*The Annals of Applied Statistics*
2020, Vol. 14, No. 2, 789–808
https://doi.org/10.1214/20-AOAS1328


# ESTIMATION AND INFERENCE IN METABOLOMICS WITH NONRANDOM MISSING DATA AND LATENT FACTORS

BY CHRIS MCKENNAN[1], CAROLE OBER[2] AND DAN NICOLAE[3]

[1]*Department of Statistics, University of Pittsburgh, chm195@pitt.edu*
[2]*Department of Human Genetics, University of Chicago, c-ober@bsd.uchicago.edu*
[3]*Department of Statistics, University of Chicago, nicolae@galton.uchicago.edu*

High-throughput metabolomics data are fraught with both nonignorable missing observations and unobserved factors that influence a metabolite's measured concentration, and it is well known that ignoring either of these complications can compromise estimators. However, current methods to analyze these data can only account for the missing data or unobserved factors, but not both. We therefore developed MetabMiss, a statistically rigorous method to account for both nonrandom missing data and latent factors in high-throughput metabolomics data. Our methodology does not require the practitioner specify a likelihood for the missing data, and makes investigating the relationship between the metabolome and tens, or even hundreds, of phenotypes computationally tractable. We demonstrate the fidelity of MetabMiss's estimates using both simulated and real metabolomics data and prove their asymptotic correctness when the sample size and number of metabolites grows to infinity.

**1. Introduction.** Metabolomics is the study of tissue- or body fluid-specific small molecule metabolites and has the potential to lead to new insights into the origin of human disease (Finkelstein et al. (2015), Reinke et al. (2017)) and drug metabolism (Dubuis, Ortmayr and Zampieri (2018)). Recent advances in both liquid chromatography (LC) and untargeted mass spectrometry (MS) have made it possible to identify and quantify hundreds to thousands of metabolites per sample (Liu, Ser and Locasale (2014)). Similar to high-throughput gene expression and DNA methylation data, these data contain systematic technical and biological variation whose sources are not observed by the practitioner (Salerno Stephen et al. (2017)). However, what makes untargeted LC-MS metabolomic data particularly challenging is the vast amount of missing data, nearly all of which is missing not at random due to an unknown, metabolite-specific missingness mechanism in which more abundant and ionizable analytes are more likely to be observed (Do et al. (2018)). For instance, 22% of all #metabolites $\times$ #samples $= 1138 \times 533$ observations were missing from our data example in Section 8, where Figure 1 shows that only analytes with the strongest signals were likely to be quantified in all technical replicates.

There are several methods that attempt to account for either latent factors (De Livera et al. (2012), De Livera et al. (2015), Salerno Stephen et al. (2017)) or nonrandom missing data (Chen et al. (2017), O'Brien et al. (2018), Wang et al. (2019)) when trying to infer the relationship between the metabolome and a variable of interest. However, these are not amenable to real, untargeted metabolomic data because the former set of methods cannot accommodate nonignorable missing data, and the latter set ignores latent factors that can bias estimators. Surprisingly, to the best of our knowledge, Wehrens et al. (2016) is the only work to even acknowledge the challenge of accounting for both. However, they propose imputing missing

Received September 2019; revised February 2020.
*Key words and phrases.* Metabolomics, latent factors, batch variables, generalized method of moments, missing not at random (MNAR).

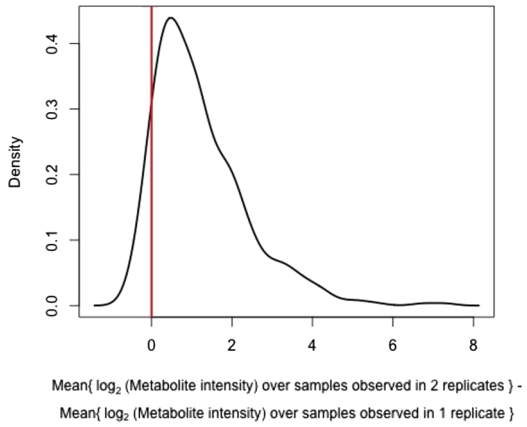


FIG. 1. *A density plot of the differences in mean observed metabolite* $\log_2$*-intensity between samples with observations in both technical replicates and those from samples with only one observation among the two replicates. Replicate pairs were obtained by running* 20 *biological samples from our motivating data example twice on the same mass spectrometer.*

data with an arbitrary limit of detection, and require knowledge of control metabolites whose concentrations are unrelated to the variable of interest to estimate latent factors.

Given the paucity of methods to analyze untargeted metabolomic data, we developed MetabMiss, the first method to account for both latent factors and nonrandom missing data that does not rely on erringly imputing missing data. It leverages both key properties of the missing data in untargeted mass spectrometry experiments, as well as the fact that most of the variation in metabolomic data can be explained by a few, possibly latent, factors, which we treat as instrumental variables to identify and estimate each metabolite's missingness mechanism. Additionally, our method to account for the latent factors uses the natural sparsity in the relationship between the metabolome and the covariate(s) of interest, which obviates the need for control metabolites and internal standards. Our method also offers the following advantages:

(a) It does not require the user specify a likelihood for the missing data.
(b) It can accommodate both discrete and continuous covariates, and learns the degree to which missingness mechanisms are shared between metabolites.
(c) It is modularized so that each metabolite-dependent missingness mechanism is estimated only once per dataset, which makes computation on the order of a phenome wide association study tractable.

And, while we assume the functional form of the missing data mechanism is known, we provide a method to access the veracity of said function for each metabolite.

Property (b) is contrary to other methods designed to account for missing mass spectrometry data, which are only applicable to case-control studies or cannot flexibly learn the similarity between analyte-specific missingness mechanisms (Chen et al. (2017), O'Brien et al. (2018), Wang et al. (2019)). Further, as far as we are aware, our estimates for the missingness mechanisms are the first that satisfy Property (c) and do not depend on the covariate(s) of interest. This makes analyzing modern metabolomic data tractable, as practitioners often want to use the wealth of information available to investigate the relationship between metabolite abundance and many different covariates. We discuss this further in Section 3.1.

The remainder of the manuscript is organized as follows: we give a mathematical description of the data in Section 2 and give an overview of our method in Section 3. We describe how we estimate the metabolite-dependent missingness mechanisms, estimate the coefficients of interest in a linear model and recover latent factors in Sections 4, 5 and 6. We conclude by

illustrating how our method performs in simulated and real metabolomic data in Sections 7 and 8. We provide exact statements and proofs of all theory in the Supplementary Material. An R package implementing our method, together with instructions and code to reproduce the simulations from Section 7, are available from github.com/chrismckennan/MetabMiss.

## 2. Notation and problem set-up.

2.1. *Notation.* Let $n > 0$ be an integer. We let $\mathbf{1}_n \in \mathbb{R}^n$ be the vectors of all ones, $I_n \in \mathbb{R}^{n\times n}$ be the identity matrix, $[n] = \{1, \ldots, n\}$ and $\boldsymbol{x}_i$ be the $i$th element of $\boldsymbol{x} \in \mathbb{R}^n$. For $\boldsymbol{M} \in \mathbb{R}^{n\times m}$, we let $\boldsymbol{M}_{ij} \in \mathbb{R}$, $\boldsymbol{M}_{*j} \in \mathbb{R}^n$ and $\boldsymbol{M}_{i*} \in \mathbb{R}^m$ be the $(i, j)$th element, $j$th column and $i$th rows of $\boldsymbol{M}$, respectively, and define $P_M$ and $P_M^{\perp}$ to be the orthogonal projection matrices onto $\mathrm{im}(\boldsymbol{M}) = \{\boldsymbol{M}\boldsymbol{v} : \boldsymbol{v} \in \mathbb{R}^m\}$ and the null space of $\boldsymbol{M}^{\mathrm{T}}$. We also let $\boldsymbol{X} \dot{\sim} (\boldsymbol{\mu}, \boldsymbol{G})$ if $\mathbb{E}(\boldsymbol{X}) = \boldsymbol{\mu}$ and $\mathbb{V}(\boldsymbol{X}) = \boldsymbol{G}$, $\boldsymbol{X} \sim MN_{m\times n}(\boldsymbol{\mu}, \boldsymbol{V}, \boldsymbol{U})$ if $\boldsymbol{X} \in \mathbb{R}^{m\times n}$ and $\mathrm{vec}(\boldsymbol{X}) \sim N_{mn}(\mathrm{vec}(\boldsymbol{\mu}), \boldsymbol{U} \otimes \boldsymbol{V})$ and, lastly, define $F_\nu(x)$ to be the cumulative distribution function of the t-distribution with $\nu > 0$ degrees of freedom.

2.2. *A description of and model for the data.* Let $y_{gi}$ be the observed or unobserved log-transformed integrated metabolite intensity for metabolite $g \in [p]$ in sample $i \in [n]$, where the mass spectrometer intensity, integrated over time and mass-to-charge ratio, is proportional to a metabolite's concentration (Karpievitch et al. (2010)). Let $\boldsymbol{X} = (\boldsymbol{x}_1 \cdots \boldsymbol{x}_n)^{\mathrm{T}} \in \mathbb{R}^{n\times d}$ and $\boldsymbol{C} = (\boldsymbol{c}_1 \cdots \boldsymbol{c}_n)^{\mathrm{T}} \in \mathbb{R}^{n\times K}$ be observed and unobserved covariates (i.e., latent factors), where the former may contain biological factors like disease status or technical factors like observed batch variables. We assume

$$y_{gi} = \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}_g + \boldsymbol{c}_i^{\mathrm{T}}\boldsymbol{\ell}_g + e_{gi}, \quad e_{gi} \dot{\sim} (0, \sigma_g^2), g \in [p]; i \in [n], \tag{2.1}$$

where our goal is to estimate $\boldsymbol{\beta}_g$. The unobserved covariates $\boldsymbol{c}_i$ can confound the relationship between $\boldsymbol{x}_i$ and $y_{gi}$, and also induce systematic dependencies between metabolites. We assume that $\boldsymbol{c}_1, \ldots, \boldsymbol{c}_n$ are independent and are independent of $\{e_{gi}\}_{g\in[p],i\in[n]}$, where $\{e_{gi}\}_{i\in[n]}$ are independent and identically distributed for each $g \in [p]$ and $\{e_{gi}\}_{g\in[p]}$ are independent for each $i \in [n]$. The latter assumption is standard in models for metabolomic data that account for latent covariates (De Livera et al. (2015), Salerno Stephen et al. (2017)), and we explore our method's robustness to this assumption through simulation in Section 7.

We next define the indicator variable $r_{gi} = I(y_{gi}$ is observed) and assume that for some known cumulative distribution function $\Psi(x)$,

$$\mathbb{P}(r_{gi} = 1 \mid y_{gi}) = \Psi\{\alpha_g(y_{gi} - \delta_g)\}, \quad g \in [p]; i \in [n]. \tag{2.2}$$

The metabolite-dependent scale and location parameters satisfy $\alpha_g > 0$ and $\delta_g \in \mathbb{R}$, where $\alpha_g \searrow 0$ implies the mechanism is missing completely at random (MCAR) and $\alpha_g \nearrow \infty$ implies $y_{gi}$ is left-censored at $\delta_g$. Model (2.2) is consistent with Figure 1 and is a classic model for missing data in untargeted mass spectrometry experiments (Chen et al. (2017), O'Brien et al. (2018), Wang et al. (2019)). It reflects the observation that nearly all missing data in LC-MS metabolomic experiments are a technical artifact of the mass spectrometer, which can only analyze metabolites whose mass spectrometer-determined intensity is above a limit of detection determined by the ambient background noise and the mass spectrometer's sensitivity (Do et al. (2018), Wang et al. (2006)). Typical choices for $\Psi$ include the logistic function (Wang et al. (2019)) and the probit function (O'Brien et al. (2018)). However, we observed in simulations that $\Psi(x) = F_4(x)$ is a more robust option, since its heavy tails make it less sensitive to outliers. This has previously been used as a robust alternative to logistic and probit functions (Kang and Schafer (2007)).

The above model makes no parametric assumptions on the likelihoods $\text{pr}(\boldsymbol{c}_i \mid \boldsymbol{x}_i)$ and $\text{pr}(y_{gi} \mid \boldsymbol{x}_i)$ and, consequently, avoids placing assumptions on the likelihood for the missing data, $\text{pr}(y_{gi} \mid \boldsymbol{x}_i, r_{gi} = 0)$. Implicit in (2.2) is the important assumption that conditional on $\boldsymbol{Y} = (y_{gi})_{g\in[p],i\in[n]} \in \mathbb{R}^{p\times n}$, $\{r_{gi}\}_{g\in[p],i\in[n]}$ are independent, where the distribution of $r_{gi}$ only depends on $y_{gi}$. This may be only approximately true in practice if other intense analytes preclude MS/MS fragmentation, if the mass spectrometer's sensitivity changes over time or if the abundance of exogenous contaminants contributing to the background noise is significantly different between samples. However, properly tuning the dynamic exclusion time, suitably calibrating the mass spectrometer with performance standards and proper sample handling can substantially mitigate these sources bias (Bouhifd et al. (2015), Broadhurst et al. (2018), Johnson et al. (2013)).

**3. A road map of our methodology.** Here, we provide a compendious description of our method to estimate the metabolite-dependent missingness mechanisms, recover $\boldsymbol{C}$ and estimate $\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_p$. We delineate these steps in more detail in Sections 4, 5 and 6.

3.1. *IV-GMM to estimate $\alpha_g$ and $\delta_g$.* We first estimate $\alpha_g$ and $\delta_g$ for metabolites $g$ with missing data. Unlike existing methods, our estimates do not depend on the user-specified $\boldsymbol{X}$ and only need to be estimated once per data matrix $\boldsymbol{Y}$. This makes analyzing modern datasets tractable, as practitioners typically collect a wealth of covariate information for each sample $i$, and therefore will need to infer the relationship between $\boldsymbol{Y}$ and $\boldsymbol{X}$ for many different covariate matrices $\boldsymbol{X}$.

Estimating $\alpha_g$ and $\delta_g$ is challenging for two reasons. First, the population parameters $\alpha_g$ and $\delta_g$ may be difficult to identify because the observed data are not representative of the overall population when $y_{gi}$ is missing not at random (MNAR). Second, the parametric form for the likelihood $\text{pr}(y_{gi} \mid \boldsymbol{x}_i)$ is unknown because $\text{pr}(\boldsymbol{c}_i \mid \boldsymbol{x}_i)$ and $\text{pr}(e_{gi})$ are unknown. We resolve these issues by building upon Wang, Shao and Kim (2014) and use instrumental variable generalized method of moments (IV-GMM) to estimate $\alpha_g$ and $\delta_g$. The instrumental variables, which are always observed, act as a proxy for missing $y_{gi}$ and, therefore, help to identify $\alpha_g$ and $\delta_g$. Further, the moment condition obviates specifying a likelihood for $\text{pr}(y_{gi} \mid \boldsymbol{x}_i)$, which can be difficult or impossible to justify in the presence of nonrandom missing data. To describe the procedure, fix a $g \in [p]$, and let $\boldsymbol{A}_1, \ldots, \boldsymbol{A}_n \in \mathbb{R}^s$ be random vectors such that $r_{gi} \perp\!\!\!\perp \boldsymbol{A}_i \mid y_{gi}$ for all $i \in [n]$. We consider the following observable $s+1$ dimensional function for metabolite $g$:

$$\boldsymbol{h}\{(y_{gi}, r_{gi}, \boldsymbol{A}_i), (\alpha, \delta)\} = (1\boldsymbol{A}_i^{\mathrm{T}})^{\mathrm{T}}\big(1 - r_{gi}[\Psi\{\alpha(y_{gi} - \delta)\}]^{-1}\big), \quad i \in [n], \tag{3.1}$$

$$\mathbb{E}[\boldsymbol{h}\{(y_{gi}, r_{gi}, \boldsymbol{A}_i), (\alpha_g, \delta_g)\}] = \boldsymbol{0}, \quad i \in [n], \tag{3.2}$$

where (3.2) follows from (2.2). In order for (3.2) to identify $\alpha_g$ and $\delta_g$, we require $y_{gi}$ to depend on $\boldsymbol{A}_i$ and, therefore, act as an instrumental variable for $r_{gi}$ (Wang, Shao and Kim (2014)). Otherwise, $\alpha_g$ and $\delta_g$ would not be identifiable because (3.2) would hold at an infinite number of points $(\alpha, \delta)$.

Unfortunately, as is the case with nearly all biological data, $\boldsymbol{Y}$ is typically only weakly dependent on the observed covariates $\boldsymbol{X}$, meaning viable instruments $\boldsymbol{A}_i$ are almost never observed in metabolomic data. Instead, we leverage the fact that the majority of the variation in high-throughput metabolomic data, like nearly all high-throughput biological data, can be be explained by a relatively small number of potentially latent factors (De Livera et al. (2015), Leek and Storey (2007)). For example, applying principal components analysis to the metabolites with complete data from our motivating data example revealed that only 10 components were necessary to explain nearly 50% of the variation in those fully observed data. This fact forms the basis of our method to estimate each metabolite's missingness mechanism, which we briefly describe in Algorithm 3.1.

ALGORITHM 3.1. *Fix* $\epsilon_{\text{miss}} \in [0, 1)$, $K_{\text{miss}} \geq 2$, *and let* $\mathcal{S} = \{g \in [p] : n^{-1} \sum_{i=1}^{n} (1 - r_{gi}) \leq \epsilon_{\text{miss}}\}$ *and* $\mathcal{M} = \{g \in [p] : n^{-1} \sum_{i=1}^{n} (1 - r_{gi}) > \epsilon_{\text{miss}}\}$ *be the metabolites with (nearly) complete and missing data, respectively*:

(1) *Use* $\boldsymbol{Y}_{\mathcal{S}} = (y_{gi})_{g \in \mathcal{S}, i \in [n]}$ *to generate* $K_{\text{miss}}$ *n-dimensional factors that explain most of the variation in* $\boldsymbol{Y}_{\mathcal{S}}$.

(2) *For each* $g \in \mathcal{M}$, *select two out of the* $K_{\text{miss}}$ *factors estimated in Step* (1) *to act as instruments for the missingness indicators* $r_{g1}, \ldots, r_{gn}$.

(3) *For each* $g \in \mathcal{M}$, *use IV-GMM with* (3.1) *and instruments obtained from Step* (2) *to compute estimates for* $\alpha_g$ *and* $\delta_g$, $\hat{\alpha}_g^{(\text{GMM})}$ *and* $\hat{\delta}_g^{(\text{GMM})}$.

(4) *Identify metabolites* $g \in \mathcal{M}$ *whose missing data patterns may not follow Model* (2.2) *using* $\hat{\alpha}_g^{(\text{GMM})}, \hat{\delta}_g^{(\text{GMM})}$ *and the Sargan–Hansen J statistic*.

(5) *Obtain estimates for* $\alpha_g, \delta_g$ *and the weights* $w_{gi} = r_{gi} / \Psi\{\alpha_g(y_{gi} - \delta_g)\}$ *for* $g \in \mathcal{M}$ *and* $i \in [n]$ *using* ***H****ierarchical* ***B****ayesian* ***G****eneralized* ***M****ethod of* ***M****oments* (*HB-GMM*).

We set $\epsilon_{\text{miss}} = 0.05$ in practice because simulations show that trace amounts of missing data have negligible effects on the bias in our downstream estimators for $\boldsymbol{\beta}_g$. We explain how we choose $K_{\text{miss}}$ in Section 4.2. Algorithm 3.1 tends to perform well because the estimated factors from Step (1) will be approximately the columns of $(\boldsymbol{X}\boldsymbol{C})$ from Model (2.1) that explain much of the variance in $\boldsymbol{Y}$. And, since they are not estimated using metabolites with missing data, they will be approximately independent of $r_{g1}, \ldots, r_{gn}$ conditional on $y_{g1}, \ldots, y_{gn}$ and, therefore, auspicious instruments for $r_{g1}, \ldots, r_{gn}$ for $g \in \mathcal{M}$. Step (4) helps flag metabolites whose missingness mechanisms cannot be reliably estimated, and Step (5) utilizes the output from Step (3) to determine sample weights used to estimate $\boldsymbol{\beta}_g$. We detail and provide concise, intuitive explanations of Steps (1)–(5) in Sections 4.1–4.5 below. We also justify Algorithm 3.1 in Sections S8–S10 of the Supplementary Material (McKennan, Ober and Nicolae (2020)), where we study the asymptotic properties of the estimators from each step when $\epsilon_{\text{miss}} = 0$ and $n, p \to \infty$.

3.2. *Recovering latent factors and estimating coefficients of interest.* Estimating $\boldsymbol{C}$ is challenging because $\boldsymbol{C}$ is not expected to be orthogonal to $\boldsymbol{X}$. For example, diet, an important source of variation in metabolomic studies that is typically unobserved by the researcher (O'Sullivan, Gibney and Brennan (2010)), is often correlated with $\boldsymbol{X}$ (Afshin et al. (2019)). This could potentially complicate estimators, since one would need to be careful to avoid attributing variation due to $\boldsymbol{C}$ as arising from $\boldsymbol{X}$ or vice versa. Therefore, we partition $\boldsymbol{C}$ as $\boldsymbol{C} = P_X^{\perp}\boldsymbol{C} + P_X\boldsymbol{C}$ and note that, since variation due to $P_X^{\perp}\boldsymbol{C}$ is unequivocally distinguishable from that due to $\boldsymbol{X}$, we first estimate $P_X^{\perp}\boldsymbol{C}$ and subsequently use said estimate, as well as the natural sparsity of $(\boldsymbol{\beta}_1 \cdots \boldsymbol{\beta}_p)$, to carefully recover $P_X\boldsymbol{C}$. We then plug-in our estimate for $\boldsymbol{C}$ when estimating $\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_p$.

The aforementioned procedure is computational fast and only depends on the metabolite-dependent missingness mechanisms through the output of Algorithm 3.1. This ensures that Algorithm 3.1 only has to be run once per data matrix $\boldsymbol{Y}$, and makes analyzing the relationship between $\boldsymbol{Y}$ and tens, or even hundreds, of different $\boldsymbol{X}$'s computationally tractable.

## 4. Estimating the missingness mechanisms using Algorithm 3.1.

4.1. *Estimating the instruments in Step* (1). We define the factors from Step (1) of Algorithm 3.1 to be $\hat{\boldsymbol{C}}_{\text{miss}} \in \mathbb{R}^{n \times K_{\text{miss}}}$, where $\hat{\boldsymbol{C}}_{\text{miss}}$ is the maximum likelihood estimator for $\tilde{\boldsymbol{C}} \in \mathbb{R}^{n \times K_{\text{miss}}}$ in the model

$$\boldsymbol{Y}_{\mathcal{S}} \sim MN_{p_s \times n}(\tilde{\boldsymbol{\mu}}\boldsymbol{1}_n^{\mathrm{T}} + \tilde{\boldsymbol{L}}\tilde{\boldsymbol{C}}^{\mathrm{T}}, \tilde{\sigma}^2 I_{p_s}, I_n), \tag{4.1}$$

where $p_s = |\mathcal{S}|$, $\tilde{\boldsymbol{C}}^{\mathrm{T}}\mathbf{1}_n = \mathbf{0}$, $n^{-1}\tilde{\boldsymbol{C}}^{\mathrm{T}}\tilde{\boldsymbol{C}} = I_{K_{\mathrm{miss}}}$, $\tilde{\boldsymbol{L}}^{\mathrm{T}}\tilde{\boldsymbol{L}}$ is diagonal with nonincreasing elements and any missing data are MCAR. If $\epsilon_{\mathrm{miss}} = 0$, $\hat{\boldsymbol{C}}_{\mathrm{miss}}$ is a scalar multiple of the first $K_{\mathrm{miss}}$ right singular vectors of $\boldsymbol{Y}_{\mathcal{S}} P_{1_n}^{\perp}$. When $\epsilon_{\mathrm{miss}} > 0$, the columns of $\hat{\boldsymbol{C}}_{\mathrm{miss}}$ are still ordered by decreasing average effect on the log intensities of metabolites with nearly complete data. Further, by McDiarmid's Inequality, $\mathbb{P}(g \in \mathcal{S}) \leq e^{-2\eta^2 n}$ if $n^{-1}\sum_{i=1}^{n}\mathbb{E}(1 - r_{gi}) \geq \epsilon_{\mathrm{miss}} + \eta$ for $\eta > 0$, meaning it suffices to assume $\hat{\boldsymbol{C}}_{\mathrm{miss}} \perp\!\!\!\perp r_{gi} \mid y_{gi}$ if $g \in \mathcal{M}$ for sufficiently large $n$. That is, for $\boldsymbol{h}$ defined in (3.1) and $\hat{\boldsymbol{c}}_i$ the $i$th row of $\hat{\boldsymbol{C}}_{\mathrm{miss}}$, we assume $\mathbb{E}[\boldsymbol{h}\{(y_{gi}, r_{gi}, \hat{\boldsymbol{c}}_i), (\alpha_g, \delta_g)\}] = \mathbf{0}$ for $g \in \mathcal{M}$ and $i \in [n]$.

The columns of $\hat{\boldsymbol{C}}_{\mathrm{miss}}$ are the factors that explain the most variation in $\boldsymbol{Y}_{\mathcal{S}}$. While we expect most of them to derive from $\boldsymbol{C}$, some may be related to $\boldsymbol{X}$ if $\{\boldsymbol{\beta}_g\}_{g\in\mathcal{S}}$ are large enough. Note that $\hat{\boldsymbol{C}}_{\mathrm{miss}}$ is invariant of the choice of $\boldsymbol{X}$.

4.2. *Instrument selection in Step* (2). It is critical that $y_{gi}$ be dependent on the instruments chosen in Step (2) of Algorithm 3.1. Otherwise, the moment condition in (3.1) will not identify the parameters $\alpha_g$ and $\delta_g$. We therefore use Algorithm 4.1 to only select the instruments $\hat{\boldsymbol{U}}_g \in \mathbb{R}^{n\times 2}$ that influence metabolite $g$'s intensity.

ALGORITHM 4.1. *Let* $\hat{\boldsymbol{C}}_{\mathrm{miss}} = (\hat{\boldsymbol{C}}_1 \cdots \hat{\boldsymbol{C}}_{K_{\mathrm{miss}}})$ *and* $\boldsymbol{y}_g = (y_{gi})_{i\in[n]}$:

(1) *For each* $g \in \mathcal{M}$ *and* $k \in [K_{\mathrm{miss}}]$, *use ordinary least squares* (*OLS*) *to regress* $\boldsymbol{y}_g$ *onto* $(\mathbf{1}_n \hat{\boldsymbol{C}}_k) \in \mathbb{R}^{n\times 2}$, *where missing data are treated as MCAR. Let* $p_{g,k}$ *be the OLS P value for the null hypothesis that* $\hat{\boldsymbol{C}}_k$ *is independent of* $\boldsymbol{y}_g$.

(2) *For each* $k \in [K_{\mathrm{miss}}]$, *use* $\{p_{g,k}\}_{g\in\mathcal{M}}$ *to determine the corresponding q-values* $\{q_{g,k}\}_{g\in\mathcal{M}}$.

(3) *For each* $g \in \mathcal{M}$, *let* $q_{g,g_1} \leq \cdots \leq q_{g,g_{K_{\mathrm{miss}}}}$ *be the* $K_{\mathrm{miss}}$ *ordered q-values. Define* $\hat{\boldsymbol{U}}_g = (\hat{\boldsymbol{u}}_{g1} \cdots \hat{\boldsymbol{u}}_{gn})^{\mathrm{T}} = (\hat{\boldsymbol{C}}_{g_1} \hat{\boldsymbol{C}}_{g_2})$.

Step (1) is justified by Theorem S8.1 in Section S8 of the Supplementary Material (McKennan, Ober and Nicolae (2020)), which states that under technical assumptions on the distributions of $\boldsymbol{Y}_{\mathcal{S}}$ and $\boldsymbol{y}_g$ for $g \in \mathcal{M}$, $p_{g,k}$ is asymptotically uniform under the null hypothesis $H_0^{(g,k)}$ that $\hat{\boldsymbol{C}}_k$ is independent of $\boldsymbol{y}_g$. The q-value $q_{g,k}$ in Step (2), defined as the minimum false discovery rate necessary to reject $H_0^{(g,k)}$, is the multiple testing analogue of the $P$ value $p_{g,k}$ and is estimated using Storey et al. (2015). A small q-value $q_{g,k}$ therefore implies $\hat{\boldsymbol{C}}_k$ is a viable instrument for $r_{gi}$. We also use Algorithm 4.1 to choose $K_{\mathrm{miss}}$. If $f(k)$ is the fraction of metabolites $g \in \mathcal{M}$ such that $q_{g,g_2} \leq 0.05$ assuming $K_{\mathrm{miss}} = k$, we set $K_{\mathrm{miss}} = \min\{k \in \{2, \ldots, K_{PA}\} : f(k) \geq 0.9\}$, where $K_{PA}$ is parallel analysis' (Buja and Eyuboglu (1992)) estimate for $K$ under Model (4.1) with $\epsilon_{\mathrm{miss}} = 0$. The estimate $K_{\mathrm{miss}}$ is typically much smaller than $K_{PA}$ in practice. For example, $K_{\mathrm{miss}} = 10$ and $K_{PA} = 20$ in our motivating data example. We show that our results are robust to the choice of $K_{\mathrm{miss}}$ in Section 7.

Evidently, this selection step implies $\hat{\boldsymbol{u}}_{gi}$ is not strictly independent of $r_{gi}$ conditional on $y_{gi}$. However, we show in Section S8 of the Supplementary Material that this dependence is asymptotically negligible under weak assumptions (McKennan, Ober and Nicolae (2020)). Therefore, we assume that the indices $g_1, g_2$ are known and $\hat{\boldsymbol{u}}_{gi} \perp\!\!\!\perp r_{gi} \mid y_{gi}$ for the remainder of Section 4.

4.3. *IV-GMM in Step* (3). Fix a $g \in \mathcal{M}$ and define

$$\boldsymbol{h}_{gi}(\alpha, \delta) = \boldsymbol{h}\{(y_{gi}, r_{gi}, \hat{\boldsymbol{u}}_{gi}), (\alpha, \delta)\} \in \mathbb{R}^3, \qquad \bar{\boldsymbol{h}}_g(\alpha, \delta) = n^{-1} \sum_{i=1}^{n} \boldsymbol{h}_{gi}(\alpha, \delta). \tag{4.2}$$

We let $\{\hat{\alpha}_g^{(\mathrm{GMM})}, \hat{\delta}_g^{(\mathrm{GMM})}\}$ be the two-step generalized method of moments estimator

$$\{\hat{\alpha}_g^{(\mathrm{GMM})}, \hat{\delta}_g^{(\mathrm{GMM})}\} = \operatorname*{arg\,min}_{\alpha>0, \delta \in \mathbb{R}} \{\bar{\boldsymbol{h}}_g(\alpha, \delta)^{\mathrm{T}} \boldsymbol{W}_g \bar{\boldsymbol{h}}_g(\alpha, \delta)\}, \tag{4.3}$$

where, for $\{\hat{\alpha}_g^{(1)}, \hat{\delta}_g^{(1)}\} = \arg\min_{\alpha>0, \delta \in \mathbb{R}} \{\bar{\boldsymbol{h}}_g(\alpha, \delta)^{\mathrm{T}} \bar{\boldsymbol{h}}_g(\alpha, \delta)\}$, the weight matrix $\boldsymbol{W}_g$ is

$$\boldsymbol{W}_g = \boldsymbol{W}_g\{\hat{\alpha}_g^{(1)}, \hat{\delta}_g^{(1)}\} = \left[ n^{-1} \sum_{i=1}^{n} \boldsymbol{h}_{gi}\{\hat{\alpha}^{(1)}, \hat{\delta}^{(1)}\} \boldsymbol{h}_{gi}\{\hat{\alpha}^{(1)}, \hat{\delta}^{(1)}\}^{\mathrm{T}} \right]^{-1}. \tag{4.4}$$

The properties of this estimator when $\hat{\boldsymbol{U}}_g$ is observed and not estimated and the triplets $\{(r_{gi}, y_{gi}, \hat{\boldsymbol{u}}_{gi})\}_{i \in [n]}$ are independent are well understood (Hansen (1982), Wang, Shao and Kim (2014)). We extend these results in Theorem S10.1 in the Supplementary Material (McKennan, Ober and Nicolae (2020)) to account for the uncertainty in $\hat{\boldsymbol{U}}_g$ and prove that under similar regularity conditions as those considered in Wang, Shao and Kim (2014), both $|\hat{\alpha}_g^{(\mathrm{GMM})} - \alpha_g|$ and $|\hat{\delta}_g^{(\mathrm{GMM})} - \delta_g|$ are $O_P(n^{-1/2})$ and for $\boldsymbol{\Gamma}_g(\alpha, \delta) = \nabla_{\alpha,\delta} \bar{\boldsymbol{h}}_g(\alpha, \delta) \in \mathbb{R}^{3 \times 2}$,

$$n^{1/2} \hat{\boldsymbol{V}}_g^{-1/2} \big[ \{\hat{\alpha}_g^{(\mathrm{GMM})}, \hat{\delta}_g^{(\mathrm{GMM})}\} - (\alpha_g, \delta_g) \big] \xrightarrow{d} N_2(\boldsymbol{0}, I_2), \tag{4.5a}$$

$$\hat{\boldsymbol{V}}_g = \big[ \boldsymbol{\Gamma}_g\{\hat{\alpha}_g^{(\mathrm{GMM})}, \hat{\delta}_g^{(\mathrm{GMM})}\}^{\mathrm{T}} \boldsymbol{W}_g \boldsymbol{\Gamma}_g\{\hat{\alpha}_g^{(\mathrm{GMM})}, \hat{\delta}_g^{(\mathrm{GMM})}\} \big]^{-1} \tag{4.5b}$$

as $n, p \to \infty$. This result is analogous to Theorem 2 in Wang, Shao and Kim (2014), and we use (4.5) to refine our estimates for $\alpha_g$ and $\delta_g$ in Section 4.5.

4.4. *The Sargan–Hansen J statistic in Step* (4). The accuracy of downstream estimates for $\boldsymbol{\beta}_g$ is contingent on the missing data model being approximately correct. Therefore, we leverage the fact that we use three moment conditions to estimate two parameters and use the Sargan–Hansen $J$ statistic, which is routinely used to test moment restrictions in generalized method of moment estimators (Baum, Schaffer and Stillman (2003), Davidson and MacKinnon (2003), Hansen (1982)), to flag metabolites whose missingness mechanisms may not follow Model (2.2).

A consequence of (4.5) is that, under the null hypothesis $H_{0,g}$ that Model (2.2) is correct for metabolite $g \in \mathcal{M}$ and the assumptions necessary to prove (4.5) hold, the statistic $J_g = n \bar{\boldsymbol{h}}_g\{\hat{\alpha}_g^{(\mathrm{GMM})}, \hat{\delta}_g^{(\mathrm{GMM})}\}^{\mathrm{T}} \boldsymbol{W}_g \bar{\boldsymbol{h}}_g\{\hat{\alpha}_g^{(\mathrm{GMM})}, \hat{\delta}_g^{(\mathrm{GMM})}\}$ is asymptotically $\chi_1^2$ as $n, p \to \infty$, which is analogous to Lemma 4.2 in Hansen (1982). One could then use $J_g$ to test $H_{0,g}$. However, it has been repeatedly observed that using said asymptotic distribution to do inference with $J_g$ is anticonservative in data with moderate and even large sample sizes (Brown and Newey (2002), Hall and Horowitz (1996), Hansen and West (2002)). To circumvent this, we followed Brown and Newey (2002) and developed an empirical likelihood-derived bootstrap null distribution for $J_g$ to determine the $P$ value for $H_{0,g}$. We subsequently use Storey et al. (2015) to estimate $lfdr_g = \mathbb{P}(H_{0,g} \mid J_g)$ and flag any metabolites with an $lfdr_g$ smaller than a user-specified value, which defaults to 0.8 in our software. Section S2 in the Supplementary Material describes the details of the bootstrap procedure (McKennan, Ober and Nicolae (2020)).

4.5. *HB-GMM in Step* (5). So far we have estimated all $|\mathcal{M}|$ missingness mechanisms independently of one another. While the mechanisms are almost certainly not identical, one might expect them to be relatively similar, meaning one should be able to design a better estimator by pooling information across metabolites. Further, constructing an informative prior for the missingness mechanisms allows one to better explore the possible multimodal objective function in (4.3) (Franks, Airoldi and Rubin (2016)). We therefore developed Hierarchical Bayesian Generalized Method of Moments (HB-GMM), a Bayesian method to estimate $\alpha_g$, $\delta_g$ and the weights $w_{gi} = r_{gi}/\Psi\{\alpha_g(y_{gi} - \delta_g)\}$ for each $g \in \mathcal{M}$ and $i \in [n]$. The weights play an important role in estimating $\boldsymbol{C}$ in Section 6.

Our method extends Bayesian generalized method of moments (Kim (2002), Li and Jiang (2016), Yin (2009)) by both incorporating estimated instruments and estimating an informative prior from the data. Define $\mathcal{D} = \{(y_{gi}, r_{gi}, \hat{\boldsymbol{u}}_{gi})\}_{g\in\mathcal{M}, i\in[n]}$. By Bayes' rule and assuming $\{(\alpha_g, \delta_g)\}_{g\in\mathcal{M}}$ are independently drawn from some prior distribution,

$$\operatorname{pr}\big[\{(\alpha_g, \delta_g)\}_{g\in\mathcal{M}} \mid \mathcal{D}\big] \propto \operatorname{pr}\big[\mathcal{D} \mid \{(\alpha_g, \delta_g)\}_{g\in\mathcal{M}}\big] \prod_{g\in\mathcal{M}} \operatorname{pr}(\alpha_g, \delta_g). \tag{4.6}$$

However, the likelihood $\operatorname{pr}[\mathcal{D} \mid \{(\alpha_g, \delta_g)\}_{g\in\mathcal{M}}]$ is unknown because the distribution of $y_{gi}$ is unknown. Nevertheless, we do know that, under Model (2.2) and assuming $\hat{\boldsymbol{u}}_{gi} \perp\!\!\!\perp r_{gi} \mid y_{gi}$ for all $g \in \mathcal{M}$ and $i \in [n]$, $\bar{\boldsymbol{h}}_g(\alpha_g, \delta_g)$ and $\bar{\boldsymbol{h}}_s(\alpha_s, \delta_s)$ are uncorrelated for $g \neq s \in \mathcal{M}$. Further, since $\bar{\boldsymbol{h}}_g(\alpha_g, \delta_g)$ is an average of $n$ approximately independent random variables, $\bar{\boldsymbol{h}}_g(\alpha_g, \delta_g)$ has the following asymptotic distribution under the same assumptions used to prove (4.5):

$$\begin{aligned} &n^{1/2}\{\hat{\boldsymbol{\Sigma}}_g(\alpha_g, \delta_g)\}^{-1/2}\bar{\boldsymbol{h}}_g(\alpha_g, \delta_g) \xrightarrow{d} N_3(\boldsymbol{0}, I_3) \quad \text{as } n, p \to \infty, g \in \mathcal{M}, \\ &\hat{\boldsymbol{\Sigma}}_g(\alpha_g, \delta_g) = n^{-1}\sum_{i=1}^{n}\{\boldsymbol{h}_{gi}(\alpha_g, \delta_g) - \bar{\boldsymbol{h}}_g(\alpha_g, \delta_g)\}\{\boldsymbol{h}_{gi}(\alpha_g, \delta_g) - \bar{\boldsymbol{h}}_g(\alpha_g, \delta_g)\}^{\mathrm{T}}. \end{aligned} \tag{4.7}$$

These facts help to justify replacing the likelihood in (4.6) with the pseudolikelihood

$$q\big[\mathcal{D} \mid \{(\alpha_g, \delta_g)\}_{g\in\mathcal{M}}\big] = \prod_{g\in\mathcal{M}} \mathcal{N}\{\bar{\boldsymbol{h}}_g(\alpha_g, \delta_g) \mid \boldsymbol{0}, n^{-1}\hat{\boldsymbol{\Sigma}}_g(\alpha_g, \delta_g)\},$$

where $\mathcal{N}(\cdot \mid \boldsymbol{a}, \boldsymbol{b})$ is the likelihood of a normal distribution with mean $\boldsymbol{a}$ and variance $\boldsymbol{b}$. The form that the pseudolikelihood takes is computationally convenient because it implies we can sample from the pseudoposterior

$$q\big[\{(\alpha_g, \delta_g)\}_{g\in\mathcal{M}} \mid \mathcal{D}\big] \propto \prod_{g\in\mathcal{M}} \big[\mathcal{N}\{\bar{\boldsymbol{h}}_g(\alpha_g, \delta_g) \mid \boldsymbol{0}, n^{-1}\hat{\boldsymbol{\Sigma}}_g(\alpha_g, \delta_g)\} \operatorname{pr}(\alpha_g, \delta_g)\big]$$

with Markov chain Monte Carlo using $|\mathcal{M}|$ parallel chains, which we use to obtain:

$$\hat{\alpha}_g = \mathbb{E}(\alpha_g \mid \mathcal{D}), \qquad \hat{\delta}_g = \mathbb{E}(\delta_g \mid \mathcal{D}), \quad g \in \mathcal{M}, \tag{4.8a}$$

$$\hat{w}_{gi} = \mathbb{E}(w_{gi} \mid \mathcal{D}) = r_{gi}\mathbb{E}\big[1/\Psi\{\alpha_g(y_{gi} - \delta_g)\} \mid \mathcal{D}\big], \quad g \in \mathcal{M}; i \in [n], \tag{4.8b}$$

$$\hat{v}_{gi} = \mathbb{E}(w_{gi}^2 \mid \mathcal{D}) = r_{gi}\mathbb{E}\big[1/\Psi\{\alpha_g(y_{gi} - \delta_g)\}^2 \mid \mathcal{D}\big], \quad g \in \mathcal{M}; i \in [n]. \tag{4.8c}$$

This technique of replacing the likelihood with the pseudolikelihood in (4.6) is standard in Bayesian GMM when $|\mathcal{M}| = 1$ and $\hat{\boldsymbol{U}}_g$ is observed (Kim (2002), Li and Jiang (2016), Yin (2009)).

It remains to specify the prior for $(\alpha_g, \delta_g)$. We assume that $(\log(\alpha_g), \delta_g)^{\mathrm{T}} \mid (\boldsymbol{\mu}, \boldsymbol{U}) \sim N_2(\boldsymbol{\mu}, \boldsymbol{U})$ for all $g \in \mathcal{M}$, where we log-transform $\alpha_g$ to make it amenable to a normal prior.

We first estimate $\boldsymbol{\mu}$ as $\hat{\boldsymbol{\mu}} = |\mathcal{M}|^{-1}\sum_{g\in\mathcal{M}}(\log\{\hat{\alpha}_g^{(\mathrm{GMM})}\}, \hat{\delta}_g^{(\mathrm{GMM})})^{\mathrm{T}}$. Assuming (4.5) is approximately correct, we then use empirical Bayes and define our estimate for $\boldsymbol{U}$, $\hat{\boldsymbol{U}}$, as the maximizer of the following objective over $\boldsymbol{U} \succ \mathbf{0}$:

$$\prod_{g\in\mathcal{M}}\int \mathcal{N}\big[(\log\{\hat{\alpha}_g^{(\mathrm{GMM})}\}, \hat{\delta}_g^{(\mathrm{GMM})})^{\mathrm{T}} \mid (\eta_g, \delta_g)^{\mathrm{T}}, \hat{\boldsymbol{R}}_g\big]\mathcal{N}\big\{(\eta_g, \delta_g)^{\mathrm{T}} \mid \hat{\boldsymbol{\mu}}, \boldsymbol{U}\big\}\, d\eta_g\, d\delta_g,$$

where $\hat{\boldsymbol{R}}_g = \mathrm{diag}\{1/\hat{\alpha}_g^{(\mathrm{GMM})}, 1\}\hat{\boldsymbol{V}}_g\,\mathrm{diag}\{1/\hat{\alpha}_g^{(\mathrm{GMM})}, 1\}$ for $\hat{\boldsymbol{V}}_g$ defined in (4.5). We estimate $\boldsymbol{U}$ using the product of marginal likelihoods because, under the assumptions used to prove (4.5), the estimates $(\hat{\alpha}_g^{(\mathrm{GMM})}, \hat{\delta}_g^{(\mathrm{GMM})})$ and $(\hat{\alpha}_s^{(\mathrm{GMM})}, \hat{\delta}_s^{(\mathrm{GMM})})$ are asymptotically independent for $g \neq s \in \mathcal{M}$. See Section S10 in the Supplementary Material for more details (McKennan, Ober and Nicolae (2020)).

**5. Estimating coefficients when $\boldsymbol{C}$ is known.** Here, we describe our method for estimating $\boldsymbol{\beta}_g$ and $\boldsymbol{\ell}_g$ in Model (2.1) when $\boldsymbol{C}$ is known, which is based on inverse probability weighting (Liang and Qin (2000)). This methodology is used in Section 6 to recover $\boldsymbol{C}$, and is also our default method to perform inference on the coefficients of interest because estimates are consistent, it obviates specifying a probability model for the missing data and computation is fast enough to perform a metabolite phenome wide association study. For simplicity, we rewrite (2.1) as

$$y_{gi} = \boldsymbol{z}_i^{\mathrm{T}}\boldsymbol{\eta}_g + e_{gi}, \quad e_{gi} \dot{\sim} (0, \sigma_g^2),\ g \in [p];\ i \in [n]$$

for the remainder of Section 5. Since estimation is trivial when there is little missing data, our goal is to estimate $\boldsymbol{\eta}_g$ for all $g \in \mathcal{M}$ when $\boldsymbol{Z} = (\boldsymbol{z}_1 \cdots \boldsymbol{z}_n)^{\mathrm{T}}$ is observed.

5.1. *Point estimates.* Fix a $g \in \mathcal{M}$, and, for all $i \in [n]$, define the score function $\boldsymbol{s}_{gi}(\boldsymbol{\eta}) = \boldsymbol{z}_i(y_{gi} - \boldsymbol{z}_i^{\mathrm{T}}\boldsymbol{\eta})$, $\gamma_{gi} = \mathbb{P}(r_{gi} = 1 \mid \boldsymbol{Z})$ and the inverse probability weighted estimating equation $\boldsymbol{f}_g(\boldsymbol{\eta}) = \sum_{i=1}^n \hat{\gamma}_{gi}\hat{w}_{gi}\boldsymbol{s}_{gi}(\boldsymbol{\eta})$, where $\hat{w}_{gi}$ is defined in (4.8b) and $\hat{\gamma}_{gi}$ is an estimate of $\gamma_{gi}$. If $\hat{w}_{gi} = w_{gi}$ and $\hat{\gamma}_{gi} = \gamma_{gi}$ for all $i \in [n]$, then

$$\mathbb{E}\{\boldsymbol{f}_g(\boldsymbol{\eta}_g) \mid \boldsymbol{Z}\} = \sum_{i=1}^n \gamma_{gi}\mathbb{E}\{\mathbb{E}(w_{gi} \mid y_{gi})\boldsymbol{s}_{gi}(\boldsymbol{\eta}_g) \mid \boldsymbol{Z}\} = \sum_{i=1}^n \gamma_{gi}\mathbb{E}\{\boldsymbol{s}_{gi}(\boldsymbol{\eta}_g) \mid \boldsymbol{Z}\} = \mathbf{0}.$$

The above equality can be shown to hold in the more general case when $\gamma_{gi} \perp\!\!\!\perp \boldsymbol{y}_g \mid \boldsymbol{Z}$ for all $i \in [n]$, meaning the root of $\boldsymbol{f}_g$ will be an accurate estimate of $\boldsymbol{\eta}_g$ if $\hat{w}_{gi}$ is consistent for $w_{gi}$ and $\hat{\gamma}_{gi}$ is only weakly dependent on $\boldsymbol{y}_g$. Since $\hat{\gamma}_{gi}$ will tend to be small if $\hat{w}_{gi}$ is large, including $\hat{\gamma}_{gi}$ in $\boldsymbol{f}_g$ has the effect of stabilizing potentially large weights $\hat{w}_{gi}$, thereby reducing the variance of our downstream estimates. This method of stabilized inverse probability weighting has been successfully applied to data that are missing at random (Xu et al. (2010a)), and we estimate $\gamma_{gi}$ using a logistic regression with the estimated instruments $\hat{\boldsymbol{U}}_g$. We then define our estimate for $\boldsymbol{\eta}_g$ as the root of $\boldsymbol{f}_g$,

$$\hat{\boldsymbol{\eta}}_g = (\boldsymbol{Z}^{\mathrm{T}}\hat{\boldsymbol{W}}_g\boldsymbol{Z})^{-1}\boldsymbol{Z}^{\mathrm{T}}\hat{\boldsymbol{W}}_g\boldsymbol{y}_g, \qquad \hat{\boldsymbol{W}}_g = \mathrm{diag}(\hat{w}_{g1}\hat{\gamma}_{g1}, \ldots, \hat{w}_{gn}\hat{\gamma}_{gn}). \tag{5.1}$$

Note $\boldsymbol{s}_{gi}$ in $\boldsymbol{f}_g$ can be redefined to be any M-estimator, like Huber's or Tukey's robust estimators, provided $\mathbb{E}\{\boldsymbol{s}_{gi}(\boldsymbol{\eta}_g) \mid \boldsymbol{Z}\} = \mathbf{0}$.

5.2. *Quantifying uncertainty.* Fix a $g \in \mathcal{M}$. Here, we describe our estimator for $\mathbb{V}(\hat{\boldsymbol{\eta}}_g)$, which we use to both recover $\boldsymbol{C}$ in Section 6 and perform inference on $\boldsymbol{\eta}_g$. Our estimator is a novel finite sample-corrected sandwich variance estimator that also accounts for the uncertainty in the estimated weights $\hat{w}_{gi}$.

Suppose, for simplicity, that $\hat{w}_{gi} = w_{gi}$ and $\hat{\gamma}_{gi} = \gamma_{gi}$. Then,

$$n^{1/2}(\boldsymbol{\eta}_g - \hat{\boldsymbol{\eta}}_g) = (n^{-1}\boldsymbol{Z}^{\mathrm{T}}\hat{\boldsymbol{W}}_g\boldsymbol{Z})^{-1}\left(n^{-1/2}\sum_{i=1}^{n}\gamma_{gi}w_{gi}e_{gi}\boldsymbol{z}_i\right).$$

Since $\{(y_{gi}, r_{gi})\}_{i\in[n]}$ are mutually independent and $\mathbb{E}(\gamma_{gi}w_{gi}e_{gi} \mid \boldsymbol{Z}) = 0$,

$$n\mathbb{V}(\hat{\boldsymbol{\eta}}_g \mid \boldsymbol{Z}) \approx (n^{-1}\boldsymbol{Z}^{\mathrm{T}}\hat{\boldsymbol{W}}_g\boldsymbol{Z})^{-1}\left(n^{-1}\sum_{i=1}^{n}\gamma_{gi}^2w_{gi}^2e_{gi}^2\boldsymbol{z}_i\boldsymbol{z}_i^{\mathrm{T}}\right)(n^{-1}\boldsymbol{Z}^{\mathrm{T}}\hat{\boldsymbol{W}}_g\boldsymbol{Z})^{-1}.$$

Therefore, we need only to approximate the middle term to estimate $\mathbb{V}(\hat{\boldsymbol{\eta}}_g \mid \boldsymbol{Z})$. Simply plugging in $\hat{w}_{gi}^2$ for $w_{gi}^2$ will tend to underestimate $\mathbb{V}(\hat{\boldsymbol{\eta}}_g \mid \boldsymbol{Z})$, since the uncertainty in $\hat{w}_{gi}$ increases as $w_{gi}$ increases. Further, plugging in $\hat{e}_{gi} = y_{gi} - \boldsymbol{z}_i^{\mathrm{T}}\hat{\boldsymbol{\eta}}_g$ for $e_{gi}$ will also underestimate $\mathbb{V}(\hat{\boldsymbol{\eta}}_g \mid \boldsymbol{Z})$, since this ignores the uncertainty in $\hat{\boldsymbol{\eta}}_g$. We circumvent the former by replacing $w_{gi}^2$ with $\hat{v}_{gi}$, defined in (4.8c), where $\hat{v}_{gi} \geq \hat{w}_{gi}^2$ such that $\hat{v}_{gi} = \hat{w}_{gi}^2$ if and only if $\mathbb{V}(w_{gi} \mid \mathcal{D}) = 0$. That is, $\hat{v}_{gi}$ helps account for the uncertainty in our estimate for $w_{gi}$. Lastly, we show how we estimate $e_{gi}^2$ in Section S4 of the Supplementary Material (McKennan, Ober and Nicolae (2020)), which leads to the following estimate for $\mathbb{V}(\hat{\boldsymbol{\eta}}_g \mid \boldsymbol{Z})$:

$$\hat{\mathbb{V}}(\hat{\boldsymbol{\eta}}_g \mid \boldsymbol{Z}) = (\boldsymbol{Z}^{\mathrm{T}}\hat{\boldsymbol{W}}_g\boldsymbol{Z})^{-1}\left\{\sum_{i=1}^{n}(1-\hat{h}_{gi})^{-2}\hat{\gamma}_{gi}^2\hat{v}_{gi}\hat{e}_{gi}^2\boldsymbol{z}_i\boldsymbol{z}_i^{\mathrm{T}}\right\}(\boldsymbol{Z}^{\mathrm{T}}\hat{\boldsymbol{W}}_g\boldsymbol{Z})^{-1}. \tag{5.2}$$

The term $(1-\hat{h}_{gi})^{-2}$ is a finite sample correction, where $\hat{h}_{gi}$ is the $i$th leverage score of $\hat{\boldsymbol{W}}_g^{1/2}\boldsymbol{Z}$ for $i \in [n]$. This resembles the $(1-\hat{h}_{gi})^{-1}$ inflation term commonly used to correct the sandwich variance estimator (Wang et al. (2016)). The difference arises because the residuals $e_{g1}, \ldots, e_{gn}$ are dependent on the design matrix $\hat{\boldsymbol{W}}_g^{1/2}\boldsymbol{Z}$ when data are MNAR. As far as we are aware, this is the first such finite sample variance correction for inverse probability weighted estimators derived from data that are MNAR.

**6. Recovering $\boldsymbol{C}$ when data are MNAR.** In this section we return to using the notation of Model (2.1) and describe our estimator for $\boldsymbol{C}$. Let $\boldsymbol{X} = (\boldsymbol{X}_{\mathrm{int}}\boldsymbol{X}_{\mathrm{nuis}})$, where $\boldsymbol{X}_{\mathrm{int}}$ contains the covariates of interest, like disease status, and $\boldsymbol{X}_{\mathrm{nuis}}$ contains observed nuisance covariates, like the intercept and technical factors. We assume, for simplicity of presentation, that $\boldsymbol{X} = \boldsymbol{X}_{\mathrm{int}}$ and detail the simple extension when $\boldsymbol{X} = (\boldsymbol{X}_{\mathrm{int}}\boldsymbol{X}_{\mathrm{nuis}})$ in Section S3 in the Supplementary Material (McKennan, Ober and Nicolae (2020)).

Once we obtain $\hat{\boldsymbol{C}}$, our estimator for $\boldsymbol{C}$, our estimate for $\boldsymbol{\eta}_g = (\boldsymbol{\beta}_g^{\mathrm{T}}, \boldsymbol{\ell}_g^{\mathrm{T}})^{\mathrm{T}}$ is

$$\hat{\boldsymbol{\eta}}_g = \begin{cases}(\boldsymbol{Z}^{\mathrm{T}}\boldsymbol{R}_g\boldsymbol{Z})^{-1}\boldsymbol{Z}^{\mathrm{T}}\boldsymbol{R}_g\boldsymbol{y}_g & \text{if } g \in \mathcal{S},\\ (5.1) & \text{if } g \in \mathcal{M},\end{cases} \tag{6.1a}$$

$$\hat{\mathbb{V}}(\hat{\boldsymbol{\eta}}_g) = \begin{cases}\{\mathrm{Tr}(\boldsymbol{R}_g) - d - K\}^{-1}\|\boldsymbol{R}_g(\boldsymbol{y}_g - \boldsymbol{Z}\hat{\boldsymbol{\eta}}_g)\|_2^2(\boldsymbol{Z}^{\mathrm{T}}\boldsymbol{R}_g\boldsymbol{Z})^{-1} & \text{if } g \in \mathcal{S},\\ (5.2) & \text{if } g \in \mathcal{M},\end{cases} \tag{6.1b}$$

where $\boldsymbol{Z} = (\boldsymbol{X}\hat{\boldsymbol{C}})$ and $\boldsymbol{R}_g = \mathrm{diag}(r_{g1}, \ldots, r_{gn})$. Note that when $g \in \mathcal{S}$, (6.1) gives the OLS estimator obtained when missing data are treated as MCAR. Since the estimators for $\boldsymbol{\beta}_g$ and $\mathbb{V}(\hat{\boldsymbol{\beta}}_g)$ in (6.1) depend solely on $\mathrm{im}(\hat{\boldsymbol{C}})$, $\hat{\boldsymbol{C}}$ need only satisfy $\mathrm{im}(\hat{\boldsymbol{C}}) \approx \mathrm{im}(\boldsymbol{C})$, which is quite auspicious given that $\boldsymbol{C}$ itself is not identifiable in Model (2.1). If $\boldsymbol{B} = (\boldsymbol{\beta}_1 \cdots \boldsymbol{\beta}_p)^{\mathrm{T}} = \boldsymbol{0}$ and $\boldsymbol{L} = (\boldsymbol{\ell}_1 \cdots \boldsymbol{\ell}_p)^{\mathrm{T}} \in \mathbb{R}^{p\times K}$ is full rank, then $\mathrm{im}(\boldsymbol{C}) = \mathrm{im}(\boldsymbol{C}\boldsymbol{L}^{\mathrm{T}}) = \mathrm{im}\{\mathbb{E}(\boldsymbol{Y}^{\mathrm{T}} \mid \boldsymbol{C})\}$ is identifiable and, therefore, estimable. When $\boldsymbol{B} \neq \boldsymbol{0}$, $\boldsymbol{B}$ being adequately sparse is a sufficient condition

to ensure $\mathrm{im}(\boldsymbol{C})$ is identifiable, since the aforementioned argument is applicable to the submatrix of $\boldsymbol{Y}$ containing only the rows $g$ with $\boldsymbol{\beta}_g = \boldsymbol{0}$ (McKennan and Nicolae (2018)). While there are other, more general, assumptions one may place on $\boldsymbol{B}$ to ensure $\mathrm{im}(\boldsymbol{C})$ is identifiable, the sparsity condition is satisfactory because $\boldsymbol{B}$ is typically sparse in practice (Koeman et al. (2019)). Therefore, we assume that $\mathrm{im}(\boldsymbol{C})$ is identifiable and estimable throughout this section.

Let $\boldsymbol{y}_g = (y_{gi})_{i\in[n]}$ and $\boldsymbol{e}_g = (e_{gi})_{i\in[n]}$ for each $g \in [p]$. As discussed in Section 3.2, we separately estimate $P_X^{\perp}\boldsymbol{C}$ and $P_X\boldsymbol{C} = \boldsymbol{X}\{(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X})^{-1}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{C}\}$ to ensure we accurately identify the components of the variation in $\boldsymbol{y}_g$ that are attributable to $\boldsymbol{C}$ and $\boldsymbol{X}$. To do so, we see that Model (2.1) in vector form can be expressed as

$$\text{(6.2a)}\qquad \boldsymbol{y}_g = \boldsymbol{X}\tilde{\boldsymbol{\beta}}_g + \boldsymbol{C}_2\boldsymbol{\ell}_g + \boldsymbol{e}_g, \quad \tilde{\boldsymbol{\beta}}_g = \boldsymbol{\beta}_g + \boldsymbol{\Omega}\boldsymbol{\ell}_g, \quad \boldsymbol{e}_g \dot{\sim} (\boldsymbol{0}, \sigma_g^2 I_n), \quad g \in [p],$$

$$\text{(6.2b)}\qquad \boldsymbol{C}_2 = P_X^{\perp}\boldsymbol{C}, \qquad \boldsymbol{\Omega} = (\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X})^{-1}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{C},$$

where $\boldsymbol{X}\boldsymbol{\Omega} = P_X\boldsymbol{C}$ and $\boldsymbol{C} = \boldsymbol{C}_2 + \boldsymbol{X}\boldsymbol{\Omega}$. We estimate $\boldsymbol{C}_2$ and $\boldsymbol{\Omega}$ in Sections 6.1 and 6.2 below and define

$$\hat{\boldsymbol{C}} = \hat{\boldsymbol{C}}_2 + \boldsymbol{X}\hat{\boldsymbol{\Omega}}. \tag{6.3}$$

6.1. *Estimating latent factors that are orthogonal to the design.* We first describe our estimators for $\tilde{\boldsymbol{\beta}}_g$, $\boldsymbol{\ell}_g$ and $\boldsymbol{C}_2$, which we also use in Section 6.2 to estimate $\boldsymbol{\Omega}$. Let $\mathcal{M}_1 = \{g \in \mathcal{M} : q_{g1} \le 0.05, lfdr_g \ge 0.8\}$ be the set of metabolites with missing data, at least one viable instrument and whose missingness mechanisms appear to follow (2.2), where $q_{g1}$ and $lfdr_g$ were defined in Sections 4.2 and 4.4. We estimate $\tilde{\boldsymbol{\beta}}_g$, $\boldsymbol{\ell}_g$ and $\boldsymbol{C}_2$ by solving the following optimization problem:

$$\begin{aligned}\{\{\hat{\tilde{\boldsymbol{\beta}}}_g, \hat{\boldsymbol{\ell}}_g\}_{g\in\mathcal{S}\cup\mathcal{M}_1}, \hat{\boldsymbol{C}}_2\} = &\underset{\substack{\tilde{\boldsymbol{\beta}}_g^*\in\mathbb{R}^d, \boldsymbol{\ell}_g^*\in\mathbb{R}^K \\ \boldsymbol{C}_2^*\in\mathbb{R}^{n\times K}, \boldsymbol{X}^{\mathrm{T}}\boldsymbol{C}_2^*=\boldsymbol{0}}}{\arg\min} \left\{\sum_{g\in\mathcal{S}\cup\mathcal{M}_1} m_g(\tilde{\boldsymbol{\beta}}_g^*, \boldsymbol{\ell}_g^*, \boldsymbol{C}_2^*)\right\}, \\ m_g(\tilde{\boldsymbol{\beta}}_g^*, \boldsymbol{\ell}_g^*, \boldsymbol{C}_2^*) = &\begin{cases} \|\boldsymbol{R}_g\{\boldsymbol{y}_g - (\boldsymbol{X}\tilde{\boldsymbol{\beta}}_g^* + \boldsymbol{C}_2^*\boldsymbol{\ell}_g^*)\}\|_2^2 & \text{if } g \in \mathcal{S}, \\ \|\hat{\boldsymbol{W}}_g^{1/2}\{\boldsymbol{y}_g - (\boldsymbol{X}\tilde{\boldsymbol{\beta}}_g^* + \boldsymbol{C}_2^*\boldsymbol{\ell}_g^*)\}\|_2^2 & \text{if } g \in \mathcal{M}_1, \end{cases}\end{aligned} \tag{6.4}$$

where the "*" distinguishes $\tilde{\boldsymbol{\beta}}_g^*$, $\boldsymbol{\ell}_g^*$ and $\boldsymbol{C}_2^*$ from the true parameters $\tilde{\boldsymbol{\beta}}_g$, $\boldsymbol{\ell}_g$ and $\boldsymbol{C}_2$. If each $y_{gi}$ is observed, minimizing the above optimization problem is equivalent to performing principal components analysis on $\boldsymbol{Y}P_X^{\perp}$, where $\mathrm{im}(\hat{\boldsymbol{C}}_2)$ is the span of the first $K$ right singular vectors of $\boldsymbol{Y}P_X^{\perp}$ and is known to accurately estimate $\mathrm{im}(\boldsymbol{C}_2)$ (McKennan and Nicolae (2018)). When there are missing data, assume for simplicity of explanation that metabolites $g \in \mathcal{S}$ have no missing data and $\hat{\gamma}_{gi} = 1$ for all $g \in \mathcal{M}_1$ and $i \in [n]$. If we ignore the uncertainty in $\hat{\boldsymbol{W}}_g$, Model (2.2) implies the expected loss, conditional on the observed and unobserved data, is

$$\mathbb{E}\{m_g(\tilde{\boldsymbol{\beta}}_g^*, \boldsymbol{\ell}_g^*, \boldsymbol{C}_2^*) \mid \boldsymbol{y}_g, \boldsymbol{X}, \boldsymbol{C}\} = \|\boldsymbol{y}_g - \boldsymbol{X}\tilde{\boldsymbol{\beta}}_g^* - \boldsymbol{C}_2^*\boldsymbol{\ell}_g^*\|_2^2, \quad g \in \mathcal{S}\cup\mathcal{M}_1.$$

That is, the loss in (6.4) is expected to behave like the loss when each $y_{gi}$ is observed, which is known to accurately estimate $\mathrm{im}(\hat{\boldsymbol{C}}_2)$. Further, since $\boldsymbol{X}$ is observed and $\boldsymbol{X}^{\mathrm{T}}\boldsymbol{C}_2^* = \boldsymbol{0}$, we can disentangle $\boldsymbol{X}\tilde{\boldsymbol{\beta}}_g$ from the mean $\boldsymbol{X}\tilde{\boldsymbol{\beta}}_g + \boldsymbol{C}_2\boldsymbol{\ell}_g$, which helps to identify $\tilde{\boldsymbol{\beta}}_g$. Like we did in Section 5.1, we include $\hat{\gamma}_{gi}$ in $\hat{\boldsymbol{W}}_g$ to stabilize large weights, since $\hat{\gamma}_{gi}$ is likely to be small if $\hat{w}_{gi}$ is large. We provide additional intuition as to why (6.4) is expected to accurately recover $\mathrm{im}(\boldsymbol{C}_2)$ and $\{\tilde{\boldsymbol{\beta}}_g\}_{g\in\mathcal{S}\cup\mathcal{M}_1}$ in Section S5 of the Supplementary Material (McKennan, Ober and Nicolae (2020)).

6.2. *Estimating latent factors in the image of the design.* We now describe how we estimate $\boldsymbol{\Omega}$. To simplify the notation, we assume that $\mathcal{S} \cup \mathcal{M}_1 = [p]$ for the remainder of the section, but note that, like $\boldsymbol{C}_2$, $\boldsymbol{\Omega}$ is estimated using only metabolites $g \in \mathcal{S} \cup \mathcal{M}_1$. Define $\hat{\boldsymbol{L}} = (\hat{\boldsymbol{\ell}}_1 \cdots \hat{\boldsymbol{\ell}}_p)^{\mathrm{T}}$, $\boldsymbol{B} = (\boldsymbol{\beta}_1 \cdots \boldsymbol{\beta}_p)^{\mathrm{T}}$ and $\hat{\tilde{\boldsymbol{B}}} = (\hat{\tilde{\boldsymbol{\beta}}}_1 \cdots \hat{\tilde{\boldsymbol{\beta}}}_p)^{\mathrm{T}}$. Since $(\hat{\tilde{\boldsymbol{\beta}}}_g^{\mathrm{T}}, \hat{\boldsymbol{\ell}}_g^{\mathrm{T}})^{\mathrm{T}}$ can be expressed as (6.1a) using the design matrix $\boldsymbol{Z} = (\boldsymbol{X}\hat{\boldsymbol{C}}_2)$, (6.2a) suggests an approximate model for $\hat{\tilde{\boldsymbol{B}}}_{*j}$ is

$$\hat{\tilde{\boldsymbol{B}}}_{*j} \dot{\sim} (\boldsymbol{B}_{*j} + \hat{\boldsymbol{L}}\boldsymbol{\Omega}_{j*}, \hat{\boldsymbol{\tau}}_j), \qquad \hat{\boldsymbol{\tau}}_j = \operatorname{diag}(\hat{\tau}_{1,j}, \ldots, \hat{\tau}_{p,j}), \tag{6.5}$$

where $\hat{\tau}_{g,j}$ is the $j$th diagonal element of $\hat{\mathbb{V}}(\hat{\boldsymbol{\eta}}_g)$ defined in (6.1b) for $\boldsymbol{Z} = (\boldsymbol{X}\hat{\boldsymbol{C}}_2)$. If $\boldsymbol{B}_{*j}$ is sparse, (6.5) suggests we can regress $\hat{\tilde{\boldsymbol{B}}}_{*j}$ onto $\hat{\boldsymbol{L}}$ for each $j = 1, \ldots, d$ to estimate $\boldsymbol{\Omega}$. This is detailed in Algorithm 6.1 below, which extends existing methodology in Gagnon-Bartsch, Jacob and Speed (2013), McKennan and Nicolae (2019), Wang et al. (2017) to data with nonignorable missing observations.

ALGORITHM 6.1 (Estimating $\boldsymbol{\Omega}$). *Let $\epsilon_q \in [0, 1]$ and $R \geq 1$ be an integer*:

(0) *For $j \in [d]$, let $\boldsymbol{V}_j^{(0)} = \hat{\boldsymbol{\tau}}_j^{-1}$, $\hat{\boldsymbol{\Omega}}_j^{(0)} = (\hat{\boldsymbol{L}}^{\mathrm{T}}\boldsymbol{V}_j^{(0)}\hat{\boldsymbol{L}})^{-1}\hat{\boldsymbol{L}}^{\mathrm{T}}\boldsymbol{V}_j^{(0)}\hat{\tilde{\boldsymbol{B}}}_{*j}$ and $\hat{\boldsymbol{\Omega}}^{(0)} = (\hat{\boldsymbol{\Omega}}_1^{(0)} \cdots \hat{\boldsymbol{\Omega}}_d^{(0)})^{\mathrm{T}}$. Define $\hat{\boldsymbol{C}}^{(0)} = \boldsymbol{X}\hat{\boldsymbol{\Omega}}^{(0)} + \hat{\boldsymbol{C}}_2$.*

(1) *Let $\hat{\boldsymbol{C}}^{(r)}$ be given. Regress $\boldsymbol{y}_g$ onto $(\boldsymbol{X}\hat{\boldsymbol{C}}^{(r)})$ using (6.1) with $\boldsymbol{Z} = (\boldsymbol{X}\hat{\boldsymbol{C}}^{(r)})$, and define $z_{g,j}$ to be the z-score corresponding to the $j$th coefficient from the regression. Let $p_{g,j} = \mathbb{P}\{|z_{g,j}| \geq |N_1(0, 1)|\}$ be the corresponding P value.*

(2) *For all $j \in [d]$, obtain the q-values $\{q_{g,j}\}_{g\in[p]}$ using the P values $\{p_{g,j}\}_{g\in[p]}$. Repeat Step (0) with $\boldsymbol{V}_j^{(0)}$ replaced with $\boldsymbol{V}_j^{(r+1)} = \operatorname{diag}\{\hat{\tau}_{1,j}^{-1} I(q_{1,j} > \epsilon_q), \ldots, \hat{\tau}_{p,j}^{-1} I(q_{p,j} > \epsilon_q)\}$ to obtain $\hat{\boldsymbol{\Omega}}^{(r+1)}$ and $\hat{\boldsymbol{C}}^{(r+1)}$. Update $r \leftarrow r + 1$.*

(3) *Repeat Steps (1) and (2) for $r = 0, 1, \ldots, R - 1$, and return $\hat{\boldsymbol{\Omega}} = \hat{\boldsymbol{\Omega}}^{(R)}$.*

Since $q_{g,j}$ is the smallest false discovery rate necessary to reject the null hypothesis $H_0^{(g,j)} : \boldsymbol{B}_{gj} = \boldsymbol{\beta}_{g_j} = 0$, removing metabolites with small q-values in Step (2) mitigates the impact of outliers due to $\boldsymbol{B}_{*j}$ in the regression estimate for $\boldsymbol{\Omega}$ when $\boldsymbol{B}$ is only approximately sparse. Our software's default is $\epsilon_q = 0.1$ and $R = 3$.

## 7. A simulation study.

7.1. *Simulation setup.* Here, we analyze simulated metabolomic data to compare the performance of our method with other existing methods. We simulated the log-intensities of $p = 1200$ metabolites in $n = 600$ individuals, 300 of which were cases and the remaining 300 were controls. The observed design matrix was $\boldsymbol{X} = (\boldsymbol{X}_{\mathrm{int}}\boldsymbol{1}_n)$, where $\boldsymbol{X}_{\mathrm{int}} = (\boldsymbol{1}_{n/2}^{\mathrm{T}}, \boldsymbol{0}_{n/2}^{\mathrm{T}})^{\mathrm{T}} \in \mathbb{R}^n$. The parameters $p$ and $n$ were chosen to match those from our real data example in Section 8, and we include additional results when $n = 100$ and $n = 300$ in Section S6.4 of the Supplementary Material (McKennan, Ober and Nicolae (2020)). We set $K = 10$, and, for some constant $a$ and appropriate $\Psi(x)$, simulated data as

$$\log(\alpha_g) \sim N_1(\mu_\alpha, 0.4^2), \quad \delta_g \sim N_1(16, 1.2^2), \quad g \in [p], \tag{7.1a}$$

$$\boldsymbol{C} = (\boldsymbol{c}_1 \cdots \boldsymbol{c}_n)^{\mathrm{T}} \sim MN_{n\times K}\big((a\boldsymbol{X}_{\mathrm{int}}\boldsymbol{0}_n \cdots \boldsymbol{0}_n), I_n, I_K\big), \tag{7.1b}$$

$$\boldsymbol{\ell}_{g_k} \sim \pi_k\delta_0 + (1 - \pi_k)N_1(0, \tau_k^2), \quad g \in [p]; k \in [K], \tag{7.1c}$$

$$\mu_g \sim N_1(18, 5^2), \sigma_g^2 \sim \operatorname{Gamma}(0.2^{-2}, 0.2^{-2}), \quad g \in [p], \tag{7.1d}$$

TABLE 1
*The $\pi_k$ and $\tau_k$ values used to simulate $\boldsymbol{\ell}_1, \ldots, \boldsymbol{\ell}_p$ ($k = 1, \ldots, 10$)*

| Factor number ($k$) | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
|---|---|---|---|---|---|---|---|---|---|---|
| $\pi_k$ | 0 | 0 | 0.76 | 0.56 | 0.48 | 0.32 | 0.28 | 0.20 | 0.20 | 0.20 |
| $\tau_k$ | 0.78 | 0.57 | 0.5 | 0.5 | 0.5 | 0.5 | 0.5 | 0.5 | 0.5 | 0.5 |

$$\beta_g \sim 0.8\delta_0 + 0.2N_1(0, 0.4^2), \quad g \in [p], \tag{7.1e}$$

$$y_{gi} \sim N_1\big(\mu_g + \boldsymbol{X}_{\text{int}_i}\beta_g + \boldsymbol{c}_i^{\mathrm{T}}\boldsymbol{\ell}_g, \sigma_g^2\big), \quad g \in [p]; i \in [n], \tag{7.1f}$$

$$r_{gi} \sim \text{Bernoulli}\big[\Psi\{\alpha_g(y_{gi} - \delta_g)\}\big], \quad g \in [p]; i \in [n], \tag{7.1g}$$

where $\delta_0$ is the point mass at 0 and $\mu_\alpha$ in (7.1a) was set such that if $Z$ has cumulative distribution function $\Psi\{\exp(\mu_\alpha)x\}$, $\mathbb{V}(Z) = 1$. The constant $a$ in (7.1b) was chosen so that $\boldsymbol{C}$ explained 7.5% of the variance in $\boldsymbol{X}_{\text{int}}$ on average across all simulations, and Table 1 contains the values of $\pi_k$ and $\tau_k^2$. These were chosen so that the nonzero eigenvalues $\lambda_1, \ldots, \lambda_K$ of $\mathcal{I} = (n-1)^{-1}P_{1_n}^{\perp}\boldsymbol{C}(p^{-1}\sum_{g=1}^{p}\sigma_g^{-2}\boldsymbol{\ell}_g\boldsymbol{\ell}_g^{\mathrm{T}})\boldsymbol{C}^{\mathrm{T}}P_{1_n}^{\perp}$ were 0.61, 0.33, 0.19, 0.14, 0.12, 0.08, 0.07, 0.05, 0.05 and 0.05 on average across all simulated datasets, since these were the first 10 eigenvalues of the estimated $\mathcal{I}$ in our data example in Section 8. Similarly, the prior variances for the missingness mechanism parameters in (7.1a), as well as the mean and variance for the global mean $\mu_g$ in (7.1d), were set to their estimated equivalents from our data example in Section 8. Since we typically do not know the exact functional form of $\Psi(x)$ in practice, we set $\Psi(x) = \exp(x)/\{1 + \exp(x)\}$ and analyzed each simulated dataset assuming $\Psi(x) = F_4(x)$. The distribution of missing data is given in Table 2, which closely matched that in our real data example.

We simulated 60 datasets, and in each simulation, removed all metabolites that were missing in more than 50% of the samples, since we find that these metabolites tend to have large $J$ statistics in real data. We used Algorithm 3.1 with $\epsilon_{\text{miss}} = 0.05$ and $K_{\text{miss}} = 5$ to estimate the metabolite-dependent missingness mechanism parameters $\alpha_g$ and $\delta_g$ and, subsequently, estimated $\boldsymbol{C}$ as (6.3) with $\epsilon_q = 0.1$, assuming $K = 10$ was known. We lastly estimated $\beta_g$ and said estimator's variance using (6.1), and formed 95% confidence intervals and computed $P$ values assuming $\hat{\beta}_g \sim N_1(\beta_g, \hat{\mathbb{V}}(\hat{\beta}_g))$. We refer to this procedure as "MetabMiss."

Similar to our real data example, $K_{\text{miss}}$ was such that $q_{g,g_2}$, defined in Algorithm 4.1, was less than 0.05 in at least 90% of all metabolites $g \in \mathcal{M}$ in each simulated dataset. Our results were identical when we let $K_{\text{miss}}$ be as small as 3 and as large as 10. The fact that $K$ was assumed to be known when estimating $\boldsymbol{C}$ was inconsequential, since the method of Leek et al. (2017) applied to metabolites with only complete data consistently estimated $K = 10$. We demonstrate the fidelity of MetabMiss' estimates for $\beta_1, \ldots, \beta_p$ in Section 7.2. We also illustrate the accuracy of Algorithm 3.1's estimates for $\alpha_g$ and $\delta_g$, the uniformity of the bootstrapped Sargan–Hansen $J$ statistic $P$ values, as well include additional simulation results when metabolites grouped in pathways are correlated in Sections S6.1–S6.3 of the Supplementary Material (McKennan, Ober and Nicolae (2020)).

TABLE 2
*The expected number of metabolites in each missing data bin for data simulated according to (7.1) with $\Psi(x) = \exp(x)/\{1 + \exp(x)\}$, where $f$ is the frequency of missing data*

| $f = 0$ | $0 < f \leq 0.05$ | $0.05 < f \leq 0.5$ | $0.5 < f$ |
|---|---|---|---|
| 251.6 | 233.6 | 298.3 | 416.4 |

7.2. *Simulation results*. Given the paucity of methods available to analyze metabolomics data, we could only compare our method to those that account for nonrandom missing data or $\boldsymbol{C}$, but not both. It is not interesting to compare MetabMiss to methods that only account for the former, like those proposed in Chen et al. (2017), O'Brien et al. (2018), Wang et al. (2019), because ignoring $\boldsymbol{C}$ can dramatically inflate type I and reduce type II error even when there are no missing data (McKennan and Nicolae (2019)). Instead, we compared MetabMiss to existing methods that have been used to recover $\boldsymbol{C}$ in metabolomic data but do not account for the nonrandom missing data, which included IRW-SVA (Leek and Storey (2008)), dSVA (Lee et al. (2017)), RUV-2 (Gagnon-Bartsch and Speed (2012)), RUV-4 (Gagnon-Bartsch, Jacob and Speed (2013)), and when $\boldsymbol{C}$ is assumed to be known. We do not report results from the method of Wang et al. (2017), as it performed nearly identically to dSVA in all simulation scenarios. Since none of the of the aforementioned methods can accommodate missing data, we estimated $\boldsymbol{C}$ with each using only metabolites with complete observations and computed confidence intervals and $P$ values using OLS with the estimated design matrix $(\boldsymbol{X}\hat{\boldsymbol{C}})$, assuming missing data were MCAR. Seeing that RUV-2 and RUV-4 require prior knowledge of control metabolites with $\beta_g = 0$, we selected 20 metabolites uniformly at random from the set of all metabolites $g$ with no missing data and $\beta_g = 0$ to act as control metabolites when applying RUV-2 and RUV-4. This, in relative terms, is substantially more than the 30 control genes used when analyzing the simulated expression of $p = 5000$ genomic units in Wang et al. (2017), which outlines the theoretical properties of RUV-4. We remark that we could not analyze these simulated data with the methods proposed in De Livera et al. (2012) or Salerno Stephen et al. (2017) because both methods rely on a random effects model whose estimators are not amenable to any missing data.

We first evaluated each method's ability to identify metabolites with $\beta_g \neq 0$ while controlling the false discovery rate (FDR) at a nominal level. The results are given in Figure 2, where the only methods to suitably control the FDR were MetabMiss and when $\boldsymbol{C}$ was known. However, as seen in Figure S5 in Section S6.4 in the Supplementary Material (McKennan, Ober and Nicolae (2020)), the latter could not control the FDR in simulations with $n = 300$ or $n = 100$, indicating that accounting for both $\boldsymbol{C}$ and nonrandom missing data is critical for ensuring reliable FDR control across experimental settings. The fact that MetabMiss is slightly underpowered compared to the other methods that estimate $\boldsymbol{C}$ is to be expected, as anticonservative inference is typically more powerful. We also evaluated the confidence interval coverage for $\beta_g$ for each method in Figure 3, which illustrates both the consequences of performing inference on estimators that do not properly account for the missing data, as well as the fidelity of our finite sample-corrected estimator for the variance defined in (5.2).

**8. Data analysis.** We used blood plasma metabolomic data measured in $n = 533$ 6-year-old Danish children enrolled in the Copenhagen Prospective Studies of Asthma in Children cohort (Bisgaard et al. (2013)) to demonstrate the importance of accounting for both missing data and unobserved covariates in untargeted metabolomic data. Table 3 provides an overview of the extent of the missing data in each of the $p = 1138$ measured metabolites. We excluded metabolites that were missing in more than 50% of the samples and set $\epsilon_{\text{miss}} = 0.05$ and $K_{\text{miss}} = K_{PA}/2 = 10$ when estimating the missingness mechanisms with Algorithm 3.1, where $K_{PA}$ was parallel analysis' (Buja and Eyuboglu (1992)) estimates for $K$. $K_{\text{miss}}$ was chosen using the procedure outlined in Section 4.2.

Once we estimated the missingness mechanisms, we could easily assess the relationships between the quantified metabolome and the many recorded phenotypes using MetabMiss. We were particularly interested in phenotypes related to asthma, and present the results for specific airway resistance ($\text{sR}_{\text{AW}}$), which measures airway resistance to flow (Kaminsky (2012)). Using the design matrix $\boldsymbol{X} = (\boldsymbol{X}_{\text{int}}\boldsymbol{1}_n)$, where $\boldsymbol{X}_{\text{int}} \in \mathbb{R}^n$ was each individual's measured

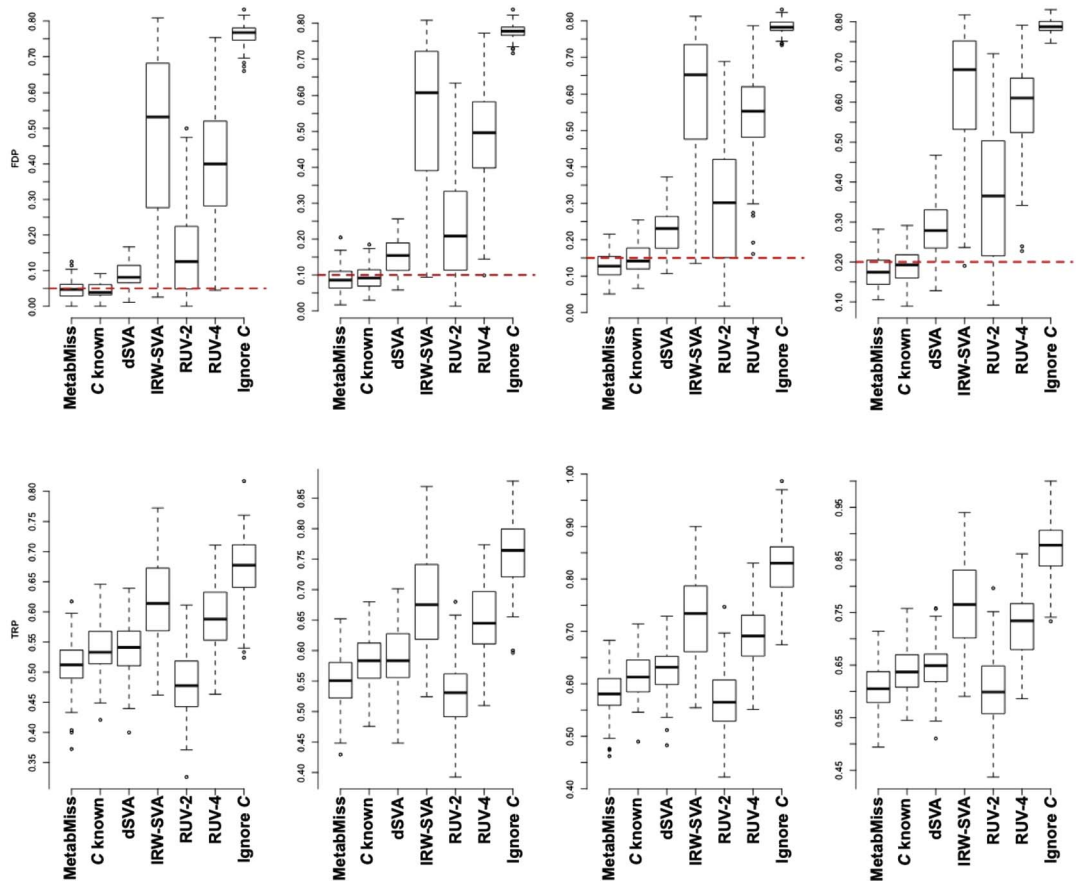


FIG. 2. *From left to right*: *The false discovery proportion*, *FDP* (*top*), *and true recovery proportion*, *TRP* (*bottom*), *for metabolites with q-values* $\leq 0.05$, 0.1, 0.15 *and* 0.2, *determined using Storey et al.* (2015). *The TRP is the fraction of metabolites with* $\beta_g \neq 0$ *identified at a given q-value cutoff.*

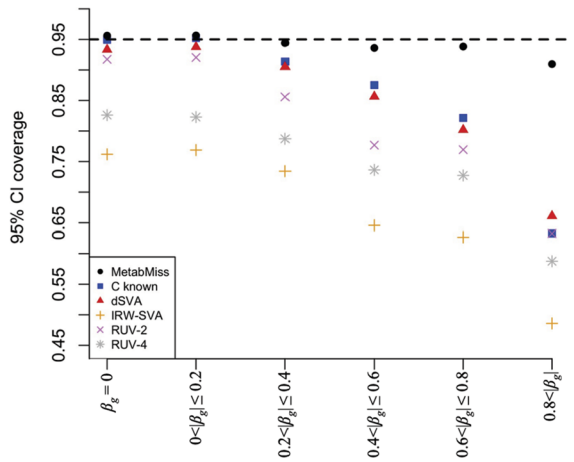


FIG. 3. *The fraction of effects of interest* $\{\beta_g\}_{g\in\mathcal{M}}$ *in all* 60 *simulated datasets that lie in their respective* 95% *confidence intervals* $\hat{\beta}_g \pm 1.96\{\hat{\mathbb{V}}(\hat{\beta}_g)\}^{1/2}$, *stratified by* $|\beta_g|$. *The coverage when* $\boldsymbol{C}$ *was ignored was uniformly less than IRW-SVA's.*

TABLE 3
*The number of metabolites in each missing data bin in the blood plasma metabolomic data, where $f$ is as defined in Table 2*

| $f = 0$ | $0 < f \leq 0.05$ | $0.05 < f \leq 0.5$ | $0.5 < f$ |
|---|---|---|---|
| 400 | 256 | 300 | 182 |

$sR_{AW}$ value, we estimated $K$ with the method of Leek et al. (2017) using the metabolites with complete data and regressed the quantified metabolites onto $\boldsymbol{X}$ using MetabMiss. We present the Q–Q plots of $P$ values in Figure 4.

Figure 4 shows that MetabMiss not only corrects the minor $P$ value inflation, but also empowers the analysis by reducing the residual variance. While the analysis with dSVA only identified a single metabolite, MetabMiss identified six additional metabolites at a a q-value threshold of 0.2: two sphingolipids, a benzoate derivative, pyruvate and three derivatives of piperine, which is an alkaloid found in black pepper. A reduction in sphingolipid synthesis was associated with increased airway hyperractivity in children (Ono, Worgall and Worgall (2015)), which is congruent with the estimated sign of the $sR_{AW}$ effect on the intensity of the two sphingolipid metabolites. Benzoate preservatives have been linked to lung function-related phenotypes (Balatsinou et al. (2004), Pacor et al. (2004)), and pyruvate and lactate (q-value = 0.23) levels have previously been associated with asthma (Ostroukhova et al. (2012), Xu et al. (2010b)).

The three derivatives of piperine were particularly interesting in the context of our methodology because all three had between 12% and 48% missing data with J-test $P$ values between 0.77 and 0.99 (see Section 4.4), suggesting that Model (2.2) is a reasonable model for their missingness mechanisms. We found that higher concentrations of these three metabolites were associated with increased airway resistance. This corroborates the known biological impact of piperine, as it has been shown that piperine has a strong affinity for and activates TRPV1 cation channels on the ends of somatic and visceral parasympathetic nervous system sensory fibers (Premkumar (2014)). This triggers mast cells, bronchial epithelial cells and immune cells to release proinflammatory cytokines (Frias and Merighi (2016)), which ultimately causes bronchoconstriction (Choi et al. (2018), Jia and Lee (2007)).

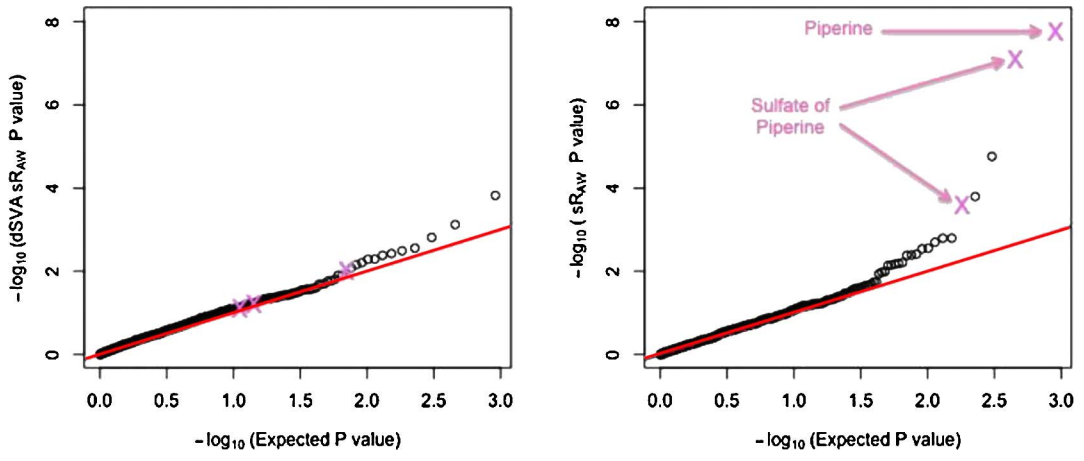


FIG. 4. *A Q–Q plot of $P$ values for the null hypotheses $H_{0,g} : \beta_g = 0$ when $\boldsymbol{C}$ is estimated with dSVA using metabolites with complete data and the missing data are treated as MCAR (left) and using MetabMiss (right). The x-axis is the expected ordered $P$ value under the null hypothesis, assuming all tests are independent. The three quantified derivatives of piperine are each labeled with a violet "×."*

An interesting feature of Figure 4 is that the ordering of the $P$ values changes when one accounts for the missing data. In fact, we would not have identified the association between $sR_{AW}$ and piperine concentration if one used existing latent factor correction methods. This is, in part, because metabolites with missing data had a slightly different latent factor signature than those with complete data, which is why MetabMiss is able to better estimate $\boldsymbol{C}$ than existing methods.

**9. Discussion.** We have presented, to the best of our knowledge, the first method to simultaneously account for latent factors and nonignorable missing data in untargeted metabolomic data. Our method simplifies this complex problem by modularizing the estimation of each metabolite-dependent missingness mechanism and latent factors, and does so without assuming a likelihood for the missing data. This modularization also makes modern metabolomic data analysis tractable, since our estimators for the missingness mechanism only depend on $\boldsymbol{Y}$ and are invariant to the choice of model matrix $\boldsymbol{X}$.

An important assumption we made was that the missingness mechanism was only a function of $y_{gi}$ and did not further depend on the sample $i$, which as described in Section 2.2, reflects the nature of the missing data. However, there might be scenarios where Model (2.2) is incorrect. For example, the missingness mechanism might depend on $i$ in experiments where there is a significant amount of time between the analysis of batches of samples or in experiments whose samples are run on different mass spectrometers. These considerations may inspire new and interesting research.

**Acknowledgments.** We thank Hans Bisgaard, Klaus Bønnelykke and the other COPSAC investigators for providing the data to make this research possible. We also thank Morten Arendt Rasmussen, Daniela Rago and Donata Vercelli for comments and suggestions that have substantially improved this work.

The first author is supported in part by NIH Grant R01 HL129735.

## SUPPLEMENTARY MATERIAL

**Supplementary material for "Estimation and inference in metabolomics with nonrandom missing data and latent factors"** (DOI: 10.1214/20-AOAS1328SUPPA; .pdf). This supplemental file contains the details of our bootstrap procedure to estimate $J$ statistic $P$ values from Section 4.4, an extension of Section 6 when $\boldsymbol{X}$ contains nuisance covariates, additional intuition regarding the estimation of $\boldsymbol{C}_2$ from Section 6.1, additional simulation results and a theoretical justification for Algorithms 3.1 and 4.1.

**MetabMiss** (DOI: 10.1214/20-AOAS1328SUPPB; .zip). This supplemental file contains the R package implementing our method, as well as instructions and code to reproduce the simulations from Section 7.

## REFERENCES

AFSHIN, A., SUR, P. J., FAY, K. A., CORNABY, L., FERRARA, G., SALAMA, J. S., MULLANY, E. C., ABATE, K. H., ABBAFATI, C. et al. (2019). Health effects of dietary risks in 195 countries, 1990–2017: A systematic analysis for the Global Burden of Disease Study 2017. *Lancet* **393** 1958–1972.

BALATSINOU, L., GIOACCHINO, G. D., SABATINO, G., CAVALLUCCI, E., CARUSO, R., GABRIELE, E., RAMONDO, S., GIAMPAOLO, L. D., VERNA, N. et al. (2004). Asthma worsened by benzoate contained in some antiasthmatic drugs. *International Journal of Immunopathology and Pharmacology* **17** 225–226.

BAUM, C., SCHAFFER, M. and STILLMAN, S. (2003). Instrumental variables and GMM: Estimation and testing. *Stata Journal* **3** 1–31.

BISGAARD, H., VISSING, N. H., CARSON, C. G., BISCHOFF, A. L., FØLSGAARD, N. V., KREINER-MØLLER, E., CHAWES, B. L. K., STOKHOLM, J., PEDERSEN, L. et al. (2013). Deep phenotyping of the unselected COPSAC2010 birth cohort study. *Clinical and Experimental Allergy* **43** 1384–1394.

BOUHIFD, M., BEGER, R., FLYNN, T., GUO, L., HARRIS, G., HOGBERG, H., KADDURAH-DAOUK, R., KAMP, H., KLEENSANG, A. et al. (2015). Quality assurance of metabolomics. *ALTEX* **32** 319–326.

BROADHURST, D., GOODACRE, R., REINKE, S. N., KULIGOWSKI, J., WILSON, I. D., LEWIS, M. R. and DUNN, W. B. (2018). Guidelines and considerations for the use of system suitability and quality control samples in mass spectrometry assays applied in untargeted clinical metabolomic studies. *Metabolomics* **14** 72. https://doi.org/10.1007/s11306-018-1367-3

BROWN, B. W. and NEWEY, W. K. (2002). Generalized method of moments, efficient bootstrapping, and improved inference *Journal of Business & Economic Statistics* **20** 507–517. MR1945606 https://doi.org/10.1198/073500102288618649

BUJA, A. and EYUBOGLU, N. (1992). Remarks on parallel analysis. *Multivar. Behav. Res.* **27** 509–540.

CHEN, L. S., WANG, J., WANG, X. and WANG, P. (2017). A mixed-effects model for incomplete data from labeling-based quantitative proteomics experiments. *The Annals of Applied Statistics* **11** 114–138. MR3634317 https://doi.org/10.1214/16-AOAS994

CHOI, J. Y., LEE, H. Y., HUR, J., KIM, K. H., KANG, J. Y., RHEE, C. K. and LEE, S. Y. (2018). TRPV1 blocking alleviates airway inflammation and remodeling in a chronic asthma murine model. *Allergy, Asthma & Immunology Research* **10** 216–224.

DAVIDSON, R. and MACKINNON, J. G. (2003). *Econometric Theory and Methods*. Oxford Univ. Press, London.

DE LIVERA, A. M., DIAS, D. A., DE SOUZA, D., RUPASINGHE, T., PYKE, J., TULL, D., ROESSNER, U., MCCONVILLE, M. and SPEED, T. P. (2012). Normalizing and integrating metabolomics data. *Analytical Chemistry* **84** 10768–10776.

DE LIVERA, A. M., SYSI-AHO, M., JACOB, L., GAGNON-BARTSCH, J. A., CASTILLO, S., SIMPSON, J. A. and SPEED, T. P. (2015). Statistical methods for handling unwanted variation in metabolomics data. *Analytical Chemistry* **87** 3606–3615.

DO, K. T., WAHL, S., RAFFLER, J., MOLNOS, S., LAIMIGHOFER, M., ADAMSKI, J., SUHRE, K., STRAUCH, K., PETERS, A. et al. (2018). Characterization of missing values in untargeted MS-based metabolomics data and evaluation of missing data handling strategies. *Metabolomics* **14** 128–128.

DUBUIS, S., ORTMAYR, K. and ZAMPIERI, M. (2018). A framework for large-scale metabolome drug profiling links coenzyme A metabolism to the toxicity of anti-cancer drug dichloroacetate. *Communications Biology* **1** 101. https://doi.org/10.1038/s42003-018-0111-x

FINKELSTEIN, J. L., PRESSMAN, E. K., COOPER, E. M., KENT, T. R., BAR, H. Y. and O'BRIEN, K. O. (2015). Vitamin D status affects serum metabolomic profiles in pregnant adolescents. *Reproductive Sciences* **22** 685–695.

FRANKS, A. M., AIROLDI, E. M. and RUBIN, D. B. (2016). Non-standard conditionally specified models for non-ignorable missing data.

FRIAS, B. and MERIGHI, A. (2016). Capsaicin, nociception and pain. *Molecules* (*Basel*, *Switzerland*) **21** 797.

GAGNON-BARTSCH, J. A., JACOB, L. and SPEED, T. P. (2013). Removing unwanted variation from high dimensional data with negative controls. Technical Report, UC Berkeley.

GAGNON-BARTSCH, J. A. and SPEED, T. P. (2012). Using control genes to correct for unwanted variation in microarray data. *Biostatistics* **13** 539–552.

HALL, P. and HOROWITZ, J. L. (1996). Bootstrap critical values for tests based on generalized-method-of-moments estimators. *Econometrica* **64** 891–916. MR1399222 https://doi.org/10.2307/2171849

HANSEN, L. P. (1982). Large sample properties of generalized method of moments estimators. *Econometrica* **50** 1029–1054. MR0666123 https://doi.org/10.2307/1912775

HANSEN, B. E. and WEST, K. D. (2002). Generalized method of moments and macroeconomics. *Journal of Business & Economic Statistics* **20** 460–469. MR1973798 https://doi.org/10.1198/073500102288618603

JIA, Y. and LEE, L.-Y. (2007). Role of TRPV receptors in respiratory diseases. *Biochimica et Biophysica Acta* (*BBA*)*—Molecular Basis of Disease* **1772** 915–927.

JOHNSON, D., BOYES, B., FIELDS, T., KOPKIN, R. and ORLANDO, R. (2013). Optimization of data-dependent acquisition parameters for coupling high-speed separations with LC-MS/MS for protein identifications. *Journal of Biomolecular Techniques* **24** 62–72. https://doi.org/10.7171/jbt.13-2402-003

KAMINSKY, D. A. (2012). What does airway resistance tell us about lung function? *Respiratory Care* **57** 85–96; discussion 96–99. https://doi.org/10.4187/respcare.01411

KANG, J. D. Y. and SCHAFER, J. L. (2007). Demystifying double robustness: A comparison of alternative strategies for estimating a population mean from incomplete data. *Statistical Science* **22** 523–539. MR2420458 https://doi.org/10.1214/07-STS227

KARPIEVITCH, Y. V., POLPITIYA, A. D., ANDERSON, G. A., SMITH, R. D. and DABNEY, A. R. (2010). Liquid chromatography mass spectrometry-based proteomics: Biological and technical aspects. *The Annals of Applied Statistics* **4** 1797–1823.

KIM, J.-Y. (2002). Limited information likelihood and Bayesian analysis. *Journal of Econometrics* **107** 175–193.

KOEMAN, M., ENGEL, J., JANSEN, J. and BUYDENS, L. (2019). Critical comparison of methods for fault diagnosis in metabolomics data. *Scientific Reports* **9** 1123. https://doi.org/10.1038/s41598-018-37494-7

LEE, S., SUN, W., WRIGHT, F. A. and ZOU, F. (2017). An improved and explicit surrogate variable analysis procedure by coefficient adjustment. *Biometrika* **104** 303–316. MR3698255 https://doi.org/10.1093/biomet/asx018

LEEK, J. T. and STOREY, J. D. (2007). Capturing heterogeneity in gene expression studies by surrogate variable analysis. *PLoS Genetics* **3** 1724–1735.

LEEK, J. T. and STOREY, J. D. (2008). A general framework for multiple testing dependence. *Proceedings of the National Academy of Sciences* **105** 18718–18723.

LEEK, J. T., JOHNSON, W. E., PARKER, H. S., FERTIG, E. J., JAFFE, A. E., STOREY, J. D., ZHANG, Y. and TORRES, L. C. (2017). sva: Surrogate Variable Analysis. R package version 3.26.0.

LI, C. and JIANG, W. (2016). On oracle property and asymptotic validity of Bayesian generalized method of moments. *Journal of Multivariate Analysis* **145** 132–147. MR3459943 https://doi.org/10.1016/j.jmva.2015.12.009

LIANG, K.-Y. and QIN, J. (2000). Regression analysis under non-standard situations: A pairwise pseudolikelihood approach. *Journal of the Royal Statistical Society*: *Series B* **62** 773–786. MR1796291 https://doi.org/10.1111/1467-9868.00263

LIU, X., SER, Z. and LOCASALE, J. W. (2014). Development and quantitative evaluation of a high-resolution metabolomics technology. *Analytical Chemistry* **86** 2175–2184.

MCKENNAN, C. and NICOLAE, D. (2018). Estimating and accounting for unobserved covariates in high dimensional correlated data. Available at arXiv:1808.05895v1.

MCKENNAN, C. and NICOLAE, D. (2019). Accounting for unobserved covariates with varying degrees of estimability in high-dimensional biological data. *Biometrika* **106** 823–840. MR4031201 https://doi.org/10.1093/biomet/asz037

MCKENNAN, C., OBER, C. and NICOLAE, D. (2020). Supplement to "Estimation and inference in metabolomics with nonrandom missing data and latent factors." https://doi.org/10.1214/20-AOAS1328SUPPA, https://doi.org/10.1214/20-AOAS1328SUPPB

O'BRIEN, J. J., GUNAWARDENA, H. P., PAULO, J. A., CHEN, X., IBRAHIM, J. G., GYGI, S. P. and QAQISH, B. F. (2018). The effects of nonignorable missing data on label-free mass spectrometry proteomics experiments. *Ann. Appl. Stat.* **12** 2075–2095. MR3875693 https://doi.org/10.1214/18-AOAS1144

O'SULLIVAN, A., GIBNEY, M. J. and BRENNAN, L. (2010). Dietary intake patterns are reflected in metabolomic profiles: Potential role in dietary assessment studies. *Am. J. Clin. Nutr.* **93** 314–321.

ONO, J. G., WORGALL, T. S. and WORGALL, S. (2015). Airway reactivity and sphingolipids-implications for childhood asthma. *Molecular and Cellular Pediatrics* **2** 13–13.

OSTROUKHOVA, M., GOPLEN, N., KARIM, M. Z., MICHALEC, L., GUO, L., LIANG, Q. and ALAM, R. (2012). The role of low-level lactate production in airway inflammation in asthma. *Am. J. Physiol., Lung Cell. Mol. Physiol.* **302** L300–L307.

PACOR, M. L., DI LORENZO, G., MARTINELLI, N., MANSUETO, P., RINI, G. B. and CORROCHER, R. (2004). Monosodium benzoate hypersensitivity in subjects with persistent rhinitis. *Allergy* **59** 192–197.

PREMKUMAR, L. S. (2014). Transient receptor potential channels as targets for phytochemicals. *ACS Chemical Neuroscience* **5** 1117–1130.

REINKE, S. N., GALLART-AYALA, H., GÓMEZ, C., CHECA, A., FAULAND, A., NAZ, S., KAMLEH, M. A., DJUKANOVIĆ, R., HINKS, T. S. C. et al. (2017). Metabolomics analysis identifies different metabotypes of asthma severity. *The European Respiratory Journal* **49** 1601740.

SALERNO STEPHEN, J., MEHRMOHAMADI, M., LIBERTI, M. V., WAN, M., WELLS, M. T., BOOTH, J. G. and LOCASALE, J. W. (2017). RRmix: A method for simultaneous batch effect correction and analysis of metabolomics data in the absence of internal standards. *PLoS ONE* **12** e0179530.

STOREY, J. D., BASS, A. J., DABNEY, A. and ROBINSON, D. (2015). qvalue: Q-value estimation for false discovery rate control. R package version 2.10.0.

WANG, S., SHAO, J. and KIM, J. K. (2014). An instrumental variable approach for identification and estimation with nonignorable nonresponse. *Statist. Sinica* **24** 1097–1116. MR3241279

WANG, P., TANG, H., ZHANG, H., WHITEAKER, J., PAULOVICH, A. G. and MCINTOSH, M. (2006). Normalization regarding non-random missing values in high-throughput mass spectrometry data. *Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing* 315–326.

WANG, M., KONG, L., LI, Z. and ZHANG, L. (2016). Covariance estimators for generalized estimating equations (GEE) in longitudinal analysis with small samples. *Stat. Med.* **35** 1706–1721. MR3513479 https://doi.org/10.1002/sim.6817

WANG, J., ZHAO, Q., HASTIE, T. and OWEN, A. B. (2017). Confounder adjustment in multiple hypothesis testing. *The Annals of Statistics* **45** 1863–1894. MR3718155 https://doi.org/10.1214/16-AOS1511

WANG, J., WANG, P., HEDEKER, D. and CHEN, L. S. (2019). Using multivariate mixed-effects selection models for analyzing batch-processed proteomics data with non-ignorable missingness. *Biostatistics* **20** 648–665. MR4019723 https://doi.org/10.1093/biostatistics/kxy022

WEHRENS, R., HAGEMAN, J. A., VAN EEUWIJK, F., KOOKE, R., FLOOD, P. J., WIJNKER, E., KEURENTJES, J. J. B., LOMMEN, A., VAN EEKELEN, H. D. L. M. et al. (2016). Improved batch correction in untargeted MS-based metabolomics. *Metabolomics* **12** 88–88.

XU, S., ROSS, C., RAEBEL, M. A., SHETTERLY, S., BLANCHETTE, C. and SMITH, D. (2010a). Use of stabilized inverse propensity scores as weights to directly estimate relative risk and its confidence intervals. *Value Health* **13** 273–277.

XU, Y.-D., CUI, J.-M., WANG, Y., YIN, L.-M., GAO, C.-K., LIU, Y.-Y. and YANG, Y.-Q. (2010b). The early asthmatic response is associated with glycolysis, calcium binding and mitochondria activity as revealed by proteomic analysis in rats. *Respiratory Research* **11** 107.

YIN, G. (2009). Bayesian generalized method of moments. *Bayesian Anal.* **4** 191–207. MR2507358 https://doi.org/10.1214/09-BA407