples to be marked as safe by the test in step 5. We will discuss optimization of the optimal number of components more thoroughly below.

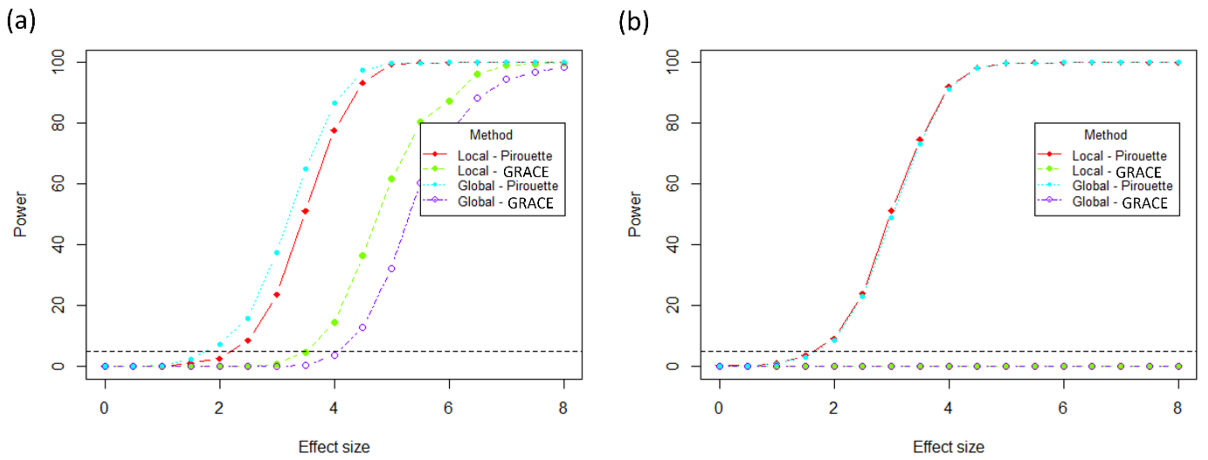

*Figure 2: power versus effect size of global and local SIMCA models based on the Pirouette or GRACE test statistic for simulated multivariate data with (a) 20 and (b) 40 reference varieties. The reference observations were drawn from a 20 dimensional normal distribution with a blocked correlation structure where groups of 10 variables had a pairwise correlation of 0.8 amongst themselves and a correlation of 0 to all other variables. GM samples were drawn from the sample distribution, but the mean of one variable was shifted as indicated above. For each effect size the power is defined as the percentage of significant outcomes over 1000 simulated GMs.*

In van Dijk et al (2014) use was made of the Pirouette software [Pirouette 4.5 software manual; van Dijk et al 2014], in later work in the GRACE project [Kok et al., subm., Corujo et al. subm.] own R code was used with slightly different choices for the calculation of the thresholds. The average squared class distance $s_0^2$ was now calculated as the average of $\hat{s}_j^2 = \frac{\hat{e}_j\hat{e}_j^T}{m-n-k}$, therefore applying a different expression for the denominator. We will refer to this algorithm as the GRACE method.  The one-class model is similar to the difference tests discussed in section 2.3 with respect to the hypothesis that are being tested. In the literature, many statistical tests have been proposed for difference testing in the orthogonal subspace. Unfortunately, no references are provided that detail the mathematical (statistical) foundations of both of these tests. Therefore, it is difficult to verify the statistical validity of the approach taken in van Dijk (2014) and in the more recent GRACE papers. The difference between the Pirouette and GRACE tests is a larger critical value for the GRACE test, e.g. a significant difference between a GM and the references is declared less quickly. An example is shown in figure 2a. As shown in figure 2b, the GRACE test used in Kok et al and Corujo et al is not applicable when the number of reference varieties is larger than the number of endpoints measured. This is due to the fact that negative values for the test statistic obtained in this case (the estimated squared class distance $s_0^2$ is zero). Additionally, as shown in figures 2 -5 we observe that the Pirouette and GRACE test are often quite conservative. **Given these issues, we question the statistical validity of these approaches and further research is required in this respect.** *Note that all results labelled 'Pirouette' in this report were obtained by our own R code rather than the Pirouette software package itself.*

A global model is defined as the one-class PCA model that is based on all training samples and optimal number of PCs as determined by double-cross validation. Local models are defined as one-class PCA models based on a subset of the training samples as used in the inner-validation loop in the double cross-validation procedure. Van Dijk et al discuss the use of one global versus multiple local models with respect to detecting differences between a GM and references [van Dijk et al 2014]. They choose the strategy of using multiple models following the reasoning of Westerhuis et al. (2008) that no accepted criteria are available for the way to choose an overall model. Figure 2 compares the power of local and global models for simulated data with 25 and 40 reference varieties and 40 end points, i.e. 25 and 40 local models vs a single global model. Although local modelling seems to somewhat improve the performance of the GRACE test this is not the case for the Pirouette test. Also in other simulations **we did not observe a clear advantage of the "local modelling" strategy and the effects seemed to diminish with increasing dimension of the data. Therefore, we leave this for further research and only use the global one-class PCA model with Pirouette test below.** Note that similar as was observed in section 2, typically the power of the tests became lower when either the number of reference values decreased or the dimension of the data increased (see figures 3 and 4).

Figures 3 and 4 compare the power of the SIMCA model of van Dijk to that of another PCA-based one-class model that is often used in the field of multivariate statistical process control (MSPC) [Qin 2003]. Here, the so-called *Q*-statistic is used to detect differences between the group means in the subspace orthogonal to the selected PCs. We observe that the *Q*-statistic often seems to have much greater power for detection of unintended effects compared to the Pirouette test used in SIMCA. An example is shown in figure 3 where *Q*-statistic clearly outperforms the GRACE test, especially for larger sample sizes.

In MSPC is has been realized that only specific types of faults (unintended effects) can be detected in the orthogonal subspace. By constraining the analysis to the subspace orthogonal to the *k* PCs, only unintended effects that are orthogonal to the normal variation between the reference varieties can be detected. As mentioned above, these can e.g. be the presence of metabolites in the GM profile that absent in the reference profiles and / or patterns of metabolite abundances that do not match the between-metabolite correlation structure in the references. In MSPC, Hoteling's $T^2$ is often also applied to detect differences in the PC-space, i.e. differences in the directions in multivariate space along which also a large amount of variance between the reference varieties is observed [Qin 2003]. *Note that this is a different test compared to Hoteling's $T^2$ from section 2.4, which is applied to all endpoints rather*. A combined index has also been defined, which allows for detection of unintended effects in both subspaces simultaneously [Qin 2003]. In contrast to the tests used by Pirouette and GRACE, a wealth of literature is available on the use of dimension reduction by PCA and subsequent application of Hoteling's $T^2$, *Q*, or the combined statistic for difference testing, also when applied to high-dimensional data. For the unintended effect shown in figure 3 (a shift in a single endpoint) the *Q*-statistic and Pirouette test clearly outperform Hoteling's $T^2$ and the combined index. This is expected since this is an unintended effect in a direction that is orthogonal to the first principal components. As shown in figure 5, however, for other unintended effects the power of the PCA-based one-class model is greatly improved by also considering Hoteling's $T^2$ and the combined statistic. **The use of alternative test statistics such as *Q* or $T^2$ from fields such as MSPC therefore may offer an interesting route towards further improvement of the one-class**

**model.** We would like to remark that further research on the use of the combined index for high-dimensional data is required as shown in figure 3, the attained significance level of this test was often much higher than the nominal level. Therefore, this test is not considered further for now.

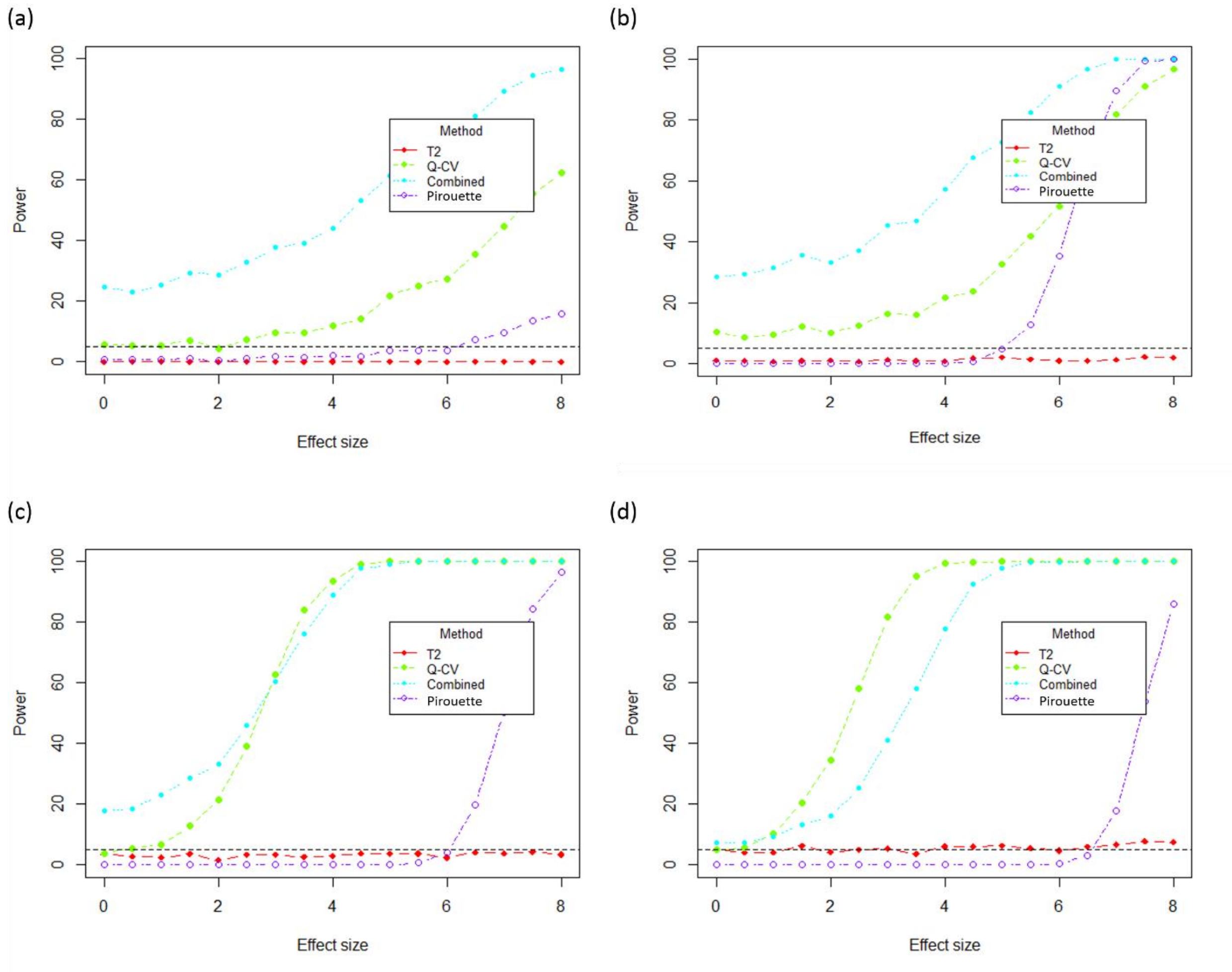

*Figure 3: power versus effect size of the SIMCA model (Pirouette test statistic) and another PCA-based one-class model using Hoteling's $T^2$,Q, and combined-statistics for simulated data with (a) 25, (b) 50, (c) 100, and (d) 500 reference varieties. The reference observations were drawn from a 100 dimensional normal distribution with a blocked correlation structure where groups of 10 variables had a pairwise correlation of 0.8 amongst themselves and a correlation of 0 to all other variables. GM samples were drawn from the same distribution, but the mean of one variable was shifted as indicated above. For each effect size the power is defined as the percentage of significant outcomes over 1000 simulated GMs.*

We would like to remark that in figure 3 (and 4 and 5) the number of principal components for the PCA-based one class model ($T^2$, *Q*, and combined statistics) was not based on double cross-validation. Instead, the number of components was determined by a bootstrapping procedure called NUMFACT [Henry et al 1999]. In essence, the principal components for each resampling were compared for changes. Components which changed greatly from one resampling to the next were probably due to noise and not selected. Note that this method is much less computationally expensive compared to double cross-validation. Additionally, during the simulation studies it was observed that a similar classification accuracy (for marking reference varieties as safe) was obtained by double cross-validation for different numbers of selected components. Moreover, the nature of the classification accuracies observed was very discrete for low numbers of references varieties. **This suggests that it would be interesting to compare different methods to optimize the number of PCs, e.g. NUMFACT, to double cross-validation, both with respect computational complexity and power to detect of different types of unintended effects of the one-class model.**

### 3.3 Methods based on shrinkage

The covariance matrix is a central aspect of many multivariate data analysis methods [Kent et al 2006; Engel et al 2017]. It essentially determines the direction, shape and volume of the confidence ellipse in Hoteling's $T^2$ test (figure 1) for testing for differences between a GM and a set of references. However, the test breaks down when the number of endpoints measured is larger than the number of reference varieties (e.g. omics data). As mentioned in section 3.1, this can be seen when considering the eigenvalue structure of the sample estimator for the covariance matrix. Typically, in high dimensional data the largest eigenvalues are overestimated while the smaller eigenvalues are underestimated [Engel et al 2017; Bun et al 2017; Paul and Aue 2014]. This has several undesirable consequences as shown in figure 1a. The eigenvalues are proportional to the variance of the reference varieties along the axes of the confidence ellipse. The estimation error of the eigenvalues means that the between-variety variance can be over- or underestimated. Let's assume that the eigenvalue corresponding to the major axes of the confidence ellipse in figure 1a is biased upwards and the eigenvalue of the minor axis is biased downwards. This means that the Hoteling's $T^2$ test will underestimate the difference between a GM and the reference varieties for unintended effects along the major axis of the confidence ellipse (the direction marked PC1 in figure 1b). In contrast, differences due to an unintended effect along the minor axis (marked E in figure 1b) are overestimated. Finally, when the number of reference varieties is smaller than the number of endpoints measured, some of the eigenvalues will be zero and Hoteling's $T^2$ cannot be estimated. These issues may potentially be resolved by using modern regularized estimators of the covariance matrix rather than the sample estimator.

Shrinkage of eigenvalues is the oldest approach to regularization of the covariance matrix [Ledoit and Wolf 2004; Engel et al 2017; Bun et al 2017; Paul and Aue 2014]. The term shrinkage refers to the fact that these methods essentially pull (shrink) the sample eigenvalues towards a central value thereby correcting their bias. An example is shown in figure 1c. Note that the eigenvalues are indeed adjusted as indicated by the different shape of the ellipse. A popular method is the linear shrinker introduced by Ledoit and Wolf (LW) where each sample eigenvalue is shrunk towards the average sample eigenvalue [Ledoit and Wolf 2004]. The shrunken eigenvalues are typically much closer to the population eigenvalues compared to those of the sample covariance matrix, resulting in a better estimate of the between-reference covariance (figure 1c). The LW estimator of the covariance matrix can be easily combined with the Hoteling's $T^2$-statistic for testing for a difference between the mean of the GM and reference population (hypothesis 5 in section 2) [Ullah et al 2017]. Similar to the PCA-based models discussed above, the resulting regularized Hoteling's $T^2$-test is applicable to high-dimensional data even when the number of endpoints is much larger than the number of reference varieties. An advantage of this approach is that no parameter optimization is required as for the PCA-based models (optimization of number of components). Also, similar to the combined index (section 3.2) the test considers unintended effects in any possible direction in multivariate space. Unfortunately, the null distribution of the shrinkage-based test statistic is unknown and is typically approximated empirically using cross-validation or a bootstrap procedure [Ullah et al 2017]. Geometrically, the test based on linear shrinkage of the eigenvalues can be seen as a combination of the difference tests discussed in sections 2. Without shrinkage, the oriented confidence region of Hoteling's $T^2$ is obtained. This test is not applicable to high-dimensional data. With increasing amounts of shrinkage the confidence region becomes more circular similar to the Euclidean distance-based tests (section 2.3), i.e. the correlations between the endpoints are less taken into account. The Euclidean distance-based tests are applicable to high-dimensional data, but possibly have lower power for detecting unintended effects because the correlations between the endpoints are ignored. The regularized Hoteling's $T^2$ based on LW shrinkage has a confidence region that takes into account the correlations between the end points as much as the data allows. This way, the test is applicable to high-dimensional data while typically having greater power than Euclidean-distance based tests.

It is well known that linear shrinkage captures almost all possible improvement (over the use of the sample eigenvalues) when the number of endpoints is very large compared to the number of reference varieties and / or when the between-reference variance in the population confidence ellipse (figure 1) is similar along most directions. However, when these variances are dispersed (e.g. much larger variation can be observed along a few axis of the ellipse), linear shrinkage only improves upon the sample estimator of the covariance matrix slightly [Engel et al 2017; Ledoit and Wolf 2012; Ledoit and Wolf 2017; Ledoit and Wolf 2018]. Note that this is similar to the low-dimensional case that was considered in the previous subsection. Random matrix theory shows that eigenvalue shrinkage is fundamentally a nonlinear problem of which linear shrinkage is only an approximation [Ledoit and Wolf 2012]. Therefore, nonlinear shrinkage of the sample eigenvalues offers a great route to possibly further improve the Hoteling's $T^2$-test for high-dimensional data. Note that whereas linear shrinkage estimators shrink all eigenvalues uniformly this is not the case for non-linear shrinkage approaches. Therefore, a non-linear shrinkage Hoteling's $T^2$-test cannot be seen as a weighted average of the traditional Hoteling's $T^2$-test and Euclidean-distance based tests, as was the case for linear shrinkage. Unfortunately, most nonlinear shrinkage estimators are computationally quite expensive and therefore not applicable to data of moderate to high-dimensionality. Recently, however, Ledoit and Wolf introduced a nonlinear shrinkage estimator of the covariance matrix that is much faster than previous approaches without loss of numerical accuracy [Ledoit and Wolf 2017]. Here, we consider this nonlinear shrinkage approach in combination with Hoteling's $T^2$ for difference testing between a GM and a set of reference varieties.

### 3.4 Comparison of methods

Figure 4 compares the shrinkage-based Hoteling's $T^2$-tests (linear and nonlinear) to the PCA-based one-class models using the SIMCA - Pirouette test-statistic, or the $T^2$ and Q-statistic. Also, the variable-wise two-sample *t*-test in combination with Bonferroni correction from section 2.2 and the Srivastava test from 2.3 are included. As shown in figure 4a and 4b, all of these tests are applicable when the number of reference varieties is smaller than the dimension of the data (number of endpoints measured). Additionally, it can be observed that the power of the tests to detect differences decreases when the number of reference varieties goes down. Similar patterns were observed when the dimension of the data increased for a fixed number of references (not shown). Interestingly, for a moderate number of reference varieties (4b – d) the shrinkage-based Hoteling's $T^2$ approaches clearly outperform the other tests, or are very competitive. However, when only 25 reference varieties are available the variable-wise t-test had the highest power. This is ascribed to the masking effect discussed in 3.6. The difference of the unintended effect in the single endpoint is masked by relatively large size of the natural variation between the varieties in all other endpoints. Because the variable-wise *t*-test does not take correlations into account (and only considers shifts along single variables) the masking affect is much less strong compared to the multivariate approaches.

As was shown in section 2.5, the power of the various difference tests depends not only on the number of reference varieties, the number of variables in the data and the correlation between them, but also on the type of unintended effect. In figure 5, the power of the tests is compared for different unintended effects for high-dimensional data. Similar to section 2.5, it can be seen that the variable-wise TOST and Srivastava test outperform the other methods when the unintended effect lies in the direction of the major-axis of the confidence ellipse (figure 5c). Typically, the shrinkage-based Hotelling $T^2$-tests and the *Q*-statistic have higher power for detecting differences in a few variables, although the power might be reduced when the number of "different" variables is small compared to the total number of variables (masking). As shown in figure 5d, the shrinkage-based tests outperformed all other methods for unintended effect is along the minor axis of the confidence ellipse. In other words, the GMO differs from the references along a direction in multivariate space where the reference genotypes are quite similar. Such unintended effects cannot be identified well if the correlations between the endpoints are not taken into account. **Due to the limited number of simulated scenarios we do not make any recommendations regarding which method is most suitable for safety assessment on the basis of high-dimensional data. Although the use of shrinkage-based Hoteling's $T^2$-tests seems promising, the method should be evaluated carefully in more extensive simulation studies.**

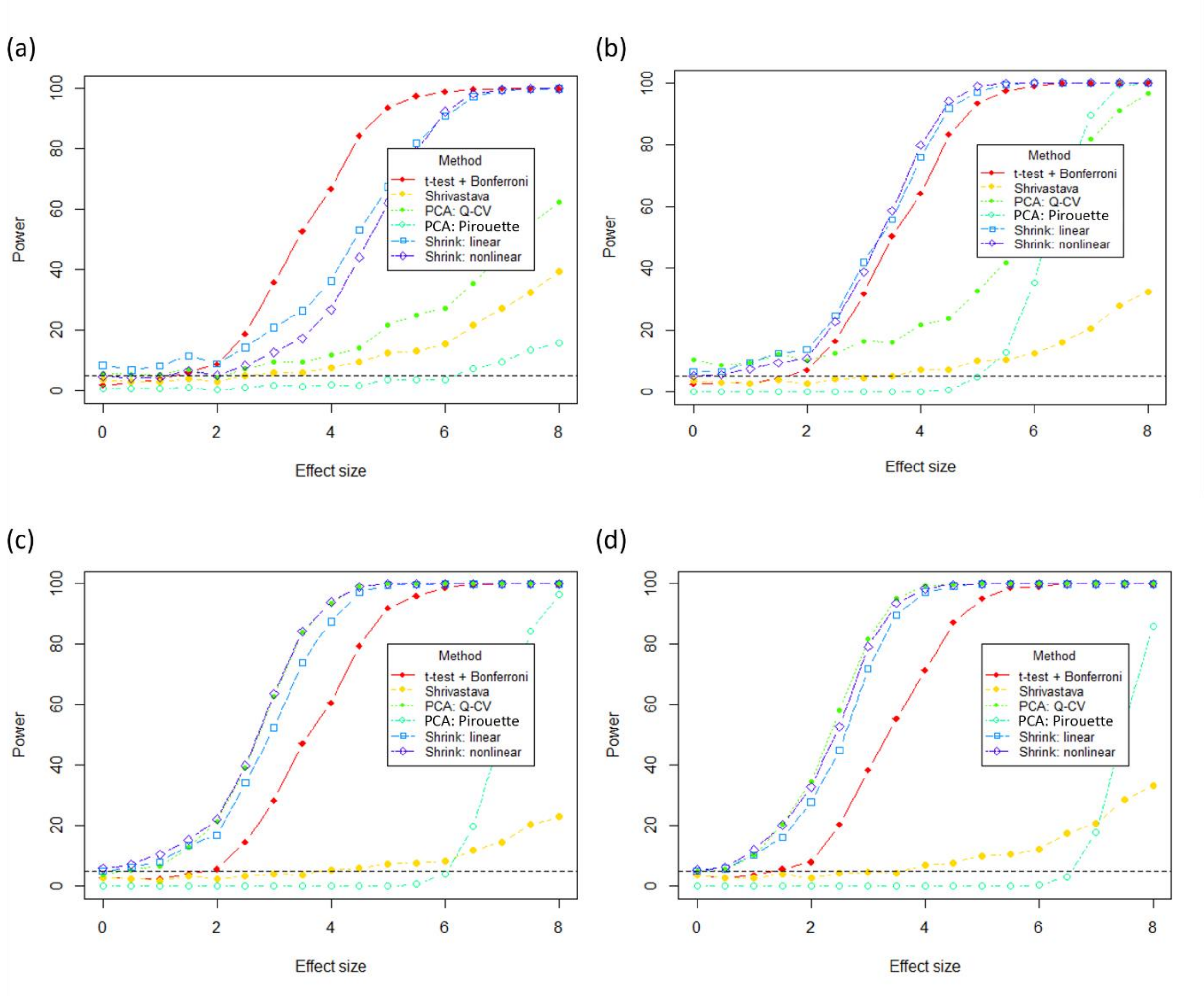

*Figure 4: power versus effect size of the variable-wise t-test with Bonferroni correction, Srivastava test, one-class PCA (Q-statistic and Pirouette test-statistic), Hotellings $T^2$ (linear and nonlinear shrinkage) for simulated data with (a) 25, (b) 50, (c) 100, and (d) 500 reference varieties. The reference observations were drawn from a 100 dimensional normal distribution with a blocked correlation structure where groups of 10 variables had a pairwise correlation of 0.8 amongst themselves and a correlation of 0 to all other variables. GM samples were drawn from the sample distribution, but the mean of one variable was shifted as indicated above. For each effect size the power is defined as the percentage of significant outcomes over 1000 simulated GMs.*

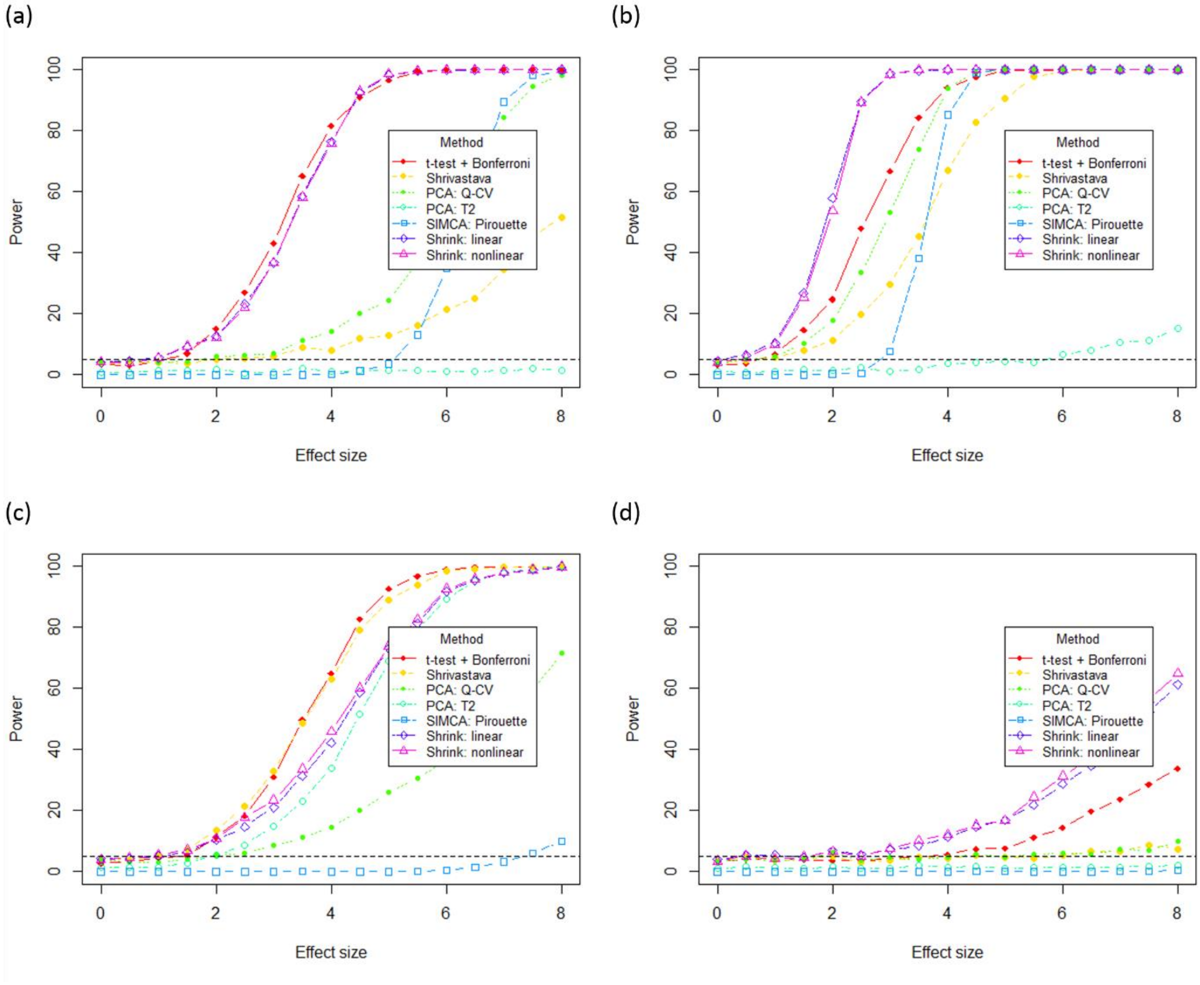

*Figure 5: power versus effect size of the variable-wise t-test with Bonferroni correction, Srivastava test, one-class PCA (T$^2$, Q, and Pirouette test-statistics), Hotellings T$^2$ (linear and nonlinear shrinkage) for simulated data with an unintended effect along (a) one variable, (b) five correlated variables, (c) along the first eigenvectors of the population correlation matrix, and (d) along the last eigenvector of the population correlation matrix. Fifty reference observations were drawn from a 100 dimensional normal distribution with a blocked correlation structure where groups of 10 variables had a pairwise correlation of 0.8 amongst themselves and a correlation of 0 to all other variables. GM samples were drawn from the sample distribution, but the mean of one variable was shifted as indicated above. For each effect size the power is defined as the percentage of significant outcomes over 1000 simulated GMs.*

### 3.5 Equivalence testing in high-dimensional data

To the best of our knowledge, multivariate equivalence testing in high-dimensional data has not been considered yet in the literature. However, given the similarity between the Hoteling's $T^2$-based difference and equivalence tests for low-dimensional data (section 2.4) it seems that the regularization strategies discussed above for difference testing can also be applied to the equivalence testing case. As mentioned in sections 1 and 2, ideally the equivalence region of the equivalence test should be based on expert-knowledge. If this knowledge is (not yet) available, possibly the equivalence region can be estimated from historical data. However, this is not straightforward for high-dimensional data given the difficulties with estimating covariance matrices for high-dimensional data discussed above.

In principle equivalence tests based on variable-wise application of TOST or equivalence tests based on Euclidean distances are directly applicable to high-dimensional data. However, further research is required to assess the power of these methods for such data.

### 3.6 Discussion and future perspectives

In this section several multivariate difference tests of high-dimensional means are presented. We specifically focus on typical omics data sets where the number of references is (much) lower than the number of endpoints measured. Such data poses specific challenges for data analysis due to the curse of dimensionality. Most notably, the traditional Hoteling's $T^2$ test is not applicable to such data. Several regularized Hotelling's $T^2$ tests are presented, including the PCA-based one-class model (the SIMCA model) of van Dijk (van Dijk et al 2014]. Simple simulations in this section show that the SIMCA model had lower power than competing approaches to detect differences between an outlying genotype and the class of reference genotypes.

Besides the regularization strategies considered in this section (dimension reduction or eigenvalue-shrinkage) other strategies are available as well. Especially, when prior-knowledge is available with respect to the population covariance structure of the reference varieties, e.g. a block-diagonal structure or a generally sparse structure. Especially the test of Cai et al seems interesting in this respect since the authors claim that the method has greater power than Hoteling's $T^2$-based tests for detecting differences when the GMO differs from the references on only a few variables [Tony Cai et al 2014]. Also in other fields such as machine learning other (not necessarily statistical) methods for multivariate outlier detection have been proposed which might be of interest. Details of many of these techniques are provided in the excellent review of Zimek et al [Zimek et al 2012].

Note that the discussion in this section solely focused on making the two-sample Hoteling $T^2$-test applicable to high-dimensional data. However, most likely for detection of outlying GMs other issues will need to be considered as well. Zimek et al review some specific aspects of the curse of dimensionality for outlier detection [Zimek et al 2012], i.e. for detecting differences between a GMO and a set of reference varieties. In particular they show that outlier detection in high-dimensional data is hampered by the so-called concentration and masking effects. More specifically, the concentration effect refers to the fact that in high-dimensional data space all observations (GMO and references) will appear to be equally dissimilar. Masking means that an unintended effect in a few variables might be unobserved due to the “noise” (natural variation between references) in all other endpoints that are considered. In other words, although the measurement of a large number of endpoints using omics techniques offers great potential for identification of unintended effects, this identification is simultaneously also hampered by the fact that so many endpoints are measured. An example of this effect appears to be visible in figure 4a. As shown by Zimek et al., these effects might be reduced by application of a dimensional reduction or variable selection step, which specifically tries to find the subspace in which the outlier genotype differs from the reference samples [Zimek et al 2012]. Also in Engel et al a sparse Hoteling’s $T^2$-test was proposed which combines estimation of Hoteling’s $T^2$ with variable selection [Engel et al 2017].

Besides potentially reducing the influence of the concentration and masking effects on outlier detection an additional advantage of variable selection is improved interpretation. Generally, understanding why a GM differs from the references is not straightforward for multivariate models. Clearly, using variable-selection the variables that are most associated with the differences between the GM and reference varieties can be identified. These might provide insight into the specific unintended effect(s) that were introduced by the GM, for example by mapping altered metabolites to biological pathways. Also for PCA-based methods such as SIMCA many approaches have been developed to identify the variables mainly associated to an outlier [Qin 2003].

# 4 Discussion and directions for future research

Theoretical considerations, review of literature and simulation studies reported here lead us to the following considerations and directions for future research:

1. The proper framework for non-targeted food safety assessments ('proofs of safety') is equivalence testing. The traditional strategy of first performing difference tests and only investigating equivalence if a significant difference has been found is incorrect, because it remains unknown if a difference of a potentially relevant size can be detected. The most relevant type of error is to declare that no difference has been found (and therefore not take any further action) when in effect there is a deviation that cannot be guaranteed to be safe. Therefore the primary test should be an equivalence test rather than a difference test.
2. There is no reason why the fundamental statistical approach should be different for data obtained from plant studies or from animal studies, therefore equivalence testing is the correct framework in both cases.
3. In G-TwYST an equivalence testing approach was proposed in the context of animal study data by fitting a bandwidth to historical non-GM reference data. This approach can be used similarly, or with a few adaptations, for plant study data.
4. There is however a difference between safety assessments for any single endpoint and safety assessments based on high-dimensional data (such as omics data). The G-TwYST equivalence testing method has been developed for single variables.
5. The SIMCA one-class model was developed in the GRACE project for high-dimensional data. The originally published version (van Dijk et al. 2014) differs sligthly from the later version as used in the final GRACE publications (Kok et al, subm., Corujo et al. subm.) Both versions of the one-class model used in GRACE follow the same underlying philosophy as the G-TwYST equivalence testing by fitting a bandwidth to historical non-GM reference data.
6. Nevertheless, the SIMCA one-class model approach is not an equivalence test. It is based on standard software for classification (Pirouette), leading to in/out classifications rather than proofs of safety.
7. In this report, some alternative statistical models are described. These methods include variable-wise (multiple-testing) approaches and multivariate approaches as applied in multivariate statistical process control (MSPC). In MSPC there is a similar situation as in food safety assessments, with historical data that are assumed to be under control, and new data that have to be tested.
8. Variable-wise approaches lead to rectangular equivalence regions with an easy interpretation. Multivariate approaches lead, under the common assumption of multinormal distributions, to ellipsoid equivalence regions which are expected to be closer to patterns observed in historical datasets than rectangular regions.

9. There are many types of multivariate approaches. The SIMCA one-class model developed in GRACE belongs to the category of PCA-based models. Alternative models are shrinkage-based models, and models based on specific assumed structures in the covariance matrix, e.g. block-wise correlation or a high frequency of zero elements. Such specific assumptions may be based on system biological modelling of adverse outcome pathways.
10. Simple simulations using rough versions of some of the models reported in this study show that in many cases the Pirouette and GRACE one-class models had lower power than some competing approaches to detect differences between an outlying genotype and the class of reference genotypes. The power to detect differences of a certain magnitude is expected to be strongly related to the power to show equivalence given a difference of the same magnitude for theoretical reasons.
11. In general, the unavoidable price of considering many endpoints at the same time (untargeted analysis) as opposed to considering just one or a few endpoints (targeted analysis) is a decrease in power for difference tests, or the need to set wider equivalence regions for equivalence tests which should have a pre-set power. With appropriate methods this effect can be less severe for strongly correlated endpoints.
12. Additional work is needed to
    a. Connect the multivariate models for MSPC, which currently seem to focus on detecting outliers with difference tests, to equivalence testing, potentially using the methods of fiducial inference that also proved to be useful for the G-TwYST equivalence testing approach.
    b. Characterise the covariance structure of omics datasets: are there strong or only minor correlations? If so, are correlations grouped in blocks of variables? Can we assume many correlations to be zero?
    c. Investigate the nature of potential unintended effects: are they supposed to be deviations in a single variable, in a small group of variables, or across large groups of variables.
    d. Investigate the behaviour of methods under different conditions of data availability (number of reference genotypes, number of variables) and internal structure (expected correlation patterns in omics datasets).
    e. Compare the different models to find an optimal approach for practical safety assessment.

## References

Altman, D. G., & Bland, J. M. (1995). Statistics notes: Absence of evidence is not evidence of absence. *Bmj*, *311*(7003), 485.

Bai, Z., & Saranadasa, H. (1996). Effect of high dimension: by an example of a two sample problem. *Statistica Sinica*, 311-329.

Berger, R. L., & Hsu, J. C. (1996). Bioequivalence trials, intersection-union tests and equivalence confidence sets. *Statistical Science*, *11*(4), 283-319.

Bun, J., Bouchaud, J. P., & Potters, M. (2017). Cleaning large correlation matrices: tools from random matrix theory. *Physics Reports*, *666*, 1-109.

Cai, T., Liu, W., & Xia, Y. (2014). Two-sample test of high dimensional means under dependence. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *76*(2), 349-372.

Chen, S. X., & Qin, Y. L. (2010). A two-sample test for high-dimensional data with applications to gene-set testing. *The Annals of Statistics*, *38*(2), 808-835.

Corujo M, Pla M, van Dijk J, Voorhuijzen M, Staats M, Slot M, Lommen A, Barros E, Nadal A, Puigdomènech P, La Paz JL, van der Voet H, Kok E (submitted). Use of extensive analytical methods in the study of genetically modified maize varieties tested in 90 days feeding trials.

Dempster, A. P. (1958). A high dimensional two sample significance test. *The Annals of Mathematical Statistics*, 995-1010.

Dempster, A. P. (1960). A significance test for the separation of two highly multivariate small samples. *Biometrics*, *16*(1), 41-50.

Engel, J., Buydens, L., & Blanchet, L. (2017). An overview of large-dimensional covariance and precision matrix estimators with applications in chemometrics. *Journal of Chemometrics*, *31*(4).

Engel, J., Blanchet, L., Engelke, U. F. H., Wevers, R. A., & Buydens, L. M. C. (2017). Sparse statistical health monitoring: A novel variable selection approach to diagnosis and follow-up of individual patients. *Chemometrics and Intelligent Laboratory Systems*, *164*, 83-93.

Friedman, J., Hastie, T., & Tibshirani, R. (2001). *The elements of statistical learning* (Vol. 1, pp. 337-387). New York: Springer series in statistics.

Goedhart, P.W. & van der Voet, H. (2017). G TwYST Study B. a 90-day toxicity study in rats fed GM maize NK603. Statistical report. Report 31.10.17, Biometris, Wageningen, The Netherlands.

Goedhart, P.W. & van der Voet, H. (2018a). G TwYST Study A. Combined chronic toxicity and carcinogenicity study in rats fed GM maize NK603. Main statistical report. Report 32.02.18, Biometris, Wageningen, The Netherlands.

Goedhart, P.W. & van der Voet, H. (2018b). G TwYST Study A. Combined chronic toxicity and carcinogenicity study in rats fed GM maize NK603. Statistical report, 3 months. Report 33.02.18, Biometris, Wageningen, The Netherlands.

Goedhart, P.W. & van der Voet, H. (2018c). G TwYST Study A. Combined chronic toxicity and carcinogenicity study in rats fed GM maize NK603. Statistical report, 6 months. Report 34.02.18, Biometris, Wageningen, The Netherlands.

Goedhart, P.W. & van der Voet, H. (2018d). G TwYST Study A. Combined chronic toxicity and carcinogenicity study in rats fed GM maize NK603. Statistical report, 12 months. Report 35.02.18, Biometris, Wageningen, The Netherlands.

Goedhart, P.W. & van der Voet, H. (2018e). G TwYST Study A. Combined chronic toxicity and carcinogenicity study in rats fed GM maize NK603. Statistical report, 24 months. Report 36.02.18, Biometris, Wageningen, The Netherlands.

Goedhart, P.W. & van der Voet, H. (2018f). G TwYST Study C. a 90-day toxicity study in rats fed GM maize NK603. Statistical report. Report 37.03.18, Biometris, Wageningen, The Netherlands.

Goeman, J. J., & Solari, A. (2014). Multiple hypothesis testing in genomics. *Statistics in medicine*, *33*(11), 1946-1978.

Henry, R. C., Park, E. S., & Spiegelman, C. H. (1999). Comparing a new algorithm with the classic methods for estimating the number of factors. *Chemometrics and Intelligent Laboratory Systems*, *48*(1), 91-97.

Hoffelder, T., Gössl, R., & Wellek, S. (2015). Multivariate equivalence tests for use in pharmaceutical development. *Journal of biopharmaceutical statistics*, *25*(3), 417-437.

Hong, B.; Du, Y.Z.; Mukerji, P.; Roper, J.M.; Appenzeller, L.M. Safety assessment of food and feed from GM crops in Europe: Evaluating EFSA’s alternative framework for the rat 90-day feeding study. J. Agric. Food Chem. 2017, 65, 5545-5560.

Infometrix, Inc. Pirouette Multivariate Data analysis Software version 4.5.

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An introduction to statistical learning* (Vol. 112). New York: springer.

Kang, Q.; Vahl, C.I. Statistical analysis in the safety evaluation of genetically-modified crops: equivalence tests. Crop Sci. 2014, 54, 2183-2200.

Kent, J. T., Bibby, J. M., & Mardia, K. V. (2006). Multivariate analysis (probability and mathematical statistics).

Kok E, van Dijk J, Voorhuijzen M, Staats M, Slot M, Lommen A, Venema A, Pla M, Corujo M, Barros E, Hutten R, Jansen J, van der Voet H (submitted). Omics analyses of potato plant materials using an improved one class classification tool to identify aberrant compositional profiles in risk assessment procedures.

Lakens, D. (2017). Equivalence tests: a practical primer for t tests, correlations, and meta-analyses. *Social Psychological and Personality Science*, *8*(4), 355-362.

Ledoit, O., & Wolf, M. (2004). A well-conditioned estimator for large-dimensional covariance matrices. *Journal of multivariate analysis*, *88*(2), 365-411.

Ledoit, O., & Wolf, M. (2012). Nonlinear shrinkage estimation of large-dimensional covariance matrices. *The Annals of Statistics*, *40*(2), 1024-1060.

Ledoit, O., & Wolf, M. (2017). Direct nonlinear shrinkage estimation of large-dimensional covariance matrices.

Ledoit, O., & Wolf, M. (2018). Optimal estimation of a large-dimensional covariance matrix under Stein's loss. *Bernoulli*, *24*(4B), 3791-3832.

Limentani, G. B., Ringo, M. C., Ye, F., Bergquist, M. L., & MCSorley, E. O. (2005). Beyond the t-test: statistical equivalence testing.

Meinshausen, N. (2008). Hierarchical testing of variable importance. *Biometrika*, *95*(2), 265-278.

Munk, A., & Pflüger, R. (1999). 1—ɑ Equivariant Confidence Rules for Convex Alternatives are ɑ/2–Level Tests—With Applications to the Multivariate Assessment of Bioequivalence. *Journal of the American Statistical Association*, *94*(448), 1311-1319.

Pallmann, P., & Jaki, T. (2017). Simultaneous confidence regions for multivariate bioequivalence. *Statistics in medicine*, *36*(29), 4585-4603.

Paul, D., & Aue, A. (2014). Random matrix theory in statistics: A review. *Journal of Statistical Planning and Inference*, *150*, 1-29.

Qin, S.J. (2003). Statistical process monitoring: basics and beyond. *Journal of Chemometrics*, *17*(8-9), 480-502.

Saccenti, E., Hoefsloot, H. C., Smilde, A. K., Westerhuis, J. A., & Hendriks, M. M. (2014). Reflections on univariate and multivariate analysis of metabolomics data. *Metabolomics*, *10*(3), 361-374.

Schäfer, J., & Strimmer, K. (2005). A shrinkage approach to large-scale covariance matrix estimation and implications for functional genomics. *Statistical applications in genetics and molecular biology*, *4*(1).

Srivastava, M. S., & Du, M. (2008). A test for the mean vector with fewer observations than the dimension. *Journal of Multivariate Analysis*, *99*(3), 386-402.

Ullah, I., Pawley, M. D., Smith, A. N., & Jones, B. (2017). Improving the detection of unusual observations in high-dimensional settings. *Australian & New Zealand Journal of Statistics*, *59*(4), 449-462.

Vahl, C. I., & Kang, Q. (2016). Equivalence criteria for the safety evaluation of a genetically modified crop: a statistical perspective. *The Journal of Agricultural Science*, *154*(3), 383-406.

van Dijk JP, Souza de Mello C, Voorhuijzen MM, Hutten RCB, Maisonnave Arisi AC, Jansen JJ, Buydens LMC, van der Voet H, Kok EJ (2014). Safety assessment of plant varieties using transcriptomics profiling and a one-class classifier. *Regulatory Toxicology and Pharmacology*, 70: 297-303. http://dx.doi.org/10.1016/j.yrtph.2014.07.013

van der Voet H (2018). Safety assessments and multiplicity adjustment: comments on a recent paper. Journal of Agricultural and Food Chemistry, 66: 2194-2195. https://doi.org/10.1021/acs.jafc.7b03686

van der Voet, H., Perry, J. N., Amzal, B., & Paoletti, C. (2011). A statistical assessment of differences and equivalences between genetically modified and reference plant varieties. *BMC Biotechnology*, *11*(1), 15. https://dx.doi.org/10.1186/1472-6750-11-15

van der Voet H, Goedhart PW, Schmidt K (2017). Equivalence testing using existing reference data: an example with genetically modified and conventional crops in animal feeding studies. *Food and Chemical Toxicology*, 109: 472-485. https://doi.org/10.1016/j.fct.2017.09.044

Wellek, S. (2010). *Testing statistical hypotheses of equivalence and noninferiority*. CRC Press.

Wellek, S. (2011). On easily interpretable multivariate reference regions of rectangular shape. *Biometrical Journal*, *53*(3), 491-511.

Westerhuis JA et al. (2008). Assessment of PLSDA cross validation. Metabolomics, 4:81-89.

Zhang, J., & Pan, M. (2016). A high-dimension two-sample test for the mean using cluster subspaces. *Computational Statistics & Data Analysis*, *97*, 87-97.

Zimek, A., Schubert, E., & Kriegel, H. P. (2012). A survey on unsupervised outlier detection in high-dimensional numerical data. *Statistical Analysis and Data Mining: The ASA Data Science Journal*, *5*(5), 363-387.