Statistics and Its Interface Volume 4 (2011) 443–450

Spin–spin coupling information is crucial for unbiased NMR analysis in metabonomics

Hong-Dar Isaac Wu* and Hsieh Fushing†

We address the critical use of spin–spin coupling information for extracting unbiased concentration information of metabolites based on NMR spectroscopy. This coupling information reveals a truncating status due to binning and baseline correction, termed as vertical and horizontal truncations, which are typical operations needed in bin integration for area under spectral peaks. The likelihood function incorporating truncation status of the involved peak areas is analytically derived. We demonstrate that the maximum likelihood estimation (MLE) provides unbiased estimates of metabolite concentrations. When the information of truncation status is neglected, the extent of resultant bias is substantial. These results bear fundamental implications on reliability and validity of most of the popular statistical methodology in metabolomics, including the analysis of variance (ANOVA), principal component analysis (PCA) and multiple testings. For a multiple-response one-way ANOVA, a test statistic is proposed and implemented through a numerical study.

Keywords and phrases: Baseline correction, Peak area, Binning, Truncation, Maximum likelihood estimation, Mixture model, Heterogeneity, Random effect, ANOVA.

1. INTRODUCTION

In the process of maintaining homeostasis, the metabolic profile of an organism is constantly fluctuating, especially in response to pathophysiological stimuli such as a toxin or drug. Thereby the quantitative measurement of small-molecule metabolites is the "snapshot" of an organism's metabolic dynamics. Metabonomics is an emerging area of analytical chemistry that is designated to detect, identify, quantify and catalog such time-varying metabolic changes in an integrated biological system (Lindon, et al., 2001, 2004; Viant et al., 2003). The metabonomic information is mostly extracted effectively and reliably from high-resolution NMR spectroscopy of biofluid, cells and tissues. Ideally the resultant profile of NMR spectra provides the complete list of all metabolites that are actual variables of interest, while each spectral peak reveals the concentration information of corresponding metabolite.

*Supported by grant: NSC 98-2118-M-005-002 of Taiwan.
†Corresponding author. Supported by NSF grant: DMS-1007219.

Conversely the collection of metabolites' concentration information is thought to generate the observed NMR spectra via an underlying physical mixture model. Hence the pattern recognition in chemometric statistical analysis are based directly on the concentrations of all metabolites in the sample. Since a spectroscopic profile of an unknown compound usually contain several thousands of spectral peaks, it is a very high dimensional mixture. Besides the high dimensionality, the sensitivity of NMR analysis to the chemical environment and sampling experiment brings challenges in the extraction of concentration information. Specifically the matrix effects, such as differences in pH, ionic strength, temperature, etc., between samples could cause variations of peak position and line width in $^1H-\text{NMR}^1$ spectra. In addition, baseline or other spectral distortions could also arise even within the same experiment. Consequently multivariate statistical analysis for pattern recognition purposes in metabonomics are often hampered by the combination of these effects (Weljie, et al., 2006). Because each metabolite is differentially sensitive to the aforementioned effects, it is acknowledged that global correction becomes infeasible for such a high dimensional mixture of metabolites.

Various techniques have been proposed to reduce dimensionality and sensitivity problems to a certain extent. For example, the "spectral binning" reduces NMR spectral profile to a few hundreds variables upon fixed or variable bin-width (Gartland, et al., 1991; Anthony, et al., 1994). When the chemical compounds of interest are defined, the method of 'targeted profiling' is proposed as a mixture analysis of the involving NMR resonances which are mathematically modeled based on extensive collection of pure compound spectra. These methods, as do many other multivariate statistical analysis, critically rely on the assumption that each constructed variable contain the same latent chemical information across different samples. However such an assumption is likely violated due to the overlapping peaks, which is further complicated by differences of peak position, line width and baseline correction. Hence bin integration does not necessarily reflect the true changes in spectral areas (Crockford, et al., 2005; Potts, et al., 2001); the concentration information is incoherently evaluated, and is potentially biased.

In this paper, we propose a resolution to this fundamental issue of how to coherently measure the concentration information of a metabolite at a designated peak position of NMR spectroscopy. The heuristic idea behind our proposal

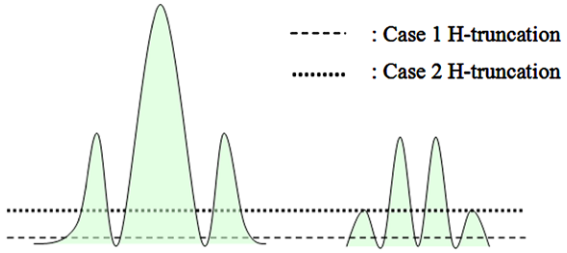

Figure 1. Two chemical shifts (CS) are horizontally truncated. For the CS at the right, case 1 illustrates an H-truncation that partly masks the minor sub-peaks (dash line), and case 2 H-truncation has these two sub-peaks being fully masked (dotted line).

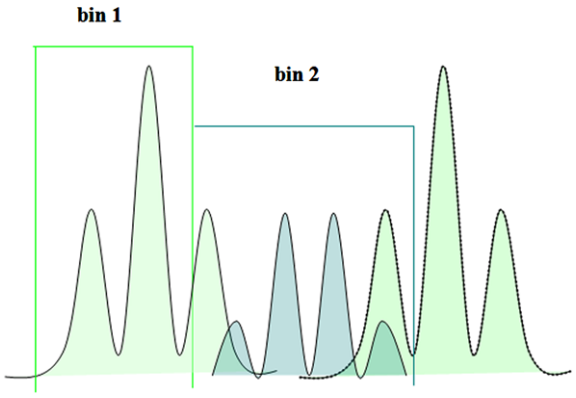

Figure 2. Schematic example of V-truncations with three neighboring major peaks and two imaginary bins. In bin 1, the bin-integration on the left major peak misses part of one sub-peak, and it does not include the area of the overlapping central major peak. However, in 'bin 2', the binning keeps all four sub-peaks of the central major peak, and it also includes part of minor peaks belonging to two adjacent spectral peaks on both sides. This automatically creates serial dependence among adjacent bins for area computation.

is that for each peak area integration, spin–spin coupling information is used to clarify the overlapping spectra content within the chosen bin and a random baseline correction is parametrically modeled, and then the concentration is computed via the maximum likelihood estimation (MLE). Such a MLE-computed concentration is guaranteed to reflect the true change in spectral area. Further, a collection of MLE of the concentrations computed from the spectroscopic profile of all metabolites in the biofluid ideally will serve as the basis for metabonomics.

The spin–spin coupling information is available from the second dimension of $^1H-$NMR[2]. With this information, the content of overlapping peaks within any bin can be explicitly counted. Here binning is referred to as a *vertical truncation*, while the baseline correction is referred to as the *horizontal truncation*. Operationally these two truncations are necessarily involved in any procedure of peak integration in NMR spectroscopy. Pictorial expressions of these two truncations are given in Figures 1 and 2.

To illustrate these two types of truncations, we consider an example structure of ethanol. In high-resolution ^1H NMR spectroscopy, the pure ethanol ($CH_3 - CH_2 - OH$) has three major spectral peaks located at three chemical shifts corresponding to the three structurally different environments of hydrogen (H). The Methyl (CH_3) peak splits into three sub-peaks (a triplet) with height proportional to $(\frac{1}{4}, \frac{2}{4}, \frac{1}{4}) \times 3$, while Methylene ($CH_2$)-peak splits into a quartet with height proportional to $(\frac{1}{8}, \frac{3}{8}, \frac{3}{8}, \frac{1}{8}) \times 2$. The adjacent sub-peaks in the triplet and the quartet is J (Hertz) apart on the chemical-shift coordinate, where J is the *coupling constant*. The information that a major spectral peak located at a particular chemical shift splits into several sub-peaks is specifically termed as the spin–spin coupling information that constitutes the second dimension of 2D-^1HNMR spectra. When ethanol is mixed with many other chemicals or metabolites, the ^1H NMR gives rise to a metabolite profile on the chemical shift dimension and spin–spin coupling information on the other dimension. Methylene's specific spin–spin coupling information at the designated chemical shift becomes the signature of ethanol's existence within the mixture. Similarly Methyl's (CH_3) spin–spin coupling information can identify the existence of ethanol as well. On the metabolite profile, the peak area of Methylene (CH_2), for example, is calculated as the sum of spectral heights of sub-peaks contained in the categorized bin multiplied by the coupling constant J. The bin might contain only a subset of a quartet. This potentially truncated peak area is then used to infer the concentration information of ethanol, so does peak area of Methyl (CH_3).

Although there is a wide spectrum of potential truncations that could occur on either Methylene (CH_2) or Methyl (CH_3), it is noted that Methylene (CH_2) is more susceptible than Methyl (CH_3) due to that it has less number of hydrogens and a larger number of sub-peaks. Specifically the outer two sub-peaks, $\frac{1}{8} \times 2$, of the splitting $(\frac{1}{8}, \frac{3}{8}, \frac{3}{8}, \frac{1}{8}) \times 2$ are the shortest ones among the 7 sub-peaks. Thus they are most likely being truncated by either baseline correction or binning due to possible overlapping with sub-peaks of adjacent major spectral peak.

Intuitively the peak area of Methylene (CH_2) missing two sub-peaks would contain *imbalanced* ethanol's concentration information in contrast to that contained in a peak area of Methyl (CH_3) missing no sub-peaks. It is reasonable to anticipate that statistical inference based on such kind of imbalanced information produces biased results. Further, correction is often subjectively carried out by choosing a uniform baseline for all heterogeneously based major peaks in NMR spectra. Thus this approach possibly truncates many small peaks and causes the resultant peak areas being only meaningful for making intra-metabolite profiling comparison, not inter-profiling comparison.

Despite a peak area subject to these two types of truncation mechanisms, the truncation status can be identified from the spin–spin coupling information. The goal of this paper is to show that, by incorporating such information, the truncated peak area can still coherently provide unbiased estimation of metabolite's concentration. The subsequent statistical analysis, such as ANOVA, PCA analysis and multiple testings, based on the unbiased concentration information render coherent inferences in metabonomics. We organize this paper as follows. Truncation mechanism are rigorously depicted in Section 2. Construction of the likelihood function accommodating the spin–spin coupling information is derived in Section 3, along with computational illustrations of the maximum likelihood estimate. Based on the proposed approach, a possible application to the analysis of variance is sketched in Section 4. A simple numerical study is conducted in Section 5.

2. TRUNCATION MECHANISM AND SPIN–SPIN COUPLING INFORMATION

Let spin–spin coupling information be represented and denoted by $\{K_m(\cdot)\}_1^M$, which is a known set of positive and *discrete* kernels located and centered at M different chemical shifts, such that $\sum_h K_m(h) = 1$. For instance, the kernel for the peak of Ethanol's Methyl (CH_3) is the triplet $(\frac{1}{4}, \frac{2}{4}, \frac{1}{4})$, which means $K_m(1) = \frac{1}{4} = K_m(3)$ and $K_m(2) = \frac{1}{2}$. Similarly, the kernel for Ethanol's Methylene (CH_2) is the quartet $(\frac{1}{8}, \frac{3}{8}, \frac{3}{8}, \frac{1}{8})$. An ideal peak area measured at the m-th chemical shift without truncation is calculated as

$$C \cdot H_m \cdot \sum_{h=1}^{s_m} K_m(h)\Delta,$$

where C denotes the possibly random concentration of interest, H_m the known magnitude of the m-th peak (or the m-th kernel $K_m(\cdot)$), s_m is the total number of sub-peaks, and Δ is the coupling constant (as denoted by J in analytical chemistry literature). Here H_m is simply the number of *equivalent hydrogens*. For Ethanol, H_m equals 3 and 2 respectively for methyl and methylene.

The random effect embedded in the concentration variable C is designed to reflect heterogeneity of peak areas in NMR experiments. In the following, we firstly discuss the horizontal truncation mechanism, and then proceed to two versions of *horizontal plus vertical* truncations.

I. [H-truncation] In the mechanism of horizontal (H) truncation, the observable attenuated peak area is assumed to have the model structure:

$$Y_m = \sum_{h=1}^{s_m} [C \cdot H_m \cdot K_m(h) - T]^+ \Delta + \varepsilon_m,$$

For expository simplicity we assume that both random variable C and measurement error ε_m are independent and normally distributed as $N(\mu, \tau^2)$ and $N(0, \sigma^2)$ respectively, and the truncation random variable T is independent of C and ε_m. In a later context, C is extended to be a multivariate variable. We denote the density function of T by $f_T(t)$, and the vector of measurements by $\mathbf{Y} = (Y_1, \ldots, Y_M)$.

Below we describe two versions of *horizontal plus vertical* ($H + V$) truncation mechanism on spin–spin coupling information.

II. [H+V-truncation]^0 The observed peak area is an integration of a single kernel subject to horizontal and vertical truncations involving no overlapping sub-peaks of adjacent major spectral peaks on the left and right-hand sides. This version gives

$$Y_m = \sum_{h \in I_m} [C \cdot H_m \cdot K_m(h) - T]^+ \Delta + \varepsilon_m,$$

where I_m is a subset of $\{1, 2, \ldots, s_m\}$ found within a designated bin.

[H+V-truncation]^1 The observed peak area is an integration of a major kernel subject to horizontal and vertical truncation, in combination of several overlapping sub-peaks belonging to adjacent major peaks. The two nearest adjacent major peaks are most likely involved: one on the left and the other on the right-hand sides of the major peak of interest. In such a case, the peak area has the following mixed structure:

$$\begin{aligned} Y_{m(m',m^*)} = & \sum_{h \in I_m} [C \cdot H_m \cdot K_m(h) - T]^+ \Delta \\ & + \sum_{h' \in I'_{m'}} [C' \cdot H'_{m'} \cdot K'_{m'}(h') - T']^+ \Delta \\ & + \sum_{h^* \in I^*_{m^*}} [C^* \cdot H^*_{m^*} \cdot K^*_{m^*}(h^*) - T^*]^+ \Delta + \varepsilon_m, \end{aligned}$$

where C' and C^* are concentration variables corresponding to the left and right adjacent major peaks, respectively. Accordingly $I'_{m'}$ is a subset of $\{1, 2, \ldots, s_{m'}\}$ pertaining to the kernel $K'_{m'}(\cdot)$ on the left, while $I^*_{m^*}$ is a subset of $\{1, 2, \ldots, s_{m^*}\}$ pertaining to the kernel $K^*_{m^*}(\cdot)$ on the right.

3. LIKELIHOOD CONSTRUCTION FOR CONCENTRATION INFORMATION

In this section, we consider likelihood construction for concentration information with the objective of evaluating exact concentration of a known metabolite. That is, the proposed statistical inferences are based on *likelihood* function. In particular, the likelihood pertaining to the [H-truncation] mechanism is fully developed. For the [H+V-truncation]^0 mechanism the likelihood can be derived similarly, but notice that the peak areas are calculated in restricted regions due to vertical truncation.

However the likelihood function for the [H+V-truncation]^1 mechanism would not be addressed explicitly here due to its complicated *serial dependence* structure. This serial dependence could be seen from that $Y_{m(m',m^*)}$, given in the previous section, depends on both C' and C^* concentration variables, which could further depend on other concentration variables of chemicals in the chain fashion toward the left and right directions along the chemical shift axis. The serial dependence is expected to potentially and critically affect the validity of statistical inferences when series of peak areas are computed for PCA or multiple testing purposes. This is the reason why the likelihood construction pertaining to [H+V-truncation]^1 mechanism is deferred to a separate future study, while we focus on the bias issues stemming from the other two relatively simple truncation mechanisms.

3.1 Concentration of a known metabolite

Suppose that a known metabolite gives rise to multiple major peaks with spin–spin coupling information specified by the kernels $\{K_m(\cdot)\}_1^M$. (For instance, the Ethanol has $M = 3$ major peaks.) Based on the j-th metabolite profiling, a M-vector of peak areas $\mathbf{Y}_j = (Y_{j1}, \ldots, Y_{jM})'$ is measured under the [H-truncation] via a random variable T_j. We accordingly denote the truncation status of spin–spin coupling information pertaining to peak area Y_{jm} as the subset $M_u(jm)$ of the full domain $S_m = \{1, \ldots, s_m\}$ of the kernel $K_m(\cdot)$, which is integrated to produce Y_{jm}. The complement of subset $M_u(jm)$ is denoted as $M_c(jm) = S_m - M_u(jm)$. If $M_u(jm)$ is empty (that is, $Y_{jm} = 0$), this is called a *complete truncation* and is denoted by $\delta_{jm} = 0$; otherwise $\delta_{jm} = 1$.

The concentration random variables $c_m (m = 1, \ldots, M)$ manifested through kernel $K_m(\cdot)$ are assumed to have identical marginal normal distribution; they are possibly correlated. Denote the multivariate normal distribution of the vector $C = (c_1, \ldots, c_m, \ldots, c_M)$ by $N(\mu \times \tilde{1}, \Sigma)$, where μ is the common mean and $\tilde{1}$ is a M-dimensional 1-vector and Σ is the covariance matrix. For simplicity, in the sequel, let Σ have its diagonal terms $\sigma_m^2 \equiv \tau^2, \forall m$ when σ_m denotes the marginal standard deviation of c_m; and the off-diagonal terms $\rho\sigma_m\sigma_l \equiv \rho\tau^2$ with ρ denotes the correlation between c_m and c_l $(m \neq l)$. Then the likelihood based on the *marginal* distributions of $\{c_m\}_1^M$ contributed by the j-th measurement is

$$(1)\quad L_j(\mu, \tau, \sigma | T = t) = \int_{R_{j1}} \cdots \int_{R_{jM}} \prod_m \phi_\sigma(Y_{jm} - \sum_{h=1}^{s_m} [c_m H_m K_m(h) - t]^+ \Delta) \cdot \phi_\tau(c_m - \mu_j) dc_m,$$

where $\phi_b(y) = 1/\sqrt{2\pi}b \cdot \exp\{-(y)^2/b^2\}$ is a normal density, μ is the common mean of main interest, σ is the standard deviation of the error terms $\{\varepsilon_m\}$ (defined in Sec. 2). The integration regions of c_m in j-th measurement are specified as:

$$\begin{aligned} R_{jm} &= \delta_{jm} R_{jm}^+ \bigcup (1 - \delta_{jm}) R_{jm}^-, \\ R_{jm}^+ &= \left\{ \frac{t}{\min_{h \in M_u(jm)}[H_m K_m(h)]} < c \right. \\ &\qquad \left. < \frac{t}{\max_{h \in M_c(jm)}[H_m K_m(h)]} \right\}, \\ R_{jm}^- &= \left\{ \frac{t}{\max_{h \in M_c(jm)}[H_m K_m(h)]} < c \right\}, \end{aligned}$$

where $\delta_{jm} R_{jm}^+$ is an empty set if $\delta_{jm} = 0$, and $\bigcup$ is the union set operation. If $M_c(jm)$ is an empty set, then $t/\max_{h \in M_c(jm)}[H_m K_m(h)]$ is taken as infinity.

The likelihood (1) is complex and computationally time-consuming. This likelihood can be simplified by dropping the multiple integration as follows. First, conveniently define $1_{mh} = 1\{c_m H_m K_m(h) \geq t\}$ where $1\{A\}$ is the indicator of event A. If the likelihood is constructed through the *joint* distribution of $C = (c_1, \ldots, c_M)$, then the observed $\{Y_{jm}\}$ should be used:

$$Y_{jm} = \sum_{h=1}^{s_m} 1_{mh}(c_m H_m K_m(h) - t)\Delta + \varepsilon_{jm}.$$

For a given j, assume $\{\varepsilon_{jm}\}_{m=1}^M$ to be independent. Thus the joint likelihood is:

$$(2)\quad L_j(\mu, \Sigma, \sigma | T = t) = \int \cdots \int \Phi_{j,\Sigma}(c_1 - \mu, \ldots, c_M - \mu) \prod_m \phi_\sigma(\varepsilon_{jm}) d\varepsilon_{jm},$$

where $\Phi_{j,\Sigma}$ is a *multivariate* Gaussian distribution with variance-covariance matrix Σ, and

$$(3)\quad c_m = \frac{(Y_{jm} - \varepsilon_{jm})/\Delta + (\sum_h 1_{mh})t}{H_m \sum_h 1_{mh} K_m(h)},\ m = 1, \ldots, M.$$

Second, by this expression for $\{c_m\}$, we calculate the simple case when εs are ignored:

$$(4)\qquad Y_{jm} = \sum_{h=1}^{s_m} 1_{mh}(c_m H_m K_m(h) - t)\Delta.$$

The simplification ignoring ε leads to the following likelihood which involves no multiple integrals, and is easily expressed as

$$(5)\qquad L_j(\mu, \Sigma | T = t) = \Phi_{j,\Sigma}(c_1 - \mu, c_2 - \mu, \ldots, c_M - \mu)$$

with

$$(6)\qquad c_m = \frac{Y_{jm}/\Delta + (\sum_h 1_{mh})t}{H_m \sum_h 1_{mh} K_m(h)},\ m = 1, \ldots, M.$$

We adopt (4) and (5) in the next subsection to demonstrate the computational aspects through an artificial example. In Section 5, the same settings ((4) and (5)) are used to report the issues of estimation and an ANOVA-type test. The problem of ignoring *measurement errors* (εs) will be discussed in Section 6.

3.2 Computational illustrations

By considering a simple but realistic setting, we illustrate the construction of the integration region $R\{t|\mathbf{Y_{ij}}\}$ stated in subsection 3.1. Numerical studies are also reported in a later context to confirm potential bias in the MLE estimation if the information of spin–spin coupling and the existence of truncation variable T are ignored.

Let us consider a hypothetical setting having two major chemical shifts ($M = 2$) with number of hydrogen $H_1 = 5, H_2 = 2$ and two kernels of spin–spin coupling $\{K_m(\cdot)\}_1^M$ specified by $K_1(\cdot) = (\frac{1}{4}, \frac{1}{2}, \frac{1}{4})$ and $K_2(\cdot) = (\frac{1}{32}, \frac{5}{32}, \frac{10}{32}, \frac{10}{32}, \frac{5}{32}, \frac{1}{32})$. Consider the case when the first chemical shift exhibits 3 sub-peaks, which are completely observed; and the second chemical shift has a probability (say, 30%) that the two smallest sub-peaks (with $K_2(1) = \frac{1}{32} = K_2(6)$) are fully truncated, while the rest of sub-peaks ($K_2(2) = \frac{5}{32} = K_2(5)$ or $K_2(3) = \frac{10}{32} = K_2(4)$) are always seen. Accordingly potential values of $T = t$ satisfy the following two requirements:

(i) Because $Y_{i1} = (\frac{5}{4}C_1 - t)^+ + (\frac{5}{2}C_1 - t)^+ + (\frac{5}{4}C_1 - t)^+$, we need the truncation variable to be $T < \frac{5}{4}C_1$. If this is the case, we have $Y_{i1} + 3t = 5C_1$ or $C_1 = (Y_{i1} + 3t)/5$. The above requirement of $T < \frac{5}{4}C_1$ is then equivalent to

$$t < \frac{5}{4}\frac{Y_{i1} + 3t}{5} \text{ or just } t < Y_{i1}.$$

(ii) Because $Y_{i2} = (\frac{2}{32}C_2 - t)^+ + (\frac{10}{32}C_2 - t)^+ + (\frac{20}{32}C_2 - t)^+ + (\frac{20}{32}C_2 - t)^+ + (\frac{10}{32}C_2 - t)^+ + (\frac{2}{32}C_2 - t)^+$, for those cases with only "$\frac{1}{32}$" sub-peaks being truncated, we observed $Y_{i2} = 0 + (\frac{10}{32}C_2 - t)^+ + (\frac{20}{32}C_2 - t)^+ + (\frac{20}{32}C_2 - t)^+ + (\frac{10}{32}C_2 - t)^+ + 0 = \frac{15}{8}C_2 - 4t$, or $C_2 = (8Y_2 + 32t)/15$. Due to the requirement that $t > \frac{2}{32}C_2$, we have $t > \frac{1}{26}Y_{i2}$; and because the higher sub-peaks (above $\frac{5}{32}$) are not truncated, it is necessary that $t < \frac{10}{32}C_2 = Y_{i2}/6 + 2t/3$, which leads to $t < Y_{i2}/2$. By combining these arguments, we arrive at the following integration interval for T:

$$\frac{Y_{i2}}{26} < t < \min\left(Y_{i1}, \frac{Y_{i2}}{2}\right).$$

Hence $R\{t|\mathbf{Y_{ij}}\} = (Y_{i2}/26, \min(Y_{i1}, Y_{i2}/2))$.

By the above reasoning, we can then construct the integration region $R\{t|\mathbf{Y_{ij}}\}$ under more general and complex settings. That is, based on (5), the likelihood function obtained by integrating out the distribution of T can be explicitly calculated.

4. COMPARISON OF A METABOLITE'S CONCENTRATIONS CROSS SEVERAL TREATMENTS: ONE-WAY ANOVA WITH MULTIVARIATE RESPONSES

In this section we offer an application of the proposed approach to comparing the concentrations of metabolites across several treatments. Statistical inference involves the one-way analysis of variance (ANOVA) with multivariate responses.

Consider $I + 1$ treatments in an experiment of one-way layout with collected data set $\{\mathbf{Y}_{ij}|i = 0, \ldots, I;\ j = 1, \ldots, n_i\}$, where n_i is the sample size in the $i-$th treatment. Again the horizontal truncation random variables $\{T_{ij}\}$ are assumed to be independent with known distributions. The concentration variables $C_{i1}, \ldots, C_{in_i}$ are independent and identically distributed (*i.i.d*) as $N(\mu_i \times \tilde{1}, \Sigma)$ in each treatment, where $\{C_{ij} = (c_{ij1}, \ldots, c_{ijm}, \ldots, c_{ijM})\}$. Besides, under the [H-truncation] mechanism, the truncation status of spin–spin coupling information $\{M_u(ijm)|m = 1, \ldots, M\}$ is available for all of the (i, j) replicates. Finally, all measurement errors across all treatments are *i.i.d* $N(0, \sigma^2)$ as in the classic ANOVA setting. The equal-covariance matrix (Σ) and equal-variance (σ^2) assumptions are not essential within our likelihood framework. They are used here only for ease of exposition.

In order to test the null hypothesis $H_0 : \mu_0 = \mu_1 = \mu_2 \cdots = \mu_I$ of equal random-effect means among the $I + 1$ treatments, the likelihood contributed by the i-th treatment is

$$L_i(\mu_i, \Sigma) = \prod_{j=1}^{n_i} \int_{R\{t|\mathbf{Y_{ij}}\}} L_{ij}(\mu_i, \Sigma|T_{ij} = t) f_{T_{ij}}(t) dt \tag{7}$$

with $R\{t|\mathbf{Y_{ij}}\}$ being the integration region of t given the observed M-vector of peak area $\mathbf{Y_{ij}}$. The integrand $L_{ij}(\mu_i, \Sigma|T = t)$ is the likelihood function contributed by $\mathbf{Y_{ij}}$ based on (5) and (6) in the above section.

For the purpose of making an ANOVA-type of inference, re-parameterization on $\{\mu_0, \mu_1, \ldots, \mu_I\}$ can be done as follows:

$$\bar{\mu} = \sum_{i=0}^{I} \mu_i/(I + 1);\ \mu_i = \bar{\mu} + \alpha_i;$$

$$\mu_0 = \bar{\mu} - \sum_{i=1}^{I} \alpha_i,\ \alpha_0 = -\sum_{i=1}^{I} \alpha_i.$$

By denoting $\tilde{\alpha} = (\alpha_1, \ldots, \alpha_I)$, the likelihood function of $\theta = (\tilde{\alpha}, \bar{\mu}, \Sigma)$ is computed as:

$$L(\theta) = \prod_{i=0}^{I} L_i(\bar{\mu} + \alpha_i, \Sigma) \tag{8}$$

Table 1. Several basic statistics (including the mean, variance, and some p-th quantiles (q_p)) characterizing the empirical distribution of the realized statistic T_H under null hypothesis $H_0 : \mu_1 = \mu_2 = \mu_3$, and compared to χ^2 distributions with 2 and 3 degrees of freedom

	distribution	mean	variance	$q_{0.25}$	$q_{0.50}$	$q_{0.75}$	$q_{0.90}$	$q_{0.95}$	$q_{0.99}$
parameter	χ^2_2	2	4	0.575	1.386	2.773	4.605	5.991	9.210
condition	χ^2_3	3	6	1.213	2.366	4.108	6.251	7.815	11.347
(1)	$F_n^{(1)}$	2.162	5.047	0.592	1.510	2.884	5.148	6.361	10.587
(2)	$F_n^{(2)}$	2.238	5.055	0.627	1.469	3.212	5.069	6.693	10.863
(3)	$F_n^{(3)}$	2.309	4.835	0.625	1.890	3.145	5.291	6.840	10.798

Based on the above likelihood function, we can solve the MLE for θ. Further denote $\theta(\tilde{\alpha}) = (\bar{\mu}, \Sigma)$ and let the Fisher information of θ, $\mathcal{I}_{\theta\theta}$, be decomposed into

$$\mathcal{I}_{\theta\theta} = \begin{pmatrix} \mathcal{I}_{\tilde{\alpha}\tilde{\alpha}} & \mathcal{I}_{\tilde{\alpha}:\theta(\tilde{\alpha})} \\ \mathcal{I}_{\theta(\tilde{\alpha}):\tilde{\alpha}} & \mathcal{I}_{\theta(\tilde{\alpha})|\theta(\tilde{\alpha})} \end{pmatrix}.$$

The net *sample Fisher information* matrix of $\tilde{\alpha}$, $\hat{\mathcal{I}}_{\tilde{\alpha}|\theta(\tilde{\alpha})}$, is then calculated as

$$\hat{\mathcal{I}}_{\tilde{\alpha}|\theta(\tilde{\alpha})} = \hat{\mathcal{I}}_{\tilde{\alpha}\tilde{\alpha}} - \hat{\mathcal{I}}_{\tilde{\alpha}:\theta(\tilde{\alpha})}\hat{\mathcal{I}}^{-1}_{\theta(\tilde{\alpha})|\theta(\tilde{\alpha})}\hat{\mathcal{I}}_{\theta(\tilde{\alpha}):\tilde{\alpha}}. \tag{9}$$

The proposed statistic for testing the null hypothesis H_0 is

$$T_H = \widehat{\tilde{\alpha}}'\hat{\mathcal{I}}_{\tilde{\alpha}|\theta(\tilde{\alpha})}\widehat{\tilde{\alpha}} \tag{10}$$

which should be asymptotically chi-square distributed with degrees of freedom I (χ^2_I) under H_0. For small or moderate sample size, the computation of the exact degrees of freedom might be very involved, depending on the number of M, I, the sample sizes, and the truncation status of spin–spin coupling information.

5. A NUMERICAL STUDY

In order to perform a numerical experiment, we only consider the case of [H-truncation] in this study. The main goal is to present the MLE estimation under the following two scenarios:

S1 : all information of spin–spin coupling and truncation variable T are considered in the likelihood function;
S2 : information of T is disregarded.

Consider the computational example illustrated in Section 3.2. According to Scenario S1 (H-truncation is considered) and Scenario S2 (H-truncation is ignored), we report in this section: (i) the empirical distribution, under null hypothesis based on 500 replicates, of the proposed test statistic T_H; (ii) the 'mean' and 'standard error' of the estimates of relevant parameters (mainly, μ); and (iii) the test of ANOVA when H-truncation is ignored, compared to the test of T_H which incorporates the truncation.

For (i), notice that $M = 2$ and there are $I + 1 = 3$ treatment ($I = 2$). We compare the empirical distribution of T_H to χ^2_2 and χ^2_3. For simplicity, let

$$C_i = \begin{pmatrix} c_{1i} \\ c_{2i} \end{pmatrix} \sim \left(\begin{pmatrix} \mu_i \\ \mu_i \end{pmatrix}, \tau^2 \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix} \right), \text{ for } i = 1, 2, 3.$$

Here we only report the results corresponding to the *second chemical shift* because the sub-peaks of it tend to be easily truncated since $H_2 = 2$. The following six conditions are analyzed. Throughout the 6 simulations, we let $\tau = 0.5$, $\rho = 0.7$; sample sizes are 20 for all groups. And, when the truncation variable t_i is designed for the i-th treatment, $t_i \sim$ exponential(λ_i) with p.d.f. $f_i(t) = \lambda_i \exp(-\lambda_i \cdot t)$.

(1) Under $H_0, \mu_1 = \mu_2 = \mu_3 (= 2)$, $t_i = 0$ (or, $\lambda_i = \infty$) for $i = 1, 2, 3$.
(2) Under $H_0, \mu_1 = \mu_2 = \mu_3 (= 2)$, $\lambda_1 = 8$, $\lambda_2 = 7$, $\lambda_3 = 6$.
(3) Under $H_0, \mu_1 = \mu_2 = \mu_3 (= 2)$, $\lambda_1 = 8$, $\lambda_2 = 5$, $\lambda_3 = 2$.
(4) Under $H_a, \mu_1 = 2, \mu_2 = 2.25, \mu_3 = 2.5$, $t_i = 0$ (or, $\lambda_i = \infty$) for $i = 1, 2, 3$.
(5) Under $H_a, \mu_1 = 2, \mu_2 = 2.25, \mu_3 = 2.5$, $\lambda_1 = 8$, $\lambda_2 = 7$, $\lambda_3 = 6$.
(6) Under $H_a, \mu_1 = 2, \mu_2 = 2.25, \mu_3 = 2.5$, $\lambda_1 = 8$, $\lambda_2 = 5$, $\lambda_3 = 2$.

For conditions (1) and (4), $t_i = 0$ implies *there is no truncation*. For others, a smaller λ corresponds to a severe truncation so that a larger part of peak area in the chemical shift will not be observed. Thus, for conditions (2) and (5), the truncations only get slightly severe for $i = 2$ and $i = 3$ compared to $i = 1$. However, for conditions (3) and (6), truncations get more severe (in particular for $i = 3$).

Let the empirical distributions of the realized 500 T_Hs be denoted as $F_n^{(k)}$ for $k = 1, 2, 3$, representing the first three simulation conditions (under H_0). The means, standard errors, and p-th quantiles of the 500 realizations of T_H are tabulated in Table 1 with comparisons to χ^2_2 and χ^2_3 for $p = 0.25, 0.50, 0.75, 0.90, 0.95$, and 0.99. Here the p-th quantile, denoted as x_p, of a right-continuous cumulative distribution $F_X(x)$ is defined as $x_p = \inf\{x : p \le F_X(x)\}$.

For point estimation, we observe from Table 2 and Table 3 that Scenario S1 have very small bias in the estimates of $\mu_i, i = 1, 2, 3$ for all six conditions. The biases of Scenario S2 are substantial. Specifically, in Table 2, the estimates of $\mu_2 (= 2)$ and $\mu_3 (= 2)$ are 1.802 and 1.780 for condition

Table 2. Means and standard errors (s.e.) of $\hat{\mu}_j$ of the three treatment groups under $H_0 : \mu_1 = \mu_2 = \mu_3$ for the two scenarios S1 and S2. The ANOVA F-test focuses on the second major peaks (or chemical shift) when H-truncation is neglected (i.e., S1), with $\hat{p}(F)$ denoting the p-value. For the proposed T_H-test, p-values are reported when the null-distribution is assumed to be (i) χ^2_2, or (ii) reference $F_n^{(\cdot)}$, denoted respectively as $\hat{p}$(i) and $\hat{p}$(ii). All tests are based on type I error = 0.05

(H_0)		μ_1=2	μ_2=2	μ_3=2	F-test	T_H-test(i)	T_H-test(ii)
estimate (s.e.)	scenario	$\hat{\mu}_1$	$\hat{\mu}_2$	$\hat{\mu}_3$	$\hat{p}(F)$	$\hat{p}$(i)	$\hat{p}$(ii)
condition (1)	**S1**	2.082	2.079	2.085		0.070	×
		(0.109)	(0.106)	(0.102)			
	S2	2.002	1.994	1.993	0.050		
		(0.107)	(0.103)	(0.097)			
condition (2)	**S1**	2.012	2.011	2.003		0.070	×
		(0.107)	(0.108)	(0.114)			
	S2	1.808	1.802	1.780	0.048		
		(0.110)	(0.102)	(0.101)			
condition (3)	**S1**	2.009	2.003	1.915		0.068	×
		(0.099)	(0.112)	(0.127)			
	S2	1.813	1.727	1.403	0.234		
		(0.107)	(0.121)	(0.126)			

Table 3. Means and standard errors (s.e.) of $\hat{\mu}_j$ for the three treatment groups under $H_a : \mu_i \neq \mu_j$ for some $i \neq j$

(H_a)		μ_1=2	μ_2=2.25	μ_3=2.5	F-test	T_H-test(i)	T_H-test(ii)
estimate (s.e.)	scenario	$\hat{\mu}_1$	$\hat{\mu}_2$	$\hat{\mu}_3$	$\hat{p}(F)$	$\hat{p}$(i)	$\hat{p}$(ii)
condition (4)	**S1**	2.072	2.346	2.590		0.898	0.864
		(0.114)	(0.112)	(0.106)			
	S2	1.980	2.260	2.501	0.798		
		(0.101)	(0.103)	(0.120)			
condition (5)	**S1**	2.010	2.232	2.503		0.868	0.822
		(0.098)	(0.112)	(0.119)			
	S2	1.815	2.038	2.261	0.740		
		(0.114)	(0.122)	(0.105)			
condition (6)	**S1**	2.004	2.238	2.448		0.646	0.614
		(0.121)	(0.117)	(0.123)			
	S2	1.825	1.952	1.868	0.278		
		(0.114)	(0.110)	(0.138)			

(2), and 1.727 and 1.403 for condition (3). In Table 3, the estimates of $\mu_2(= 2.25)$ and $\mu_3(= 2.5)$ are 2.038 and 2.261 for condition (5), and 1.952 and 1.868 for condition (6). We only report the estimate of μ_is. The other estimates ($\hat{\tau}_i, \hat{\rho}_i$, and $\hat{\lambda}_i$) are omitted.

The simulations thus designed have the following purposes. First, under the null hypothesis (H_0), unequal truncations for different treatments torture the observed means to be unequal. This is particularly true for condition (3), where the nominal level inflates to 0.234 for $\alpha = 0.05$ when the H-truncation is neglected (Table 2). Second, we have a reverse situation for the alternative hypothesis (H_a). For conditions (5) and (6), larger means will be truncated to produce lower values, in particular for (6) where the power of traditional ANOVA ignoring truncation lower down to 0.278 (Table 3). However, the proposed statistic T_H retains reasonable nominal levels under H_0 (Table 2): 0.070, 0.070, and 0.068 respectively for conditions (1), (2) and (3).

In order to make the comparisons more reasonable, we report the power of T_H for conditions (4), (5), and (6) using $F_n^{(1)}$, $F_n^{(2)}$, and $F_n^{(3)}$ (in addition to χ^2_2) as the *null distribution* because, except for the μs, the pairs (1) and (4), (2) and (5), and (3) and (6) have the same parameter values in τ, ρ, and λ. Table 3 shows that T_H also has satisfactory power: 0.864, 0.822 and 0.614 for conditions (4), (5), and (6) when $F_n^{(\cdot)}$ is conservatively taken as the null distribution. The corresponding test is written as T_H-test(ii). If χ^2_2 is treated as the null distribution, the powers of the test (T_H-test(i)) elevate to 0.898, 0.868, and 0.646 respectively for the three conditions. Note that there is no parallel rejection probability for T_H-test(ii) under H_0. So, in Table 2, we have the symbol ×.

6. DISCUSSION

In this study we demonstrate that the spin–spin coupling information plays an important role in statistical inferences

for extracting chemical concentration based on peak areas calculated from metabolite profile in NMR spectroscopic analysis. The MLE estimation is shown to give unbiased estimation only when such spin–spin coupling information is properly accommodated; otherwise biased inferences are expected. Based on our numerical study, we also show that more efficient statistical inferences on a metabolite's concentration can be derived from multivariate peak areas corresponding to those chemical shifts that identify the same metabolite.

As spin–spin coupling information being available along with metabolite profile in $^1H-$NMR[2], it clearly indicates how many sub-peaks are involved within a particular bin of chemical shift. This overlapping content of sub-peaks becomes a vital basis for resolving the sensitivity issue due to differences of peak positions and line width, and baseline distortion in metabonomics. This is of practical importance since operational procedures of horizontal truncation and vertical truncation are necessary and unavoidable. Thus, most of the computed peak areas are indeed truncated data. From this perspective, the likelihood-based statistical inference presented here for extracting chemical concentration is fundamental in metabonomics.

Another immediate implication of our development is that the widely used methodologies, such as principal component analysis (PCA) and multiple testing on multi-dimensional peak areas in metabonomics, are prone to be biased if the truncation status of spin–spin coupling information are not properly incorporated into the statistical frameworks. Unfortunately this seems to be the case in the real world application of NMR spectroscopy.

It is understood that the truncation mechanism **[H+V-truncation]**[1] induces serial dependence among the peak areas calculated along the entire metabolite profile. This dependence structure is expected to further complicate the applications of PCA and statistical multiple testings. Therefore research on how to resolve this dependence issue by properly accommodating spin–spin coupling information is surely critical to the successes of these two statistical methods for handling high dimensionality in metabonomics.

Our modeling on the correlation of multivariate peak areas is also useful to help identify highly correlated concentrations. In the numerical study, the maximum likelihood estimates of ρ (not reported in this context) are quite consistent. Thus our MLE approach can reliably point out which pair of peak areas has high correlation, and which two chemical shifts correspond to the same chemical or metabolite. This kind of inference could be useful for the identification of unknown metabolite(s), or for the construction of metabolomic pathway. For the truncation variable T, an adequate modeling of its distribution would shed light on the reliability of peak area calculations. Finally, assuming the existence of measurement error (ε) is reasonable but difficult to check. The resultant likelihood (5) adopted in this study ignores the errors. If the variance of ε (σ^2) is substantial, the performance of the proposed test could retain the empirical Type I error under H_0, but the power under H_a could be lower than what has been observed in Table 3 because of the effect of *attenuation* due to the measurement errors.

ACKNOWLEDGEMENTS

The authors are grateful to the Referees for their very helpful comments which largely improve the presentation of this study.

Received 23 March 2010

REFERENCES

[1] Anthony, M. L., Sweatman, B. C., Beddell, C. R., Lindon, J. C. and Nicholson, J. K. (1994). Pattern recognition classification of the site of nephrotoxicity based on metabolic data derived from proton nuclear magnetic resonance spectra of urine. *Molecular Pharmacology* **46** 199–211.

[2] Crockford, D. J., Keun, H. C., Smith, L. M., Holmes, E. and Nicholson, J. K. (2005). Curve-fitting method for direct quantitation of compounds in complex biological mixtures using ^{1}H NMR: application in metabonomic toxicology studies. *Anal. Chem.* **77** 4556–4562.

[3] Gartland, K. P., Beddell, C. R., Lindon, J. C. and Nicholson, J. K. (1991). Application of pattern recognition methods to the analysis and classification of toxicological data derived from proton nuclear magnetic resonance spectroscopy of urine. *Molecular Pharmacology* **39** 629–642.

[4] Lindon, J. C., Holmes, E. and Nicholson, J. K. (2001). Pattern recognition methods and applications in biomedical magnetic resonance. *Progress in NMR Spectroscopy* **39** 1–40.

[5] Lindon, J. C., Holmes, E., Bollard, M. E., Stanley, E. G. and Nicholson, J. K. (2004). Metabonomics technologies and their applications in physiological monitoring, drug safety assessment and disease diagnosis. *Biomarkers* **9** 1–31.

[6] Potts, B.C., Deese, A. J., Stevens, G. J., Reily, M. D., Robertson, D. G. and Theiss, J. (2001). NMR of biofluids and pattern recognition: assessing the impact of NMR parameters on the principal component analysis of urine from rat and mouse. *J. Pharm. Biomed. Anal.* **26** 463–476.

[7] Viant, M. R., Rosenblum E. S. and Tjeerdema, R. S. (2003). NMR-based metabolomics: a powerful approach for characterizing the effects of environmental stressors on organism health. *Environ. Sci. Technol.* **37** 4982–4989.

[8] Weljie, A. M., Newton, J., Mercier, P., Carlson, E. and Slupsky, C. M. (2006). Targeted profiling: quantitative analysis of ^{1}H NMR metabolomics data. *Anal. Chem.* **78** 4430–4442.

Hong-Dar Isaac Wu
Department of Applied Mathematics
National Chung-Hsing University
Taiwan
E-mail address: honda@amath.nchu.edu.tw

Hsieh Fushing
Department of Statistics
Univ. of California at Davis, CA95616
USA
E-mail address: fushing@wald.ucdavis.edu