Multiple two-sample testing under arbitrary covariance dependency with an application in imaging mass spectrometry

Vladimir Vutov | Thorsten Dickhaus

Abstract
Large-scale hypothesis testing has become a ubiquitous problem in high-dimensional statistical inference, with broad applications in various scientific disciplines. One relevant application is constituted by imaging mass spectrometry (IMS) association studies, where a large number of tests are performed simultaneously in order to identify molecular masses that are associated with a particular phenotype, for example, a cancer subtype. Mass spectra obtained from matrix-assisted laser desorption/ionization (MALDI) experiments are dependent, when considered as statistical quantities. False discovery proportion (FDP) estimation and control under arbitrary dependency structure among test statistics is an active topic in modern multiple testing research. In this context, we are concerned with the evaluation of associations between the binary outcome variable (describing the phenotype) and multiple predictors derived from MALDI measurements. We propose an inference procedure in which the correlation matrix of the test statistics is utilized. The approach is based on multiple marginal models. Specifically, we fit a marginal logistic regression model for each predictor individually. Asymptotic joint normality of the stacked vector of the marginal regression coefficients is established under standard regularity assumptions, and their (limiting) correlation matrix is estimated. The proposed method extracts common factors from the resulting empirical correlation matrix. Finally, we estimate the realized FDP of a thresholding procedure for the marginal p-values. We demonstrate a practical application of the proposed workflow to MALDI IMS data in an oncological context.

Keywords
false discovery proportion, imaging mass spectrometry, logistic regression, matrix-assisted laser desorption/ionization, multiple marginal models, multiple testing, simultaneous statistical inference
Imaging mass spectrometry (IMS) is a technique that acquires spatially resolved mass spectral information of small to large molecules. Provided a thin tissue section, mass spectra are collected in a spatially orientated pattern within the tissue. This produces an image, where each discrete spot represents a mass spectrum. Mass spectra associate molecular masses to their relative molecular abundances. Hence, this provides insights into the chemical decomposition of a unique and specific region in the tissue. A promising technology that has evolved over the recent years is matrix-assisted laser desorption/ionization (MALDI) IMS, also known as MALDI imaging. This technology allows for analyzing a wide range of analytes (e.g., proteins, peptides, lipids) from many types of biological samples. MALDI imaging is a versatile tool and has the advantage of combining spatial and molecular information from biological samples. This makes the technology interesting for biomedical and cancer research (for more applications in pathology, see Aichler & Walch, 2015; Kriegsmann et al., 2015, among others). The latter is possible by virtue of its applicability to analyzing formalin-fixed paraffin-embedded (FFPE) tissue samples. One of the key benefits of utilizing MALDI IMS on fixed samples is that multiple FFPE core biopsies can be arranged in a single tissue microarray (TMA) block (see, Boskamp et al., 2017; Poté et al., 2013). In other words, multiple tumor cores from different patients can be examined simultaneously (for more details, see Behrmann et al., 2018).

As pointed out by Boskamp et al. (2017), modern MALDI-IMS instruments manage to acquire molecular information with a small signal-to-noise ratio at short time measurements.

A challenging task for advanced bioinformatics tools, as acknowledged by Alexandrov (2012), is stable feature extraction or, in other words, extracting biologically meaningful evidence out of a huge amount of spectra. Statistically, we model each spectrum individually as it is measured from small tissue core regions with slight fluctuating structure within a single core. This is a standard practice in modeling MALDI data (cf. Behrmann et al., 2018; Boskamp et al., 2017; Leuschner et al., 2019).

Large-scale multiple testing is a widely used methodology in the analysis of high-dimensional data and has a variety of applications in scientific fields like, for example, genomics, proteomics, and brain–computer interfacing (for more life science applications, cf. chapters 9–12 in Dickhaus, 2014). Starting with the highly influential work by Benjamini and Hochberg (1995), control of the expected proportion of false positive findings, called false discovery rate (FDR), has become a standard type I error criterion in large-scale multiple testing. Another well-known technique to control the FDR has been proposed by Storey (2002), and is often referred to as Storey’s procedure. Its main idea is to fix a rejection threshold value $t$ for the marginal $p$-values, then to estimate the FDR of the resulting thresholding procedure, and finally to choose $t$ such that the estimated FDR is lower than or equal to the predefined FDR level $\alpha$. Early FDR research has mainly established FDR control of the aforementioned procedures in the case of independent test statistics. However, high-dimensional studies seldom involve the analysis of independent variables. In contrast, most studies involve many related variables simultaneously (cf., among many others, Friguet et al., 2009; Stange et al., 2016). Similarly, MALDI-IMS data consist of a couple of thousands of variables, and many of them are related. Explicitly taking into account these dependencies can increase the power of the multiple test, cf. Dickhaus et al. (2021) for an overview of so-called multivariate multiple tests.

There are multivariate multiple tests, which are based on block structures in the data. For instance, Stange et al. (2016) have proposed to control the family-wise error rate (FWER) in blocks of adjacent genetic markers; see also section 5 in Stange et al. (2016). Likewise, Stevens et al. (2017) have reported an extensive study to compare different controlling methods based on the assumption of block-correlation positively dependent tests. However, there is no evidence that MALDI-IMS data can be grouped straightforwardly into adjacent blocks. Other methods utilize a multifactor model in order to describe the dependencies among the test statistics, meaning that the latter dependency structure may be explained by latent factors.

In addition to modeling the dependencies, a further task is to integrate the correlation effects in the decision process; see, for example, Efron (2007, 2010) and Leek and Storey (2008). Fan et al. (2012) have introduced a general setting for approximating the false discovery proportion (FDP). They have assumed that the test statistics are (approximately) following a multivariate normal distribution with an arbitrary and known covariance matrix. The idea of their approach is to carry out a spectral decomposition of the covariance matrix of the test statistics, and then to subtract the principal factors that cause the strong dependency across the z-values before evaluating the FDP. This method is called principal factor approximation (PFA). Fan and Han (2017) have established a fully data-driven process to estimate the FDP, where the authors adopted a POET estimator (see Fan et al., 2013) to estimate an unknown covariance matrix, and subsequently to compute the realized FDP. Recently, Fan et al. (2019) have addressed the problem when the assumption of normality
is violated, for instance, in the context of multiple testing under arbitrary dependency and heavy-tailed data. The method utilizes a robust covariance estimator and constructs factor-adjusted test statistics.

Since the FDR is the expected value of the FDP, FDP control and FDR control are related concepts. In the present work, we explore the problem of two-sample multiple hypotheses testing under arbitrary correlation dependency under the scope of multiple marginal logistic regression models by making use of PFA. We propose to estimate the realized value of the FDP of a thresholding procedure in this context, and to choose the rejection threshold such that this estimated FDP value equals a given constant $\alpha \in (0, 1)$. Furthermore, we apply our proposed method to MALDI imaging data.

One common approach for extracting meaningful variables in the MALDI context is based on discovering significant signal peaks, which is also known as peak detection (cf. Yang et al., 2009). These peaks are anticipated to help distinguish spectra from different cancerous classes. For example, Wijetunge et al. (2015) introduced a peak detection method that carries out dual-tree complex wavelet transformation. Moreover, a novel spatial approach has been proposed by Lieb et al. (2020). Namely, the latter study proposed to incorporate the neighboring information (an isotope pattern) around the selected peaks, which can (potentially) enhance the peak picking process (a so-called “spatially-aware approach”). In the isotope context, for example, in Slawski et al. (2012), a methodology has been proposed, which discovers this type of pattern in MALDI data. For the aforementioned studies, the (column) indexes of the peaks (i.e., the indexes of explanatory variables) are important for their estimation procedures. In methods of that type, based on selected peaks, further analysis is performed in order to find a subset of those peaks that are statistically related to the response variable. This has been pointed out by, for instance, Boskamp et al. (2017).

On the other hand, there are also other methods for MALDI data analysis, which use “more of the intrinsic information hidden in the data by exploring statistical correlations” (Boskamp et al., 2017; cf. also Deininger et al., 2010; Hanselmann et al., 2008; Jones et al., 2011). Since the spectral data are nonnegative, a popular approach is the usage of nonnegative matrix factorization. Methods of that type aim at representing the original data in a lower dimensional space by combining “characteristic spectral patterns” (cf. Boskamp et al., 2017; Leuschner et al., 2019) given the nonnegativity constraint. The latter studies proposed so-called “automated tumour typing”: After mapping the original data into a lower dimensional space, supervised classification methods are carried out on the resulting feature vectors in order to classify observational units into tumor subtypes. These methods are invariant with respect to the permutation of the indexes of the explanatory variables, but rather they exploit similarities among these variables. Our proposed methodology is more in the spirit of these latter methods, which are invariant with respect to the indexes of the explanatory variables and instead use statistical similarities (in our case quantified by correlations) to reduce the (effective) dimensionality of the feature space.

The remainder of the work is structured as follows. In Section 2, we describe the proposed two-sample multiple testing framework. Section 3 is devoted to computer simulations. We demonstrate the practical application to a MALDI-IMS study in Section 4, and we conclude with a discussion in Section 5.

2 | PROPOSED METHODOLOGY

In MALDI imaging related studies, data are commonly stored in an $n \times p$ matrix $X = (x_{ij})_{1 \leq i \leq n, 1 \leq j \leq p}$, where spectra are stored as rows and columns correspond to mass-to-charge ratio (m/z) values (in the context of MALDI interpreted as molecular masses, cf. Alexandrov, 2012). Usually, both $n$ and $p$ are in thousands. We address the biological question of the association between m/z values and a cancerous status by testing multiple hypotheses. More specifically, we test for each $j$ the null hypothesis $H_{0j}$, which states there is no association between a particular m/z value $X_j$ and the cancer subtype.

2.1 | Marginal modeling

The first step of the proposed framework is to model the marginal associations, for each $j$ separately. Therefore, let $j$ be arbitrary, but fixed throughout this section. The motivation for this marginal modeling is that the number $p$ of potential regressors is very large in our context, such that the available sample size $n$ does not allow for fitting a multivariate model, which incorporates all regressors in one model. Screening for potentially relevant regressors by means of marginal logistic regression models is common practice as remarked, for instance, in the section titled “Variables inclusion and selection” of Sperandei (2014). Marginal logistic regression models have also been considered and analyzed by Pepe et al. (1999). Let $Y$ denote the (random) binary outcome and let $X_{j}$ denote the random variable describing the $j$-th m/z value. Thus, the tuple
(X_j, Y) takes its values in $\mathbb{R} \times \{0, 1\}$. We are interested in the conditional distribution $P(Y \mid X_j)$. To this end, we assume a (marginal) binary regression model with the canonical (logit) link function. This model has two parameters, namely, the intercept $\alpha_j$ and the regression coefficient $\beta_j$. We denote the observational units for the $j$-th marginal regression problem by $(X^{(i)}_j, Y^{(i)})_{1 \leq i \leq n}$, and we assume that they are independent copies of $(X_j, Y)$. Letting, for a given $i \in \{1, \ldots, n\}$, $\pi^{(i)}_j = P(Y^{(i)} = 1 \mid X^{(i)}_j)$, the model equation for the $j$-th binary logistic regression models is given by

$$
g(\pi^{(i)}_j) : = \log \left( \frac{\pi^{(i)}_j}{1 - \pi^{(i)}_j} \right) = \alpha_j + X^{(i)}_j \beta_j, \quad (1)$$

where $g = \text{logit}$ is the canonical link function mentioned before. The unknown parameters $(\alpha_j, \beta_j)$ are estimated by the principle of the maximum (log-) likelihood. The log-likelihood function pertaining to the model in (1) is given by

$$
l(\alpha_j, \beta_j) = \sum_{i=1}^{n} Y^{(i)} \left[ \log \pi^{(i)}_j - \log(1 - \pi^{(i)}_j) \right] + \log(1 - \pi^{(i)}_j). \quad (2)$$

By substituting

$$
\pi^{(i)}_j = \frac{\exp(\alpha_j + X^{(i)}_j \beta_j)}{1 + \exp(\alpha_j + X^{(i)}_j \beta_j)} \quad \text{as well as} \quad 1 - \pi^{(i)}_j = \frac{1}{1 + \exp(\alpha_j + X^{(i)}_j \beta_j)}
$$

in (2), we obtain that

$$
l(\hat{\alpha}_j, \hat{\beta}_j) = \max_{(\alpha_j, \beta_j)} \sum_{i=1}^{n} Y^{(i)} \left( \alpha_j + X^{(i)}_j \beta_j \right) - \log \left( 1 + \exp(\alpha_j + X^{(i)}_j \beta_j) \right), \quad (3)$$

where the estimation is performed conditionally to the actually observed values $X^{(i)}_j = x^{(i)}_j$ for $1 \leq i \leq n$.

In this study, we are concerned with simultaneous testing of the pairs of hypotheses

$$
H_{0j} : \beta_j = 0 \text{ versus } H_{1j} : \beta_j \neq 0, j = 1, \ldots, p. \quad (4)
$$

This means, that we test a family $H_p = \{H_{01}, \ldots, H_{0p}\}$ of $p$ null hypotheses, and a binary decision (rejection or nonrejection) is made for each of these $p$ null hypotheses on the basis of the study data at hand. The result of the data analysis therefore is a binary decision vector $d \in \{0, 1\}^p$. The $j$-th entry $d_j$ of this decision vector encodes the decision referring to the (marginal) test problem $H_{0j}$ versus $H_{1j}$ for $1 \leq j \leq p$, where $d_j = 1$ means (by convention) that $H_{0j}$ is rejected in favor of $H_{1j}$, while $d_j = 0$ means that $H_{0j}$ is not rejected. Biologically speaking, we aim at discovering the most distinctive m/z values for a cancer association.

### 2.2 Multiple marginal models

The second step of the proposed procedure is to combine all $p$ marginal models and to approximate the joint null distribution of all estimators. To this end, we follow the framework described by Pipper et al. (2012) for jointly estimating multiple marginal association parameters, and apply this framework to the marginal models described in the previous section. Notice that we assume that regression coefficients are unique to one model $j$ and not shared between any two models $j_1 \neq j_2$. Furthermore, the intercepts $(\alpha_j)_{1 \leq j \leq p}$ are nuisance parameters in the sense that the hypotheses in (4) only refer to the $\beta_j$s.

The main goal of this section is to establish a central limit theorem for the vector $\hat{\beta} = (\hat{\beta}_1, \ldots, \hat{\beta}_p)^T$, which is achieved by stacking the score contributions of the $\hat{\beta}_j$s across all $p$ marginal models. Following Pipper et al. (2012), we consider the
asymptotic \((n \to \infty)\) expansion

\[
(\hat{\beta}_j - \beta_j) \sqrt{n} = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \Psi_{ij} + o_p(1),
\]

which follows from \((5)\) under standard regularity assumptions like, for instance, finiteness of the Fisher information and nonvanishing (limiting) proportion of data points corresponding to \(Y = 1\) and \(Y = 0\), respectively. The left-hand side of \((7)\) converges in distribution, by the multivariate central limit theorem, to a \(p\)-variate normal distribution, that is,

\[
(\hat{\beta} - \beta) \sqrt{n} \overset{d}{\to} N_p(0, \Sigma).
\]

The limiting variance–covariance matrix \(\Sigma\) can be estimated in a consistent manner, namely, by

\[
\hat{\Sigma} = \frac{1}{n} \sum_{i=1}^{n} \hat{\Psi}_i \hat{\Psi}_i^\top,
\]

whence \(\hat{\Psi}_i\) are yielded by plugging the parameter estimates from all marginal models into \(\Psi_i\).

In our study, we are genuinely interested in the effect of \(\beta_j\) for \(j \in \{1, \ldots, p\}\). However, the intercepts \((\alpha_j)_{1 \leq j \leq p}\) contribute to the estimation and standardization of the \(\beta_j\)’s. Specifically, for the logit model described in the previous section, \(\hat{\Psi}_{ij}\) is given by the second coordinate of the bivariate vector

\[
\left\{ \pi_j(1 - \pi_j)(1, X_j(1) - (1, X_j^T(1))^{-1}(1, X_j^T(1)) (Y(i) - \pi_j^0) \right\}
\]

where \(\pi_j^0 = \frac{\exp(\hat{\alpha}_j + X_j^T(\hat{\beta}))}{1 + \exp(\hat{\alpha}_j + X_j^T(\hat{\beta}))}\) and \(\hat{\alpha}_j, \hat{\beta}_j\) are as in \((3)\).
TABLE 1 Decision pattern of the multiple test, which thresholds the marginal p-values at a given value $t \in [0,1]$

| Number         | Number accepted | Number rejected | Overall |
|----------------|-----------------|-----------------|---------|
| True nulls     | $U(t)$          | $V(t)$          | $p_0$   |
| False nulls    | $T(t)$          | $S(t)$          | $p_1$   |
| All nulls      | $p - R(t)$      | $R(t)$          | $p$     |

Next, we denote by $Z_1, \ldots, Z_p$ the Studentized versions of $\hat{\beta}_1, \ldots, \hat{\beta}_p$, meaning that

$$Z_j = \frac{\hat{\beta}_j}{\sqrt{\text{Var}(\hat{\beta}_j)}}, \quad j = 1, \ldots, p,$$

where $\sqrt{\text{Var}(\hat{\beta}_j)}$ is the square root of the $j$-th diagonal element of $\hat{\Sigma}$, divided by $\sqrt{n}$. Then, we have that

$$(Z_1, Z_2, \ldots, Z_p)^T \overset{\text{approx.}}{\sim} N_p((\mu_1, \mu_2, \ldots, \mu_p)^T, \hat{\Sigma}^*),$$

where $\mu_j = \beta_j / \sqrt{\text{Var}(\hat{\beta}_j)}$ for $1 \leq j \leq p$, $\hat{\Sigma}^* = \text{diag}[\hat{\Sigma}]^{-1/2} \cdot \hat{\Sigma} \cdot \text{diag}[\hat{\Sigma}]^{-1/2}$ is the correlation matrix pertaining to $\hat{\Sigma}$, and the notation $\sim$ indicates the approximate distribution for large $n$. Assuming that $\sqrt{\text{Var}(\hat{\beta}_j)}$ is positive for all $j \in \{1, \ldots, p\}$, $\hat{\beta}_j = 0$ if and only if $\mu_j = 0$. Thus, the family of hypotheses from (4) can then equivalently be expressed as

$$H_{0j} : \mu_j = 0 \text{ versus } H_{1j} : \mu_j \neq 0, j = 1, \ldots, p,$$

although $\mu_j$ depends on the data via $\sqrt{\text{Var}(\hat{\beta}_j)}$ and is therefore not a statistical parameter.

2.3 Approximation of the FDP

Throughout this paper, we consider the multiple test problem, which is given by the $p$ pairs of null and alternative hypotheses specified in (11). Let $p_0 = \#\{j : \mu_j = 0\}$ denote the number of true null hypotheses and $p_1 = \#\{j : \mu_j \neq 0\}$ the number of false null hypotheses, such that $p = p_0 + p_1$. Throughout the remainder, we make the following sparsity assumption.

Assumption 2.1. The number $p_1$ of false null hypotheses is very small compared to the number $p$ of all null hypotheses.

In an asymptotic setting where $p = p(n)$ tends to infinity as $n$ tends to infinity, Assumption 2.1 can be formalized as $p_0(n)/p(n) \rightarrow 1$. In the present work, however, we rely only on asymptotics in the sample size $n$ and regard $p$ as fixed. For the calibration of a multiple test with respect to type I error control, we proceed similarly as in Storey’s method (see Storey, 2002). Namely, for a (data-dependent) threshold $t$, we will reject the null hypotheses, which correspond to those $p$-values that are not exceeding $t$. This approach has been broadly used in practice (e.g., see, Efron, 2007, 2010; Fan et al., 2012; Fan & Han, 2017; Storey, 2002). The aim of the proposed method is to estimate the realized FDP for any fixed $t$ in the multiple testing setting given by (11), based on the $Z$-statistics (10) under an arbitrary structure of $\Sigma$.

To this end, we consider empirical processes given by

$$V(t) = \#\{\text{true null } P_j : P_j \leq t\},$$
$$S(t) = \#\{\text{false null } P_j : P_j \leq t\},$$
$$R(t) = \#\{P_j : P_j \leq t\},$$

where $t$ ranges in $[0,1]$. For a given value of $t$, the null hypothesis $H_{0j}$ is rejected if and only if its corresponding $p$-value $p_j$ does not exceed $t$. This decision rule leads to the decision pattern, which is displayed in Table 1. The random variables $V(t)$, $S(t)$, and $R(t)$ are the number of false discoveries (i.e., false rejections), the number of true discoveries,
and the total number of discoveries, respectively. Clearly, \( R(t) = V(t) + S(t) \). The latter random variables depend on the test statistics \( Z_1, Z_2, \ldots, Z_p \), because every \( p \)-value \( P_j \) is a transformation of the corresponding \( Z \)-statistic \( Z_j \), \( 1 \leq j \leq p \), as we will describe below. Furthermore, \( V(t) \) and \( S(t) \) are both unobservable, whereas \( R(t) \) is observable. We recall here the definition of the FDP, namely, \( \text{FDP}(t) = V(t) / \max\{R(t), 1\} \). Table 1 provides a summary of the relevant quantities.

2.4 PFA

The next step of the analysis is to model and to utilize the dependency structure of the test statistics in an approximation of FDP \( t \) for a given \( t \). The proposed technique relies on an approximation of a normally distributed random vector with a factor model involving weakly dependent, normally distributed random errors. In our case, we use the factor model as a tool to approximate the correlation matrix \( \hat{\Sigma}^* \) with a reduced number of parameters, without actually assuming that latent factors are involved in the data-generating process. To this end, we first employ a spectral decomposition of the correlation matrix \( \hat{\Sigma}^* \) (cf. Fan et al., 2012). Namely, \( \hat{\Sigma}^* \) is represented in terms of its eigenvalue–eigenvector pairs \((\lambda_j, \gamma_j)_{1 \leq j \leq p} \), where \( \lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_p \geq 0 \). The representation can be written as

\[
\hat{\Sigma}^* = \lambda_1 \gamma_1 \gamma_1^T + \lambda_2 \gamma_2 \gamma_2^T + \cdots + \lambda_p \gamma_p \gamma_p^T.
\]

For a fixed integer \( k \geq 1 \), we let \( A_k = \sum_{j=k+1}^p \lambda_j \gamma_j \gamma_j^T \), and we note that

\[
\|A_k\|^2 = \lambda_{k+1}^2 + \cdots + \lambda_p^2,
\]

where \( \| \cdot \| \) is the Frobenius norm. We further let \( L_k = (\sqrt{\lambda_1} \gamma_1, \sqrt{\lambda_2} \gamma_2, \ldots, \sqrt{\lambda_k} \gamma_k) \), which presents a \( p \times k \) matrix. Thus, \( \hat{\Sigma}^* \) can be written as

\[
\hat{\Sigma}^* = L_k L_k^T + A_k.
\]

Respectively, \( Z_1, \ldots, Z_p \) can be approximated by

\[
\mu_j + b_j^T W + K_j = \mu_j + \eta_j + K_j, \quad j = 1, \ldots, p,
\]

where \( b_j = (b_{j1}, \ldots, b_{jk})^T \) and \( (b_{1j}, \ldots, b_{pk})^T = \sqrt{\lambda_j} \gamma_j \). The vector \( W = (W_1, \ldots, W_k)^T \sim N_k(0, I_k) \) is called the vector of common factors, and these factors are stochastically independent of each other. The random vector \( (K_1, \ldots, K_p)^T \sim N_p(0, A_k) \) is called the vector of random errors, and it is assumed that factors and random errors are stochastically independent. We can think of (13) as an approximation of the data-generating process for \( Z_1, \ldots, Z_p \). In this interpretation, \( \mu_j = 0 \) corresponds to the true null hypotheses, and \( \mu_j \neq 0 \) corresponds to the false null hypotheses.

It is essential to choose the number \( k \) of factors carefully. On the one hand, it is important to choose \( k \) large enough to capture most of the dependencies among \( Z_1, \ldots, Z_p \). On the other hand, a small \( k \) stabilizes the computations, both from a numerical and from a statistical point of view. Fan et al. (2012) have discussed one way to determine a suitable value of \( k \). Concretely, the authors proposed to choose the smallest \( k \) such that

\[
\frac{\sqrt{\lambda_{k+1}^2 + \cdots + \lambda_p^2}}{\lambda_1 + \cdots + \lambda_p} < \varepsilon,
\]

where \( \varepsilon \) is some small number, for example, 0.01. It has been pointed out by Fan and Han (2017) that an overestimation of \( k \) does not invalidate the approximation of the FDP, as long as the factor approximation of \( \hat{\Sigma}^* \) can still be estimated with a reasonable accuracy.
Based on the aforementioned derivations, we consider the “principal factor” FDP estimator from Proposition 2 in Fan et al. (2012), which is given by

$$
\hat{FDP}(t) = \min \left\{ \frac{\sum_{j=1}^{p} [\Phi(a_j(z_{t/2} + \hat{\eta}_j)) + \Phi(a_j(z_{t/2} - \hat{\eta}_j))]}{R(t)} \right\} / R(t)
$$

whenever $R(t) \neq 0$, and $\hat{FDP}(t) = 0$ in the case of $R(t) = 0$. In (15), $a_j = (1 - \sum_{h=1}^{k} b_{jh}^2)^{-1/2}$ and $R(t) = \{ j : 2\Phi(-|Z_j|) \leq t \}$ is the (total) number of rejections for a given $t$, where $\Phi$ and $z_{t/2} = \Phi^{-1}(t/2)$ are the cumulative distribution function and the lower $t/2$-quantile of the standard normal distribution on $\mathbb{R}$, respectively. The (unadjusted) two-sided (random) $p$-value corresponding to $Z_j$ is given by $P_j = 2\Phi(-|Z_j|) = 2(1 - \Phi(|Z_j|))$, and this $p$-value is thresholded at $t$ for every $j \in \{1, ..., p\}$ when computing $R(t)$. Furthermore, $\hat{\eta}_j = \sum_{h=1}^{k} b_{jh} \hat{w}_j$ is an estimator for $\eta_j = b_j^\top w$, where $w = (w_1, ..., w_k)^\top$ denotes the value of $W = (W_1, ..., W_k)^\top$.

The estimator in (15) relies on the intuition that large $|\mu_j|$s tend to generate large $|z_j|$s, meaning that false null hypotheses tend to produce large $Z$-statistics (in absolute value). Furthermore, the estimator in (15) relies on Assumption 2.1, namely, that the number $p_0$ of true null hypotheses is close to $p$. This assumption justifies the summation over all $j$ from one to $p$ in (15). Different FDP estimators have been compared by Schwartzman (2012), and under sparsity in the aforementioned sense, the author has proposed to use the estimator from (15). There are several reasons why the assumption of sparsity is plausible in our study. First, due to high sensitivity during sample preparation and acquisition, there is evidence of a small signal-to-noise ratio. Second, a reasonable assumption is that solely a tiny fraction of molecular masses are distinctive for a cancer association. In fact, we have applied the proposed method to real MALDI data, where there have been characterized five biomarkers (i.e., biologically meaningful covariates) out of a couple of thousands of measured covariates.

In order to evaluate (15) in practice, it remains to specify the estimator of $w$. Fan et al. (2012) have proposed to construct this estimator by means of $L_2$-regression or by means of $L_1$-regression, respectively. For the former, the authors proposed to include only the 95% smallest $|z_j|$s in the regression fit. Specifically, the estimator based on $L_2$-regression is given by

$$
\hat{w} = \arg \min_{v \in \mathbb{R}^k} \left( \sum_{j=1}^{0.95p} (Z_j - b_j^\top v)^2 \right),
$$

where we assume that the $Z_j$s in (16) are ordered from small to large according to their absolute values. This estimator has been used in our simulation study. The estimator based on $L_1$-regression is given by

$$
\hat{w} = \arg \min_{v \in \mathbb{R}^k} \sum_{j=1}^{p} |Z_j - b_j^\top v|.
$$

We adopted $L_1$-regression rather than $L_2$-regression, because it is more robust to outliers. The consistency of $\hat{w}$ has been discussed by Fan et al. (2012) under model assumptions, which are similar to ours.

Finally, the dependency-adjusted (random) $p$-values corresponding to the $Z_j$s are given by

$$
\tilde{P}_j = 2\Phi(-|a_j(Z_j - b_j^\top \hat{w})|).
$$

The null hypothesis $H_{0j}$ from (11) gets rejected based on the observed data, if $\tilde{P}_j \leq t, 1 \leq j \leq p$. In this, the data-dependent rejection threshold is chosen as the largest value $t = t_\alpha \in [0, 1]$ such that $\hat{FDP}(t_\alpha)$ is not exceeding a predefined level $\alpha$. In practice, a (grid) search algorithm can be employed to find the value $t_\alpha$ for a given level $\alpha$.

2.5 | Schematic description of the entire data analysis workflow

Algorithm 1 provides a step-by-step description of the proposed data analysis workflow.
The Logit-PFA method

1. Fit the marginal logistic regression model with the logit link function for each $j \in \{1, ..., p\}$ separately on the basis of $(X^{(i)}_j : 1 \leq i \leq n)$ and $(Y^{(i)} : 1 \leq i \leq n)$.

2. Find the maximum-likelihood estimates for $\hat{\beta}_j$ and $\hat{\alpha}$.

3. Calculate the standardized score contributions $\hat{\Psi}_j$, based on $\hat{\beta}_j$, for $i \in \{1, ..., n\}$ and stack them on top of each other to build a vector $\hat{\Psi}_i$.

4. Calculate the estimated covariance matrix $\hat{\Sigma}$, given in (8), based on $\hat{\Psi}_i : 1 \leq i \leq n$, and obtain the correlation matrix pertaining to $\hat{\Sigma}$. Calculate the Z-statistics given in (9).

5. Choose a grid $\mathcal{G} \subset [0, 1]$ of candidate values for the rejection threshold $t$.

6. Based on the Z-statistics, evaluate $R(t)$ for each $t \in \mathcal{G}$.

7. Apply singular value decomposition to the correlation matrix pertaining to $\hat{\Sigma}$, and determine an appropriate number of factors $k$.

8. Obtain the estimate $\hat{\omega}$ of the values of the common factors by means of regression; cf. (16) and (17), respectively. Plug these factor estimates into (15), and obtain the estimate $\hat{\Psi}_i$ for each $i \in \mathcal{G}$.

9. For a given value of $\alpha \in (0, 1)$, find the largest value $t_\alpha \in \mathcal{G}$ fulfilling $\text{FDP}(t_\alpha) \leq \alpha$.

10. Obtain adjusted p-values according to (18).

11. Threshold the adjusted p-values at $t_\alpha$.

### 3 COMPUTER SIMULATIONS

In this section, we illustrate the performance of the proposed approach based on simulated data under different data-generating processes. Specifically, we consider the sample size $n = 400$, the number of false null hypotheses $p_1 = 10$, and the total number of hypotheses $p \in \{500, 1000\}$. For each combination of these parameters, 1000 simulation runs have been performed. For a given value of $p_1$, we assume without loss of generality that $\hat{\beta}_j \neq 0$ for $j \in \{1, ..., p_1\}$, while the $p_0$ true nulls with $\hat{\beta}_j = 0$ correspond to the coordinates $j \in \{p_1 + 1, ..., p\}$. We employed the least-squares estimator, defined in (16), for the estimation of the values of the common factors. For each observational unit $i \in \{1, ..., n\}$, the simulation data have been generated according to the model

$$P_\beta(Y^{(i)} = 1|X^{(i)}) = \frac{\exp \left( \sum_{j=1}^{p_1} \hat{\beta}_j X^{(i)}_j \right)}{1 + \exp \left( \sum_{j=1}^{p_1} \hat{\beta}_j X^{(i)}_j \right)}$$

for the response variable $Y^{(i)}$ given the covariate vector $X^{(i)} = (X^{(i)}_1, ..., X^{(i)}_p)$. Moreover, we have set $\beta_j = 1$ for all $j \in \{1, ..., p_1\}$. As remarked, for instance, in Section 3.2 of Wang et al. (2017), the marginal logistic regression model given by (1) is in general incorrect if the true model is a multiple logistic regression model as in (19) with continuous covariates. However, the regression coefficient pertaining to covariate $X_j$ is zero in both models if the distribution of $Y$ does not depend on $X_j$ and $X_j$ is stochastically independent of all those covariates that have an influence on $Y$. In this sense, screening for potentially relevant regressors by marginal models is still meaningful under the multiple logistic regression model given in (19). The considered data-generating distributions for the vector $X = (X_1, ..., X_p)^T$ are provided in Model 1.

**Model 1.** Scenario 1: $X_1, ..., X_p$ are stochastically independent and identically $N(0, 1)$-distributed random variables.

Scenario 2: $X_1, ..., X_p$ are jointly normally distributed on $\mathbb{R}^p$. The parameters of their $p$-variate joint normal distribution have been chosen such that each $X_j$ is marginally $N(0, 1)$-distributed, $j = 1, ..., p$. Furthermore, the correlation coefficient $\text{Corr}(X_{j1}, X_{j2}) = \rho$ for all $1 \leq j_1 < j_2 \leq p$, as well as for all $p_1 + 1 \leq j_1 < j_2 \leq p$ (Gaussian equi-correlation model). The subvector $(X_j : 1 \leq j \leq p_1)$ is stochastically independent of the subvector $(X_j : p_1 + 1 \leq j \leq p)$, to avoid spurious effects of covariates $X_j$ with $p_1 + 1 \leq j \leq p$ on the response variable, which arise from confounding of covariates $X_j$ with $1 \leq j \leq p_1$.

Scenario 3: As Scenario 2, but now $(X_j : 1 \leq j \leq p_1)$ are stochastically independent and identically $N(0, 1)$-distributed random variables.

For the simulation of correlated independent variables, we have used the function `rmvnorm` from the R (see R Development Core Team, 2021) package `mvtnorm`. Furthermore, we have used the function `pfa.test()` from Fan et al. (2016) for implementing the PFA method. In Tables 2–7, we report summaries (over the 1000 simulation runs) of $\text{FDP}(t)$, $R(t)$, and
TABLE 2  Simulation results under Scenario 1 (I). The total number of hypotheses equals $p = 500$ in the first row and $p = 1000$ in the second row; the number of factors equals $k = 10$; the rejection threshold equals $t = 10^{-4}$, except for the last column.

| Median of FDP$(t)$ | Standard error of FDP$(t)$ | Average of R$(t)$ | Standard error of R$(t)$ | Average of S$(t)$ | Standard error of S$(t)$ | Median of $t_{0.05}$ |
|---------------------|--------------------------|-------------------|-------------------------|-----------------|------------------------|---------------------|
| 0.004144            | 0.000918                 | 6.930             | 1.247                   | 6.892           | 1.236                  | 1.24e-03            |
| 0.009116            | 0.002175                 | 6.965             | 1.270                   | 6.898           | 1.228                  | 6.6e-04             |

TABLE 3  Simulation results under Scenario 1 (II). The total number of hypotheses equals $p = 500$ in the first row and $p = 1000$ in the second row; the number of factors equals $k = 10$; the rejection threshold equals $t = 0.005$.

| Median of FDP$(t)$ | Standard error of FDP$(t)$ | Average of R$(t)$ | Standard error of R$(t)$ | Average of S$(t)$ | Standard error of S$(t)$ |
|---------------------|--------------------------|-------------------|-------------------------|-----------------|------------------------|
| 0.158180            | 0.022796                 | 11.774            | 1.664                   | 9.519           | 0.632                  |
| 0.277878            | 0.045845                 | 14.062            | 2.194                   | 9.517           | 0.650                  |

TABLE 4  Simulation results under Scenario 2 (I). The total number of hypotheses equals $p = 500$; the number of factors equals $k = 1$; the rejection threshold equals $t = 10^{-4}$, except for the last column.

| $\rho$ | Median of FDP$(t)$ | Standard error of FDP$(t)$ | Average of R$(t)$ | Standard error of R$(t)$ | Average of S$(t)$ | Standard error of S$(t)$ | Median of $t_{0.05}$ |
|--------|---------------------|--------------------------|-------------------|-------------------------|-----------------|------------------------|---------------------|
| 0.2    | 0.001395            | 0.009750                 | 10.030            | 0.182                   | 2.41e-03        |                        |
| 0.5    | 0.000131            | 0.024349                 | 10.025            | 0.250                   | 7.41e-03        |                        |
| 0.8    | 0.000099            | 0.018323                 | 10.021            | 0.572                   | 3.49e-02        |                        |

$S(t)$ for fixed values of $t$, and we report the median value of $t_{0.2}$ for the common choice of $\alpha = 0.05$.

Tables 2 and 3 summarize our simulation results under Scenario 1. Here, due to joint independence of the test statistics, $t_{0.05}$ is rather small, because the “effective number of tests” (in the sense of section 3.4 in Dickhaus et al., 2021, and the references therein) equals $p$ under joint independence of the test statistics, meaning that a rather strong multiplicity correction is required. On the other hand, the standard error of FDP$(t)$ is rather small under Scenario 1, too, because the FDP concentrates well around its expectation (the FDR) under joint independence of all $p$ test statistics. The results given in Tables 4 and 5 refer to Scenario 2 of Model 1, and Tables 6 and 7 refer to Scenario 3. Under Scenarios 2 and 3, the effective number of tests is smaller than $p$ whenever $\rho > 0$, and it decreases with increasing $\rho$. Thus, $t_{0.05}$ increases with $\rho$, too. Under our Scenario 2, the considered multiple test always rejected all 10 false null hypotheses. For this reason, we do not report summaries of $S(t)$ in Tables 4 and 5. The reason for the high power of the multiple test under Scenario 2 is that the correlation among $(X_j : 1 \leq j \leq p_1)$ amplifies the signal strength for each $j \in \{1, \ldots, p_1\}$. Under Scenario 3, where the

TABLE 5  Simulation results under Scenario 2 (II). The total number of hypotheses equals $p = 1000$; the number of factors equals $k = 1$; the rejection threshold equals $t = 10^{-4}$, except for the last column.

| $\rho$ | Median of FDP$(t)$ | Standard error of FDP$(t)$ | Average of R$(t)$ | Standard error of R$(t)$ | Average of S$(t)$ | Standard error of S$(t)$ | Median of $t_{0.05}$ |
|--------|---------------------|--------------------------|-------------------|-------------------------|-----------------|------------------------|---------------------|
| 0.2    | 0.002842            | 0.014190                 | 10.082            | 0.306                   | 1.28e-03        |                        |
| 0.5    | 0.000167            | 0.028907                 | 10.059            | 0.400                   | 4.69e-03        |                        |
| 0.8    | 0.000099            | 0.024967                 | 10.028            | 0.447                   | 2.74e-02        |                        |

TABLE 6  Simulation results under Scenario 3 (I). The total number of hypotheses equals $p = 500$; the number of factors equals $k = 1$; the rejection threshold equals $t = 10^{-4}$, except for the last column.

| $\rho$ | Median of FDP$(t)$ | Standard error of FDP$(t)$ | Average of R$(t)$ | Standard error of R$(t)$ | Average of S$(t)$ | Standard error of S$(t)$ | Median of $t_{0.05}$ |
|--------|---------------------|--------------------------|-------------------|-------------------------|-----------------|------------------------|---------------------|
| 0.2    | 0.002292            | 0.012646                 | 6.915             | 1.256                   | 6.902           | 1.240                  | 2.41e-03            |
| 0.5    | 0.000221            | 0.026264                 | 6.920             | 1.313                   | 6.898           | 1.240                  | 7.41e-03            |
| 0.8    | 0.000142            | 0.021452                 | 6.907             | 1.295                   | 6.895           | 1.237                  | 3.49e-02            |
relevant regressors are stochastically independent, the power of the multiple test is smaller than under Scenario 2, such that on average only approximately 7 of the 10 false null hypotheses can be rejected by the multiple test considered in Tables 6 and 7.

### 4 | REAL DATA ANALYSIS

#### 4.1 | Description of the data set

We applied the proposed multiple testing approach to a MALDI IMS data frame introduced by Kriegsmann et al. (2016). Kriegsmann et al. (2016) have characterized five biomarkers. Broadly speaking, biomarkers are biologically meaningful molecules indicative of a distinct biological state or condition (cf. Schwamborn, 2012). Statistically speaking, biomarkers are well-identified predictors that can be used to accurately predict relevant clinical outcomes, and also, they are an apt starting point for an evaluation of any statistical model. The aforementioned data frame has been reanalyzed by several researchers; cf. Boskamp et al. (2017), Leuschner et al. (2019), and Behrmann et al. (2018). Therefore, we refer to the aforementioned references for an extensive description of the data frame. Here, we only give a brief overview of sample acquisition, data preparation, measurement, and data processing.

FFPE lung tumor tissues samples, for this study, were provided by the bank of the National Center for Tumour Diseases (NCT, Heidelberg, Germany). Cylindrical tissue cores of non–small cell lung cancer (NSCLC) were taken from 304 patients, where 168 patients were associated with primary lung adenocarcinoma (ADC), and 136 patients were associated with primary squamous cell carcinoma (SqCC). Cylindrical tissue cores of all tissue samples were collected in eight TMA blocks in total. Lung cancer is the leading reason for cancer-related deaths worldwide, with around 1.59 million reported deaths in 2012; for more concrete numbers, see, for example, Kriegsmann et al. (2016) or Reck et al. (2013). Two primary lung cancer categories are determined, namely, small cell lung cancer and NSCLC, whence the latter constituted around 85% of all cases. The two most fatal histological NSCLC entities are ADC and SqCC, compromising of approx. 50% and approx. 40% of all lung cancers, respectively. Differentiation of these two subtypes is critical for the choice of chemotherapy regimens and further test strategies.

Tissue sections were cut from all TMA blocks and treated in accordance with a previously published protocol for tryptic peptide imaging; cf. Casadonte and Caprioli (2011). MALDI data were obtained through an Autoflex speed MALDI-TOF instrument (Bruker Daltonik) in positive ion reflector mode. Spectra were measured in the mass range 500–5000 m/z at 150 μm spatial resolution using 1600 laser shots. Tumor status and typing for all cores were confirmed by standard histopathological examination; cf. Boskamp et al. (2017). Afterwards, the raw spectral data were loaded into SCiLS Lab (version 2016b, Bruker Daltonik), the standard baseline correction was performed (convolution method of 20), and total-ion-count (TIC) normalization was employed. The normalizing step is crucial in order to reduce the laboratory variation resulting from day-to-day instrument fluctuations or biological artifacts coming from sample preparation. Finally, spectral smoothing was performed to intervals of 0.4 Da (dalton) width (cf. Senko et al., 1995), and the remaining 4669 spectra were pruned to the mass range of 800 – 2500 (outside of this interval m/z values were not considered), resulting in 1699 m/z channels (columns).

In summary, we worked on an MALDI data set where all data-processing steps are based on standard protocols. It is out of the scope of this paper to compare different data-processing steps, like normalization and smoothing.

#### 4.2 | Results of the data analysis

Figure 1 displays two mass spectra from both cancer subtypes, where the m/z values are illustrated on the horizontal axes, while the vertical axes refer to the relative abundances (intensities values) of ionizable molecules. These two graphs...
FIGURE 1 Two exemplary MALDI spectra. Two unique and specific spots within a tissue. Each of these spots represents a mass spectrum.

TABLE 8 Number of rejections and estimated FDP for several plausible rejection thresholds

| Threshold t | R(t) | FDP(t) |
|------------|-----|-------|
| 8.23e-03   | 405 | 0.2304|
| 3.70e-03   | 352 | 0.1524|
| 2.48e-03   | 322 | 0.1263|
| 1.36e-03   | 291 | 0.0926|
| 9.12e-04   | 275 | 0.0746|
| 6.11e-04   | 262 | 0.0596|

represent unique and specific spots within a patient’s tissue, and correspond to two mass spectra. Therefore, we have modeled each pixel marginally to identify which m/z values (based on 0.4 DA) are distinctive for a particular cancer subtype. We refer to Behrmann et al. (2018) (see their Figure 1) for a more detailed illustration of the pipeline from a tissue to a single spectrum.

As discussed by Efron (2007) and Efron (2010), the density of the empirical distribution of all Z-values does in general not coincide with the density of the standard normal distribution on R, even if almost all p null hypotheses are true. The reason for this phenomenon, which can also be observed on our data (see Figure 2), is the presence of dependencies among the Z-statistics, as well as the presence of some extreme outliers, which presumably correspond to strong effect sizes. In particular, these effects lead to an inflation of the variance of the null distribution of the Z-statistics. However, we nevertheless have that the Z-statistics of the previously identified biomarkers lie in tails of the distribution. Namely, their Z-statistics are large in absolute value, and might be declared as statistically significant. Note that we consider an absolute value for the Z-values, since we wish to find distinctive m/z values for either cancer subtype.

To disentangle the two sources for variance inflation (correlations among the Z-statistics and extreme outliers), we employed the method described in Section 5 (specifically, around eqs. (53)–(55)) of Efron (2007). This method suggested for our data a spread of approximately 2.63 in the central part (which presumably corresponds to true null hypotheses) of the empirical Z-score distribution. Thus, following the recommendation of Efron (2007) to “empirically correct” the null distribution, we divided all Z-scores by 2.63 prior to the following steps of data analysis. Notice that this rescaling of the Z-scores is not an (essential) part of our proposed methodology, but has been applied to our real data for the sake of a better comparability with the simulation results presented in Section 3. If we would omit the rescaling, the p-values reported in Table 9 as well as the plausible rejection thresholds listed in Table 8 would both be given on a smaller scale. The test decisions, however, would stay the same of all j, because the rescaling does not alter the ordering of the Z-scores.
FIGURE 2  The empirical distribution and fitted normal density curve of the Z-values for the MALDI data. Due to dependencies among the Z-values, they are not following the theoretical $N(0, 1)$ distribution. Instead, a closer look at the empirical distribution reveals that it can best be approximated by $N(-0.03455, 6.86^2)$. Consequently, the nonadjusted $p$-values have a lot of mass around zero.

FIGURE 3  The approximated number of false discoveries as well as the approximated FDP as functions of the total number of rejections. Each curve corresponds to a different choice of the number $k$ of common factors, where $k \in \{6, 7, 8, 9, 10\}$ has been considered.

The next step of the data analysis has been to determine an appropriate number $k$ of common factors. To this end, we performed the proposed data analysis workflow described in Algorithm 1 over a range of different candidate values for $k$ and compared the results. As documented in Figure 3, the estimated number of false discoveries as well as the estimated FDP are rather stable for $k \in \{6, \ldots, 9\}$. Based on this and for stability reasons, we chose $k = 6$ for our actual data analysis.

The main results of our real data analysis are illustrated in Figure 4. It is evident that $R(t), \hat{V}(t),$ and $\hat{FDP}(t)$ are increasing in the rejection threshold $t$. For $t \in [6 \times 10^{-4}, 8 \times 10^{-3}]$, the estimated FDP lies between approximately 6% and approxi-
mately 23%. This indicates that most of the smallest p-values correspond to false nulls (leading to true discoveries). Table 8 lists the total number of rejections as well as the estimated FDP for several plausible choices of $t$.

The Logit-PFA method indicates, as highly significant, m/z values that are closely related to the five biomarkers identified by Kriegsmann et al. (2016), for all considered thresholds $t$. These findings were confirmed by the dependency-adjustment method and also by the original Z-statistics with a fixed threshold value. Table 9 lists the 20 top-ranked (i.e., most significant) null hypotheses (m/z values) for both cancer subtypes and thereby illustrates the overall significance of the five previously identified biomarkers, which are indicated by stars in Table 9. For a comparison to the findings published previously by Kriegsmann et al. (2016), Boskamp et al. (2017), and Leuschner et al. (2019), we attribute the values m/z = 1410.7 Da and m/z = 1411.7 Da to the peak of a peptide of the CK5 protein and its second isotopic peak. In addition, the values m/z = 1877.9 Da and m/z = 1905.9 Da appearing in Table 9 are likely to be attributable to peptides of the proteins CK15 and HSP27. The m/z value = 1406.7 Da, indicating a negative direction, is likely associated to a peptide of the CK7 protein, which indicates to be efficient for an identification for ADC in the lung. The biomarker at m/z = 1821.9 can be attributed to a peptide CK15 protein distinctive for SqCC.

5 DISCUSSION

From the statistical perspective, we have proposed an inferential framework for two-sample comparisons in high-dimensional settings when the test statistics have an arbitrary correlation structure. The major assumptions underlying the proposed methodology are (i) sparsity in the sense of Assumption 2.1, (ii) asymptotic normality of the vector of test statistics, and (iii) that the dependency structure among the test statistics can be described accurately by a factor model. To account for the high multiplicity of the considered applications, our criterion for type I error control is to keep the estimated value of the FDP at a given value of $\alpha$. In the presence of strong dependencies among test statistics, the FDP is typically not well concentrated around its mean (the FDR), and hence many authors have considered FDP control as the more appropriate criterion than FDR control under strong dependencies; see, for example, Blanchard et al. (2014) and the references therein. Under slightly different model assumptions than ours, Fan et al. (2012) have provided conditions under which the approximated FDP value is close to the true value of the FDP when high probability if PFA is applied. Whenever such conditions are fulfilled, (approximate) FDP control can be achieved with the proposed methodology.
TABLE 9  The top 20 ranked m/z values based on their rescaled Z-scores for both cancer subtypes. Those m/z values that are presumably related to the five previously identified biomakers are indicated by the symbol * in both subtables

| m/z Values | Z-scores | p-Values | m/z Values | Z-scores | p-Values |
|------------|----------|----------|------------|----------|----------|
| (a) The most sign. m/z-values for ADC | (b) The most sign. m/z-values for SqCC | | | | |
| 1407.7*    | −9.52    | < 10⁻⁶   | 1410.7*    | 12.31    | < 10⁻⁶   |
| 1406.7     | −9.05    | < 10⁻⁶   | 1411.7     | 11.9     | < 10⁻⁶   |
| 1234.6     | −7.76    | < 10⁻⁶   | 865.43     | 11.44    | < 10⁻⁶   |
| 975.48     | −7.58    | < 10⁻⁶   | 810.4      | 10.5     | < 10⁻⁶   |
| 1813.9     | −7.56    | < 10⁻⁶   | 878.43     | 10.15    | < 10⁻⁶   |
| 1293.6     | −7.12    | < 10⁻⁶   | 1877.9*    | 9.88     | < 10⁻⁶   |
| 1476.7     | −7.04    | < 10⁻⁶   | 1821.9*    | 9.73     | < 10⁻⁶   |
| 1516.8     | −7.03    | < 10⁻⁶   | 1878.9     | 9.58     | < 10⁻⁶   |
| 1277.6     | −6.80    | < 10⁻⁶   | 1439.7     | 9.37     | < 10⁻⁶   |
| 2247.1     | −6.80    | < 10⁻⁶   | 1822.9     | 9.24     | < 10⁻⁶   |
| 2246.1     | −6.59    | < 10⁻⁶   | 1437.7     | 8.98     | < 10⁻⁶   |
| 1292.6     | −6.46    | < 10⁻⁶   | 1412.7     | 8.96     | < 10⁻⁶   |
| 2248.1     | −6.45    | < 10⁻⁶   | 1879.9     | 8.92     | < 10⁻⁶   |
| 1812.9     | −6.35    | < 10⁻⁶   | 1438.7     | 8.42     | < 10⁻⁶   |
| 1838.9     | −6.30    | < 10⁻⁶   | 1425.7     | 8.42     | < 10⁻⁶   |
| 1641.8     | −6.06    | < 10⁻⁶   | 866.43     | 8.27     | < 10⁻⁶   |
| 1814.9     | −5.95    | < 10⁻⁶   | 1905.9*    | 8.19     | < 10⁻⁶   |
| 1738.9     | −5.93    | < 10⁻⁶   | 1823.9     | 8.13     | < 10⁻⁶   |
| 1705.8     | −5.85    | < 10⁻⁶   | 1906.9     | 8.07     | < 10⁻⁶   |
| 1512.7     | −5.79    | < 10⁻⁶   | 879.44     | 8.01     | < 10⁻⁶   |

From the application perspective, we have applied the proposed method to an MALDI imaging data frame with a large number of covariates (m/z values). The results derived with the proposed method are consistent with already reported insights about this data frame. Reliable statistical modeling of MALDI data is a challenging task; cf. for example, von Schroeder (2020). Our approach based on multiple marginal models (MMM) does not rely on heavy assumptions. Essentially, it is assumed that the (binary) phenotype of interest is associated with certain m/z-values, and that this association can be described by a (marginal) logistic regression model for each m/z-value separately. These assumptions are well established in the statistical theory of modeling binary data; see, for example, Agresti (2002).

There are several possible directions for future research: First, it may be interesting to consider other supervised statistical learning models (for instance, neural networks with more than one layer) instead of the logistic regression model proposed in this work. Second, it is of interest to quantify the uncertainty about the realized FDP for different threshold values, with the goal of providing a confidence region for this realized FDP, in addition to a mere point estimate. Third, it is of interest to analyze the statistical properties of MALDI data in a more detailed manner, which may allow for a joint modeling (after potential dimension reduction) instead of the MMM-based approach presented here. Finally, it will be worthwhile to consider categorical response variables with more than two categories.

ACKNOWLEDGMENTS
We thank the editor-in-chief, the associate editor, and two anonymous reviewers for their detailed reading of the paper and for their constructive suggestions, which have improved the presentation. We acknowledge the aid of Jonathan von Schroeder for a highly detailed and kind introduction to the MALDI data as well as for many meaningful and fruitful discussions related to the applicability of multiple testing to the MALDI context. We thank Johannes Leuschner for his constructive criticism of the paper. Financial support by the German Research Foundation via the RTG 2224, titled “π³: Parameter Identification - Analysis, Algorithms, Implementations,” is gratefully acknowledged.

Open Access funding enabled and organized by Projekt DEAL.

CONFLICT OF INTEREST
The authors have declared no conflict of interest.
DATA AVAILABILITY STATEMENT
The data that support the findings of this study are openly available at https://gitlab.informatik.uni-bremen.de/digipath/Supervised_NMF_Methods_for_MALDI.git

OPEN RESEARCH BADGES
This article has earned an Open Data badge for making publicly available the digitally-shareable data necessary to reproduce the reported results. The data is available in the Supporting Information section.

This article has earned an open data badge “Reproducible Research” for making publicly available the code necessary to reproduce the reported results. The results reported in this article could fully be reproduced.

ORCID
Vladimir Vutov https://orcid.org/0000-0002-4758-5822
Thorsten Dickhaus https://orcid.org/0000-0003-3084-3036

REFERENCES
Agresti, A. (2002). Categorical data analysis (2nd ed.). Wiley Series in Probability and Statistics. Wiley-Interscience, John Wiley & Sons.
Aichler, M., & Walch, A. (2015). MALDI imaging mass spectrometry: Current frontiers and perspectives in pathology research and practice. Laboratory Investigation, 95, 422–431.
Alexandrov, T. (2012). MALDI imaging mass spectrometry: Statistical data analysis and current computational challenges. BMC Bioinformatics, 13(Suppl 16), S11.
Behrmann, J., Etmann, C., Boskamp, T., Casadonte, R., Kriegsmann, J., & Maaß, P. (2018, 04). Deep learning for tumor classification in imaging mass spectrometry. Bioinformatics, 34(7), 1215–1223.
Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. Journal of the Royal Statistical Society. Series B: Statistical Methodology, 57(1), 289–300.
Blanchard, G., Dickhaus, T., Roquain, E., & Villers, F. (2014). On least favorable configurations for step-up-down tests. Statistica Sinica, 24(1), 1–23.
Boskamp, T., Lachmund, D., Oetjen, J., Cordero Hernandez, Y., Trede, D., Maass, P., Casadonte, R., Kriegsmann, J., Warth, A., Dienemann, H., Weichert, W., & Kriegsmann, M. (2017). A new classification method for MALDI imaging mass spectrometry data acquired on formalin-fixed paraffin-embedded tissue samples. Biochimica et Biophysica Acta—Proteins and Proteomics, 1865(7), 916–926.
Casadonte, R., & Caprioli, R. (2011). Proteomic analysis of formalin-fixed paraffin-embedded tissue by MALDI imaging mass spectrometry. Nature Protocols, 6, 1695–1709.
Deininger, S.-O., Becker, M., & Suckau, D. (2010). Tutorial: Multivariate statistical treatment of imaging data for clinical biomarker discovery. In S. S. Rubakhin & J. V. Sweedler (Eds.), Mass spectrometry imaging: Principles and protocols (pp. 385–403). Humana Press.
Dickhaus, T. (2014). Simultaneous statistical inference with applications in the life sciences. Springer.
Dickhaus, T., Neumann, A., & Bodnar, T. (2021). Multivariate multiple test procedures. In Cui, X., Dickhaus, T., Ding, Y., & J. C. Hsu (Eds.), Handbook of multiple comparisons (Chapter 3). Chapman & Hall / CRC Press.
Efron, B. (2007). Correlation and large-scale simultaneous significance testing. Journal of the American Statistical Association, 102(477), 93–103.
Efron, B. (2010). Correlated z-values and the accuracy of large-scale statistical estimates. Journal of the American Statistical Association, 105(491), 1042–1055.
Fan, J., & Han, X. (2017). Estimation of the false discovery proportion with unknown dependence. Journal of the Royal Statistical Society. Series B: Statistical Methodology, 79(4), 1143–1164.
Fan, J., Han, X., & Gu, W. (2012). Estimating false discovery proportion under arbitrary covariance dependence. Journal of the American Statistical Association, 107(499), 1019–1035.
Fan, J., Ke, T., Li, S., & Xia, L. (2016). PFA: Estimates false discovery proportion under arbitrary covariance dependence. R package version 1.1. https://CRAN.R-project.org/package=pfa
Fan, J., Ke, Y., Sun, Q., & Zhou, W.-X. (2019). FarmTest: Factor-adjusted robust multiple testing with approximate false discovery control. Journal of the American Statistical Association, 114(528), 1880–1893.
Fan, J., Liao, Y., & Mincheva, M. (2013). Large covariance estimation by thresholding principal orthogonal complements. Journal of the Royal Statistical Society. Series B: Statistical Methodology, 75(4), 603–680. With 33 discussions by 57 authors and a reply by Fan, Liao and Mincheva.
Friguet, C., Kloareg, M., & Causeur, D. (2009). A factor model approach to multiple testing under dependence. Journal of the American Statistical Association, 104(488), 1406–1415.
Hanselmann, M., Kirchner, M., Renard, B. Y., Amstalden, E. R., Glunde, K., Heeren, R. M., & Hamprecht, F. A. (2008). Concise representation of mass spectrometry images by probabilistic latent semantic analysis. Analytical Chemistry, 80(24), 9649–9658.
Jones, E. A., van Remoortere, A., van Zeijl, R. J. M., Hogendoorn, P. C. W., Bovée, J. V. M. G., Deelder, A. M., & McDonnell, L. A. (2011). Multiple statistical analysis techniques corroborate intratumor heterogeneity in imaging mass spectrometry datasets of myxofibrosarcoma. PLOS One, 6(9), Article No. e24913.
Kriegsmann, J., Kriegsmann, M., & Casadonte, R. (2015). MALDI TOF imaging mass spectrometry in clinical pathology: A valuable tool for cancer diagnostics (review). International Journal of Oncology, 46(3), 893–906.

Kriegsmann, M., Casadonte, R., Kriegsmann, J., Dienemann, H., Schirmacher, P., Hendrik Kobarg, J., Schwamborn, K., Stenzinger, A., Warth, A., & Weichert, W. (2016). Reliable entity subtyping in non-small cell lung cancer by matrix-assisted laser desorption/ionization imaging mass spectrometry on formalin-fixed paraffin-embedded tissue specimens. Molecular and Cellular Proteomics, 15(10), 3081–3089.

Leek, J. T., & Storey, J. D. (2008). A general framework for multiple testing dependence. The Proceedings of the National Academy of Sciences USA, 105(48), 18718–18723.

Leuschner, J., Schmidt, M., Fernsel, P., Lachmund, D., Boskamp, T., & Maass, P. (2019). Supervised non-negative matrix factorization methods for MALDI imaging applications. Bioinformatics, 35(11), 1940–1947.

Lieb, F., Boskamp, T., & Stark, H. G. (2020). Peak detection for MALDI mass spectrometry imaging data using sparse frame multipliers. Journal of Proteomics, 225, 103852.

Pepe, M. S., Whitaker, R. C., & Seidel, K. (1999). Estimating and comparing univariate associations with application to the prediction of adult obesity. Statistics in Medicine, 18(2), 163–173.

Pipper, C. B., Ritz, C., & Bisgaard, H. (2012). A versatile method for confirmatory evaluation of the effects of a covariate in multiple models. Journal of the Royal Statistical Society. Series C: Applied Statistics, 61(2), 315–326.

Poté, N., Alexandrov, T., Le Faouder, J., Laouirem, S., Léger, T., Mebarki, M., Belghiti, J., Camadro, J. M., Bedossa, P., & Paradis, V. (2013). Imaging mass spectrometry reveals modified forms of histone H4 as new biomarkers of microvascular invasion in hepatocellular carcinomas. Hepatology, 58(3), 983–994.

R Development Core Team (2021). R: A language and environment for statistical computing. http://www.R-project.org

Reck, M., Heigener, D. F., Mok, T., Soria, J. C., & Rabe, K. F. (2013). Management of non-small-cell lung cancer: Recent developments. Lancet, 382(9893), 709–719.

Schwamborn, K. (2012, Aug). Imaging mass spectrometry in biomarker discovery and validation. Journal of Proteomics, 75(16), 4990–4998.

Schwartzman, A. (2012). Comment: FDP vs FDR and the effect of conditioning. Journal of the American Statistical Association, 107(499), 1039–1041.

Senko, M. W., Beu, S. C., & McLafferty, F. W. (1995). Determination of monoisotopic masses and ion populations for large biomolecules from resolved isotopic distributions. Journal of the American Society for Mass Spectrometry, 6(4), 229–233.

Slawski, M., Hussong, R., Tholey, A., Jakoby, T., Gregorius, B., Hildebrandt, A., & Hein, M. (2012, Nov). Isotope pattern deconvolution for peptide mass spectrometry by non-negative least squares/least absolute deviation template matching. BMC Bioinformatics, 13, Article No. 291.

Sperandei, S. (2014). Understanding logistic regression analysis. Biochemia Medica, 24(1), 12–18.

Stange, J., Dickhaus, T., Navarro, A., & Schunk, D. (2016). Multiplicity- and dependency-adjusted $p$-values for control of the family-wise error rate. Statistics and Probability Letters, III, 32–40.

Stange, J., Loginova, N., & Dickhaus, T. (2016). Computing and approximating multivariate chi-square probabilities. Journal of Statistical Computation and Simulation, 86(6), 1233–1247.

Stevens, J. R., Al Masud, A., & Suyundikov, A. (2017). A comparison of multiple testing adjustment methods with block-correlation positively-dependent tests. PLoS One, 12(4), e0176124.

Storey, J. D. (2002). A direct approach to false discovery rates. Journal of the Royal Statistical Society. Series B: Statistical Methodology, 64(3), 479–498.

von Schroeder, J. (2020). Stable feature selection with applications to MALDI imaging mass spectrometry data. https://arxiv.org/abs/2006.15077

Wang, H., Peng, J., Wang, B., Lu, X., Zheng, J. Z., Wang, K., Tu, X. M., & Feng, C. (2017). Inconsistency between univariate and multiple logistic regressions. Shanghai Archives of Psychiatry, 29(2), 124–128.

Wijetunge, C. D., Saeed, I., Boughton, B. A., Roessner, U., & Halgamuge, S. K. (2015). A new peak detection algorithm for MALDI mass spectrometry data based on a modified asymmetric pseudo-voigt model. BMC Genomics, 16(Suppl 12), Article No. S12.

Yang, C., He, Z., & Yu, W. (2009, Jan). Comparison of public peak detection algorithms for MALDI mass spectrometry data analysis. BMC Bioinformatics, 10, Article No. 4.

SUPPORTING INFORMATION
Additional supporting information can be found online in the Supporting Information section at the end of this article.

How to cite this article: Vutov, V., & Dickhaus, T. (2023). Multiple two-sample testing under arbitrary covariance dependency with an application in imaging mass spectrometry. Biometrical Journal, 65, 2100328. https://doi.org/10.1002/bimj.202100328