

VARIANCE FUNCTION ESTIMATION IN QUANTITATIVE MASS SPECTROMETRY WITH APPLICATION TO ITRAQ LABELING

BY MICHA MANDEL, MANOR ASKENAZI, YI ZHANG
AND JARROD A. MARTO

Hebrew University of Jerusalem, Dana-Farber Cancer Institute, Harvard Medical School and Hebrew University of Jerusalem, Dana-Farber Cancer Institute, and Dana-Farber Cancer Institute and Harvard Medical School

This paper describes and compares two methods for estimating the variance function associated with iTRAQ (isobaric tag for relative and absolute quantitation) isotopic labeling in quantitative mass spectrometry based proteomics. Measurements generated by the mass spectrometer are proportional to the concentration of peptides present in the biological sample. However, the iTRAQ reporter signals are subject to errors that depend on the peptide amounts. The variance function of the errors is therefore an essential parameter for evaluating the results, but estimating it is complicated, as the number of nuisance parameters increases with sample size while the number of replicates for each peptide remains small. Two experiments that were conducted with the sole goal of estimating the variance function and its stability over time are analyzed, and the resulting estimated variance function is used to analyze an experiment targeting aberrant signaling cascades in cells harboring distinct oncogenic mutations. Methods for constructing conservative p -values and confidence intervals are discussed.

1. Introduction. Improvements in mass spectrometer resolution, accuracy and sensitivity coupled with the development of increasingly sophisticated algorithms for protein identification from spectra have resulted in mass spectrometry (MS) becoming the tool of choice in large scale proteomics research. Typically, the mass spectrometer is used to measure short portions of the proteins called peptides. These are subjected to a process called tandem mass spectrometry (MS/MS) which ultimately yields a mass spectrum containing peaks which correspond to the primary amino acid sequence and enable the identification of peptides. When samples are labeled with iTRAQ (isobaric tag for relative and absolute quantitation) stable isotope reagents, the MS/MS spectra also contain peaks at predefined masses, whose intensities provide a relative measure of the peptide abundance in a set of samples. A single experiment can yield tens of thousands of spectra identifying thousands of peptides belonging to thousands of proteins. Analysis of samples from different sources, for example, cells expressing different mutations in

Received October 2011; revised May 2012.

Key words and phrases. Heteroscedasticity, iTRAQ, mixture model, nuisance parameter, proteomics.

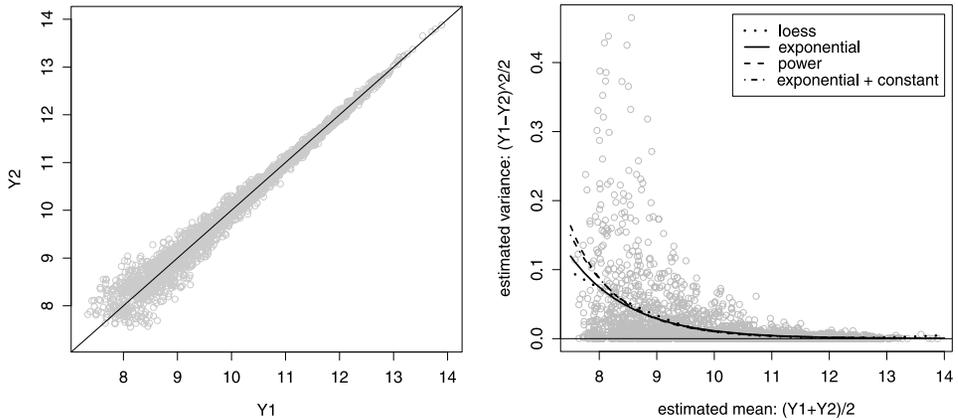


FIG. 1. *Left: scatterplots of Y_{i1} versus Y_{i2} in the March experiment. Right: pointwise estimates of mean and variance and different models for the variance.*

a known oncogene versus the wild-type (e.g., nontransforming) counterpart can provide insight as to the mechanisms by which specific genetic lesions manifested in the same protein drive a malignant phenotype. A workflow diagram and a brief explanation of the iTRAQ technique are given in Part A of the supplemental article [Mandel et al. (2013)]; for a detailed discussion on MS techniques see Eckel-Passow et al. (2009).

In an experiment conducted in March 2009 and described below, Zhang et al. (2010) applied the iTRAQ protocol using two different labels for the same biological sample. The experiment yielded 2174 pairs of measurements corresponding to the amounts of 2174 peptides in the sample. The left panel of Figure 1 depicts the data¹ (on a logarithmic scale) and clearly shows that the variance of peak intensity measurements is nonconstant and depends on the mean. This variance should be estimated for better understanding of the results of the MS analysis and to enable statistical inference about the peptide amounts. The right panel of Figure 1 displays the mean of each pair against its variance together with several estimates of the variance function described in Section 2.3.

The purpose of the current paper is twofold: first, to estimate the variance function related to observations obtained by the common technique of iTRAQ stable isotope labeling [Ross et al. (2004), Aggarwal, Choe and Lee (2006)], which provides measurements of the relative amounts of peptides from two different biological samples in a single experimental run; and second, to construct confidence intervals for the abundance of a given peptide in a biological sample and to the ratio of two abundances under different conditions (e.g., cancer and wild type cells)

¹Pairs with exactly the same realized value were excluded from the analysis; there were two such peptides.

using the estimated variance function, and to calculate p -values for the hypothesis of equality in peptide amounts in two independent samples.

Models that define the variance as a function of the mean have been studied intensively in the framework of heteroscedastic regression, where different estimation techniques have been suggested [e.g., [Davidian and Carroll \(1987\)](#)]. However, in MS, the variance function depends on the unknown peptide amount and the estimation problem is much more involved. A similar problem arises in certain immunoassay studies where few or no standard concentrations are available [[Raab \(1981\)](#), [Sadler and Smith \(1986\)](#), [O'Malley, Smith and Sadler \(2008\)](#)], and in the evolving area of microarray mRNA expression analysis [[Carroll and Wang \(2008\)](#), [Wang, Ma and Carroll \(2009\)](#), [Fan, Feng and Niu \(2010\)](#)]. Unlike immunoassay and microarray, labeled MS data can contain as few as two measurements of each peptide relative quantity and, therefore, analysis requires a special experiment for estimating the variance function.

In a previous study, [Zhang et al. \(2010\)](#) suggest a novel controlled experiment for the estimation of the variance function in the iTRAQ protocol. Using the standard workflow, they held the sample constant by generating pairs of iTRAQ intensities from identical biological samples. Under this experiment, it is reasonable to assume that the iTRAQ labels are interchangeable in their error characteristics since labeled samples are mixed before processing and, therefore, the difference within pairs of such measurements are entirely attributable to the measurement error of the instrument itself. Two controlled experiments were conducted in January and March of 2009 with the sole goal of estimating the variance function [[Zhang et al. \(2010\)](#)]. A separate experiment was conducted to explore differences observed between wild-type and cancer cells expressing distinct mutations of the same oncogenic kinase (FLT3); this experiment relied on the variance function estimated in the controlled study. The motivation for our current study arises from these past experiments, and below we describe in detail the mathematical problem, suggest statistical methods to tackle it and apply them to the three data sets mentioned above.

In Section 2 we present a naive method for estimating the variance function employed by [Zhang et al.](#) and explore its validity. We show that the method works well when the error terms are typically small, as is the case in the instrument explored by [Zhang et al. \(2010\)](#), but may yield biased estimators for the variance function in other situations. We then suggest an alternative mixture model approach for estimating the variance function and prove its consistency. In Section 3 we use the estimated variance function for interval estimation of the ratio of peptides across two different biological conditions, and we apply the method to iTRAQ-based MS analysis that is intended to decipher the oncogenic potential of two different clinically relevant mutations in the FLT3 receptor tyrosine kinase. Section 4 discusses testing of the hypothesis that the amounts of peptides in two biological samples are equivalent. The properties of the estimation approaches are investigated in Section 5 by simulation. Section 6 completes the paper with some remarks.

2. Estimation of the variance function.

2.1. *The model.* Consider a controlled iTRAQ experiment that quantifies N peptides. Let Y_{i1}, Y_{i2} denote two measures of intensity of peptide i (on a logarithmic scale) having an unknown mean μ_i . Assume that (Y_{i1}, Y_{i2}) ($i = 1, \dots, N$) are independent following the model:

$$(1) \quad Y_{ij} \sim N(\mu_i, h(\theta, \mu_i)), \quad j = 1, 2, \quad \text{independent,}$$

where θ is a vector of unknown variance parameters, and $h(\theta, \mu)$ is a known positive function, such as the power function $\theta_1 \mu^{\theta_2}$ or the exponential function $e^{\theta_1 + \theta_2 \mu}$. In problems where μ_1, \dots, μ_N are known or are modeled by a small number of auxiliary variables, standard techniques for estimating heteroscedastic regression models apply [e.g., [Davidian and Carroll \(1987\)](#)]. However, this is not the case in MS data where μ_1, \dots, μ_N are unknown nuisance parameters.

[Klawonn, Hundertmark and Jansch \(2006\)](#) and [Hundertmark et al. \(2009\)](#) developed an EM algorithm that maximizes over μ_1, \dots, μ_N and θ the likelihood corresponding to N independent observations, (y_{i1}, y_{i2}) , from model (1):

$$(2) \quad \prod_{i=1}^N \{2\pi h(\theta, \mu_i)\}^{-1} \exp\left\{-\frac{(y_{i1} - \mu_i)^2 + (y_{i2} - \mu_i)^2}{2h(\theta, \mu_i)}\right\}.$$

However, since the number of nuisance parameters increases with sample size, the maximum likelihood approach may provide biased estimators. The bias is readily seen in the classical example by [Neyman and Scott \(1948\)](#) of the one-parameter homoscedastic model $h(\theta, \mu) = \theta$. The maximum likelihood estimator for θ under this model is $N^{-1} \sum_i (Y_{i1} - Y_{i2})^2 / 4$ having expectation $\theta/2$, hence it converges to half the true variance. Thus, although the model is parametric, alternatives to the maximum likelihood technique should be employed.

2.2. *The MACL approach of Sadler and Smith.* Motivated by immunoassay data, [Raab \(1981\)](#) suggests to modify the likelihood (2) by multiplying the contribution of each of the pairs (Y_{i1}, Y_{i2}) by $h^{1/2}(\theta, \mu_i)$, and then to maximize the modified likelihood with respect to μ_1, \dots, μ_N and θ . Raab shows by simulation that the standard maximum likelihood estimator is biased, but the modified estimator performs reasonably well. Raab's method is computer intensive, as it requires the estimation of all nuisance parameters.

[Sadler and Smith \(1986\)](#) estimate θ by maximizing Raab's modified likelihood at the point $\mu_i = \bar{Y}_i := (Y_{i1} + Y_{i2})/2$ ($i = 1, \dots, N$):

$$(3) \quad \hat{\theta} = \arg \max_{\theta} \prod_{i=1}^N \frac{1}{4\pi^2 \sqrt{h(\theta, \bar{Y}_i)}} \exp\{-S_i^2 / 2h(\theta, \bar{Y}_i)\},$$

where $S_i^2 := (Y_{i1} - \bar{Y}_i)^2 + (Y_{i2} - \bar{Y}_i)^2 = (Y_{i1} - Y_{i2})^2 / 2$. This approach, called maximum approximate conditional likelihood (MACL) by Sadler and Smith, reduces

the estimation task to solving a small set of nonlinear equations. The resulting estimating equations under the normal model are therefore

$$\sum_{i=1}^N \frac{\partial h(\theta, \bar{Y}_i) / \partial \theta_j}{h^2(\theta, \bar{Y}_i)} \{S_i^2 - h(\theta, \bar{Y}_i)\} = 0.$$

For example, for the model $h(\theta, \mu) = \exp(\theta_1 + \theta_2 \mu)$, the estimating equations reduce to

$$(4) \quad 1 - N^{-1} \sum_{i=1}^N S_i^2 \exp(-\theta_1 - \theta_2 \bar{Y}_i) = 0,$$

$$(5) \quad N^{-1} \sum_{i=1}^N \bar{Y}_i - N^{-1} \sum_{i=1}^N \bar{Y}_i S_i^2 \exp(-\theta_1 - \theta_2 \bar{Y}_i) = 0,$$

and can be easily solved by standard optimization algorithms using, for example, the R function *optim* [R Development Core Team (2011)]. As pointed out by Sadler and Smith, the solution for (3) can be obtained by an iterative reweighted least squares algorithm.

We applied the MACL approach to the January ($N = 2144$) and March ($N = 2174$) experiments described in Section 1 using the functional form $h(\theta, \mu) = \exp(\theta_1 + \theta_2 \mu)$ and obtained the estimates $\hat{\theta} = (4.89, -0.935)$ and $\hat{\theta} = (4.89, -0.925)$, respectively.² The similarity of the two estimated variance functions is remarkable, suggesting that the between-study variability is small. This is a very important finding, as the variance function can be estimated in a control experiment and be applied to data obtained in independent experiments on the same instrument. We finally pooled the data together and obtained an overall MACL estimate of $\hat{\theta} = (4.86, -0.927)$.

Neither Raab nor Sadler and Smith provide sound theoretical justification for their methods, but explore them in several special relevant cases. In general, the expectation of the estimating equations differs from 0, hence, the estimators of the variance function are, in general, inconsistent. This is shown in Appendix A and is illustrated by simulation in Section 5. However, Appendix A suggests that the bias is small when the variance (for all μ_i) is small, because in such circumstances \bar{Y}_i is a good estimator for μ_i , even though it is based on only two observations. Recently, Wang, Ma and Carroll (2009) studied variance functions for microarray experiments and showed a similar inconsistency problem of estimators obtained by the method of moments.

²Data and R codes can be accessed as project syn310406 on the Sage Bionetworks Synapse system (<http://synapse.sagebase.org>).

2.3. *A mixture model.* A possible strategy to deal with the inconsistency of the MACL approach is to impose additional reasonable assumptions on the nuisance parameters. We consider the model

$$(6) \quad \begin{aligned} Y_{ij} | \mu_i &\sim N(\mu_i, h(\theta, \mu_i)), & j = 1, 2, & \text{independent,} \\ \mu_i &\sim G_0, & i = 1, \dots, N, & \text{independent,} \end{aligned}$$

where (i) the support of G_0 is in the segment $[a, b]$, that is, $P(a \leq \mu_i \leq b) = 1$, (ii) the variance is bounded, that is, $\alpha < h(\theta, \mu) < \beta$ for all μ in the support of G_0 for some $0 < \alpha < \beta < \infty$, (iii) $h(\theta, \mu)$ is continuous on $[a, b]$ and identifies θ , that is, knowing $h(\theta, \mu)$ on the support of G_0 implies knowledge of θ .

The assumptions on h are satisfied by most practical models. The reason for bounding G_0 and the choice of the values a and b are discussed in Section 3. To see the importance of the identifiability assumption, consider the model $h(\theta, \mu) = \exp(\theta_1 + \theta_2\mu)$ and a degenerate G_0 that assigns all the mass to some μ_0 . In such a model, $\exp(\theta_1 + \theta_2\mu_0)$ is identifiable, but the pair (θ_1, θ_2) is not.

THEOREM 1. *Under model (6) and the assumptions following it, the maximum likelihood estimator of (θ, G_0) is consistent.*

The proof, which is sketched in Appendix B, is based on the seminal paper by Kiefer and Wolfowitz (1956) who prove the consistency of the maximum likelihood estimator in mixture models such as (6). A recent application of mixture models for variance estimation in microarray analysis can be found in Wang, Ma and Carroll (2009), though they suggest a different estimation strategy for θ .

Several algorithms for deriving the maximum likelihood estimator of a mixture model have been suggested in the literature [see, e.g., Böhning (1999)]. Here we estimate θ and G_0 by employing the EM algorithm, treating (Y_{i1}, Y_{i2}, μ_i) as the complete data on the i th peptide. One strategy for estimating G_0 is to restrict the search to distributions supported on a fine grid and to find the maximum likelihood among these distributions. In the current problem, the variance becomes small for large values of μ and using a simple grid may lead to data points which are too far (in terms of standard deviations) from all support points. We found that defining the support points of G_0 as a function of the variance performed better than using a simple grid. Thus, we first obtained an initial estimate of the variance function using, for example, the MACL approach, and then restricted the distance between two support points to be at most d standard deviations according to the estimated function. Specifically, let $\tilde{\theta}$ be an initial estimate for θ , and define the maximal support point to be $\mu_J = b$, the second to maximal to be $\mu_{J-1} = \mu_J - d\sqrt{h(\tilde{\theta}, \mu_J)}$ and recursively $\mu_{j-1} = \mu_j - d\sqrt{h(\tilde{\theta}, \mu_j)}$. The selected points of support depend on the initial estimate of the variance function and can be updated as part of the algorithm, though in our experience, this update made no significant improvement

when using $d = 1/4$. Once the support points for G_0 are determined, the EM algorithm is applied to estimate θ and G_0 . The algorithm is quite standard and its description is detailed in Appendix C.

We fit the following three forms for $h(\theta, \mu)$ that have been suggested in the literature³: $\exp(\theta_1)\mu^{\theta_2}$, $\exp(\theta_1 + \theta_2\mu)$ and $\exp(\theta_1 + \theta_2\mu) + \exp(\theta_3)$. The right panel of Figure 1 presents estimates of these three functions on a scatter diagram of $\bar{Y}_i = (Y_{i1} + Y_{i2})/2$ against $S_i^2 = (Y_{i1} - Y_{i2})^2/2$. The X-axis is a naive estimate of the mean of each pair, μ_i , and the Y-axis is a naive estimate of the variance. All three models give similar results in most of the range with some deviation for very small values of μ . A loess fit, presented in the figure by the dotted line, is quite close to the functional form $h(\theta, \mu) = \exp(\theta_1 + \theta_2\mu)$ originally suggested by Zhang et al. (2010). This latter model is used throughout this paper.

Applying the EM approach using the functional form $h(\theta, \mu) = \exp(\theta_1 + \theta_2\mu)$, we obtained the estimates $\hat{\theta} = (4.91, -0.929)$ and $\hat{\theta} = (4.96, -0.944)$ based on the January and March experiments, respectively. The estimates are very similar to the estimates obtained by the MACL approach. We then pooled the data from the two experiments together and obtained our final estimate, $\hat{\theta} = (4.84, -0.927)$. The corresponding estimate of G_0 is displayed in the supplemental article [Mandel et al. (2013)]. Initial values for the EM algorithm were obtained by the MACL approach. Starting the algorithm from different points resulted in essentially the same estimate (details are provided in the supplementary materials).

3. Confidence intervals. Having estimated the variance function, confidence intervals for various parameters of interest can be constructed based on data from a new experiment that compares different biological samples. We construct frequentist confidence intervals that are based on the estimate $\hat{\theta}$ of θ under model (6), but do not use the estimate of the mixing distribution G_0 . We take this approach since the variance function is a stable characteristic of the mass spectrometer, whereas G_0 depends on the biological sample and may differ from sample to sample.

3.1. *Confidence intervals for μ .* It is of interest to attach a measure of uncertainty to the observed intensity or to report a range rather than one value for each peptide. This section discusses the construction of $1 - \alpha$ confidence intervals for the peptide amount, μ , based on one observation Y from the model $Y \sim N(\mu, h(\theta, \mu))$; the next section deals with the construction of confidence sets for the relative abundance of a peptide in two different samples.

Similar to the MACL approach for estimation, construction of confidence intervals can be simplified by plugging an estimate of μ in $h(\theta, \mu)$. Thus, a naive confidence interval is constructed by $Y \pm z_{1-\alpha/2} \sqrt{h(\hat{\theta}, Y)}$, where z_α is the α quantile of

³Data and R codes can be accessed as project syn310406 on the Sage Bionetworks Synapse system (<http://synapse.sagebase.org>).

the standard normal distribution. This method is expected to perform reasonably well only in cases where the variance is small and changes slowly with μ .

An exact $1 - \alpha$ confidence set for μ can be constructed using the pivotal quantity $g_Y(\mu) = (Y - \mu)^2 / h(\theta, \mu)$. This interval is defined as $C_\alpha = \{\mu : g_Y(\mu) \leq \chi_{1,1-\alpha}\}$, where $\chi_{df,\alpha}$ is the α quantile of the chi-squared distribution with df degrees of freedom.

For the model $h(\theta, \mu) = \exp(\theta_1 + \theta_2\mu)$, this set can be easily found by a bisection search using the following observations:

- (1) $g_Y(Y) = 0$ is a local minimum.
- (2) For $\theta_2 < 0$, $\mu^* = Y + 2/\theta_2$ is a local maximum of $g_Y(\mu)$, with $g_Y(\mu^*) = 4\theta_2^{-2}e^{-(2+\theta_1+\theta_2Y)}$.

We thus obtain the following properties of the confidence procedure:

- If $g_Y(\mu^*) \leq \chi_{1-\alpha}$, then the confidence set is a one-sided interval of the form $(-\infty, r_1)$.
- If $g_Y(\mu^*) > \chi_{1-\alpha}$, then the confidence set is a union of two intervals, $(-\infty, r_1) \cup (l_2, r_2)$, where $r_1 < \mu^* < l_2 < Y < r_2$.

This nonstandard shape of the confidence set reflects the fact that a realization y is likely either when μ is close to y or when μ is much smaller than y and the variance of the measurement is very large. Often, the range of μ is a priori bounded so that values smaller than r_1 do not belong to the parameter space, and the confidence set is always an interval.

Since the parameter θ is unknown, a consistent estimator based on the mixture approach is plugged in to generate confidence intervals with an approximate level $1 - \alpha$.

3.2. Confidence intervals for $\mu_1 - \mu_2$. Let $Y_1 \sim N(\mu_1, h(\theta, \mu_1))$ and $Y_2 \sim N(\mu_2, h(\theta, \mu_2))$ be the log intensities of the same peptide obtained by the iTRAQ protocol under two different conditions, and assume that Y_1 and Y_2 are independent conditionally on μ_1 and μ_2 . We consider the construction of a confidence set for $\mu_1 - \mu_2$, which is the parameter of primary interest in many quantitative MS studies.

As in the one-parameter case, a naive confidence interval for $\mu_1 - \mu_2$ can be obtained by plugging Y_1 and Y_2 into the variance term: $Y_1 - Y_2 \pm z_{1-\alpha/2} \sqrt{h(\hat{\theta}, Y_1) + h(\hat{\theta}, Y_2)}$. However, such intervals may be anti-conservative, as the estimator Y_i for μ_i is inconsistent, and hence so is the estimator $h(\hat{\theta}, Y_i)$ for $h(\theta, \mu_i)$.

A direct and a relatively simple way of calculating conservative intervals is by Bonferroni correction, that is, by first constructing $1 - \alpha/2$ intervals for μ_1 and for μ_2 , as described in the previous section, and then calculating the minimum and

maximum differences of the two intervals. However, a more direct construction uses the reparametrization $v_1 = \mu_1 - \mu_2$ and $v_2 = \mu_1 + \mu_2$.

Let

$$g_{Y_1 - Y_2}(v_1, v_2) = \frac{Y_1 - Y_2 - v_1}{\sqrt{h(\theta, (v_2 - v_1)/2) + h(\theta, (v_2 + v_1)/2)}}$$

and

$$g_{Y_1 + Y_2}(v_1, v_2) = \frac{Y_1 + Y_2 - v_2}{\sqrt{h(\theta, (v_2 - v_1)/2) + h(\theta, (v_2 + v_1)/2)}}$$

then $(g_{Y_1 - Y_2}(v_1, v_2), g_{Y_1 + Y_2}(v_1, v_2))$ has a bivariate standard normal distribution with correlation

$$\rho(v_1, v_2) = \frac{\{h(\theta, (v_2 + v_1)/2) - h(\theta, (v_2 - v_1)/2)\}}{\{h(\theta, (v_2 - v_1)/2) + h(\theta, (v_2 + v_1)/2)\}}.$$

Thus,

$$g_{Y_1, Y_2}(v_1, v_2) = (g_{Y_1 - Y_2}(v_1, v_2), g_{Y_1 + Y_2}(v_1, v_2)) \\ \times \begin{pmatrix} 1 & \rho(v_1, v_2) \\ \rho(v_1, v_2) & 1 \end{pmatrix}^{-1} \begin{pmatrix} g_{Y_1 - Y_2}(v_1, v_2) \\ g_{Y_1 + Y_2}(v_1, v_2) \end{pmatrix}$$

has a $\chi_{(2)}^2$ distribution and can serve as a pivot for constructing confidence regions. Specifically,

$$(7) \quad C_{Y_1, Y_2}(v_1, v_2) = \{(v_1, v_2) : g_{Y_1, Y_2}(v_1, v_2) \leq \chi_{2, 1-\alpha}\}$$

is an exact $1 - \alpha$ confidence region for (v_1, v_2) and, therefore, $\{v_1 : (v_1, v_2) \in C_{Y_1, Y_2}(v_1, v_2), -\infty < v_2 < \infty\}$ is a conservative confidence set for v_1 .

3.3. An application to the iTRAQ protocol. In this section we analyze data from the two control experiments mentioned in Section 1. As the two biological samples in these experiments were identical, $\mu_{i1} - \mu_{i2} = 0$ for all peptides $i = 1, \dots, N$. In order to evaluate the performance of our confidence intervals, we used the parameters' estimates from one experiment to construct 95% level confidence intervals for the difference of peptide abundances in the other experiment, and calculated the proportion of intervals that did not cover 0, the true difference.

As in the uni-parameter case, the confidence set for (v_1, v_2) is not necessarily a connected set and the resulting confidence set for v_1 is not always an interval. Figure 2 demonstrates the shape of the confidence region for the pairs $(Y_1, Y_2) = (8, 9)$ and $(Y_1, Y_2) = (7.5, 8)$. The X and Y axes are, respectively, v_1 and v_2 , and the confidence sets are the shaded areas. The figure reveals that without restricting the parameter space of μ (hence the parameter space of v_1), the confidence interval for v_1 comprises all of the real line and is noninformative. We therefore assumed that $\mu \in [7.3, 13.9]$, where the limits were determined from typically observed

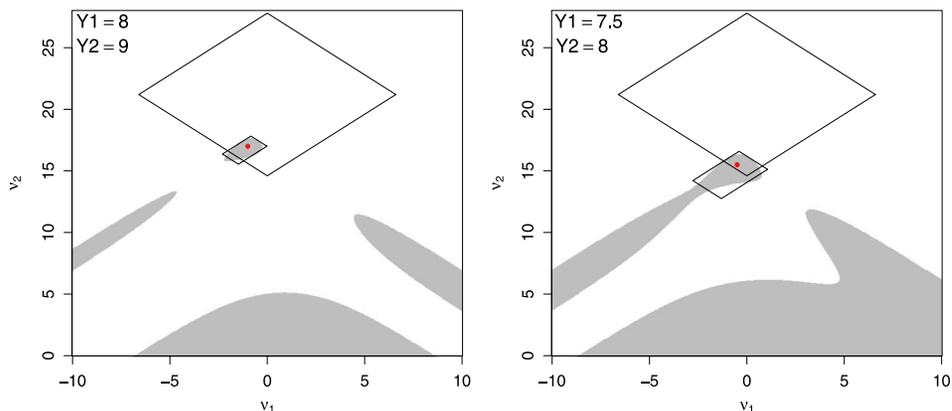


FIG. 2. 95% confidence regions for the pair (v_1, v_2) constructed for two data points: $(8, 9)$ (left) and $(7.5, 8)$ (right).

values as well as the expected range of the intensity values. This decision restricts the values of v_1 and v_2 to the large parallelograms depicted in Figure 2, and enables the construction of informative confidence intervals. The smaller parallelograms represent confidence sets obtained by Bonferroni correction applied to univariate confidence intervals for μ_1 and for μ_2 .

Using estimates from the January data, we calculated confidence intervals for v_1 for peptides in the March experiment by (7), Bonferroni correction and the naive approach, and found that, respectively, 18 (0.8%), 1 (0.05%) and 86 (3.96%) of the 2174 intervals did not include the true parameter $v_1 = 0$. The corresponding numbers for the 2144 peptides in the January experiments using estimates from the March data are 47 (2.19%), 8 (0.4%) and 106 (4.94%). This exercise suggests that interval (7) is better than the Bonferroni interval, but is still conservative. The naive approach performs surprisingly well in our experiment, but, in general, its theoretical coverage probability is not controlled. Further research is needed to understand this phenomenon.

3.4. *An application to cancer phosphoproteomics.* Zhang et al. (2010) used iTRAQ labeling to study a key modification to proteins called phosphorylation, which is important in cell signaling, and is often deregulated in cancer cells. Their study focused on aberrant signaling arising from oncogenic FLT3 mutations in acute myeloid leukemia. In particular, they monitored a key, subcomponent of signal transduction, namely, protein tyrosine phosphorylation, in order to obtain a global understanding of the oncogenic potential of two clinically identified FLT3 mutants (FLT3-D835Y and FLT3-ITD). Both FLT3 mutants induce constitutive signaling even in the absence of proper external cues, which results in uncontrolled cell proliferation, a hallmark of cancer development [Blume-Jensen and Hunter (2001)]. Since receptor tyrosine kinases are often constitutively active in

TABLE 1

95% confidence intervals for the ratio of phosphopeptide quantity across two experimental conditions using the conservative and the naive approaches

Phosphopeptide	FLT3-D835Y	FLT3-ITD	CI	CI naive
VLPQDKEpYYK	10.21	10.78	(0.41, 0.76)	(0.44, 0.72)
GQESEpYGNITYPPAVR	13.62	11.89	(5.05, 6.36)	(5.11, 6.22)
HKEEVpYENVHVK	11.19	9.92	(2.66, 5.05)	(2.76, 4.59)
pYKNILPFDHSR	10.83	9.80	(2.03, 4.10)	(2.12, 3.69)
AVDGpYVKPQIK	11.45	13.36	(0.13, 0.17)	(0.13, 0.17)

cancer cells, it is not surprising that a large proportion of downstream phosphorylation events diverge from a 1:1 value (measured relative to a control cell line), effectively eliminating the possibility of deriving a variance function from the experimental data themselves.

Based on the mixture model results of the pooled control experiments, we calculated 95% confidence intervals for the ratio of phosphopeptides mentioned explicitly in the figures of Zhang et al.; these are reported in Table 1. For comparison, we calculated confidence intervals based on a naive approach mentioned in Section 3.2. Using the intensity-dependent variance function described in this paper, as opposed to a constant cutoff point often used in the literature, subtle changes in phosphorylation levels can be found (e.g., peptide VLPQDKEpYYK). Such ratios can be considered statistically significant despite being smaller than the typical cutoff of 2:1. This in turn suggests that experiments can be designed to explore phosphorylation under physiological conditions, unlike many current experimental designs where artificial or extreme environments are used in order to amplify the observed ratios.

4. Hypothesis testing.

4.1. *Calculation of p-values.* For a given peptide, consider testing the hypothesis $H_0: \mu_1 = \mu_2 = \mu$ versus the two-sided alternative $H_1: \mu_1 \neq \mu_2$. Theoretically, this can be done by inverting the confidence intervals discussed in the previous section. However, for calculating p -values, this inversion is computationally difficult and the current section explores an alternative approach.

As before, let $Y_1 \sim N(\mu_1, h(\theta, \mu_1))$ and $Y_2 \sim N(\mu_2, h(\theta, \mu_2))$ be independent, then $Y_1 - Y_2 \sim N(\mu_1 - \mu_2, h(\theta, \mu_1) + h(\theta, \mu_2))$ and a reasonable test may compare

$$(8) \quad \frac{(Y_1 - Y_2)^2}{2h(\theta, \mu)}$$

to the chi-squared distribution with one degree of freedom. However, (8) contains the unknown parameters θ and μ , and hence is not a legitimate test statistic. As

in the construction of confidence intervals, a naive p -value can be calculated by replacing the denominator of (8) with $2h(\hat{\theta}, (Y_1 + Y_2)/2)$. Although it is reasonable to replace θ with its consistent estimator $\hat{\theta}$, the average $(Y_1 + Y_2)/2$ is an inconsistent estimator for μ , possibly leading to an anti-conservative p -value.

As in other statistical problems involving nuisance parameters, an asymptotically conservative p -value is obtained by

$$\sup_{\mu \in [a, b]} P(Z^2 > (y_1 - y_2)^2 / 2h(\hat{\theta}, \mu)),$$

where y_1 and y_2 are the realized values and Z^2 is a $\chi_{(1)}^2$ random variable. This approach is simple and easy to implement but results in an overly conservative p -value. We therefore suggest to employ the approach of Berger and Boos (1994), where the p -value is calculated by maximization over a confidence interval for the nuisance parameter. Specifically, let C_β be a $1 - \beta$ level confidence interval for μ , obtained in a way similar to that presented in Section 3.1, then we define our p -value as

$$(9) \quad \sup_{\mu \in C_\beta} P(Z^2 > (y_1 - y_2)^2 / 2h(\hat{\theta}, \mu)) + \beta.$$

The choice of β depends on the context. If a univariate hypothesis is tested with a significant level of 5%, then $\beta = 0.001$ is usually a good choice. However, if a Bonferroni correction is needed, then β must be much smaller, as p -value $> \beta$ by construction.

4.2. *Application to the iTRAQ data.* p -values for testing no difference of peptide amounts were calculated for all pairs in the pooled iTRAQ control experiments described in Section 3.3. We used the variance function $h(\theta, \mu) = \exp(\theta_1 + \theta_2 \mu)$ with θ estimated by the EM algorithm applied to the pooled data. Three approaches were compared: the naive approach that replaces μ in $h(\theta, \mu)$ with $\bar{Y} = (Y_1 + Y_2)/2$, a conservative simple approach that replaces μ with a , and the approach of Berger and Boos based on (9) with $\beta = 10^{-6}$.

Figure 3 presents scatterplots and histograms of the p -values. Scatterplots are depicted in logarithmic scale with dotted lines indicating 0.05 and $0.05/N$ significant levels. Recall that the null hypothesis holds in these experiments. Two peptides had naive p -values smaller than the Bonferroni cutoff value, for one of them the Berger and Boos p -value was also under the cutoff level. For about 40% of the experiments the Berger and Boos p -value was similar to the conservative p -value, but for the other peptides, the Berger and Boos method gives considerable smaller values. These are valid p -values and certainly worth the additional computation effort. The distribution of p -values under the naive approach is very close to the uniform distribution, whereas the distributions under the other two approaches are stochastically larger. This and the simulation presented in the next section suggest that the naive approach for testing differences of peptide amounts

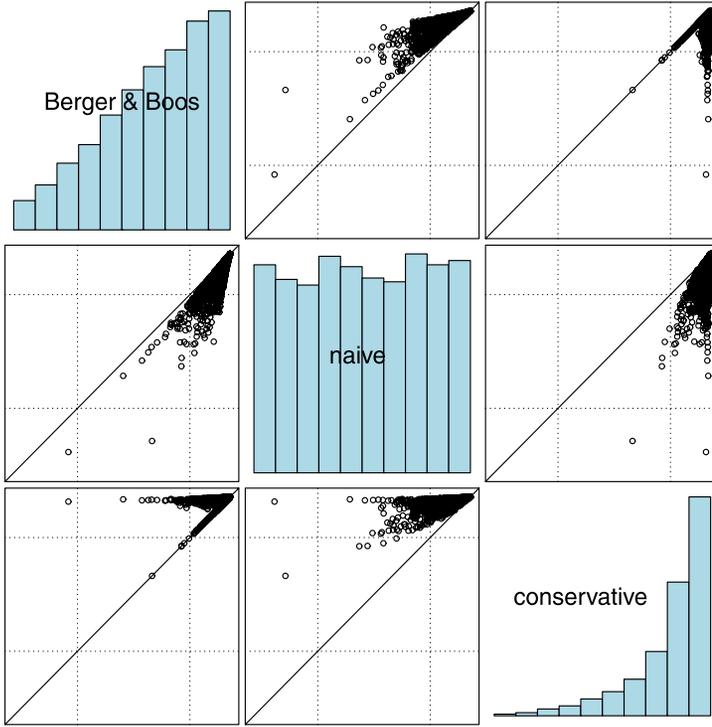


FIG. 3. Scatterplots and histograms of p -values for the pooled control experiments obtained by different approaches. Scatterplots are in log scale with dotted lines indicating significant level of 5% with and without Bonferroni correction.

may be only slightly anti-conservative, though a more extensive study is required before the naive approach can be recommended.

As a second application, we calculated p -values for the data described in Section 3.4 that contrast two mutants found often in certain types of cancer. There are $N = 205$ peptides in the data, the p -values of 141, 123 and 9 of them were smaller than $0.05/N$ according to the naive, the Berger and Boos and the conservative approach, respectively. These peptides are then prioritized for in-depth functional characterization.

5. Simulation.

5.1. *Estimating the variance function.* The performances of the two estimation approaches, the MACL and the mixture model, were tested by simulation under various conditions. The first set of simulations aimed at testing the performance of the MACL approach. For each of the scenarios described below, we generated values for the nuisance parameters and then generated independent pairs of observations from the corresponding normal distributions. We repeated this process 1000 times for different sample sizes ($N = 200, 500, 1000$ and 2000) and for

TABLE 2
Simulation results for the MACL method under the observations fixed scenario

N	θ_1			θ_2		
	θ_1	bias	std	θ_2	bias	std
200	5	-0.061	0.830	-1	0.006	0.081
500	5	-0.051	0.508	-1	0.005	0.049
1000	5	-0.020	0.360	-1	0.002	0.036
2000	5	-0.002	0.251	-1	0.000	0.025
200	5	-1.238	0.759	-0.5	0.127	0.074
500	5	-1.175	0.484	-0.5	0.121	0.047
1000	5	-1.156	0.319	-0.5	0.119	0.031
2000	5	-1.164	0.238	-0.5	0.120	0.023

two different variance functions of the form $\exp(\theta_1 + \theta_2\mu)$: $(\theta_1, \theta_2) = (5, -1)$, which is similar to the values obtained in our data, and $(\theta_1, \theta_2) = (5, -0.5)$, which reflects observations with a much larger variance. The following scenarios were considered:

- Observations fixed: a set of μ_i values was sampled from the observed \bar{Y}_i 's (with replacement) and the same values were used in all 1000 replications.
- Observations random: a different set of μ_i values was sampled from the observed \bar{Y}_i 's (with replacement) for each simulation.
- $U(8, 12)$: the μ_i values were generated from the continuous uniform distribution over $(8, 12)$.
- $U\{8, 9, \dots, 12\}$: the μ_i values were generated from the discrete uniform distribution over $\{8, 9, \dots, 12\}$.

The results of the simulation are summarized in Table 2 and in more details in the supplemental article [Mandel et al. (2013)]. There seems to be almost no difference between the scenarios considered (see Table 1 of the supplementary materials), and this suggests that the approach is insensitive to modest changes in the distribution of the nuisance parameters. Both the variance and the bias decrease with sample size for the model $(\theta_1, \theta_2) = (5, -1)$, and the overall performance of the MACL approach for this case is satisfactory. However, the MACL estimators are biased for the case $(\theta_1, \theta_2) = (5, -0.5)$, and the bias did not decrease with sample size (Table 2). Thus, unless the variance is very small, the approach is problematic and is not recommended.

In order to test the mixture model approach, we generated μ_i by the observations fixed scenario under the two variance functions described above. For each sample size, we simulated 200 data sets and calculated the empirical biases and

TABLE 3

Simulation results for the mixture model method under the observations fixed scenario

N	θ_1			θ_2		
	θ_1	bias	std	θ_2	bias	std
200	5	0.580	0.914	-1	-0.071	0.088
500	5	0.423	0.534	-1	-0.049	0.052
1000	5	0.274	0.328	-1	-0.030	0.032
2000	5	0.173	0.227	-1	-0.019	0.022
200	5	0.070	1.026	-0.5	-0.009	0.099
500	5	0.019	0.608	-0.5	-0.003	0.059
1000	5	0.027	0.397	-0.5	-0.003	0.039
2000	5	-0.013	0.291	-0.5	0.001	0.028

standard deviations. The results are listed in Table 3. As expected, the bias and variance of the estimators decrease with sample size for both models.

Figures 4 and 5 display the performance of the estimators for the variance, that is, the performance of $\exp(\hat{\theta}_1 + \hat{\theta}_2\mu)$ as a function of μ . The gray lines are estimated variance functions from 200 simulated data sets and the true variance function is depicted in black. The figures demonstrate that the mixture model approach is as good as the MACL in the low variance case and performs better in the large variance scenario, especially when the sample size is large.

5.2. *Confidence intervals for μ .* Intervals for μ based on one observation use the pivot $(Y - \mu)^2/h(\theta, \mu)$ and are exact. Here we study the performance of the corresponding $1 - \alpha$ naive confidence intervals that replace μ with Y in the variance function $h(\theta, \mu)$; see Section 3.1.

Figure 6 presents the coverage probability of the naive intervals as a function of the mean ($\mu = 7, 7.1, \dots, 14$), the coverage probability ($1 - \alpha = 0.9, 0.95, 0.99$) and the variance function $[\exp(5 - \mu), \exp(5 - 0.5\mu)]$. For each μ, α and variance function $h(\theta, \mu)$, we estimated the coverage probability by simulating 100,000 replications from the model $N(\mu, h(\theta, \mu))$, constructing naive confidence intervals of level $1 - \alpha$, and calculating the proportion of intervals covering μ . The performance depends on the variance at μ , where the true coverage for small μ (i.e., a large variance) could be much lower than the aimed coverage, especially for large values of $1 - \alpha$ and for the model $h(\theta, \mu) = \exp(5 - 0.5\mu)$ represented by dashed lines.

5.3. *p-values.* A third simulation study was conducted with the aim of understanding the properties of p -values obtained by the different approaches. For a selected set of values for μ , we generated 10,000 independent pairs of observations (Y_1, Y_2) such that $Y_1 \sim N(\mu, e^{\theta_1 + \theta_2\mu})$ and $Y_2 \sim N(\mu_k, e^{\theta_1 + \theta_2\mu_k})$, independently,

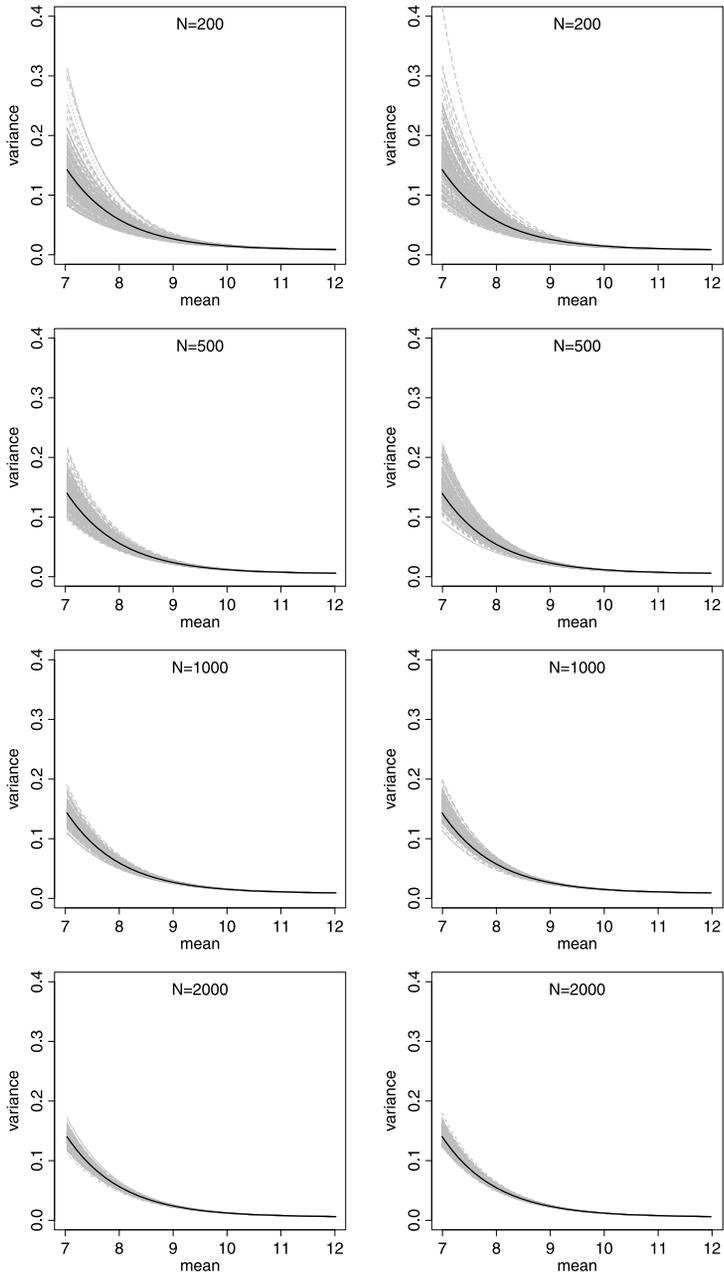


FIG. 4. Estimated variance functions (gray) and the true variance function (black) obtained in 200 simulated data sets for the case $(\theta_1, \theta_2) = (5, -1)$. The figures on the left show the results of the MACL approach and those on the right are for the mixture model approach. The simulated data sample sizes are, from top to bottom, 200, 500, 1000 and 2000.

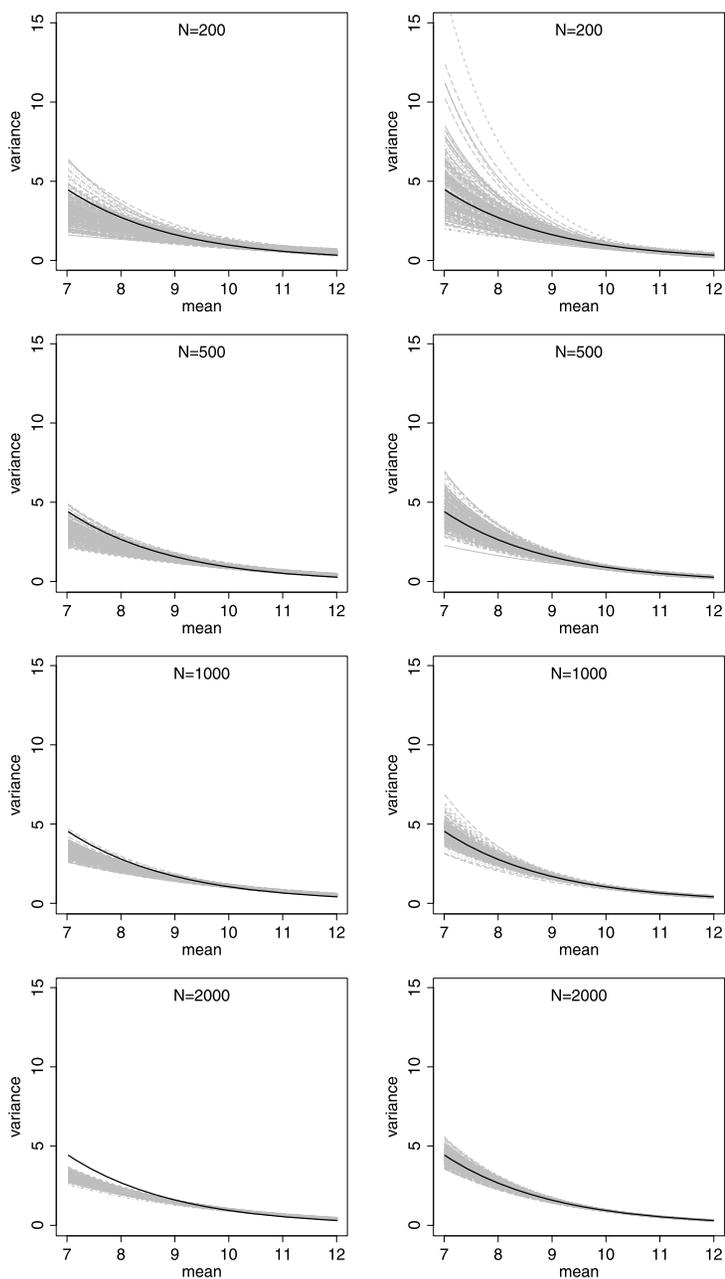


FIG. 5. Estimated variance functions (gray) and the true variance function (black) obtained in 200 simulated data sets for the case $(\theta_1, \theta_2) = (5, -0.5)$. The figures on the left show the results of the MACL approach and those on the right are for the mixture model approach. The simulated data sample sizes are, from top to bottom, 200, 500, 1000 and 2000.

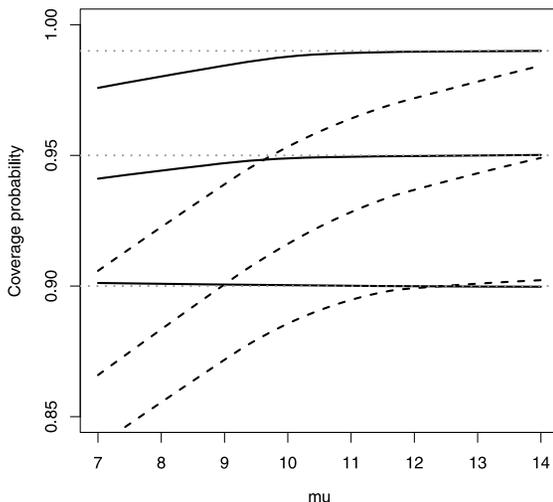


FIG. 6. Coverage probability of the naive confidence interval for μ for confidence levels 0.99, 0.95 and 0.90. Solid lines: $h(\theta, \mu) = \exp(5 - \mu)$, dashed lines: $h(\theta, \mu) = \exp(5 - 0.5\mu)$.

where $\mu_k = \mu + k \times \sqrt{e^{\theta_1 + \theta_2 \mu}}$, that is, Y_1 and Y_2 are centered about k standard deviations apart. The parameter k ranged from 0 to 3, and for (θ_1, θ_2) we studied the values $(5, -1)$ and $(5, -0.5)$. We used the value $\beta = 10^{-3}$ for the Berger and Boos p -values.

Figures 4 and 5 of the supplemental article [Mandel et al. (2013)] present the proportions of p -values that were smaller than 0.05 as a function of μ , k and θ_2 . For the case $\theta_2 = -1$ (supplemental article, Figure 5), the naive approach is only slightly anti-conservative and only for small values of μ , and its power is much larger than that of the other methods. The Berger and Boos approach works reasonably well only for very small and very large values of μ ; the conservative approach is useless. For the case $\theta_2 = -0.5$ (supplemental article, Figure 4), the naive approach is anti-conservative for small values of μ with a significant level of up to 0.08 instead of the declared level of 0.05. However, it is the only method that has a useful power function.

6. Discussion. This paper presents a protocol for estimating the variance function of a mass spectrometer when used for relative quantification in proteomic applications. Using two sets of data collected three months apart, we found that the variance function is stable over time. However, we expect to find different variance functions in different instruments and, hence, each lab should estimate the variance parameters of each spectrometer independently using the protocol described here, and update them periodically. More importantly, the variance estimated here corresponds only to the variance of the instrument itself, and does not include error

terms corresponding to the natural variability of biological samples or the processing required to solubilize proteins from cells and tissues, generate peptides, etc.

Two inference approaches are considered. The first estimates the nuisance parameters by simple averages and then maximizes a target function; the second approach assumes a mixture model and estimates the variance function by maximum likelihood. When the variance is small and changes slowly as a function of the mean, as is the case in the data we analyzed, an average of two iTRAQ reporter ions per peptide provides a reasonable estimate for the unknown μ and the first approach gives good results. However, it may yield highly biased estimators in other scenarios, as demonstrated by simulations, and we therefore recommend the routine use of the mixture model approach that has a sound theoretical justification.

When using the estimated variance function for statistical inference on one parameter μ , exact methods for constructing confidence intervals are available and are much more appropriate than intervals constructed by the anti-conservative naive approach. On the other hand, for inference on the difference or the ratio of peptide abundance measured across two biological conditions, the naive approach performs quite well, yielding a significant level only slightly larger than the aimed one.

The iTRAQ protocol is somewhat more complicated than presented here, as it involves preprocessing of the spectral data to correct for differences in total protein amount, iTRAQ label purity and instrument-specific parameters. Moreover, iTRAQ is known to suffer from contamination due to co-eluting chromatographic peaks that also share similar precursor masses (i.e., peptides which have a similar m/z value and retention time). Theoretically, these factors may induce dependence between measurements that is ignored in the current analysis.

In order to test the underlying normal assumption, Zhang et al. produced a q-q plot of $(Y_{i1} - Y_{i2})/\sqrt{h(\hat{\theta}, \bar{Y}_i)}$ that showed a very good fit. Although the graph is suggestive, it relies on \bar{Y}_i as an estimate for the nuisance parameter μ_i , hence it does not have a theoretical support. A formal approach we intend to explore requires at least three observations for each peptide. A simple transformation of each of the triplets results in variables that, under the normal model, have a Cauchy distribution, and a q-q plot or formal goodness-of-fit tests can be easily employed. This approach requires iTRAQ data from a control experiment that yields more than two independent and identically distributed measures for each peptide. We intend to conduct such an experiment in the future.

The conservative results of the exercise conducted in Section 3.2 are partially due to the inherently conservative construction of the intervals, but may also be a result of the special parametric shape for the variance function we considered. Our chosen parametric model is very simple, enabling the implementation of simple algorithms, and is supported quite well by the data. A nonparametric method has been recently suggested for the analysis of microarray data [Carroll and Wang

(2008)]; the possibility of adopting this method to MS data and of using it for goodness-of-fit testing should be further explored.

APPENDIX A: BIAS OF THE ESTIMATING EQUATIONS

Recall that $\bar{Y}_i \sim N(\mu_i, \frac{1}{2}e^{\theta_1+\theta_2\mu_i})$ and $S_i^2 \sim e^{\theta_1+\theta_2\mu_i} \chi_{(1)}^2$, and that \bar{Y}_i and S_i^2 are independent. Straightforward calculations show that $E(S_i^2) = e^{\theta_1+\theta_2\mu_i}$ and $E\{\exp(-\theta_1 - \theta_2\bar{Y}_i)\} = \exp(-\theta_1 - \theta_2\mu_i + \frac{1}{4}\theta_2^2 e^{\theta_1+\theta_2\mu_i})$ and, therefore,

$$E\left\{N^{-1} \sum_{i=1}^N S_i^2 \exp(-\theta_1 - \theta_2\bar{Y}_i)\right\} = N^{-1} \sum_{i=1}^N \exp\left(\frac{1}{4}\theta_2^2 e^{\theta_1+\theta_2\mu_i}\right)$$

so that the expectation of the first equation (4) differs from zero, unless $\theta_2 = 0$, that is, the homogeneous model of [Neyman and Scott \(1948\)](#).

For the second equation, we have $E\bar{Y}_i \exp(-\theta_2\bar{Y}_i) = -\frac{d}{dt} E \exp(-t\bar{Y}_i)|_{t=\theta_2} = -\frac{d}{dt} \exp\{-t\mu_i + t^2\frac{1}{4}e^{\theta_1+\theta_2\mu_i}\}|_{t=\theta_2} = (\mu_i - \frac{1}{2}\theta_2 e^{\theta_1+\theta_2\mu_i}) \exp\{-\theta_2\mu_i + \theta_2^2\frac{1}{4} \times e^{\theta_1+\theta_2\mu_i}\}$, so that $E\{\bar{Y}_i S_i^2 \exp(-\theta_1 - \theta_2\bar{Y}_i)\} = (\mu_i - \frac{1}{2}\theta_2 e^{\theta_1+\theta_2\mu_i}) \exp(\frac{1}{4}\theta_2^2 \times e^{\theta_1+\theta_2\mu_i})$, and the expectation of the left-hand side of (5) is

$$N^{-1} \sum_{i=1}^N \mu_i - N^{-1} \sum_{i=1}^N \left(\mu_i - \frac{1}{2}\theta_2 e^{\theta_1+\theta_2\mu_i}\right) \exp\left(\frac{1}{4}\theta_2^2 e^{\theta_1+\theta_2\mu_i}\right),$$

which again differs from 0 for $\theta_2 \neq 0$.

In general, the bias will be small if $e^{\theta_1+\theta_2\mu_i}$ is small for all i , which means that the variance of the measurements is small and, hence, the local averages are good estimators for the unknown μ_i parameters.

APPENDIX B: CONSISTENCY OF THE MLE

A generic sample point is $y = (y_1, y_2)$, where, by conditional independence,

$$(10) \quad f(y; \theta|\mu) = \{2\pi h(\theta, \mu)\}^{-1} \exp\left\{-\frac{(y_1 - \mu)^2 + (y_2 - \mu)^2}{2h(\theta, \mu)}\right\}.$$

The marginal density of $Y = (Y_1, Y_2)$ is $g(y; \theta, G_0) = \int_t f(y; \theta|t) dG_0(t)$.

The proof of consistency is based on the result of [Kiefer and Wolfowitz \(1956\)](#) (KW hereafter); the metric we use below is given in KW equation (2.2).

We complete the parameter space of (θ, G_0) by including all proper distributions functions with support in $[a, b]$.

Assumption 1 of KW trivially holds with respect to the Lebesgue measure. Next, note that $f(y; \theta|\mu)$, and hence $g(y; \theta, G_0)$, is bounded above by α^{-1} . Let $(\theta_i, G_i) \rightarrow (\theta^*, G^*)$, where (θ^*, G^*) is in the complete parameter space. In order to verify Assumption 2 of KW, we need to show that $g(y; \theta_i, G_i) \rightarrow g(y; \theta^*, G^*)$.

We have

$$\begin{aligned}
& |g(y; \theta_i, G_i) - g(y; \theta^*, G^*)| \\
&= \left| \int f(y; \theta_i | t) dG_i(t) - \int f(y; \theta^* | t) dG^*(t) \right| \\
&\leq \left| \int \{f(y; \theta_i | t) - f(y; \theta^* | t)\} dG^*(t) \right| + \left| \int f(y; \theta_i | t) d(G_i(t) - G^*(t)) \right| \\
&\leq \int |f(y; \theta_i | t) - f(y; \theta^* | t)| dG^*(t) + \alpha^{-1} \int d|G_i(t) - G^*(t)|.
\end{aligned}$$

The first term vanishes by the Dominated Convergence theorem and the second vanishes by the convergence of G_i to G^* .

For verifying Assumption 3 of KW, define $m(y; \theta^*, G^*, \rho) = \sup g(y; \theta, G)$, where the supremum is taken over all (θ, G) such that $|\theta - \theta^*| + |G - G^*| < \rho$. We need to show that m is a measurable function of y for any $\rho > 0$ and any (θ^*, G^*) in the complete parameter space. This is true for the same arguments given by KW in their first example: g is for each y continuous in (θ, G) and the parameter space is separable. To show this formally, define $A(\theta^*, G^*, \rho, c) = \{y : m(y; \theta^*, G^*, \rho) > c\}$, and let $\{(\theta_i, G_i)\}$ and $\{y_j\}$ be dense subsets in the parameter and sample space, respectively. Let $B(y, r)$ and $B(\theta, G, r)$ be balls of radius r around the corresponding points, then

$$A(\theta^*, G^*, \rho, c) = \bigcap_{n=1}^{\infty} \bigcup_j B(y_j, 1/n),$$

where the union is over $\{j : \exists(\theta_i, G_i) \in B(\theta^*, G^*, \rho) \text{ such that } g(y_j; \theta_i, G_i) > c\}$.

Assumption 4 of identification follows from Bruni and Koch [(1985), Theorem 1], that proves that G_0 and $h(\theta, \mu)$ are identifiable on the support of μ . Assumption (iii) below equation (6) ensures identifiability of θ .

To verify Assumption 5, note that our assumptions on $h(\theta, \mu)$ guarantee that $g(y; \theta, G)$ is bounded above and below so that $E \log\{g(Y; \theta, G)\} > -\infty$, where the expectation is taken with respect to $g(y; \theta_0, G_0)$, the true density of Y .

APPENDIX C: AN EM ALGORITHM

Let $f(y; \theta | \mu)$ be the bivariate normal density defined in (10), and let $a \leq \mu_1 \leq \dots, \mu_J \leq b$ be fixed scalars (support points of G_0). We approximate the likelihood of one observed pair by the following discrete mixture model:

$$(11) \quad g(y; \theta, \boldsymbol{\pi}) = \sum_{j=1}^J \pi_j f(y; \theta | \mu_j),$$

where $\boldsymbol{\pi} = (\pi_1, \dots, \pi_J)$, $\pi_j \geq 0$ and $\pi_1 + \dots + \pi_J = 1$. The unknown parameters are the π_j 's and θ .

To construct the EM algorithm, consider a Multinomial variable Δ over $1, \dots, J$ with a probability vector $\boldsymbol{\pi}$, and define $(\delta_1, \dots, \delta_J)$ by $\delta_j = I\{\Delta = j\}$, where I is the indicator function. Let $Y = (y_1, \dots, y_N)$ be data on N pairs, then the complete log likelihood can be written as

$$(12) \quad \ell(\boldsymbol{\pi}, \theta; Y) = \sum_{i=1}^N \sum_{j=1}^J \delta_{ij} \log\{f(y_i; \theta | \mu_j)\} + \sum_{i=1}^N \sum_{j=1}^J \delta_{ij} \log(\pi_j),$$

where δ_{ij} is an indicator for the (unobserved) event {pair i has mean μ_j }.

Denote by *old* the current estimates of the unknown parameters, then, using the Bayes formula, the expectation step reduces to estimating

$$\begin{aligned} E^{\text{old}}(\delta_{ij} | Y) &= E^{\text{old}}(\delta_{ij} | y_i) = P^{\text{old}}(\delta_{ij} = 1 | y_i) \\ &= \frac{\pi_j^{\text{old}} f(y_i; \theta^{\text{old}} | \mu_j)}{\sum_{k=1}^J \pi_k^{\text{old}} f(y_i; \theta^{\text{old}} | \mu_k)} =: w_{ij}^{\text{old}}. \end{aligned}$$

Note that $\sum_{j=1}^J \delta_{ij} = 1$ by definition, so the above formula can be interpreted as the current estimate of the probability that y_i was generated by the distribution having mean μ_j .

The maximization step is obtained by replacing δ_{ij} in (12) with w_{ij}^{old} and solving

$$(13) \quad \max_{\boldsymbol{\pi}, \theta} \sum_{i=1}^N \sum_{j=1}^J w_{ij}^{\text{old}} \log\{f(y_i; \theta | \mu_j)\} + \sum_{i=1}^N \sum_{j=1}^J w_{ij}^{\text{old}} \log(\pi_j),$$

which is done separately for $\boldsymbol{\pi}$ and θ . For θ , the problem is of a nonparametric regression type and can be solved by reweighted least squares, similar to the MACL approach. The mixing probabilities are simply updated by

$$\pi_j^{\text{new}} = \frac{1}{N} \sum_{i=1}^N w_{ij}^{\text{old}}.$$

Acknowledgments. We thank Yosi Rinott for helpful comments regarding the proof of Theorem 1. We thank the referees and the Associate Editor for many helpful comments and suggestions.

SUPPLEMENTARY MATERIAL

Web-based supplementary materials variance function estimation in quantitative mass spectrometry with application to iTRAQ labeling (DOI: [10.1214/12-AOAS572SUPP](https://doi.org/10.1214/12-AOAS572SUPP); .pdf). Section A: Workflow of the iTRAQ technique. Section B: Estimate of G_0 . Section C: Sensitivity of the EM algorithm to initial values. Section D: Detailed simulation results.

REFERENCES

- AGGARWAL, K., CHOE, L. H. and LEE, K. H. (2006). Shotgun proteomics using the iTRAQ isobaric tags. *Briefings in Functional Genomics and Proteomics* **5** 112–120.
- BERGER, R. L. and BOOS, D. D. (1994). P values maximized over a confidence set for the nuisance parameter. *J. Amer. Statist. Assoc.* **89** 1012–1016. [MR1294746](#)
- BLUME-JENSEN, P. and HUNTER, T. (2001). Oncogenic kinase signalling. *Nature* **411** 355–365.
- BÖHNING, D. (1999). *Computer-Assisted Analysis of Mixtures and Applications: Meta-Analysis, Disease Mapping and Others. Monographs on Statistics and Applied Probability* **81**. Chapman & Hall/CRC, Boca Raton, FL. [MR1684363](#)
- BRUNI, C. and KOCH, G. (1985). Identifiability of continuous mixtures of unknown Gaussian distributions. *Ann. Probab.* **13** 1341–1357. [MR0806230](#)
- CARROLL, R. J. and WANG, Y. (2008). Nonparametric variance estimation in the analysis of microarray data: A measurement error approach. *Biometrika* **95** 437–449. [MR2422697](#)
- DAVIDIAN, M. and CARROLL, R. J. (1987). Variance function estimation. *J. Amer. Statist. Assoc.* **82** 1079–1091. [MR0922172](#)
- ECKEL-PASSOW, J. E., OBERG, A. L., THERNEAU, T. M. and BERGEN, H. R. (2009). An insight into high-resolution mass-spectrometry data. *Biostatistics* **10** 481–500.
- FAN, J., FENG, Y. and NIU, Y. S. (2010). Nonparametric estimation of genewise variance for microarray data. *Ann. Statist.* **38** 2723–2750. [MR2722454](#)
- HUNDERTMARK, C., FISCHER, R., REINL, T., MAY, S., KLAWONN, F. and JÄNSCH, L. (2009). MS-specific noise model reveals the potential of iTRAQ in quantitative proteomics. *Bioinformatics* **25** 1004–1011.
- KIEFER, J. and WOLFOWITZ, J. (1956). Consistency of the maximum likelihood estimator in the presence of infinitely many incidental parameters. *Ann. Math. Statist.* **27** 887–906. [MR0086464](#)
- KLAWONN, F., HUNDERTMARK, C. and JÄNSCH, L. (2006). A maximum likelihood approach to noise estimation for intensity measurements in biology. In *Proceedings of the Sixth IEEE International Conference on Data Mining Workshops* 180–184. IEEE conference publications.
- MANDEL, M., ASKENAZI, M., ZHANG, Y. and MARTO, J. A. (2013). Supplement to “Variance function estimation in quantitative mass spectrometry with application to iTRAQ labeling.” DOI:[10.1214/12-AOAS572SUPP](#).
- NEYMAN, J. and SCOTT, E. L. (1948). Consistent estimates based on partially consistent observations. *Econometrica* **16** 1–32. [MR0025113](#)
- O’MALLEY, A. J., SMITH, M. H. and SADLER, W. A. (2008). A restricted maximum likelihood procedure for estimating the variance function of an immunoassay. *Aust. N. Z. J. Stat.* **50** 161–177. [MR2516873](#)
- R DEVELOPMENT CORE TEAM (2011). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0. Available at <http://www.R-project.org/>.
- RAAB, G. M. (1981). Estimation of a variance function, with application to immunoassay. *Appl. Statist.* **30** 32–40.
- ROSS, P. L., HUANG, Y. N., MARCHESE, J. N., WILLIAMSON, B., PARKER, K., HATTAN, S., KHAINOVSKI, N., PILLAI, S., DEY, S., DANIELS, S., PURKAYASTHA, S., JUHASZ, P., MARTIN, S., BARTLET-JONES, M., HE, F., JACOBSON, A. and PAPPIN, D. J. (2004). Multiplexed protein quantitation in *saccharomyces cerevisiae* using amine-reactive isobaric tagging reagents. *Molecular and Cellular Proteomics* **3** 1154–1169.
- SADLER, W. A. and SMITH, M. H. (1986). A reliable method of estimating the variance function in immunoassay. *Comput. Statist. Data Anal.* **3** 227–239.
- WANG, Y., MA, Y. and CARROLL, R. J. (2009). Variance estimation in the analysis of microarray data. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **71** 425–445. [MR2649604](#)

ZHANG, Y., ASKENAZI, M., JIANG, J., LUCKEY, C. J., GRIFFIN, J. D. and MARTO, J. A. (2010).
A robust error model for iTRAQ quantification reveals divergent signaling between oncogenic
FLT3 mutants in acute myeloid leukemia. *Mol. Cell Proteomics* **9** 780–790.

M. MANDEL
DEPARTMENT OF STATISTICS
HEBREW UNIVERSITY OF JERUSALEM
MOUNT SCOPUS, JERUSALEM
ISRAEL, 91905
E-MAIL: msmic@huji.ac.il

M. ASKENAZI
DEPARTMENTS OF CANCER BIOLOGY
AND BLAIS PROTEOMICS CENTER
DANA-FARBER CANCER INSTITUTE
AND
DEPARTMENT OF BIOLOGICAL CHEMISTRY
AND MOLECULAR PHARMACOLOGY
HARVARD MEDICAL SCHOOL
BOSTON, MASSACHUSETTS 02215-5450
USA
AND
DEPARTMENT OF BIOLOGICAL CHEMISTRY
HEBREW UNIVERSITY OF JERUSALEM
JERUSALEM
ISRAEL

Y. ZHANG
DEPARTMENTS OF CANCER BIOLOGY
AND BLAIS PROTEOMICS CENTER
DANA-FARBER CANCER INSTITUTE
BOSTON, MASSACHUSETTS 02215-5450
USA

J. A. MARTO
DEPARTMENTS OF CANCER BIOLOGY
AND BLAIS PROTEOMICS CENTER
DANA-FARBER CANCER INSTITUTE
AND
DEPARTMENT OF BIOLOGICAL CHEMISTRY
AND MOLECULAR PHARMACOLOGY
HARVARD MEDICAL SCHOOL
BOSTON, MASSACHUSETTS 02215-5450
USA