Multiple multi-sample testing under arbitrary covariance dependency

Vladimir Vutov | Thorsten Dickhaus

Institute for Statistics, University of Bremen, Bremen, Germany

Correspondence
Thorsten Dickhaus, Institute for Statistics, University of Bremen, Bibliothekstr. 1, 28359 Bremen, Germany. Email: dickhaus@uni-bremen.de

Funding information
Deutsche Forschungsgemeinschaft, Grant/Award Number: RTG 2224

Abstract
Modern high-throughput biomedical devices routinely produce data on a large scale, and the analysis of high-dimensional datasets has become commonplace in biomedical studies. However, given thousands or tens of thousands of measured variables in these datasets, extracting meaningful features poses a challenge. In this article, we propose a procedure to evaluate the strength of the associations between a nominal (categorical) response variable and multiple features simultaneously. Specifically, we propose a framework of large-scale multiple testing under arbitrary correlation dependency among test statistics. First, marginal multinomial regressions are performed for each feature individually. Second, we use an approach of multiple marginal models for each baseline-category pair to establish asymptotic joint normality of the stacked vector of the marginal multinomial regression coefficients. Third, we estimate the (limiting) covariance matrix between the estimated coefficients from all marginal models. Finally, our approach approximates the realized false discovery proportion of a thresholding procedure for the marginal p-values for each baseline-category logit pair. The proposed approach offers a sensible trade-off between the expected numbers of true and false findings. Furthermore, we demonstrate a practical application of the method on hyperspectral imaging data. This dataset is obtained by a matrix-assisted laser desorption/ionization (MALDI) instrument. MALDI demonstrates tremendous potential for clinical diagnosis, particularly for cancer research. In our application, the nominal response categories represent cancer (sub-)types.

KEYWORDS
false discovery proportion, hyperspectral imaging data, matrix-assisted laser desorption/ionization, multinomial regression, multiple marginal models

1 | INTRODUCTION

Many datasets in various scientific disciplines, such as neuroimaging, genomics, brain-computer interfacing, and others, are nowadays high-dimensional. Such datasets are frequently analyzed by means of large-scale multiple testing simultaneously for some phenotype—for example, cancer (sub-)types—on each feature among thousands of features. This approach
has multiple applications in the life sciences.\textsuperscript{1} At least since the seminal publication by Benjamini and Hochberg,\textsuperscript{2} control of the false discovery rate (FDR) is a commonly used type I error criterion in high-dimensional multiple test problems. However, applying the well-known Benjamini-Hochberg\textsuperscript{2} (henceforth, BH) procedure or Storey's procedure,\textsuperscript{3,4} which have originally been designed for stochastically independent or weakly dependent \(p\)-values, respectively, can lead to an FDR inflation under certain forms of (strong) dependencies.\textsuperscript{5-8} Under the assumption of positive regression dependency on subsets (PRDS), the BH procedure still controls the FDR,\textsuperscript{9} but checking the validity of the PRDS assumption on the basis of a sample is far from trivial and only established for a limited number of special cases like, for example, elliptically distributed vectors of test statistics.\textsuperscript{10}

For these reasons and with the goal of optimizing the statistical power of the multiple test under the constraint of (at least asymptotic) FDR control at the desired level, multivariate (FDR-controlling) multiple tests have drawn a lot of attention over recent years.\textsuperscript{11} A multivariate multiple test incorporates the dependence structure among test statistics or \(p\)-values, respectively, or an estimate thereof, explicitly in its decision rule. Large-scale multiple testing under general and strong dependency remains challenging and an active research topic in modern statistics.\textsuperscript{7,8,12,13} Among other approaches like, for instance, considering block dependencies in genetics,\textsuperscript{14-16} several multi-factor models have been proposed to model dependencies among test statistics.\textsuperscript{5-8,13} In particular, a general framework to approximate the false discovery proportion (FDP) has been introduced.\textsuperscript{7,17} It employs a so-called principal factor approximation (PFA) of the (estimated limiting) covariance matrix of test statistics, assuming that these test statistics are (at least asymptotically) jointly normally distributed. The general strategy of the approach is to apply a spectral decomposition of the aforementioned covariance matrix and then to deduct its principal factors that mostly induce the strong correlation dependency. A more recent procedure has been proposed in the context of FDR control under arbitrary (covariance) dependency.\textsuperscript{18} This approach does not heavily rely on the assumption of the (asymptotic) normality of the test statistics and can be utilized even in the case of an unknown covariance matrix among the test statistics.

Biological datasets generated by advanced high-throughput devices typically contain thousands of measured variables, many of which are related. In this work, we consider a dataset obtained by a matrix-assisted laser desorption/ionization (MALDI) imaging mass spectrometry (IMS) tool, also known as MALDI imaging. MALDI has advanced considerably and demonstrates immense potential in numerous pathological applications.\textsuperscript{19,20} MALDI-generated datasets contain thousands of measured variables (corresponding to molecular masses), many of which are (highly) related.\textsuperscript{21} Since MALDI data do not generally exhibit dependencies in blocks, it has recently been advocated to model dependencies among MALDI features by means of factor models and to employ PFA in the context of multiple two-sample comparisons based on MALDI data.\textsuperscript{22} In the present work, we extend our previous work by proposing a (multivariate) generalization to multi-sample comparisons, meaning that simultaneous inference with more than two nominal classes can be performed.

In this work, we deal with multiple multi-sample hypotheses testing under arbitrary correlation dependency, in particular in the case of more than two samples or categories, respectively. We do so under the scope of multiple marginal multinomial regression models and by applying the PFA methodology. In general, multinomial regression is a specific case of multivariate generalized linear models (GLMs), where (categorical) outcome variables are expressed as multi-dimensional vectors.\textsuperscript{23} As a result, our methodology performs multi-sample group testing simultaneously across these categorical groups. An essential step in our framework is the approximation of the joint null distributions among all marginal fits. In the present work, we exploit the framework of multiple marginal models (MMM).\textsuperscript{24} The classical setting of the latter is provided by modelling data with multiple outcome variables.\textsuperscript{25} Often, these studies consider procedures to control the family-wise error rate (FWER) with respect to type I errors.\textsuperscript{26,27} In our study, however, each “endpoint” (comparison) refers to a single independent (ie, explanatory) variable, where the number \(p\) of independent variables under consideration is in the thousands. In the context of high-dimensional studies, it is well-known that FWER-controlling procedures are conservative and typically lead to substantially less statistical power than procedures that control the proportion of false discoveries.\textsuperscript{2,28} A computationally efficient alternative to the MMM approach, based on score test statistics, has been proposed recently.\textsuperscript{16,29,30} The latter methodology needs only one fit of a GLM model. However, this methodology requires a data regime in which \(p \ll n\), where \(n\) denotes the sample size, in order to approximate the (limiting) covariance matrix with sufficient accuracy. We call our method multinomial-principal factor approximation (Multi-PFA).

The remainder of this article is structured as follows. In Section 2, we present the proposed multi-sample multiple testing approach. Section 3 is dedicated to a simulation study given different data-generating scenarios. We demonstrate
the practical application of the proposed approach to hyperspectral data in Section 4, and we conclude with a general discussion and outlook in Section 5.

2 | PROPOSED METHODOLOGY

2.1 | Data structures

MALDI imaging data are typically stored in an $n \times p$ matrix $X = (x_{ij})_{1 \leq i \leq n}$, where the entry $x_{ij}$ corresponds to an intensity value of the $i$-th mass spectrum at the $j$-th mass-to-charge ratio (m/z) value. Since MALDI data represent molecular masses of ionizable molecules, all entries of the data matrix $X$ are non-negative. However, our proposed framework makes no explicit assumption that data points have to be non-negative. Furthermore, our framework can also work with discrete covariates. We approach the biological challenge of analyzing the associations between (individual) m/z values (predictors $(X_j)_{1 \leq j \leq p}$) and several cancer (sub-)types (categorical response $Y$) by carrying out multiple statistical hypothesis tests simultaneously. Under the $j$-th null hypothesis $H_{0j}$, there is no association between $X_j$ and $Y$, and we aim at testing the $p$ null hypotheses $H_{01}, \ldots, H_{0p}$ (against their two-sided alternatives $H_{11}, \ldots, H_{1p}$) simultaneously based on one and the same dataset. Throughout the remainder, we assume that $Y$ is a (random) nominal outcome variable, meaning that the response categories are unordered. Hence, for each $j \in \{1, \ldots, p\}$, the tuple $(X_j, Y)$ takes its values in $\mathbb{R}_{\geq 0} \times \{1, 2, \ldots, q\}$, where we denote by $q$ the number of (response) categories, and $\mathbb{R}_{\geq 0}$ denotes the set of non-negative real numbers.

2.2 | Marginal modeling for categorical responses

First, we model marginal associations for each predictor $X_j$ individually. The reason for considering marginal modelling relies on the fact that we consider an inferential framework, in which the number of covariates is either quite large or even (much) larger than the number of observational units ($n \ll p$). Thus, a model that includes all predictors in a single model cannot be fitted (reliably). However, this condition is not obligatory, and the proposed approach is also applicable to other data regimes. In the context of large-scale multiple (association) testing, marginal linear regression models have been adopted to test an individual hypothesis for each predictor. Marginal GLMs, for positive outcome variables, have been considered in situations where the focus is on an inferential process around the mean of the response given a group of predictors. Also, marginal logistic regressions have been utilized to analyse child obesity, with the usage of generalized estimating equations for jointly estimating (univariate) associations.

For this section, let $j \in \{1, \ldots, p\}$ be arbitrary, but fixed. To model the conditional distribution $\mathbb{P}(Y|X_j)$, we assume a marginal multinomial regression with a canonical (logit) link function. In general, these response categories cannot be treated as a one-dimensional response, and we have to establish a dummy variable for each category. We, therefore, have a multivariate outcome variable. One (response) category is used as a baseline, and the multinomial logistic model is established by pairing each (except one) nominal response category with the baseline. Throughout this article, response category $q$ is used as the baseline. Furthermore, we mainly consider the case of $q = 3$ throughout the remainder, for concreteness and because of its relevance for the application that we will present in Section 4. For each (marginal) fit, we consider intercept terms $\alpha_c$ and slope parameters $\beta_{jc}$, for $c \in \{1, \ldots, q-1\}$. We denote the observables (regressor and response variable) for the $j$-th marginal regression fit by $(X^{(i)}_j, Y^{(i)})_{1 \leq i \leq n}$, and we assume that these tuples are stochastically independent and identically distributed bivariate random vectors, each having the same (joint) distribution as $(X_j, Y)$. For observational unit $i \in \{1, \ldots, n\}$, our model equation for the $c$-th baseline-category logit is given by

$$\log \left( \frac{\mathbb{P}(Y^{(i)}_c = 1|X^{(i)}_j) / \mathbb{P}(Y^{(i)}_q = 1|X^{(i)}_j)}{\mathbb{P}(Y^{(i)}_q = 1|X^{(i)}_j)} \right) = \alpha_c + \beta_{jc} X^{(i)}_j,$$

where $Y^{(i)}_c$ denotes the dummy indicator for category $c \in \{1, \ldots, q-1\}$ pertaining to observational unit $i \in \{1, \ldots, n\}$, and $\mathbb{P}(Y^{(i)}_c = 1|X^{(i)}_j)$ by definition. The unknown model parameters $(\alpha_c, \beta_{jc})_{1 \leq c \leq q-1}$ can be estimated by the maximum (log-) likelihood (ML) principle. To this end, we follow the derivations in section 7.1.4 of the book by Agresti. Given that $\pi^{(i)}_{jc} = 1 - \pi^{(i)}_{j1} - \pi^{(i)}_{j2} - \ldots - \pi^{(i)}_{jq-1}$ as well as $Y^{(i)}_q = 1 - Y^{(i)}_1 - Y^{(i)}_2 - \ldots - Y^{(i)}_{q-1}$, the log-likelihood contribution for the $i$-th
observational unit is under the model specified in (1) given by

\[
\log \left( \prod_{c=1}^{q} \left\{ \pi_{jc}^{(i)} \right\}^{y_{jc}^{(i)}} \right) = \sum_{c=1}^{q-1} Y_{c}^{(i)} \log \pi_{jc}^{(i)} + \left( 1 - \sum_{c=1}^{q-1} Y_{c}^{(i)} \right) \log \left[ 1 - \sum_{c=1}^{q-1} \pi_{jc}^{(i)} \right] \\
= \sum_{c=1}^{q-1} Y_{c}^{(i)} \log \frac{\pi_{jc}^{(i)}}{1 - \sum_{c=1}^{q-1} \pi_{jc}^{(i)}} + \log \left[ 1 - \sum_{c=1}^{q-1} \pi_{jc}^{(i)} \right].
\]

(2)

Substituting \( \log(\pi_{jc}^{(i)}/\pi_{jq}^{(i)}) = \alpha_{jc} + X_{jc}^{(i)} \beta_{jc} \) as well as \( \pi_{jq}^{(i)} = 1/(1 + \sum_{c=1}^{q-1} \exp(\alpha_{jc} + X_{jc}^{(i)} \beta_{jc})) \) in (2), we find that

\[
\ell' \left( \hat{\alpha}_{jc}, \hat{\beta}_{jc} \mid Y^{(i)}, X_j^{(i)} \right) = \max_{\alpha_{jc}, \beta_{jc}} \sum_{i=1}^{n} \left\{ \sum_{c=1}^{q-1} Y_{c}^{(i)} (\alpha_{jc} + X_{jc}^{(i)} \beta_{jc}) - \log \left[ 1 + \sum_{c=1}^{q-1} \exp(\alpha_{jc} + X_{jc}^{(i)} \beta_{jc}) \right] \right\},
\]

(3)

where \( \ell' \) stands for the log-likelihood function, and \( \hat{\alpha}_{jc} \) and \( \hat{\beta}_{jc} \) denote the ML estimators. In practice, the estimation is carried out conditionally to the actual observed values of the response indicators and the regressors, leading to the numerical values of the ML estimates.

In this work, we are interested in testing simultaneously two families of hypothesis-alternative pairs, which are given by

\[
H_{0jc} : \beta_{jc} = 0 \text{ versus } H_{1jc} : \beta_{jc} \neq 0, \quad j = 1, \ldots, p, \ c \in \{1, 2\},
\]

(4)

for \( q = 3 \). By means of testing the hypotheses in (4), we make binary decisions (rejection or non-rejection) for both baseline-category sets, consisting of \( p \) null hypotheses each, based on the data at hand. In the MALDI context, this means that we aim at identifying the most distinctive m/z values for the cancer association, where three different cancer types are considered in our application.

2.3 Multiple marginal models

The following (second) step of the proposed framework is to combine all marginal fits and to approximate the joint null distribution of all ML estimators for each baseline-category pair. To do so, we exploit the approach for jointly estimating the parameters of multiple marginal models developed by Pipper et al.\textsuperscript{24} and apply this approach to the marginal models explained in the previous section. Notice that our modelling approach implies that the regression parameters are unique to one model and are not shared by any two models \( j_1 \neq j_2 \). In addition to that, the intercepts \( (\alpha_{jc})_{1 \leq j \leq p, \ c \in \{1, 2\}} \) are nuisance parameters in the sense that the hypotheses in (4) solely concern the \( \beta_{jc} \)‘s. However, the intercepts \( (\alpha_{jc})_{1 \leq j \leq p, \ c \in \{1, 2\}} \) contribute to the estimation and the standardization of \( \beta_{jc} \)‘s.

The methodology by Pipper et al\textsuperscript{24} yields a central limit theorem for the two \( p \)-dimensional random vectors \( \tilde{\beta}_c = (\tilde{\beta}_{1c}, \ldots, \tilde{\beta}_{pc})^\top, \ c \in \{1, 2\} \). This is accomplished by stacking the standardised score contributions of the \( \tilde{\beta}_{jc} \)’s across all \( p \) marginal models, for each baseline-category pair. To this end, it is used that, under standard regularity assumptions (like finite variances and non-vanishing limiting relative category frequencies), each ML estimator \( \hat{\beta}_{jc}^{(i)} \) admits the asymptotic (ie, \( n \to \infty \)) representation\textsuperscript{24}

\[
(\hat{\beta}_{jc} - \beta_{jc}) \sqrt{n} = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \Psi_{jc}^{(i)} + o_p(1),
\]

(5)

where \( \Psi_{jc} = (F_{jc}^{(i)})^{-1} \Psi_{jc}, \ (F_{jc}^{(i)})^{-1} \) is the relevant row of the inverse Fisher information matrix of the model specified in (1) that corresponds to coordinate \( j \) for the \( i \)-th observational unit, \( F_{jc}^{(i)} \) is the score function pertaining to coordinate \( j \) and category \( c \) for the \( i \)-th observational unit, and \( o_p(1) \) indicates a sequence of random variables converging to zero in probability.
Elementary calculations yield that \((\Psi_{j1}, \Psi_{j2})^T\) is given by the second and the fourth coordinate of the four-variate vector
\[
\left( \left( \pi_{jc}^{(i)} \begin{bmatrix} 1 \end{bmatrix} (c = c') - \pi_{jc'}^{(i)} \right) (1, X_j^{(i)})^T (1, X_j^{(i)}) \right)_{c, c' \in \{1, 2\}}^{-1} \left( Y_1^{(i)} - \pi_{jc}^{(i)}, Y_2^{(i)} - \pi_{jc}^{(i)} \right)^T \otimes (1, X_j^{(i)})^T,
\]
where
\[
\pi_{jc}^{(i)} = \exp\left( a_{jc} + X_j^{(i)} \beta_{jc} \right) / \left( 1 + \sum_{h=1}^2 \exp\left( a_{jh} + X_j^{(i)} \beta_{jh} \right) \right),
\]
and \(\otimes\) denotes the Kronecker product. For more (computational) details regarding multinomial regression, we refer to previous literature on the topic.\(^{37,38}\)

Now, let \(\beta_c := (\beta_{1c}, \ldots, \beta_{pc})^T\), \(\hat{\beta}_c := (\hat{\beta}_{1c}, \ldots, \hat{\beta}_{pc})^T\), and \(\Psi_{ic} := (\Psi_{1ic}, \ldots, \Psi_{pic})^T\) for \(c \in \{1, 2\}\) and \(i \in \{1, \ldots, n\}\). Then, we can conclude from (5) the asymptotic expansion \((\hat{\beta}_c - \beta_c) \sqrt{n} = (1/n) \sum_{i=1}^n \Psi_{ic} + o_p(1)\), where now \(o_p(1)\) indicates a sequence of \(p\)-dimensional random vectors converging to the zero vector in probability. Since the observational units are assumed to be stochastically independent and identically distributed, the multivariate central limit theorem yields convergence in distribution (indicated by the symbol \(\rightarrow\)) to a centred \(p\)-variate normal distribution, that is, \((\hat{\beta}_c - \beta_c) \sqrt{n} \rightarrow^d N_p(0, \Sigma_c)\) for \(c \in \{1, 2\}\). For each \(c \in \{1, 2\}\), the limiting variance-covariance matrix \(\Sigma_c\) can be estimated consistently by \(\hat{\Sigma}_c = (1/n) \sum_{i=1}^n \Psi_{ic}^T \Psi_{ic}\) where each \(\Psi_{ic}\) is obtained by inserting the parameter estimates \((\hat{\alpha}_{jc}, \hat{\beta}_{jc})\) for \(c \in \{1, 2\}\) from the \(p\) marginal fits (cf. (3)).

For each \(c \in \{1, 2\}\), let \(Z_{1c}, \ldots, Z_{pc}\) be the standardised versions of \(\hat{\beta}_{1c}, \ldots, \hat{\beta}_{pc}\), which are given by
\[
Z_{jc} = \frac{\hat{\beta}_{jc}}{\sqrt{\text{Var}(\hat{\beta}_{jc})}}, \quad j = 1, \ldots, p, \quad c \in \{1, 2\},
\]
where \(\sqrt{\text{Var}(\hat{\beta}_{jc})}\) is the square root of the \(j\)-th diagonal element of \(\hat{\Sigma}_c\), divided by \(\sqrt{n}\). We have that
\[
(Z_{1c}, Z_{2c}, \ldots, Z_{pc})^T \approx \mathcal{N}_p((\mu_{1c}, \mu_{2c}, \ldots, \mu_{pc})^T, \hat{\Sigma}_c^e),
\]
where \(\mu_{jc} = \beta_{jc} / \sqrt{\text{Var}(\hat{\beta}_{jc})}\) for \(1 \leq j \leq p\) and \(c \in \{1, 2\}\), \(\hat{\Sigma}_c^e = \text{diag}[\Sigma_c]^{-1/2} \hat{\Sigma}_c \text{diag}[\Sigma_c]^{-1/2}\) denotes the correlation matrix pertaining to \(\hat{\Sigma}_c\), and the notation \(\approx\) indicates the approximate distribution for large \(n\). Therefore, since \(\sqrt{\text{Var}(\hat{\beta}_{jc})} > 0\) holds true with probability one under our regularity assumptions, the families of hypotheses from (4) can be equivalently written as
\[
H_{0jc} : \mu_{jc} = 0 \quad \text{versus} \quad H_{1jc} : \mu_{jc} \neq 0, \quad j = 1, \ldots, p, \quad c \in \{1, 2\}.
\]
For the sake of presentation, we use henceforth without loss of generality \(\hat{\Sigma}_c^e\) and \(Z_{1c}, \ldots, Z_{pc}\) only for the first baseline-category (ie, for \(c = 1\)) and omit the index \(c\), for convenience of notation. However, we consider the same inference procedure simultaneously for both baseline-category pairs. In order to address the fact that two (or, in general, \(q - 1\)) families of hypotheses are considered together, one may adjust the significance level \(\alpha\) occurring in Section 2.5 below accordingly.
2.4 False discovery proportion

Considering the multiple test problem given by (9) (for \( c = 1 \)), let \( p_0 = \# \{ j : \mu_j = 0 \} \) denote the number of true null hypotheses, and \( p_1 = p - p_0 = \# \{ j : \mu_j \neq 0 \} \) the number of false null hypotheses. Throughout the remainder, we make the following sparsity assumption.

**Assumption 1.** The number \( p_1 \) of false null hypotheses is very small in comparison to the number \( p \) of all null hypotheses.

In an asymptotic context, where \( p = p(n) \) tends to infinity as \( n \) tends to infinity, Assumption 1 can be formalized as \( p_0(n)/p(n) \to 1 \). For controlling type I errors, we employ empirical process techniques, in the spirit of Storey’s method.\(^3\) In particular, we consider thresholding rules which reject a null hypothesis \( H_0 \) if and only if a corresponding \( p \)-value \( p_j \) is smaller than or equal to a (data-dependent) threshold \( t \). Such types of multiple tests have been considered widely in previous literature.\(^5,7,17\) Conceptually, the goal of the proposed method is to approximate the proportion of false discoveries among all rejections (commonly referred to as the false discovery proportion, FDP for short) for a fixed threshold \( t \) under an arbitrary correlation matrix \( \Sigma^* \). To this end, we consider the three empirical processes

\[
V(t) = \# \{ \text{true null } P_j : P_j \leq t \},
\]

\[
S(t) = \# \{ \text{false null } P_j : P_j \leq t \},
\]

\[
R(t) = \# \{ P_j : P_j \leq t \},
\]

where \( t \) ranges in \([0, 1]\) and the notation \( P_j \) is used to indicate that we consider the \( j \)-th \( p \)-value as a random variable here. For fixed \( t \in [0, 1] \), the quantities \( V(t), S(t), \) and \( R(t) \) represent the (random) number of false discoveries (ie, false rejections or, synonymously, type I errors), the (random) number of true discoveries, and the (random) total number of discoveries, respectively, hence \( R(t) = V(t) + S(t) \). The FDP is for a fixed \( t \) given by \( \text{FDP}(t) = V(t)/\max\{R(t), 1\} \), where the maximum is taken to avoid an expression of the form \( 0/0 \). All of the latter random variables depend on the \( Z \)-statistics \( Z_1, Z_2, \ldots, Z_p \), because each (random) \( p \)-value \( P_j \) will be obtained by a transformation of \( Z_j \), for \( 1 \leq j \leq p \), see below. Notice also that \( V(t) \) and \( S(t) \) are both unobservable, whereas \( R(t) \) is observable.

2.5 Principal factor approximation

The next step of our proposed framework is to incorporate the correlation effects of the \( Z \)-statistics in an approximation of \( \text{FDP}(t) \) for a fixed threshold \( t \). The procedure relies on the identification of a low-dimensional linear space of random vectors which captures most of the dependence structure of the \( Z \)-statistics.

To this end, we carry out the spectral decomposition of the correlation matrix \( \Sigma^* \). We let \( \lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_p \geq 0 \) denote the eigenvalues of \( \Sigma^* \), and \( \gamma_1, \ldots, \gamma_p \) the corresponding eigenvectors. Hence, \( \hat{\Sigma}^* \) is represented by its eigenvalue-eigenvector pairs \((\hat{\lambda}_j, \hat{\gamma}_j)_{1 \leq j \leq p} \). For a fixed number \( k \) of common factors, where \( 1 \leq k \ll p \), the factor approximation of \( \Sigma^* \) is given by

\[
\hat{\Sigma}^* = L_kL_k^T + A_k, \tag{10}
\]

where \( A_k = \sum_{j=k+1}^p \hat{\lambda}_j \hat{\gamma}_j \hat{\gamma}_j^T \) and \( L_k = (\sqrt{\hat{\lambda}_1} \hat{\gamma}_1, \sqrt{\hat{\lambda}_2} \hat{\gamma}_2, \ldots, \sqrt{\hat{\lambda}_k} \hat{\gamma}_k) \). If \( k \) is chosen appropriately, the \( p \times k \) matrix \( L_k \), corresponding to \( k \) latent variables, can describe most of the dependence structure among the \( Z \)-statistics. Accordingly, \( Z_1, \ldots, Z_p \) can be decomposed as

\[
Z_j = \mu_j + b_j^T W + K_j = \mu_j + \eta_j + K_j, \quad j = 1, \ldots, p, \tag{11}
\]

where \( b_j = (b_{j1}, \ldots, b_{jk})^T \) and \( b_j = (b_{j1}, \ldots, b_{jk})^T = \sqrt{\hat{\lambda}_j} \hat{\gamma}_j \). The vector \( W = (W_1, \ldots, W_k)^T \sim N_k(0, I_k) \) is the vector of (latent) common factors. Due to orthogonality projection, these common factors are jointly stochastically independent. 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. In our previous work,\(^{32}\) we elaborated more on the importance and the choice of \( k \).
On the basis of the above considerations, we consider the FDP estimator from Proposition 2 by Fan et al.,\(^7\) which is for a given \(t \in [0, 1]\) defined as

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

(12)

if \(R(t) > 0\), and \(\hat{\text{FDP}}(t) = 0\) if \(R(t) = 0\). In (12), \(a_j = (1 - \sum_{h=1}^{k} b_{jh}^2)^{-1/2}\) and \(R(t) = \{j : P_j \leq t\}\), where \(\Phi\) and \(Z_{j/2} = \Phi^{-1}(t/2)\) are the cumulative distribution function (cdf) and the lower \(t/2\)-quantile of the standard normal distribution on \(\mathbb{R}\), respectively. In this, the dependency-unadjusted, two-sided and random \(p\)-value corresponding to \(Z_j\) is given by \(P_j = 2\Phi(|Z_j|) = 2(1 - \Phi(|Z_j|))\), and \(\hat{\eta}_j = \sum_{h=1}^{k} b_{jh} \hat{W}_j\) is a linear estimator for \(\eta_j = b_j^T W\).

The FDP estimator given in (12) is based on a sparsity assumption.\(^7\) Specifically, it is assumed that the numbers \(p\) and \(p_0\) are large, while the number \(p_1\) of false nulls is relatively low. This assumption allows for the summation of all hypotheses (not only over “true nulls”) in the numerator of (12). In the context of MALDI modeling, this sparsity assumption is justified, for three reasons: First, several researchers who have analyzed MALDI datasets have reported a low number of m/z values that are highly associative for a cancer (sub-)type.\(^{20,39}\) Second, to the best of our knowledge, the reported number of biologically meaningful molecules, so-called biomarkers, in MALDI-related studies, is low—from three to five per study.\(^{21,40}\) Third, there is evidence for considerable noise in the data.\(^{20,21}\)

To evaluate (12) in a practical application, one needs to determine the (linear) estimator \(\hat{W} = (\hat{W}_1, \ldots, \hat{W}_k)^T\) of the common factors. The authors have proposed to estimate \(\hat{W}\) via a linear regression (\(L_2\)-regression) or by using a quantile regression (\(L_1\)-regression).\(^7\) Regarding the \(L_2\)-estimator, it has been proposed to carry out the estimation of \((\hat{W}_1, \ldots, \hat{W}_k)^T\) only on the basis of the subset of length 0.95\(p\) of the smallest (in absolute values) \(Z\)-statistics. This leads to the \(L_2\)-estimator

\[
\hat{w} = (\hat{w}_1, \ldots, \hat{w}_k)^T = \arg \min_{\hat{W} \in \mathbb{R}^k} \sum_{j=1}^{[0.95p]} (Z_j - b_j^T \hat{W})^2,
\]  

(13)

where the \(Z\)-values in (13) are sorted in ascending order based on their absolute values. We have utilized this estimator in our simulation study (see Section 3). The estimator based on the \(L_1\)-regression is given by

\[
\hat{w} = (\hat{w}_1, \ldots, \hat{w}_k)^T = \arg \min_{\hat{W} \in \mathbb{R}^k} \sum_{j=1}^{p} |Z_j - b_j^T \hat{W}|.
\]  

(14)

For the analysis of the MALDI data (Section 4), we have used the \(L_1\)-regression rather than the \(L_2\)-regression because it is more robust to highly atypical observations (ie, outliers).

Finally, the dependency-adjusted \(p\)-values related to \(Z_1, \ldots, Z_p\) are given by

\[
\tilde{P}_j = 2\Phi(-|a_j(Z_j - b_j^T \hat{W})|).
\]  

(15)

For a given threshold \(t \in [0, 1]\), the null hypothesis \(H_{0j}\) from (9) gets rejected based on the data at hand, if and only if \(\tilde{P}_j \leq t, 1 \leq j \leq p\). For the purpose of FDP control, \(t\) can be selected as the largest value \(t = t_{\alpha} \in [0, 1]\) fulfilling that \(\hat{\text{FDP}}(t_{\alpha})\) is not exceeding a pre-determined level \(\alpha \in (0, 1)\), for example, \(\alpha = 10\%\). In practice, one can carry out a grid search over a set of candidate values of \(t\) in order to find the value \(t_{\alpha}\) for a given \(\alpha\).\(^{22}\)

3 | SIMULATION STUDY

3.1 | Simulation setup

We carried out a simulation study to assess the performance of our proposed methodology under different data-generating schemes. We considered the parameter settings \(n = 500, p \in \{500, 1000\}\), and \(p_1 = 10\) (for each baseline-category pair). For each combination of the aforementioned parameter values, we performed 1000 Monte Carlo repetitions. Without
TABLE 1 Simulation results under scenario 1 (I).

| \( c \) | Median of FDP(t) | Std. error of FDP(t) | Mean of R(t) | Std. error of R(t) | Mean of S(t) | Std. error of S(t) | Median of \( t_{0.05} \) | Mean \( S(\tau_{0.05}) \) |
|-------|-----------------|------------------|-------------|-----------------|-------------|-----------------|-----------------|-----------------|
| 1     | 0.00433         | 0.00147          | 5.80        | 1.39            | 5.75        | 1.36            | 1.60e-03        | 8.37            |
| 2     | 0.00442         | 0.00134          | 5.78        | 1.36            | 5.74        | 1.35            | 1.60e-03        | 8.31            |

Note: The total number of hypotheses equals \( p = 500 \); the number of false null hypotheses equals \( p_1 = 10 \) per baseline-category pair (non-zero regression coefficients equal 1); the number of factors equals \( k = 10 \); the rejection threshold equals \( \tau = 10^{-4} \), apart from the last two columns.

TABLE 2 Simulation results under scenario 1 (II).

| \( c \) | Median of FDP(t) | Std. error of FDP(t) | Mean of R(t) | Std. error of R(t) | Mean of S(t) | Std. error of S(t) | Median of \( t_{0.05} \) | Mean \( S(\tau_{0.05}) \) |
|-------|-----------------|------------------|-------------|-----------------|-------------|-----------------|-----------------|-----------------|
| 1     | 0.00946         | 0.00325          | 5.95        | 1.35            | 5.86        | 1.33            | 5.52e-04        | 7.50            |
| 2     | 0.00943         | 0.00360          | 5.83        | 1.43            | 5.76        | 1.40            | 5.52e-04        | 7.46            |

Note: The total number of hypotheses equals \( p = 1000 \); the number of false null hypotheses equals \( p_1 = 10 \) per baseline-category pair (non-zero regression coefficients equal 1); the number of factors equals \( k = 10 \); the rejection threshold equals \( \tau = 10^{-4} \), apart from the last two columns.

TABLE 3 Simulation results under scenario 2 (I).

| \( c \) | \( \rho \) | Median of FDP(t) | Std. error of FDP(t) | Mean of R(t) | Std. error of R(t) | Mean of S(t) | Std. error of S(t) | Median of \( t_{0.05} \) | Mean \( S(\tau_{0.05}) \) |
|-------|--------|-----------------|------------------|-------------|-----------------|-------------|-----------------|-----------------|-----------------|
| 1     | 0.2    | 0.00287         | 0.0199           | 5.83        | 1.36            | 5.81        | 1.31            | 2.33e-03        | 8.63            |
| 2     | 0.2    | 0.00273         | 0.00995          | 5.90        | 1.33            | 5.89        | 1.31            | 2.33e-03        | 8.72            |
| 1     | 0.5    | 0.000274        | 0.0475           | 5.84        | 1.48            | 5.81        | 1.31            | 6.99e-03        | 9.28            |
| 2     | 0.5    | 0.000265        | 0.0168           | 5.88        | 1.33            | 5.89        | 1.31            | 6.99e-03        | 9.31            |
| 1     | 0.8    | 0.000171        | 0.0455           | 5.87        | 2.46            | 5.81        | 1.31            | 2e-02           | 9.78            |
| 2     | 0.8    | 0.000168        | 0.0141           | 5.86        | 1.31            | 5.90        | 1.32            | 2e-02           | 9.77            |

Note: The total number of hypotheses equals \( p = 500 \); the number of false null hypotheses equals \( p_1 = 10 \) per baseline-category pair (non-zero regression coefficients equal 1); the number of factors equals \( k = 1 \); the rejection threshold equals \( \tau = 10^{-4} \), apart from the last two columns.

For each observational unit \( i \in \{1, \ldots, n\} \), the simulation data have been generated according to the model

\[
\mathbb{P}_\beta(Y_1(i) = 1|X(i)) = \frac{\exp\left(\sum_{j=1}^{p_1} \beta_{j1}X_{j(i)}^{(i)}\right)}{1 + \exp\left(\sum_{j=1}^{p_1} \beta_{j1}X_{j(i)}^{(i)}\right) + \exp\left(\sum_{j=1}^{p_1} \beta_{j2}X_{j(i)}^{(i)}\right)}.
\]

\[
\mathbb{P}_\beta(Y_2(i) = 1|X(i)) = \frac{\exp\left(\sum_{j=1}^{p_1} \beta_{j1}X_{j(i)}^{(i)}\right)}{1 + \exp\left(\sum_{j=1}^{p_1} \beta_{j1}X_{j(i)}^{(i)}\right) + \exp\left(\sum_{j=1}^{p_1} \beta_{j2}X_{j(i)}^{(i)}\right)}.
\]

\[
\mathbb{P}_\beta(Y_3(i) = 1|X(i)) = \frac{1}{1 + \exp\left(\sum_{j=1}^{p_1} \beta_{j1}X_{j(i)}^{(i)}\right) + \exp\left(\sum_{j=1}^{p_1} \beta_{j2}X_{j(i)}^{(i)}\right)}.
\]

The considered dependency structures for \( X_1, \ldots, X_p \) are given in Model 1.
Simulation results

| c  | ρ   | Median of FDP(t) | Std. error of FDP(t) | Mean of R(t) | Std. error of R(t) | Mean of S(t) | Std. error of S(t) | Median of t_{0.05} | Mean S(t_{0.05}) |
|----|-----|------------------|----------------------|--------------|--------------------|--------------|--------------------|-------------------|------------------|
| 1  | 0.2 | 0.00497          | 0.0208               | 5.86         | 1.36               | 5.81         | 1.32               | 1.01e-03            | 8.00             |
| 2  | 0.2 | 0.000505         | 0.0265               | 5.84         | 1.38               | 5.79         | 1.32               | 1.01e-03            | 7.98             |
| 1  | 0.5 | 0.000299         | 0.0316               | 5.84         | 1.37               | 5.81         | 1.32               | 5.06e-03            | 9.08             |
| 2  | 0.5 | 0.000315         | 0.0496               | 5.84         | 1.63               | 5.79         | 1.32               | 5.06e-03            | 9.07             |
| 1  | 0.8 | 0.000168         | 0.0122               | 5.81         | 1.33               | 5.81         | 1.32               | 2.9e-02             | 9.73             |
| 2  | 0.8 | 0.000169         | 0.0402               | 5.84         | 2.69               | 5.79         | 1.32               | 2.9e-02             | 9.75             |

Note: The total number of hypotheses equals p = 1000; the number of false null hypotheses equals p_{1} = 10 per baseline-category pair (non-zero regression coefficients equal 1); the number of factors equals k = 1; the rejection threshold equals t = 10^{-4}, apart from the last two columns.

Model 1. Scenario 1: X_{1}, …, X_{p} are stochastically independent and identically standard normally distributed random variables. Scenario 2: X_{P_{1}+1}, …, X_{p} are jointly normally distributed on \( \mathbb{R}^{p_{1}} \), namely, according to \( N_{p_{1}}(0, \Sigma^*) \), where \( \Sigma^* \) is an equi-correlation matrix with diagonal elements equal to one, and all off-diagonal elements equal to the same given value \( \rho \). Furthermore, the subvector \( (X_{P_{1}+1}, \ldots, X_{p})^{T} \) is stochastically independent of the subvector \( (X_{1}, \ldots, X_{P_{1}})^{T} \). This is in order to avoid spurious effects of covariates \( X_{j} \) with \( P_{1} + 1 \leq j \leq p \) on the response, which would arise from confounding if the two aforementioned subvectors would be dependent. The random variables \( X_{1}, \ldots, X_{P_{1}} \) are stochastically independent and identically \( N(0, 1) \)-distributed. The equi-correlation model is a one-factor model with a dominating first eigenvalue; see, for example, Example 2.1 by Finner et al.\(^{41}\)

This simulation study has been run in \textit{R} version 4.1.1\(^{42}\) employing the \textit{R} function \texttt{rmvnorm()}\(^{43}\) to simulate correlation dependency among the independent variables. In addition, we have employed the function \texttt{vglm()}\(^{44}\) for the estimation of the marginal multinomial estimates. The function \textit{pfa.test()}\(^{45}\) has been used for applying the PFA method.

### 3.2 Simulation results

We summarize the results from our computer simulations in Tables 1–4 in terms of \( \text{FDP}(t) \), \( \text{R}(t) \), and \( \text{S}(t) \) for a fixed threshold \( t = 10^{-4} \). The estimates for \( \text{FDP}(t) \) have been derived by applying Equation (12), and the first columns of each table correspond to the respective baseline-category pair. In addition, we report the median value of \( t_{a} \) for the conventional choice of \( \alpha = 0.05 \) as well as the average (over the 1000 Monte Carlo repetitions) of \( \text{S}(t_{a}) \). Due to the choice of \( p_{1} = 10 \), the largest possible value of \( \text{S}(t) \) equals ten, for any choice of the threshold \( t \). The purpose of the summary tables is to illustrate that the proposed framework can produce meaningful results under different strengths of dependency among the test statistics. In this sense, the simulations serve as a proof of concept for the Multi-PFA approach.

Tables 1 and 2 contain our simulation results under Scenario 1 from Model 1. Because the test statistics are jointly independent (ie, joint independence among all \( X_{i} \)), \( t_{0.05} \) is rather small, because the “effective number of tests” (in the sense of section 3.4 by Dickhaus et al.\(^{11}\)) equals \( p \) under joint independence of all test statistics or \( p \)-values, respectively. This means that a rather strong multiplicity correction is necessary. However, the standard error of \( \text{FDP}(t) \) is also quite small under Scenario 1, since the FDP is well concentrated around its expectation (the FDR) under joint independence across among all test statistics or \( p \)-values, respectively. With regard to type II errors, the reported average values of \( \text{S}(t_{0.05}) \) demonstrate that on average 7–8 out of 10 active coordinates could be identified by our proposed multiple test procedure.

Simulation results under Scenario 2 (Gaussian equi-correlation model) are summarised in Tables 3 and 4. Here, the effective number of tests is smaller than \( p \) whenever \( \rho \geq 0 \), and it decreases as \( \rho \) increases. Consequently, \( t_{0.05} \) grows with \( \rho \), too. In turn, this also leads to an improved power of the proposed multiple test, which is reflected by the numbers reported for \( \text{S}(t) \), which are under Scenario 2 on average larger than the corresponding values under Scenario 1.
FIGURE 1  Examples of mass spectra for three cancer (sub-)types.

4 | REAL DATA APPLICATION: MALDI IMAGING DATA

4.1 | Description of the dataset

We applied our proposed inferential procedure to a hyperspectral dataset obtained from a MALDI imaging instrument. For a detailed description of this dataset in terms of its acquisition protocols, tissue sections, tissue blocks, and so forth.\textsuperscript{20,39}

In the aforementioned studies, the researchers analyzed this dataset by merging two lung cancer subtypes into a single category. By doing so, the authors worked with two cancer types based on two distinctive human organs (the so-called “LP task”, ie, lung vs. pancreas). In our study, however, we model this dataset with a three-class model, using pancreatic adenocarcinoma as a baseline. The resulting goal of the statistical analysis is to simultaneously perform two (sub-) tasks, namely the pancreatic versus lung adenocarcinoma and the pancreatic versus lung squamous cell carcinoma comparison.

As mentioned in the introduction, the high-throughput device that motivated this study is MALDI IMS. In general, this technology provides molecular information about a given analyte (eg, tissue) in a spatial manner. More concretely, the MALDI IMS tool measures mass spectra at multiple discrete spatial positions and yields an image for each unique spot within a provided tissue. For an illustration, Figure 1 displays instances of outputs from a MALDI experiment from three spatial spots. In Figure 1, the m/z values are plotted on the horizontal axes, and the relative abundances (intensities values) of ionizable molecules are plotted on the vertical axes. This “spatial molecular (biochemical) information can then be used for the determination of the cancer (sub-)types or the identification of the origin of the primary tumour in patients”.\textsuperscript{39} Each spot is called a mass spectrum,\textsuperscript{31,46} and it depicts relative abundances of ionizable molecules with numerous mass-to-charge (m/z) values—“ranging from several hundred up to a few tens of thousands of m/z values”\textsuperscript{31}—whereas a single m/z value in the context of MALDI is interpreted as a molecular mass. The full pipeline of a MALDI experiment, from a tissue spot to data analysis, is thoroughly illustrated in Figures 1 and C2.\textsuperscript{20,21} In the dataset analyzed here, formalin-fixed paraffin-embedded (FFPE)\textsuperscript{47} tumour samples from 445 patients were provided by the tissue bank of the
National Center for Tumor Diseases (NCT) Heidelberg, Germany. Tissue cores of three cancer (sub-)types were used, comprising lung adenocarcinoma (ADC) (168 patients), lung squamous cell carcinoma (SqCC) (136 patients) and pancreatic adenocarcinoma (141 patients). The corresponding dataset is publicly available from (https://gitlab.informatik.uni-bremen.de/digipath/Deep_Learning_for_Tumor_Classification_in_IMS–last accessed 7 February 2022).

4.2 Data preprocessing

We performed the following (further) data preprocessing steps. First, due to the high dimensionality of the feature space (27,286 m/z values), spectral filtering was carried out. Namely, we loaded the data into MATLAB R2018, and used the internal library MSClassifyLib (function MSAdaptiveResampleTrigger()) to apply the spectral filtering with the default value of 0.4 Da. Second, the mass range was pruned up to the mass range of 2100 m/z, and we used one mass spectrum from each patient. Finally, the well-established normalisation step TIC (total ion count) was employed on the chosen mass spectra range. This data preprocessing resulted in a dataset with 445 mass spectra (observational units) and 1579 m/z channels (features).

4.3 Analysis of the MALDI dataset

We have modelled each pixel (based on the averaging filter around 0.4 Da as explained in the previous section) individually to determine those m/z values that are highly associative for a certain cancer (sub-)type. Thus, the m/z values take the role of the $X_j$’s and the cancer (sub-)type takes the role of $Y$ from our general setup, where $j \in \{1, \ldots, p = 1579\}$ here. Statistically speaking, our procedure identifies significant differences in distributions across the (sub-)types and is not only identifying significant peaks in the spectra.

Figure 2 displays the empirical distributions of the Z-statistics, along with histograms of the non-adjusted $p$-values, for both tasks. Clearly, the Z-values (for both tasks) do not resemble the realization of a random sample from the theoretical null distribution, which is the standard normal distribution on $\mathbb{R}$. In particular, the empirical variances are much larger than one in both histograms displayed in Figure 2. There are two reasons for this overdispersion, namely, (i) the presence of overdispersion and non-zero effect sizes, there are many non-adjusted $p$-values which are close to zero, in both graphs.

**FIGURE 2** Empirical distributions and fitted normal density curves of the Z-values, as well as histograms of the unadjusted $p$-values, for both tasks. The fitted normal distributions are $N(0.565, 3.365^2)$ for the first task and $N(0.543, 3.394^2)$ for the second task. As a result of overdispersion and non-zero effect sizes, there are many non-adjusted $p$-values which are close to zero, in both graphs.
of correlation effects among the Z-values and (ii) the presence of extremely large (in absolute value) Z-values, which is likely caused by significant effects, peaks and isotopic patterns. In our previous study, we have observed a similar phenomenon and have described a procedure (based on the “empirical null distribution”) which can be used to account for the part of the overdispersion which is caused by correlations among the Z-statistics.

The main results of our real data analysis are presented in Figure 3, where we have chosen $k = 3$ for both tasks. According to the number $p = 1579$ and taking into account the magnitude of the observed correlation effects among the $Z$-statistics, a plausible range for the rejection threshold $t$ has been selected for being displayed on the horizontal axes. The subplots in Figure 3 illustrate the total number of rejections ($R(t)$), the estimated number of false discoveries ($V(t)$) and the estimated FDP ($FDP(t)$) over this range. Apparently, all three aforementioned quantities decrease with decreasing $t$ (notice the negative logarithmic scale of the horizontal axes in Figure 3). For both tasks, $\hat{FDP}(t)$ ranges in $[5\%, 15\%]$ for $t \in [10^{-7}, 10^{-5}]$. Table 5 tabulates $R(t)$ and $\hat{FDP}(t)$ for several threshold values $t$ from the latter range.

To the best of our knowledge, no biomarkers have been confirmed yet for this dataset (pancreas vs. lung task). However, according to the previous reports, m/z values at 836.5 Da, 852.4 Da, and 868.5 Da might be potential biomarkers. Since we have used data based on a different resolution (0.4 Da) compared to the aforementioned reports, we identified as statistically significant m/z-values which are closely related to the previously published insights. Namely, the adjacent molecules have appeared to be extremely statistically significant. In particular, for the m/z-value 836.5 Da, based on the spectral filtering used in our analysis, adjacent m/z values (836.42, 837.41, and 838.42) have been indicated as highly significant for both tasks. Similarly, for the m/z value 852.4 Da, adjacent m/z values have occurred as highly considerable (853.42 and 854.42). Similarly, we observed the m/z value at 868.50 and the adjacent m/z value at 869.43. Figure 4 illustrates the identified peaks related to the pancreatic status. The Multi-PFA method indicates m/z values around 886.42 Da and 852.42 Da as highly statistically significant. Tables 6 and 7 tabulate the 10 top-ranked m/z values for both tasks. Their corresponding Z-values lie in the very tails of the empirical distributions plotted in Figure 2. Negative signs in Tables 6 and 7 indicate those m/z values that are distinctive for the pancreatic cancer type.

Finally, we compare our methodology with two classical procedures for controlling the FDR, namely the BH procedure and the Benjamini-Yekutieli (BY) procedure, in Table 8. We used the R function p.adjust() to carry out the latter multiple tests. These comparisons are relevant since the FDR is the expected value of the FDP, implying that FDP control and FDR control are related concepts. Both the BH and the BY procedures operate on (marginal) $p$-values. To obtain these $p$-values, we consider for each $c \in \{1, 2\}$ the standardized versions $Z_{ic}$, ..., $Z_{pc}$ of $\hat{\beta}_{ic}$, ..., $\hat{\beta}_{pc}$ as given in (7). In particular, the empirical standard errors used for Studentization are taken as the square roots of the diagonal
### TABLE 5  Main results of the multi-PFA analysis for both tasks.

| Threshold $t$ | $R(t)$ | $\text{FDP}(t)$ |
|--------------|--------|-----------------|
| (a) Pancreatic vs. ADC | | |
| 1.12e-05 | 374 | 0.1502 |
| 7.50e-06 | 356 | 0.1357 |
| 4.12e-06 | 332 | 0.1154 |
| 2.76e-06 | 320 | 0.1022 |
| 5.57e-07 | 275 | 0.0609 |
| 3.06e-07 | 253 | 0.0508 |
| (b) Pancreatic vs. SqCC | | |
| 5.03e-06 | 295 | 0.1505 |
| 3.37e-06 | 289 | 0.1349 |
| 2.76e-06 | 280 | 0.1304 |
| 1.85e-06 | 262 | 0.1217 |
| 8.32e-07 | 237 | 0.1013 |
| 6.18e-07 | 168 | 0.0490 |

### FIGURE 4  A comparison of the relative intensities at each m/z value across the (sub-)types for different zooms related to very significant m/z values for the pancreatic association.
TABLE 6 Top ten ranked m/z values based on their original Z-values for the task Pancreatic vs. ADC.

| m/z values | Z-values | P-values |
|------------|----------|----------|
| (a) The most sign. m/z-values for Panc | | |
| 886.44     | -9.917   | < 10^{-6} |
| 840.42     | -9.570   | < 10^{-6} |
| 679.34     | -9.300   | < 10^{-6} |
| 899.45     | -9.176   | < 10^{-6} |
| 854.42     | -8.999   | < 10^{-6} |
| 771.38     | -8.930   | < 10^{-6} |
| 841.42     | -8.866   | < 10^{-6} |
| 898.44     | -8.528   | < 10^{-6} |
| 785.39     | -8.404   | < 10^{-6} |
| (b) The most sign. m/z-values for ADC | | |
| 1647.82    | 8.585    | < 10^{-6} |
| 1575.78    | 8.528    | < 10^{-6} |
| 535.26     | 8.300    | < 10^{-6} |
| 521.26     | 8.252    | < 10^{-6} |
| 1694.84    | 8.245    | < 10^{-6} |
| 1599.79    | 8.197    | < 10^{-6} |
| 531.26     | 8.174    | < 10^{-6} |
| 522.26     | 7.963    | < 10^{-6} |
| 532.26     | 7.844    | < 10^{-6} |
| 529.26     | 7.672    | < 10^{-6} |

elements of $\hat{\Sigma}_c$, divided by $\sqrt{n}$. Next, each $Z_{jc}$ (for $1 \leq j \leq p$) is transformed into the corresponding two-sided $p$-value $P_{jc} = 2\Phi(-|Z_{jc}|) = 2(1 - \Phi(|Z_{jc}|))$ (cf. Figure 2). The resulting numbers of rejections are provided in the columns labeled "BH" and "BY" in Table 8. It is apparent that both BH and BY are (too) liberal in this context. For example, for a nominal level of $\alpha = 0.1$, BH and BY reject approximately 70% of all null hypotheses. The reason for this liberal behavior is that the scaling of the “empirical null distribution” of $Z_{jc}$, … $Z_{pc}$ is different from the unit scale due to strong dependencies, even though each individual $Z_{jc}$ is properly Studentized. In order to further substantiate this point and to rule out the possibility that the observed liberal behavior may result from an inaccurate estimation of $\Sigma_c$, we present the columns labeled BH* and BY* in Table 8. For these columns, the $p$-values have been extracted directly from the marginal model fits (with standard errors based on the marginal Fisher information matrices), without carrying out the MMM approach. The results remain essentially the same. To diminish the variance inflation effects of the dependencies and to make the comparisons with Multi-PFA "fair" and meaningful, we applied Efron’s “empirical correction” of the null distribution. Briefly put, the idea here is to divide the Z-values by an empirical value based on their central spread, where it is assumed that the central part of the empirical distribution of the Z-values presumably corresponds to true null hypotheses. In our case, these empirical values equal 2.48 (Part (a) of Table 8) and 2.50 (Part (b) of Table 8), respectively. We denote by BH and BH* the columns where this empirical correction has been performed. Now, the variance inflation is corrected, but the generic procedures turn out to be conservative in comparison with Multi-PFA. The reason for this is, that the correlation effects among the m/z-values are not accounted for by the BH procedure. We may remark here that the aforementioned variance inflation is taken care of automatically by Multi-PFA. This can be seen by comparing the values in the columns labeled “Threshold t” in Table 5 with the value $\alpha / p$. 

TABLE 7  Top ten ranked m/z values based on their original Z-values for the task Pancreatic vs. Sqcc.

| m/z values | Z-values | P-values |
|------------|----------|----------|
| (a) The most sign. m/z-values for Panc | | |
| 771.38     | −9.285   | < 10^{-6} |
| 795.39     | −8.839   | < 10^{-6} |
| 678.34     | −8.727   | < 10^{-6} |
| 785.39     | −8.390   | < 10^{-6} |
| 840.42     | −8.334   | < 10^{-6} |
| 874.43     | −8.276   | < 10^{-6} |
| 886.44     | −8.274   | < 10^{-6} |
| 796.39     | −8.266   | < 10^{-6} |
| 841.42     | −8.202   | < 10^{-6} |
| 854.42     | −8.122   | < 10^{-6} |
| (b) The most sign. m/z-values for SqCC | | |
| 535.26     | 9.196    | < 10^{-6} |
| 1575.78    | 8.860    | < 10^{-6} |
| 521.26     | 8.664    | < 10^{-6} |
| 530.26     | 8.406    | < 10^{-6} |
| 531.26     | 8.381    | < 10^{-6} |
| 522.26     | 8.339    | < 10^{-6} |
| 1694.84    | 8.125    | < 10^{-6} |
| 1611.80    | 7.946    | < 10^{-6} |
| 565.28     | 7.917    | < 10^{-6} |
| 1764.87    | 7.910    | < 10^{-6} |

TABLE 8  The numbers of rejections for the Multi-PFA, the BH, and the BY procedures are compared for three significance levels.

| Sign. lev. | Multi-PFA | BH  | BY  | BH* | BY* | BH^E | BH^E* |
|------------|-----------|-----|-----|-----|-----|------|-------|
| (a) Pancreatic vs. ADC | | | | | | | |
| α = 0.05   | 253       | 946 | 714 | 957 | 739 | 0    | 20    |
| α = 0.10   | 320       | 1045| 785 | 1068| 812 | 18   | 35    |
| α = 0.15   | 374       | 1120| 839 | 1136| 853 | 115  | 142   |
| (b) Pancreatic vs. SqCC | | | | | | | |
| α = 0.05   | 168       | 901 | 665 | 922 | 688 | 0    | 0     |
| α = 0.10   | 237       | 1018| 735 | 1034| 764 | 18   | 65    |
| α = 0.15   | 295       | 1097| 783 | 1103| 811 | 53   | 104   |

5  DISCUSSION AND OUTLOOK

Motivated by MALDI association studies, this study’s objective has been to evaluate the strength of associations between a nominal variable of interest—describing cancer (sub-)types in our application—and a large number of measured variables (in our case given by m/z values). We have proposed an approach to screen all features and identify the most associative ones with the multi-class outcome. Our approach decomposes the outcome variable into multiple baseline-category pairs and approximates the false discovery proportion under arbitrary correlation dependency for each pair. As demonstrated on simulated data, this approach leads to a sensible balance between false and true findings. Furthermore, both our simulation study and the presented application to real data demonstrate that the proposed method can be applied even in cases
where the sample size is considerably smaller than the number of features. This makes the procedure attractive for medical screening applications. Two central assumptions underlie the proposed methodology: First, we assume sparsity in the sense that the number of active features is small. Second, we assume that the dependency structure amongst the considered $Z$-statistics can accurately be approximated by a multi-factor model. The approach of MMM, however, is flexible in the sense that it does not depend on heavy assumptions. For its applicability, we only have to assume that the multi-class outcome is associated with features (m/z values), and that this association can accurately be described by a (marginal) multinomial regression model for each m/z-value separately. A clear limitation of MMM is that the approach relies on an asymptotic setting, such that small data regimes (small sample sizes) cannot be handled with the proposed methodology. Furthermore, the computational effort of the proposed methodology is quite high, because $p$ marginal models have to be fitted.

From the application perspective, we have applied the Multi-PFA method to a MALDI imaging dataset consisting of a large number of m/z values and one spectrum from each patient. To the best of our knowledge, we are the first to address the three-class problem for this dataset. In this way, we have provided a more detailed analysis for both lung cancer subtypes than in previous studies.

There are several potential directions for further research. First, it might be captivating to consider different supervised learning methods (for instance, neural networks with more than one layer) instead of the multinomial regression model proposed in this paper. Second, it may be of interest to evaluate the uncertainty regarding the realized FDP in order to provide a confidence region for it, in addition to mere point estimation of the FDP. Namely, we aim at establishing an "exceedance control" of the FDP, where the latter is approximated by the PFA estimator. Based on this, one can construct upper confidence bounds for the FDP, ideally simultaneously over a grid of rejection thresholds. Third, disentangling correlation effects and regression effects on the empirical $Z$-statistics distribution in a more detailed manner is an interesting follow-up research topic in our context.

AUTHOR CONTRIBUTIONS
Vladimir Vutov has performed data modelling and analysis, as well as R programming. Thorsten Dickhaus has conceptualized the research project and proposed the usage of MMM. Both authors have written the manuscript together.

ACKNOWLEDGEMENTS
The authors sincerely thank Johannes Leuschner for his invaluable aid in performing the data-processing steps of MALDI data in Matlab R2018. We thank the Editor-in-Chief, the Associate Editor, and two anonymous reviewers for their detailed reading of the manuscript and for their constructive suggestions, which have improved the presentation. Open Access funding enabled and organized by Projekt DEAL.

FUNDING INFORMATION
The authors gratefully acknowledge financial support from the German Research Foundation within the framework of RTG 2224, entitled “$\pi$3: Parameter Identification–Analysis, Algorithms, Applications”.

CONFLICT OF INTEREST STATEMENT
The authors declare no potential conflict of interests.

DATA AVAILABILITY STATEMENT
R code, with which all results presented in the present manuscript can be reproduced, is provided in the data file which is available as supplementary material for this article.

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

REFERENCES
1. Dickhaus T. Simultaneous Statistical Inference with Applications in the Life Sciences. Berlin, Heidelberg: Springer; 2014. doi:10.1007/978-3-642-45182-9
2. Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. J R Stat Soc Ser B Stat Methodol. 1995;57(1):289-300. doi:10.1111/j.2517-6161.1995.tb02031.x
34. Pepe M, Whitaker R, Seidel K. Estimating and comparing univariate associations with application to the prediction of adult obesity. *Stat Med*. 1999;18(2):163-173. doi:10.1002/(SICI)1097-0258(19990130)18:2<163::AID-SIM11>3.0.CO;2-F
35. Liang KY, Zeger SL. Longitudinal data analysis using generalized linear models. *Biometrika*. 1986;73(1):13-22. doi:10.1093/biomet/73.1.13
36. Agresti A. Categorical data analysis. *Wiley Series in Probability and Statistics*. second ed. New York: Wiley-Interscience [John Wiley & Sons]; 2002. doi:10.1002/0471249688
37. Böhning D. Multinomial logistic regression algorithm. *Ann Inst Stat Math*. 1992;44(1):197-200. doi:10.1007/BF00048682
38. Hasan A, Wang Z, Mahani AS. Fast estimation of multinomial logit models: R package mnlogit. *J Stat Softw*. 2016;75(3):1-24. doi:10.18637/jss.v075.i03
39. Behrmann J, et al. Deep learning for tumor classification in imaging mass spectrometry. *Bioinform*. 2018;34(7):1215-1223. doi:10.1093/bioinformatics/btx724
40. Kriegsmann M, et al. 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. *Mol Cell Proteom*. 2016;15(10):3081-3089. doi:10.1074/mcp.m115.057513
41. Finner H, et al. Dependency and false discovery rate: asymptotics. *Ann Stat*. 2007;35(4):1432-1455.
42. R Core Team. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing; 2021 https://www.R-project.org/
43. Genz A, et al. mvtnorm: multivariate normal and t distributions. R package version 1.1-2. 2021 https://CRAN.R-project.org/package=mvtnorm
44. Thomas Y. VGAM: Vector Generalized Linear and Additive Models. R package version 1.1-5. 2021 https://CRAN.R-project.org/package=VGAM
45. Fan J, et al. pfa: Estimates False Discovery Proportion Under Arbitrary Covariance Dependence. R package version 1.1. 2016 https://CRAN.R-project.org/package=pfa
46. Alexandrov T, et al. Super-resolution segmentation of imaging mass spectrometry data: solving the issue of low lateral resolution. *J Proteom*. 2011;75(1):237-245. doi:10.1016/j.jprot.2011.08.002
47. Oetjen J, et al. An approach to optimize sample preparation for MALDI imaging MS of FFPE sections using fractional factorial design of experiments. *Anal Bioanal Chem*. 2016;408:6729-6740. doi:10.1007/s00216-016-9793-4
48. Genovese CR, Wasserman L. Exceedance control of the false discovery proportion. *J Am Stat Assoc*. 2006;101(476):1408-1417.
49. Hemerik J, Goeman JJ. False discovery proportion estimation by permutations: confidence for significance analysis of microarrays. *J R Stat Soc Ser B Stat Methodol*. 2018;80(1):137-155.

**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. Multiple multi-sample testing under arbitrary covariance dependency. *Statistics in Medicine*. 2023;42(17):2944-2961. doi:10.1002/sim.9761