Normal–Gamma–Bernoulli peak detection for analysis of comprehensive two-dimensional gas chromatography mass spectrometry data

Seongho Kim a,*, Hyejeong Jang a, Imhoi Koo b,c, Joohyoung Lee a,d, Xiang Zhang b,c

a Biostatistics Core, Karmanos Cancer Institute, Department of Oncology, School of Medicine, Wayne State University, Detroit, MI, United States
b Department of Chemistry, University of Louisville, Louisville, KY, United States
c Center for Regulatory and Environmental Analytical Metabolomics, University of Louisville, Louisville, KY, United States
d Department of Family Medicine and Public Health Sciences, School of Medicine, Wayne State University, Detroit, MI, United States

A R T I C L E   I N F O

Article history:
Received 19 June 2015
Received in revised form 13 July 2016
Accepted 23 July 2016
Available online 3 August 2016

Keywords:
Comprehensive two-dimensional gas chromatography–mass spectrometry (GC×GC–MS)
Metabolomics
Normal–Exponential–Bernoulli (NEB) model
Normal–Gamma–Bernoulli (NGB) model
Peak detection

A B S T R A C T

Compared to other analytical platforms, comprehensive two-dimensional gas chromatography coupled with mass spectrometry (GC×GC–MS) has much increased separation power for analysis of complex samples and thus is increasingly used in metabolomics for biomarker discovery. However, accurate peak detection remains a bottleneck for wide applications of GC×GC–MS. Therefore, the normal–exponential–Bernoulli (NEB) model is generalized by gamma distribution and a new peak detection algorithm using the Normal–Gamma–Bernoulli (NGB) model is developed. Unlike the NEB model, the NGB model has no closed-form analytical solution, hampering its practical use in peak detection. To circumvent this difficulty, three numerical approaches, which are fast Fourier transform (FFT), the first-order and the second-order delta methods (D1 and D2), are introduced. The applications to simulated data and two real GC×GC–MS data sets show that the NGB-D1 method performs the best in terms of both computational expense and peak detection performance.

© 2016 Elsevier B.V. All rights reserved.

1. Introduction

Metabolomics is widely employed for identification and quantification of biochemical metabolites. Multiple analytical platforms, including liquid chromatography–mass spectrometry (LC–MS), gas chromatography–mass spectrometry (GC–MS), and nuclear magnetic resonance spectroscopy (NMR), have been used for the comprehensive characterization of metabolites in biological systems. One such powerful approach in metabolomics is comprehensive two-dimensional gas chromatography coupled with mass spectrometry (GC×GC–MS). Compared with other analytical platforms, GC×GC–MS can significantly increase the separation power by 3–10 folds, depending on the sample complexity and instrument operation (Winnike et al., 2015). Therefore, GC×GC–MS has been increasingly used in metabolomics for biomarker discovery (Phillips et al., 2013; Ralston-Hooper et al., 2008; Schmidt et al., 2013; Shi et al., 2014).

* Correspondence to: Biostatistics Core, Karmanos Cancer Institute, Department of Oncology, School of Medicine, Wayne State University, Detroit, MI 48201, United States.
E-mail address: kimse@karmanos.org (S. Kim).

http://dx.doi.org/10.1016/j.csda.2016.07.015
0167-9473/© 2016 Elsevier B.V. All rights reserved.
Unlike one dimensional GC–MS, there are a few peak detection algorithms available for analysis of the GC × GC–MS data. Peters et al. (2007) used one-dimensional peak detection approach for both GC dimensions. Reichenbach et al. (2004) detected the peaks in two dimensions using the drain algorithm, a modified and inverted version of the watershed algorithm. Commercial software such as ChromaTOF is widely used to reduce the GC × GC–MS raw instrument data into metabolite peak lists, while GC Image (Reichenbach et al., 2004) first detects the blobs by the drain algorithm. Vivo-Truyols (2012) introduced a Bayesian approach for peak detection by finding the best configuration for peak merging based on a Bayesian posterior probability. Recently, Kim et al. (2014) developed a peak detection algorithm for the two-dimensional, which is an R package called msPeak. The R package msPeak first finds peak regions that have potential peaks using Normal–Exponential–Bernoulli (NEB) models and then detects peaks by mixture probability models. The comparison studies demonstrated that msPeak detects less number of peaks than others but shows a lower false discovery rate.

Plancade et al. (2012) pointed out the lack of fit of the usual Normal–Exponential (NE) model and proposed a more flexible method, Normal–Gamma (NG) model, for microarray data analysis. They showed that the NG model fits more accurate than the NE model does, but concluded that the modeling improvement does not lead to a better sensitivity. It still remains unclear which of these models provide better output for peak detection in metabolomics, however. Therefore, it is valuable to study and compare the effect of the NG model in peak detection on the two-dimensional GC–MS data.

We generalized the NEB model by combining the NG model with the Bernoulli distribution to form a novel Normal–Gamma–Bernoulli (NGB) model. The gamma distribution is a natural generalization of the exponential distribution so that it provides more flexible parametric model than the exponential distribution does. Nevertheless, unlike the NEB model, the NGB model has the disadvantage that it has no known closed-form solution, requiring numerical integration or approximation. In this study, three methods are considered to evaluate the NEB models. The developed R package msPeakGC and an example are available at http://mrr.sourceforge.net.

The structure of this paper is as follows. Section 2 reviews the previously developed peak detection algorithm msPeak briefly. In the next section, the newly developed NGB algorithms are introduced, followed by the three numerical approaches to estimating the NGB model in Section 4. In Section 5, we evaluate and compare NEB and NGB models using simulated data and then apply them to two real GC × GC–MS data. Finally, Section 6 concludes the paper.

2. Brief review of msPeak

The peak detection algorithm msPeak was developed by Kim et al. (2014). The algorithm msPeak detects peaks using total ion chromatogram (TIC) and consists of three components: (i) finding potential peak regions as well as simultaneous de-noising and baseline correction, (ii) peak picking and area calculation, and (iii) peak merging.

The first component, which is the most critical process of msPeak, is performed by a hierarchical model, Normal–Exponential–Bernoulli (NEB) model. This NEB model statistically tests to see if the observed TIC at a given position is statistically different from the background signal as well as to perform the de-noising and baseline correction simultaneously. In particular, the statistical test performs to detect an observed TIC at a given position that includes a true signal. Hereafter, the TIC at a given position that is statistically different from the noise or background signal is called the true TIC for the sake of convenience. In other words, the true TIC represents the TIC that includes the true signal generated by a chromatographic compound. For positions where the TIC includes the true signal, the following model is used:

\[ x_i = \theta_i + B, \quad \theta_i \sim \text{Exp} (\phi) \] and \[ B \sim \mathcal{N} (\mu, \sigma^2) \] \tag{1}

where \( \text{Exp} \) and \( \mathcal{N} \) stand for exponential and normal distributions, respectively, \( x_i \) is an observed TIC at the \( i \)th position, \( \theta_i \) is the true signal of the exponential distribution with \( \phi, \mu \) is the mean background or baseline with variance \( \sigma^2 \), and \( \theta_i \) and \( B \) are independent of each other. In case that no true signal is present, there is only the background signal:

\[ x_i = B, \quad B \sim \mathcal{N} (\mu, \sigma^2) \] \tag{2}

Under a Bernoulli distribution, the true TICs (i.e., \( \theta_i \neq 0 \)) of some proportion \( r \) is modeled. Then the parameters \( (\mu, \sigma^2, \phi, r) \) are estimated by maximizing the following log likelihood \( l (\mu, \sigma^2, \phi, r) \):

\[
\sum_i \left( z_i \log p_1 (x_i) + (1 - z_i) \log p_0 (x_i) + z_i \log (r) + (1 - z_i) \log (1 - r) \right),
\] \tag{3}

where \( z_i \) follows a Bernoulli distribution with the proportion of the true TICs, \( r \), and \( p_0 (x_i) \) and \( p_1 (x_i) \) are the marginal density of \( x_i \) when \( \theta_i = 0 \) and \( \theta_i \neq 0 \), respectively, as follows:

\[ p_0 (x_i) = \frac{1}{\sqrt{2\pi} \sigma} \exp \left( -\frac{1}{2\sigma^2} (x_i - \mu)^2 \right); \] \tag{4}

\[ p_1 (x_i) = \frac{1}{\phi} \exp \left( \frac{\sigma^2}{2 \phi^2} - \frac{x_i - \mu}{\phi} \right) \cdot \Phi \left( \frac{x_i - \mu - \sigma^2}{\sqrt{2} \phi^2} \right), \] \tag{5}
where $\Phi(\cdot)$ is the cumulative distribution function (cdf) of the standard normal distribution. The more detailed explanation and derivation of Eqs. (3)–(5) can be found in Kim et al. (2014).

Given the estimates $\hat{\mu}, \hat{\sigma}^2, \hat{\phi}$ of the parameters, the regions or locations where possible true peaks exist (i.e., peak regions) are detected using the following posterior odds:

$$\frac{\hat{p}_1(x_i)}{\hat{p}_0(x_i)} = \frac{\varphi\left(x_i - \left(\hat{\mu} + \frac{\hat{\sigma}^2}{\hat{\phi}}\right)\right)}{\Phi\left(x_i - \left(\hat{\mu} + \frac{\hat{\sigma}^2}{\hat{\phi}}\right)\right)}.$$  (6)

Namely, if the posterior odds of a TIC are greater than or equal to a selected cut-off value (out of 1, 10, and 100), then this TIC is considered as including a true signal (i.e., $\theta_i \neq 0$). Otherwise, the TIC will be considered as a noise by fixing it at zero (i.e., $\theta_i = 0$).

After the significant TICs (i.e., TICs including true signals) are detected by the posterior odds, baseline correction and de-noising are performed simultaneously to obtain the convoluted TIC $\hat{x}_i$ by the following equation:

$$\hat{x}_i = x_i - \left(\hat{\mu} + \frac{\hat{\sigma}^2}{\hat{\phi}}\right) + \hat{\sigma} \cdot \frac{\varphi\left(x_i - \left(\hat{\mu} + \frac{\hat{\sigma}^2}{\hat{\phi}}\right)\right)}{\Phi\left(x_i - \left(\hat{\mu} + \frac{\hat{\sigma}^2}{\hat{\phi}}\right)\right)}.$$  (7)

where $\varphi$ and $\Phi$ are the probability density and cumulative distribution functions of the standard normal distribution $N(0, 1)$, respectively.

The peak picking and area calculation, the second component of $msPeak$, is carried out by model-based approaches in conjunction with first derivative test (FDT). In particular, by considering the number of peaks detected by FDT as the maximum number of peaks that can be detected, a mixture probability model is fitted to obtain the peak position and its area. The five mixture probability models, Poisson, truncated Gaussian, Gaussian, Gamma, and exponentially modified Gaussian distributions, are incorporated in $msPeak$. Note that the main purpose of the use of the mixture probability models is to differentiate overlapped peaks. Then the peak area is estimated using the highest probability density (HPD) regions of 95% based on the fitted mixture probability model.

The last component is the peak merging to correct the multiple peaks that are likely from the same compound due to systemic variations. This process is performed by calculating the MS similarity among the peaks. Furthermore, $msPeak$ provides a trial-and-error optimization in order to select the optimal cut-off value of the Bayes factor and the optimal probability mixture model.

3. Normal–Gamma–Bernoulli (NGB) model

A generalized version of NEB model, Normal–Gamma–Bernoulli (NGB) model, is developed for finding peak regions. Because an exponential distribution is a special case of a gamma distribution, the NEB model is a special case of the NGB model. For positions where the TIC includes the true signal, the NGB model is as follows:

$$x_i = \theta_i + B, \quad \theta_i \sim \text{Gamma}(\alpha, \beta) \quad \text{and} \quad B \sim N(\mu, \sigma^2),$$  (8)

where $\text{Gamma}$ and $N$ stand for gamma and normal distributions, respectively, $x_i$ is an observed TIC at the $i$th position, $\theta_i$ is the true signal of the gamma distribution with $\alpha$ and $\beta$, $\mu$ is the mean background or baseline with variance $\sigma^2$, and $\theta_i$ and $B$ are independent of each other. In case that no true signal is present, there is only the background signal:

$$x_i = B, \quad B \sim N(\mu, \sigma^2).$$  (9)

**Lemma 3.1.** Suppose the random variable $X$ is the sum of the following two random variables

$$X = \theta + B, \quad \theta \sim \text{Gamma}(\alpha, \beta) \quad \text{and} \quad B \sim N(\mu, \sigma^2),$$  (10)

where $\theta$ and $B$ are independent of each other. Then the above model is equivalent to the following hierarchy

$$X|\theta \sim N(\mu + \theta, \sigma^2); \quad \theta \sim \text{Gamma}(\alpha, \beta).$$  (11)

**Proof.** The proof is available in the Supplementary Information (see Appendix A). □

The above lemma allows us to use the hierarchy model in Eq. (11) instead when we drive the marginal density function in Theorem 3.2 and the posterior mean in Theorem 3.3.
**Theorem 3.2.** Suppose we have the hierarchy
\[ X|\theta \sim N \left( \mu + \theta, \sigma^2 \right); \]
\[ \theta \sim \text{Gamma} \left( \alpha, \beta \right), \]
then the marginal density of \( X, f(x) \), is
\[
f(x) = \frac{1}{\Gamma(\alpha) \beta^\alpha} \exp \left( -\frac{1}{\beta} \left( x - \mu - \frac{\sigma^2}{2\beta} \right) \right) \cdot \left( 1 - \Phi \left( -\frac{x - \mu - \frac{\sigma^2}{\beta}}{\sigma} \right) \right) \cdot E \left[ \theta^{\alpha-1} \right]_{\text{TN} \left( x - \mu - \frac{\sigma^2}{\beta}, \sigma \mid \theta > 0 \right)}, \tag{13} \]

where \( \Phi(\cdot) \) is the cumulative distribution function (cdf) of the standard normal distribution and \( \text{TN} \left( x - \mu - \frac{\sigma^2}{\beta}, \sigma \mid \theta > 0 \right) \) represents a truncated normal distribution with mean \( x - \mu - \frac{\sigma^2}{\beta} \) and variance \( \sigma^2 \) when \( \theta > 0 \).

**Proof.** The joint density function of \((x, \theta)\) will be, by Bayes’ law,
\[
f(x, \theta) = f(x | \theta) \cdot g(\theta).
\]
Meanwhile, since \( X|\theta \sim N \left( \mu + \theta, \sigma^2 \right) \) and \( \theta \sim \text{Gamma} \left( \alpha, \beta \right) \),
\[
f(x | \theta) = \frac{1}{\sigma \sqrt{2\pi}} \exp \left( -\frac{1}{2\sigma^2} (x - \theta - \mu)^2 \right);
\]
g(\(\theta\)) = \[
\frac{1}{\Gamma(\alpha) \beta^\alpha} \theta^{\alpha-1} \exp \left( -\frac{\theta}{\beta} \right). \]

Therefore,
\[
f(x) = \int_0^\infty f(x, \theta) \, d\theta = \int_0^\infty \frac{1}{\sigma \sqrt{2\pi}} \exp \left( -\frac{1}{2\sigma^2} (x - \theta - \mu)^2 \right) \frac{1}{\Gamma(\alpha) \beta^\alpha} \theta^{\alpha-1} \exp \left( -\frac{\theta}{\beta} \right) \, d\theta
\]
\[
= \frac{1}{\Gamma(\alpha) \beta^\alpha} \cdot \exp \left( -\frac{1}{\beta} \left( x - \mu - \frac{\sigma^2}{2\beta} \right) \right) \cdot \left( 1 - \Phi \left( -\frac{x - \mu - \frac{\sigma^2}{\beta}}{\sigma} \right) \right)
\]
\[
\cdot \int_0^\infty \frac{\theta^{\alpha-1}}{\sigma \sqrt{2\pi}} \exp \left( -\frac{1}{2\sigma^2} \left( \theta - x + \mu + \frac{\sigma^2}{\beta} \right)^2 \right) \, d\theta
\]
\[
= \frac{1}{\Gamma(\alpha) \beta^\alpha} \cdot \exp \left( -\frac{1}{\beta} \left( x - \mu - \frac{\sigma^2}{2\beta} \right) \right) \cdot \left( 1 - \Phi \left( -\frac{x - \mu - \frac{\sigma^2}{\beta}}{\sigma} \right) \right) \cdot E \left[ \theta^{\alpha-1} \right]_{\text{TN} \left( x - \mu - \frac{\sigma^2}{\beta}, \sigma \mid \theta > 0 \right)}. \tag{14}
\]

Note that the term \( \left( 1 - \Phi \left( -\frac{x - \mu - \frac{\sigma^2}{\beta}}{\sigma} \right) \right) \) is a part of the truncated normal density function of \( \theta \). \( \blacksquare \)

Under a Bernoulli distribution, the true TICs (i.e., \( \theta_i \neq 0 \)) of some proportion \( r \) is modeled. Then the parameters \((\mu, \sigma^2, \alpha, \beta, r)\) are estimated by maximizing the following log likelihood \( l(\mu, \sigma^2, \alpha, \beta, r)\):
\[
\sum_i \left( z_i \log p_1 (x_i) + (1 - z_i) \log p_0 (x_i) + z_i \log r + (1 - z_i) \log (1 - r) \right), \tag{14}
\]
where \( z_i \) follows a Bernoulli distribution with the proportion of the true TICs, \( r \), and \( p_0 (x_i) \) and \( p_1 (x_i) \) are the marginal density of \( x_i \) when \( \theta_i = 0 \) and \( \theta_i \neq 0 \), respectively, and are derived directly by **Theorem 3.2**, as follows
\[
p_0 (x_i) = \frac{1}{\sigma \sqrt{2\pi}} \exp \left( -\frac{1}{2\sigma^2} (x_i - \mu)^2 \right) \tag{15}
\]
\[
p_1 (x_i) = \frac{1}{\Gamma(\alpha) \beta^\alpha} \cdot \exp \left( -\frac{1}{\beta} \left( x_i - \mu - \frac{\sigma^2}{2\beta} \right) \right) \cdot \left( 1 - \Phi \left( -\frac{x_i - \mu - \frac{\sigma^2}{\beta}}{\sigma} \right) \right) \cdot E \left[ \theta_i^{\alpha-1} \right]_{\text{TN} \left( x - \mu - \frac{\sigma^2}{\beta}, \sigma \mid \theta_i > 0 \right)}. \tag{16}
\]
where $TN(\zeta, \eta|\xi > a)$ represents a truncated normal distribution of a random variable $\xi$ with mean $\zeta$ and variance $\eta^2$ when $a < \xi < \infty$. Likewise to Kim et al. (2014), the EM algorithm (Dempster et al., 1977) is employed to estimate the parameters of Eq. (14) by considering the variable $z_i, i = 1, \ldots, N$, as the latent (or missing) variable. In the E-step, the variable $z_i, i = 1, \ldots, N$, is estimated, after fixing the parameters $(\hat{\mu}, \hat{\sigma}^2, \hat{\alpha}, \hat{\beta}, \hat{r})$ at the current estimates, by

$$\hat{z}_i = \text{Prob}(z_i = 1|x_i, \hat{\mu}, \hat{\sigma}^2, \hat{\alpha}, \hat{\beta}, \hat{r}) = \frac{\hat{r}p_1(x_i)}{\hat{r}p_1(x_i) + (1 - \hat{r})p_0(x_i)}. \quad (17)$$

In particular, according to Kim et al. (2014) and Newton et al. (2001), the estimate of $r$, given a Beta(2, 2) as a priori over $r$, is

$$\hat{r} = \frac{2 + \sum \hat{z}_i}{2 + 2 + N}. \quad (18)$$

After fixing $\hat{r}$ and $\hat{z}_i, i = 1, \ldots, N$, the M-step estimates the remaining parameters $(\mu, \sigma^2, \alpha, \beta)$ by maximizing the loglikelihood in Eq. (14) as follows:

$$\left(\hat{\mu}, \hat{\sigma}^2, \hat{\alpha}, \hat{\beta}\right) = \arg\max_{\mu, \sigma^2, \alpha, \beta} l(\mu, \sigma^2, \alpha, \beta). \quad (19)$$

In the next theorem, we will derive the posterior mean $E(\theta|x)$ of the hierarchy model in Eq. (12).

**Theorem 3.3.** Consider the hierarchy model in Eq. (12). Then the posterior mean of $\theta$ given $x$ is

$$E(\theta|x) = \frac{E[\theta^\alpha]_{TN\left(x-\mu-\frac{\sigma^2}{p}, \sigma \mid \theta > 0\right)}}{E[\theta^{\alpha-1}]_{TN\left(x-\mu-\frac{\sigma^2}{p}, \sigma \mid \theta > 0\right)}}. \quad (20)$$

**Proof.** The proof is available in the Supplementary Information (see Appendix A). $\blacksquare$

Using the result of the above theorem, it is straightforward to derive the posterior mean of the $i$th TIC $\theta_i$ given $x_i$, as follows:

$$E(\Theta|x_i) = \int_0^\infty \theta_i p(\theta_i|x_i) d\theta_i$$

$$= \frac{E[\theta_i^\alpha]_{TN(\xi-\mu-\frac{\sigma^2}{p}, \sigma \mid \theta_i > 0)}}{E[\theta_i^{\alpha-1}]_{TN(\xi-\mu-\frac{\sigma^2}{p}, \sigma \mid \theta_i > 0)}} \quad (21)$$

where $TN$ represents a truncated normal distribution. Therefore, the convoluted TIC $\hat{\xi}_i$ can be obtained by the following equation:

$$\hat{\xi}_i = \frac{E[\hat{\theta}_i^\hat{\alpha}]_{TN(\xi-\hat{\mu}-\frac{\hat{\sigma}^2}{p}, \hat{\sigma} \mid \theta_i > 0)}}{E[\hat{\theta}_i^{\hat{\alpha}-1}]_{TN(\xi-\hat{\mu}-\frac{\hat{\sigma}^2}{p}, \hat{\sigma} \mid \theta_i > 0)}} \quad (22)$$

where $TN$ represents a truncated normal distribution. However, due to the complexity of a gamma distribution, Eqs. (16) and (22) do not have the analytical solutions.

4. Approximation to NGB

The expectations in Eqs. (16) and (22) do not have closed-form solutions. Therefore, we use the following three approaches to dealing with them: (i) fast Fourier transform (FFT), (ii) delta method with the first order approximation (D1), and (iii) delta method with the second order approximation (D2).
An FFT was first introduced by Cooley and Tukey (1965). The FFT is a numerical algorithm to compute the discrete Fourier transform and can reduce the required number of computation. Plancade et al. (2012) employed FFT to circumvent the disadvantage of the lack of analytical solutions of the NG model in an R package NormalGamma. We adopted their FFT approach to approximating our NGB model via the R package NormalGamma. In particular, the FFT was used to numerically compute the Normal–Gamma density function in Eq. (16).

The delta method is a method to asymptotically approximate to a normal distribution and can be applied to any function. In this study, we consider the delta methods with the first-order and the second-order approximations. If the expectation of a random variable following the truncated normal distribution is analytically simplified in Eqs. (16) and (22), we have closed-form solutions for all equations. Thus, we employ the delta method to approximate the expectations of the equations. The analytical solution by the delta method with the first-order approximation is as follows:

\[
E \left[ \theta_i^a \right] \approx \left\{ x_i - \mu \frac{\sigma^2}{\beta} + \sigma \cdot \frac{\phi \left( -\frac{1}{\sigma} \cdot \left( x_i - \mu \frac{\sigma^2}{\beta} \right) \right)}{1 - \Phi \left( -\frac{1}{\sigma} \cdot \left( x_i - \mu \frac{\sigma^2}{\beta} \right) \right)} \right\}^a
\]

and that with the second-order approximation is below:

\[
E \left[ \theta_i^a \right] \approx \left\{ x_i - \mu \frac{\sigma^2}{\beta} + \sigma \cdot \frac{\phi \left( -\frac{1}{\sigma} \cdot \left( x_i - \mu \frac{\sigma^2}{\beta} \right) \right)}{1 - \Phi \left( -\frac{1}{\sigma} \cdot \left( x_i - \mu \frac{\sigma^2}{\beta} \right) \right)} \right\}^a \times \left( 1 + \frac{1}{2} \cdot V (\theta) \cdot \alpha \cdot (\alpha - 1) \right)
\]

where

\[
V (\theta) = \sigma^2 \left( \frac{1}{\sigma} - \frac{\phi \left( -\frac{1}{\sigma} \cdot \left( x_i - \mu \frac{\sigma^2}{\beta} \right) \right)}{1 - \Phi \left( -\frac{1}{\sigma} \cdot \left( x_i - \mu \frac{\sigma^2}{\beta} \right) \right)} - \frac{\phi \left( -\frac{1}{\sigma} \cdot \left( x_i - \mu \frac{\sigma^2}{\beta} \right) \right)}{1 - \Phi \left( -\frac{1}{\sigma} \cdot \left( x_i - \mu \frac{\sigma^2}{\beta} \right) \right)} \right)^2
\]

and \( \phi \) and \( \Phi \) represent the probability density function and the cumulative density function, respectively, of a standard normal distribution.

**Lemma 4.1.** Consider the hierarchy model in Eq. (12). The approximated posterior means of \( \theta \) given \( x \) using D1 is

\[
E (\theta | x) \approx x - \mu \frac{\sigma^2}{\beta} + \sigma \cdot \frac{\phi \left( \frac{1}{\sigma} \cdot \left( x - \mu \frac{\sigma^2}{\beta} \right) \right)}{\Phi \left( \frac{1}{\sigma} \cdot \left( x - \mu \frac{\sigma^2}{\beta} \right) \right)}.
\]

Furthermore, when \( \alpha = 1 \), the approximated posterior means of \( \theta \) given \( x \) using D2 is

\[
E (\theta | x) \approx \left( x - \mu \frac{\sigma^2}{\beta} + \sigma \cdot \frac{\phi \left( \frac{1}{\sigma} \cdot \left( x - \mu \frac{\sigma^2}{\beta} \right) \right)}{\Phi \left( \frac{1}{\sigma} \cdot \left( x - \mu \frac{\sigma^2}{\beta} \right) \right)} \right)^3.
\]

The proof of Lemma 4.1 is omitted because it can be easily derived using Eqs. (23) and (24) as well as the property that \( \phi (x) = \phi (-x), -\infty < x < \infty \). The above lemma shows that the convoluted TIC approximated by D1 is exactly same as that under the NEB model as shown in Eq. (7), regardless of \( \alpha \). On the other hand, the convoluted TIC approximated by D2 has a tendency of being higher when the convoluted TIC under the NEB model is greater than 1 or being lower when the NEB’s convoluted TIC is less than 1, when \( \alpha = 1 \). However, it should be noted that, although the expression of the convoluted TIC is identical between NEB and NGB with D1, it does not mean that both methods will detect the same true TICs including true signals. This is because the marginal density \( p_1(x_i) \) is still different between two methods in Eq. (17).

5. Application

We used both two simulated data sets and two experimental data sets to compare and evaluate the proposed NGB models with the NEB model in analysis of GC × GC–MS data for peak detection. The first simulated data were generated under the NEB model with several parameters and the second simulated data were produced under both NEB and NGB models with the set of parameters obtained from the real data application. The performance of each method was evaluated by comparing with the true values.
Table 1
Biases in parameters estimated by NEB and NGB using the NEB simulated data.

<table>
<thead>
<tr>
<th>True r</th>
<th>NEB</th>
<th>NEB-FFT</th>
<th>NGB-D1</th>
<th>NGB-D2</th>
<th>NGB-FFT</th>
<th>NGB-D1</th>
<th>NGB-D2</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.01</td>
<td>-0.0002</td>
<td>0.0000</td>
<td>0.0000</td>
<td>0.0010</td>
<td>-0.0001</td>
<td>-0.0000</td>
<td></td>
</tr>
<tr>
<td>0.05</td>
<td>-0.0007</td>
<td>-0.0002</td>
<td>-0.0004</td>
<td>0.0002</td>
<td>-0.0000</td>
<td>-0.0000</td>
<td></td>
</tr>
<tr>
<td>0.1</td>
<td>-0.0017</td>
<td>-0.0003</td>
<td>-0.0000</td>
<td>0.0000</td>
<td>0.0000</td>
<td>0.0000</td>
<td></td>
</tr>
<tr>
<td>0.2</td>
<td>-0.0000</td>
<td>0.0000</td>
<td>0.0000</td>
<td>0.0000</td>
<td>0.0000</td>
<td>0.0000</td>
<td></td>
</tr>
</tbody>
</table>

Note: The proportion of true TICs (signals).

Two experimental data sets generated from a comprehensive two-dimensional gas chromatography time-of-flight mass spectrometry (GC × GC–TOF MS) were used. For the first experimental data set (Data 1), the sample analyzed on GC × GC–TOF MS was a mixture of 76 compound standards (8270 MegaMix, Restek Corp., Bellefonte, PA) and C7-C40 n-alkanes (Sigma-Aldrich Corp., St. Louis, MO). The concentration of each compound in the mixture was 2.5 µg/mL. The second experimental data set (Data 2) was a mixture of MTBSTFA-derivatized amino acid compounds. A mixture of analytes (50 µM) was created by a combination of two commercially-available mixtures of amino acid standards (Cat. No. A6407 and A6282). All compounds were purchased from Sigma-Aldrich Corp., St. Louis, MO. A 50 µL of the mixture was lyophilized and derivatized with 50 µL N-(tertbutyldimethylsilyl)-N-methyltrifluoroacetamine (MTBSTFA) in acetonitrile (1:1, v/v) for 12 h in a sonicator bath at room temperature. The mixture was analyzed on a LECO Pegasus 4D GC × GC–TOF MS instrument (LECO Corporation, St. Joseph, MI, USA) equipped with a cryogenic modulator. All the statistical analyses were performed using statistical software R 3.0.3 (R Development Core Team, 2014).

5.1. Simulated data

Two simulation studies were carried out to evaluate the proposed peak detection algorithms. In the first simulation study, the data were generated by the NEB model only and both NEB and NGB models were used to generate the data for the second simulation study.

The first simulated data were generated by the NEB model. The model parameters (µ, σ, α, β) were fixed at (100, 0.01, 1, 0.5), respectively, and only the value of r was varied with four different values (0.01, 0.05, 0.1, 0.2), where r is the proportion of true TICs (i.e., TICs including true signals). For each value of r, 500 data sets were generated and then the performance of each method was evaluated. It was assumed that the total number of TICs was 200000 and therefore, that there were true TICs of 2000, 10000, 20000, and 40000, respectively, when r was 0.01, 0.05, 0.1, and 0.2.

Table 1 displays the biases in estimates of parameters after both NEB and NGB models were applied to the simulated data. In case of the NEB model, the bias in the estimates increases as the value of r increases, while the NGB models are less sensitive to the value of r in terms of the biases. Among the NGB models, NGB-FFT has the smaller bias in the estimates of r than those of other methods, while both NGB-D1 and NGB-D2 achieve the smallest biases in the estimates of other parameters (µ, σ, α, β). In particular, although NGB-D2 uses the second-order approximation, there is no difference in biases in the estimates of parameters between NGB-D1 and NGB-D2.

The performance to detect the true TICs was evaluated in terms of true positive rate (TPR), false positive rate (FPR), and positive predictive value (PPV) in Fig. 1. In the figure, the left side depicts the relationship between TPR and FPR, while the TPR vs PPV plot is in the right side. All methods tend to have better performance as the value of r increases. In particular, NEB achieves a little better PPV as well as FPR, while all NGB methods perform slightly better TPR (Supplementary Information Table S1).

Table 2 shows the results of AUC and F1 score. To find the best performed method, the Tukey’s post-hoc analysis was rendered when the F-test (one-way ANOVA) was significant at the significance level of 1%. That is, we first performed one-way ANOVA and then Tukey’s post-hoc analysis was followed only when the F-test was significant. By doing so, we found the best methods that significantly performed better than others for each interesting criterion. The best methods are listed in the last column with the adjusted significance level of 0.01%. NGB-D1 and NGB-D2 perform the best AUC among
Table 2
Area under the ROC curve (AUC) and F1 score of both NEB and NGB using the NEB simulated data.

<table>
<thead>
<tr>
<th>r</th>
<th>NEB</th>
<th>NGB-FFT</th>
<th>NGB-D1</th>
<th>NGB-D2</th>
<th>Best</th>
</tr>
</thead>
<tbody>
<tr>
<td>AUC</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.01</td>
<td>0.8822</td>
<td>0.8874</td>
<td>0.8945</td>
<td>0.8945</td>
<td>NGB-D1</td>
</tr>
<tr>
<td></td>
<td>(0.0021)</td>
<td>(0.0012)</td>
<td>(0.0011)</td>
<td>(0.0011)</td>
<td></td>
</tr>
<tr>
<td>0.05</td>
<td>0.8546</td>
<td>0.8638</td>
<td>0.8675</td>
<td>0.8675</td>
<td>NGB-D1</td>
</tr>
<tr>
<td></td>
<td>(0.0035)</td>
<td>(0.0018)</td>
<td>(0.0018)</td>
<td>(0.0018)</td>
<td></td>
</tr>
<tr>
<td>0.1</td>
<td>0.8319</td>
<td>0.8406</td>
<td>0.8418</td>
<td>0.8418</td>
<td>NGB-D1</td>
</tr>
<tr>
<td></td>
<td>(0.0047)</td>
<td>(0.0025)</td>
<td>(0.0025)</td>
<td>(0.0025)</td>
<td></td>
</tr>
<tr>
<td>0.2</td>
<td>0.7899</td>
<td>0.7968</td>
<td>0.7969</td>
<td>0.7969</td>
<td>NGB-D1</td>
</tr>
<tr>
<td></td>
<td>(0.0117)</td>
<td>(0.0059)</td>
<td>(0.0060)</td>
<td>(0.0060)</td>
<td></td>
</tr>
<tr>
<td>F1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.01</td>
<td>0.6359</td>
<td>0.6381</td>
<td>0.6380</td>
<td>0.6380</td>
<td>NGB-D1</td>
</tr>
<tr>
<td></td>
<td>(0.0335)</td>
<td>(0.0315)</td>
<td>(0.0316)</td>
<td>(0.0316)</td>
<td>NGB-D2</td>
</tr>
<tr>
<td>0.05</td>
<td>0.6768</td>
<td>0.6776</td>
<td>0.6770</td>
<td>0.6770</td>
<td>–</td>
</tr>
<tr>
<td></td>
<td>(0.0359)</td>
<td>(0.0366)</td>
<td>(0.0374)</td>
<td>(0.0374)</td>
<td>–</td>
</tr>
<tr>
<td>0.1</td>
<td>0.6961</td>
<td>0.6974</td>
<td>0.6952</td>
<td>0.6952</td>
<td>–</td>
</tr>
<tr>
<td></td>
<td>(0.0396)</td>
<td>(0.0410)</td>
<td>(0.0444)</td>
<td>(0.0444)</td>
<td>–</td>
</tr>
<tr>
<td>0.2</td>
<td>0.7203</td>
<td>0.7208</td>
<td>0.7073</td>
<td>0.7073</td>
<td>NEB</td>
</tr>
<tr>
<td></td>
<td>(0.0446)</td>
<td>(0.0474)</td>
<td>(0.0772)</td>
<td>(0.0772)</td>
<td>NGB-FFT</td>
</tr>
</tbody>
</table>

* The proportion of true TICs (signals).

Empirical mean (empirical standard deviation).

Significance level at 0.01% after one-way ANOVA and Tukey’s post-hoc analysis.

Fig. 1. TPR vs. FPR and TPR vs. PPV plots for both NEB and NGB using the NEB simulated data. r is the proportion of TICs (signals). The TPR vs. FPR and TPR vs. PPV plots are shown in the left and right side, respectively.

them regardless of the value of r. In case of F1 score, the performance is slightly affected by the value of r. When the value of r is close to 0.01, all NGB methods achieve the better F1 score, but NEB and NGB-FFT have the better performance when the value of r is close to 0.2. Nonetheless, all the methods generally have the similar performance to each other.

Kim et al. (2014) used the conditional Bayes factor (see Eq. (6)) to see if the tentative TIC (signal) is statistically a true TIC (signal). However, in this study, we used the value of z (see Eqs. (14) and (17)), which can, similar to the conditional Bayes factor, be interpreted as the probability that the tentative TIC includes the true signal. In fact, the value of z was more convenient to find an optimal cut-off value mainly because the value of z is bounded between 0 and 1. We optimized the
A cut-off value based on F1 score that is defined as the harmonic mean of TPR and PPV. Namely, we chose the cut-off value of $z$ with the highest mean of F1 score.

Fig. 2 displays the boxplots of F1 score by several different cut-off values ranging from 0.1 to 0.99. Interestingly, the highest mean of F1 score occurs when the cut-off value is less than 0.5. The F1 score-based optimal cut-off value of NEB is 0.3 (empirical mean and standard deviation of F1 score = 0.7227 ± 0.0377), while those of NGB is 0.4 (FFT = 0.7232 ± 0.0372, D1 = 0.7232 ± 0.0373, and D2 = 0.7232 ± 0.0373). These cut-off values are not affected by the value of $r$ (Supplementary Information Figure S1). Table 3 lists the computation time of all methods in Table 3. While the computation time depends on how the algorithms were implemented and the computer used, NGB-FFT takes significantly more computation time than the other methods do. As expected, NGB-D1 has less expensive computation time.

It should be noted that this limited simulation study aims to see how well the NGB model performs over the NEB model when the data follow the NEB model, in terms of detecting the TICs that include true signals. In other words, the simulation was designed up to finding potential peak regions by Eq. (17), not including further procedures, such as the peak shape fitting and picking. Therefore, this simulation result might not reflect the performance of the entire procedures of the peak detection. To achieve the comprehensive comparison for the entire peak detection procedure, more rigorous simulation will be necessary by reflecting experimental conditions, such as column bleeding and injection irregularities.
Table 3
Computation expenses for parameter estimation of both NEB and NGB using the NEB simulated data.

<table>
<thead>
<tr>
<th>$r$</th>
<th>NEB</th>
<th>NGB-FFT</th>
<th>NGB-D1</th>
<th>NGB-D2</th>
<th>Best</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.01</td>
<td>163.44</td>
<td>620.74</td>
<td>112.98</td>
<td>205.34</td>
<td>NGB-D1</td>
</tr>
<tr>
<td></td>
<td>(52.79)</td>
<td>(133.85)</td>
<td>(9.12)</td>
<td>(15.51)</td>
<td></td>
</tr>
<tr>
<td>Computation Time (in second)</td>
<td>203.49</td>
<td>641.07</td>
<td>111.83</td>
<td>196.08</td>
<td>NGB-D1</td>
</tr>
<tr>
<td>0.05</td>
<td>(22.02)</td>
<td>(130.99)</td>
<td>(9.56)</td>
<td>(19.49)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>208.98</td>
<td>651.99</td>
<td>112.90</td>
<td>198.04</td>
<td>NGB-D1</td>
</tr>
<tr>
<td></td>
<td>(30.40)</td>
<td>(146.38)</td>
<td>(9.60)</td>
<td>(21.66)</td>
<td></td>
</tr>
<tr>
<td>0.1</td>
<td>235.52</td>
<td>554.59</td>
<td>101.76</td>
<td>179.80</td>
<td>NGB-D1</td>
</tr>
<tr>
<td></td>
<td>(21.29)</td>
<td>(129.11)</td>
<td>(7.66)</td>
<td>(17.63)</td>
<td></td>
</tr>
<tr>
<td>0.2</td>
<td>203.49</td>
<td>641.07</td>
<td>111.83</td>
<td>196.08</td>
<td>NGB-D1</td>
</tr>
<tr>
<td></td>
<td>(133.85)</td>
<td>(9.12)</td>
<td>(15.51)</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

* The proportion of true TICs (signals).

b Empirical mean (empirical standard deviation).

* Significance level at 0.01% after one-way ANOVA and Tukey's post-hoc analysis.

Table 4
Parameters estimated by both NEB and NGB using real GC × GC–MS data.

<table>
<thead>
<tr>
<th></th>
<th>$\hat{\mu}$</th>
<th>$\hat{\sigma}$</th>
<th>$\hat{\alpha}$</th>
<th>$\hat{\beta}$</th>
<th>$\hat{r}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Data 1</td>
<td>NEB</td>
<td>0.4975</td>
<td>0.0003</td>
<td>-</td>
<td>0.5695</td>
</tr>
<tr>
<td></td>
<td>NGB</td>
<td>0.4972</td>
<td>0.0003</td>
<td>0.1345</td>
<td>1.4213</td>
</tr>
<tr>
<td></td>
<td>FFT</td>
<td>0.4854</td>
<td>0.0002</td>
<td>0.0182</td>
<td>1.4930</td>
</tr>
<tr>
<td></td>
<td>D1</td>
<td>0.4854</td>
<td>0.0002</td>
<td>0.0182</td>
<td>1.4930</td>
</tr>
<tr>
<td></td>
<td>D2</td>
<td>0.7560</td>
<td>0.0004</td>
<td>-</td>
<td>0.3280</td>
</tr>
<tr>
<td>Data 2</td>
<td>NEB</td>
<td>0.7539</td>
<td>0.0004</td>
<td>0.0848</td>
<td>0.7348</td>
</tr>
<tr>
<td></td>
<td>NGB</td>
<td>0.7529</td>
<td>0.0007</td>
<td>0.0047</td>
<td>2.3461</td>
</tr>
<tr>
<td></td>
<td>FFT</td>
<td>0.7529</td>
<td>0.0007</td>
<td>-</td>
<td>2.3461</td>
</tr>
<tr>
<td></td>
<td>D1</td>
<td>0.7529</td>
<td>0.0007</td>
<td>-</td>
<td>2.3461</td>
</tr>
<tr>
<td></td>
<td>D2</td>
<td>0.7529</td>
<td>0.0007</td>
<td>-</td>
<td>2.3461</td>
</tr>
</tbody>
</table>

* $\alpha$ is fixed as one due to the nature of NEB.

This first simulation study shows that in terms of AUC and F1, there is no substantial difference among the methods, although NGB-FFT appears to be the best among them in general. However, in terms of computation time, NGB-D1 takes the least amount of time and NGB-FFT takes the most amount of time.

We further validated the performance of peak detection over the real data sets by additional simulation studies. In this second simulation study, the simulated data were generated by both NEB and NGB models. The model parameters $(\mu, \sigma, \alpha, \beta)$ were employed from the estimates of the real data sets available in Table 4, which were $(0.4975, 0.0003, 1, 0.5695)$ with $r = 0.0257$ for NEB and $(0.4972, 0.0003, 1.345, 1.4213)$ with $r = 0.0770$ and $(0.4854, 0.0003, 0.0182, 1.4930)$ with $r = 0.0357$ for NGB. Likewise to the first simulation study, for each value of $r$, 500 data sets were generated and then the performance of each method was evaluated. It was assumed that the total number of TICs was 200,000 and therefore, that there were true TICs of 5140, 15400, and 7140, respectively, when $r$ was 0.0257, 0.0770, and 0.0357. All the results are available in the Supplementary Information Figures S2–S3 and Tables S2–S3.

In case of parameter estimation (Supplementary Information Table S2), the NGB-FFT method performs slightly worse than others when the data were generated under an NEB model, while the NEB method is the worst method among others under an NGB model. As we observed in Table 1, NGB-FFT has the smaller bias in the estimates of $r$ than those of other methods, while both NGB-D1 and NGB-D2 achieve the smallest biases in the estimates of other parameters $(\mu, \sigma, \alpha, \beta)$.

Supplementary Information Table S3 and Figure S2 display AUC, F1 score, computation time, TPR, FPR, and PPV of four methods using additional simulation data. When the methods were applied to the data generated under a NEB model, there is no statistical difference among the four methods in terms of AUC and F1 score. The NGB-FFT shows the highest AUC and F1 score when it was applied to the data under an NGB model. The F1 score decreases as the value of $r$ decreases under an NGB model as observed under a NEB model in Table 2. Although we can see statistically significantly different performance on the four methods, their absolute difference is minimal except for the computation time in which NGB-D1 is much less expensive than other methods do. It should be noted that the AUC and F1 score values under the NEB model in Table S3 are seemingly identical across all four methods due to rounding errors. Without rounding numbers, those values are different from each method except for those of NGB-D1 and NGB-D2. In fact, the NGB-D1 and NGB-D2 methods yield identical AUC and F1 score values under the NEB model, as observed in the first NEB simulation study in Table 2.

Similar to Fig. 2, Supplementary Information Figure S3 displays the boxplots of F1 score by several different cut-off values ranged from 0.1 to 0.99. In case that the data were simulated under an NEB model, the highest mean of F1 score occurs when the cut-off value is 0.5 for all four methods. When an NGB model was used to simulate the data, the cut-off values with the highest mean of F1 score is less than or equal to 0.3. It appears that the data under an NGB model require a smaller cut-off value than those under an NEB model.

These additional simulation studies show that there is no considerable difference in the performance among the four methods under both NEB and NGB models. However, as shown before, NGB-FFT takes the most amount of computation time and NGB-D1 requires the least amount of computation time.
5.2. Two experimental GC × GC–MS data sets

Two real experimental data sets were used to further evaluate the performance of the developed peak detection algorithms. The chromatograms of these two data sets are depicted in Supplementary Information Figure S4. We applied the developed algorithms to the data in the green boxes of the figure in order to detect peaks. Table 4 shows the estimates of parameters obtained after peak detection using each of the four methods. We can see that the estimates of $\mu$ are nearly identical across all the methods and both NGB-D1 and NGB-D2 achieve the similar parameter estimation to each other. In particular, in case of Data 1, NGB-FFT has the largest value of $r$, while NEB has the smallest estimate of $r$. This demonstrates that NGB-FFT detects the largest amount of tentative TICs (i.e., signals) and NEB finds the smallest number of tentative TICs. On the other hand, based on the estimate of $r$, NGB-D1 and NGB-D2 detect much smaller number of tentative TICs than NEB does in case of Data 2.

The detected TICs (signals) can be seen in Supplementary Information Figures S5 and S6 for Data 1 and Data 2, respectively. As expected, the behaviors of NGB-D1 and NGB-D2 are similar to each other, while most TICs detected by NEB are overlapped with those by NGB-FFT. Interestingly, in the magnified region in Figure S5, some of small TICs are identified only by NGB-D1 and NGB-D2 in Data 1. On the contrary, most of the small TICs were identified only by NGB-FFT in Data 2 as shown in the magnified region in Figure S6, although some of them were also identified by NEB but not by either NGB-D1 or NGB-D2.

Using Poisson mixture model and the cut-off values of 0.3 and 0.4, we finalized the peak detection for both Data 1 and Data 2. Note that Poisson mixture model (PMM) was chosen based on Kim et al. (2014)'s work and, unlike mPeak, we used the estimate of $z$ (see Eqs. (14) and (17)) instead of the conditional Bayes factor (see Eq. (6)) to test if the tentative TIC (signal) is statistically true. The cut-off values of 0.3 and 0.4 for NEB and NGB were respectively selected based on the simulation studies in Section 5.1 (Fig. 2).

Fig. 3 displays the Venn diagrams of the detected peaks by each of the four methods. The top row depicts the peak detection results of Data 1 and those of Data 2 are depicted in the bottom row. The detected peaks before peak merging are shown in the left column and the right column shows those after peak merging. By all four methods, a total of 563 and 2392 unique peaks are identified on Data 1 and Data 2, respectively, before merging, while these numbers are decreased to 176 and 813 for Data 1 and Data 2, respectively, after peak merging. The numbers of peaks that are present in all methods are 170 (76) and 79 (31) before (after) peak merging on Data 1 and Data 2, respectively. That is, about 30% (43%) and 3% (4%) out of the total unique peaks are commonly detected by all methods on Data 1 and Data 2, respectively, before (after) peak merging. In particular, we can see that the performance of NGB-D1 and NGB-D2 is similar to each other, although NGB-D2 is more expensive in computation time than NGB-D1 (see Tables 3 and 4). NEB shares more common peaks with NGB-FFT than those with NGB-D1 or NGB-D2. Another point of interest is that NGB-FFT detects the most amount of unique peaks for both data sets (59 (17) for Data 1 and 1100 (370) for Data 2 before (after) peak merging), followed by NEB.

The chromatograms of detected peaks are shown in Figs. 4 and 5 (before peak merging) and Supplementary Information Figures S7 and S8 (after peak merging). As for Data 1 in Fig. 4, it appears that there is no apparent difference among four methods in terms of detected peak's locations and this becomes more evident in the chromatogram after peak merging in Figure S7, although their numbers of detected peaks are different from each other. However, in case of Data 2, the chromatograms of NEB and NGB-FFT are much different from those of NGB-D1 and NGB-D2, as shown in Fig. 5, although those of NEB and NGB-FFT are similar to each other. The most different areas are located in the left-most and the bottom-most chromatograms. In these areas, NEB and NGB-FFT detected much more peaks than NGB-D1 and NGB-D2 do. Indeed, these peaks correspond to the small peaks that were detected only by NGB-FFT or NEB or both, as shown in the magnified detected peaks of Supplementary Information Figure S6 (e.g., the peaks between 235 000 and 240 000).

It is worth mentioning that, due to technical reasons, during the early elution period, there are leaking chemicals that generate the straight vertical lines, in particular, in the very beginning of the chromatogram as well as there are the horizontal lines that are generated by several reagents which are not interesting to the study (e.g., the entire chromatograms of Supplementary Information Figure S4). Data 1 does not include these areas, while these areas are present in Data 2. Interestingly, in Data 2, NEB and NGB-FFT detect both TICs and peaks that are located in these areas, while NGB-D1 and NGB-D2 refrain from detecting both TICs and peaks from these areas. In fact, the manual investigation confirmed that the small TICs, which were detected only by NEB or NGB-FFT or both and are shown in the magnified detected TICs of Supplementary Information Figure S6, belong to these areas.

As mentioned before, the simulation study in Section 5.1 does not reflect the entire procedure of peak detection so that the outcome of the simulation is rather awkward to apply for real data. Thus, we rendered more implementations using real data to evaluate the performance of each method by (i) varying the cut-off values and (ii) comparing the number of 76 compound standards detected in Data 1.

Ten peak detection implementations were more carried out for each real data set using the 10 cut-off values ranging from 0.1 to 0.99 and the number of detected peaks (before and after peak merging), the chromatogram and the Venn diagrams were displayed in Fig. 6, Supplementary Information Figures S9–S11. Note that as for Data 2, the results of NGB-FFT with the cut-off value of 0.1 are not available due to its expensive and slow computation. As a result, when the cut-off value is 0.1, there is no information on the number of detected peaks in Fig. 6, the chromatogram in Supplementary Information Figure S10, and the Venn diagram in Supplementary Information Figure S11 for NGB-FFT in Data 2.
Fig. 3. The Venn diagrams of detected peaks by both NEB and NGB using two real experimental data. The top row depicts the results from Data 1 and those from Data 2 are depicted in the bottom row. The detected peaks before peak merging are shown in the left column and the right column shows the detected peaks after peak merging.

In Fig. 6, the middle row depicts the number of detected peaks after and before peak merging using Data 1. We can see that there is a sudden increase in the number of detected peaks around the cut-off value of 0.2 in case of NGB models, which the NEB model generally preserves the similar number of detected peaks across all the cut-off values. On the other hand, the bottom row of Fig. 6 shows that NGB-D1 and NGB-D2 seemingly do not depend on the cut-off value, while the number of detected peaks increases as the cut-off value decreases in case of NEB and NGB-FFT.

Supplementary Information Figures S9 and S10 display the chromatogram with the detected peaks by 10 cut-off values. Seemingly the cut-off value does not have a significant impact on the detected peak regions for both datasets. In particular, NEB and NGB-FFT keep in detecting the straight vertical and horizontal lines regardless of the cut-off values in Supplementary Information Figure S10. Furthermore, although the number of detected peaks is significantly affected by the cut-off values, the number of commonly detected peaks (i.e., across all the methods) is relatively stable across all the cut-off values, as shown in Supplementary Figure S11.

The performance of 76 compound standards identification was considered as shown in the top row of Fig. 6. In general, as the cut-off value decreases, the number of identified 76 compound standards increases because the number of detected peaks increases. However, the difference between the minimum and maximum number of identified 76 compound standards is less than or equal to 3 (before merging) and 4 (after merging) among the four methods.

The applications to real experimental data show that each method generally finds peaks in the similar chromatographic region, although the consensus analysis in Fig. 3 demonstrates that the detected peaks are influenced by the method. The additional implementations show that the number of detected peaks clearly depends on the cut-off value, as shown in Fig. 6. However, the cut-off value does not have a significant impact on both the detected chromatographic regions (Supplementary
Fig. 4. The chromatograms of detected peaks by both NEB and NGB before peak merging using the first real experimental data. The numbers in parentheses represent the number of detected peaks.

Information Figures S9 and S10) and the number of identified 76 compound standards (Fig. 6). In general, this experiment using real data suggests that the cut-off values between 0.2 and 0.5 achieve a reasonable performance in terms of the number of identified 76 compound standards, the number of detected peaks and computation expense. Among the four methods, NGB-FFT has better ability to detect the small TICs and peaks than other methods do, but at the same time has less stability to the innate technical noise. Taking everything into account, NGB-D1 and NGB-D2 perform reasonably the best peak detection procedures for both Data 1 and Data 2.

6. Concluding remarks

We generalized the previously developed peak detection algorithm msPeak using a gamma distribution. In particular, the Normal–Exponential–Bernoulli (NEB) model was generalized into a novel Normal–Gamma–Bernoulli (NGB) model for finding true TICs that include true signals. However, due to the complexity of the NGB model, there is no closed-form analytical solution. To overcome this difficulty, we introduced three approximation methods for parameter estimation of the NGB model using fast Fourier transformation and the first- and second-order delta methods and further developed the new peak detection algorithms, NGB–FFT, NGB–D1, and NGB–D2.

In addition, the developed algorithm uses the value of $z$ instead of the conditional Bayes factor to see if a TIC (signal) includes a true signal. The main benefit of the use of $z$ is to optimize the cut-off value in terms of prediction because the value of $z$ is bounded between 0 and 1. Using the simulated data, the optimal cut-off values are set to 0.3 for the NEB model.
Fig. 5. The chromatograms of detected peaks by both NEB and NGB before peak merging using the second real experimental data. The numbers in parentheses represent the number of detected peaks.

and 0.4 for the NGB model. The implementations using real data also suggest the similar cut-off values ranging from 0.2 to 0.5.

The developed algorithms were evaluated using both two sets of simulated data and two sets of experimental data. The simulation studies show that the NGB-based methods generally perform better than the NEB method does, and the NGB-D1 method is computationally less expensive than other methods do. In particular, on the NGB simulated data, although the absolute difference is minimal, the NGB-FFT method gives better performance in terms of AUC and F1 score. It further produces a higher number of true positives (i.e., higher TPRs), but gives a higher number of false positives (i.e., higher FPRs) compared to other methods. The NGB-D1 and NGB-D2 methods are not better than the NGB-FFT method in terms of AUC and F1 score, but their AUU and F1 score values are comparable to those of NGB-FFT and they yield a higher number of true discoveries (i.e., higher PPVs) than the NGB-FFT method does.

When the methods were applied to the real data, the NGB-FFT detects the most number of unique peaks, but this advantage causes it to detect uninteresting TICs and peaks. On the other hand, although NGB-D1 and NGB-D2 detect the less number of unique peaks than those of NEB and NGB-FFT, the real data analysis demonstrates that NGB-D1 and NGB-D2 are able to detect low abundance TICs and peaks as well as to avoid detecting uninteresting TICs and peaks.

In conclusion, the data analysis with the limited data suggests that the NGB-FFT method will be used to detect peaks in order to yield a higher number of true positives, but the NGB-D1 will be the best choice of the peak detection among them in terms of both computational expense and the number of true discoveries.
Fig. 6. The number of 76 compound standards and peaks detected by different cut-off values before and after peak merging using two real experimental data. The number of detected peaks with the cut-off value of 0.1 is not available for NGB-FFT when Data 2 is applied.

Acknowledgments

We appreciate the helpful comments of the anonymous reviewers and the editor for helpful suggestions. This work is supported by NSF grant DMS-1312603. The Biostatistics Core is supported, in part, by NIH Center grant P30 CA022453 to the Karmanos Cancer Institute at Wayne State University.

Appendix A. Supplementary data

Supplementary material related to this article can be found online at http://dx.doi.org/10.1016/j.csda.2016.07.015.

References


