

Supplementary Materials

1 Supporting Lemmas for the derivation of the parameter update equations

Notation for the Lemmas: We use θ as a placeholder for SN parameters and ζ as a placeholder for the complete model parameters; i.e., for the top score skew normal mixture model it is the triple $(\alpha, \theta_c, \theta_1)$ and for the top two score skew normal mixture model it is the quintuple $(\alpha, \beta, \theta_c, \theta_1, \theta_2)$. "Equality after removing additive terms constant w.r.t. a " is denoted by $\stackrel{a}{=}$ and "equality after removing multiplicative terms constant w.r.t. a " is denoted by $\stackrel{a}{\propto}$. Abusing the notation, we use small letters to denote the random variable as well as its realization. $\text{HN}(0, 1)$ denotes the standard normal distribution truncated below 0. $\text{N}(0, 1)$ denotes the standard normal distribution. $\text{TN}(\eta, \sigma^2, \mathbb{R}_+)$ is a truncated normal distribution truncated below 0 with location and (squared) scale parameters η and σ^2 . Let

$$q(x, t, \tau, \theta) = \log \Gamma + \frac{(x - \mu)^2 - 2(x - \mu)\Delta t + (\Delta^2 + \Gamma)\tau}{\Gamma}$$

And for $T_x \sim \text{TN}(\delta/\omega(x - \mu), 1 - \delta^2, \mathbb{R}_+)$,

$$\begin{aligned} v(x, \theta) &= \mathbb{E}[T_x] \\ w(x, \theta) &= \mathbb{E}[T_x^2] \end{aligned}$$

$v(x, \theta)$ and $w(x, \theta)$ can be computed using Lemma 1. When $x \rightarrow -\infty$ the ratio $\phi(x)/\Phi(x)$ can be approximated as $|x|$.

Lemma 1. For $T \sim \text{TN}(\eta, \sigma^2, \mathbb{R}_+)$ (truncated normal distribution, truncated below 0),

$$\begin{aligned} E[T] &= s_1(\eta, \sigma^2) = \eta + \sigma \frac{\phi(\frac{\eta}{\sigma})}{\Phi(\frac{\eta}{\sigma})}, \\ E[T^2] &= s_2(\eta, \sigma^2) = \eta^2 + \sigma^2 + \eta\sigma \frac{\phi(\frac{\eta}{\sigma})}{\Phi(\frac{\eta}{\sigma})}, \end{aligned}$$

where ϕ and Φ are pdf and cdf of a standard normal distribution, respectively.

Proof. (Johnson and Kotz, 1970)(pp. 156-158) gives the following result for truncated normal distributions. For $X \sim \text{TN}(\eta, \sigma^2, (a_{1,2}))$,

$$\begin{aligned} E[X] &= \eta - \sigma \frac{\phi(\alpha_2) - \phi(\alpha_1)}{\Phi(\alpha_2) - \Phi(\alpha_1)}, \\ E[X^2] &= \eta^2 + \sigma^2 - \sigma^2 \frac{\alpha_2 \phi(\alpha_2) - \alpha_1 \phi(\alpha_1)}{\Phi(\alpha_2) - \Phi(\alpha_1)} - 2\eta\sigma \frac{\phi(\alpha_2) - \phi(\alpha_1)}{\Phi(\alpha_2) - \Phi(\alpha_1)}, \end{aligned}$$

$$\text{where } \alpha_i = \frac{\alpha_i - \eta}{\sigma}$$

It follows that

$$\begin{aligned} E[T] &= \eta - \sigma \frac{0 - \phi(-\frac{\eta}{\sigma})}{1 - \Phi(-\frac{\eta}{\sigma})} \\ &= \eta + \sigma \frac{\phi(\frac{\eta}{\sigma})}{\Phi(\frac{\eta}{\sigma})}. \end{aligned}$$

and

$$\begin{aligned} E[T^2] &= \eta^2 + \sigma^2 - \sigma^2 \frac{0 - (-\frac{\eta}{\sigma})\phi(-\frac{\eta}{\sigma})}{1 - \Phi(-\frac{\eta}{\sigma})} - 2\eta\sigma \frac{0 - \phi(-\frac{\eta}{\sigma})}{1 - \Phi(-\frac{\eta}{\sigma})} \\ &= \eta^2 + \sigma^2 - \sigma\eta \frac{\phi(\frac{\eta}{\sigma})}{\Phi(\frac{\eta}{\sigma})} + 2\eta\sigma \frac{\phi(\frac{\eta}{\sigma})}{\Phi(\frac{\eta}{\sigma})} \\ &= \eta^2 + \sigma^2 + \eta\sigma \frac{\phi(\frac{\eta}{\sigma})}{\Phi(\frac{\eta}{\sigma})} \end{aligned}$$

Lemma 2. Let $X \sim \text{SN}(\theta)$, $T \sim \text{HN}(0, 1)$ and $U \sim \text{N}(0, 1)$ be related by the probabilistic representation of a SN random variable as $X \stackrel{d}{=} \mu + \Delta T + \Gamma^{1/2}U$. Then for the joint density function $p(x, t; \theta)$,

$$\mathbb{E}[\log p(X, T; \theta) | X = x; \bar{\theta}] \stackrel{\theta}{=} -\frac{1}{2}q(x, v(x, \bar{\theta}), w(x, \bar{\theta}), \theta),$$

where the expectation is taken w.r.t. the distribution of T given $X = x$ with $\bar{\theta}$ used as the parameter for the conditional distribution.

Proof.

Distribution of T given $X = x$: From the probabilistic representation of X , the distribution of X given $T = t$ is a univariate Normal, $N(\mu + \Delta t, \Gamma)$. Now, the joint distribution of X and T is given by

$$\begin{aligned} p(x, t) &= p(x|t)p(t) \\ &= \frac{1}{\sqrt{2\pi\Gamma}} \exp\left(\frac{-((x - \mu) - \Delta t)^2}{2\Gamma}\right) \cdot \sqrt{\frac{2}{\pi}} \exp\left(-\frac{t^2}{2}\right) I(t > 0) \\ &= \frac{1}{\pi\sqrt{\Gamma}} \exp\left(\frac{-((x - \mu)^2 - 2(x - \mu)\Delta t + (\Delta^2 + \Gamma_1)t^2)}{2\Gamma}\right) I(t > 0) \end{aligned} \quad (5)$$

Let $\eta(x, \theta) = \delta/\omega(x - \mu)$. The conditional distribution of T given $X = x$ can be derived as

$$\begin{aligned} p(t|x) &\stackrel{t}{\propto} p(x, t) \\ &\stackrel{t}{\propto} \exp\left(\frac{-((\Delta^2 + \Gamma)t^2 - 2(x - \mu)\Delta t)}{2\Gamma}\right) I(t > 0) \quad (\text{from Equation 5}) \\ &\stackrel{t}{\propto} \exp\left(\frac{-(t^2 - 2t\eta(x, \theta))}{2(1 - \delta^2)}\right) I(t > 0) \quad (\text{because } \frac{\Gamma}{\Gamma + \Delta^2} = 1 - \delta^2 \text{ and } \frac{\Delta}{\Gamma + \Delta^2} = \frac{\delta}{\omega}) \\ &\stackrel{t}{\propto} \exp\left(\frac{-(t^2 - 2t\eta(x, \theta) + (\eta(x, \theta))^2)}{2(1 - \delta^2)}\right) I(t > 0) \\ &\stackrel{t}{\propto} \exp\left(\frac{-(t - \eta(x, \theta))^2}{2(1 - \delta^2)}\right) I(t > 0). \end{aligned} \quad (6)$$

Thus $t|x \sim TN(\eta(x, \theta), 1 - \delta^2, \mathbb{R}_+)$

Expectation of the log of the joint distribution: From Equation 5 the expression for $p(x, t)$ can be written as

$$p(x, t) = \frac{1}{\pi} \exp\left(-\frac{q(x, t, t^2, \theta)}{2}\right) I(t > 0).$$

Taking log and removing terms constant in θ ,

$$\log p(x, t) \stackrel{\theta}{=} -\frac{q(x, t, t^2, \theta)}{2}$$

Treating $\log p(x, t)$ as a function of random variables X and T and taking its conditional expectation given $X = x$ and $\bar{\theta}$,

$$\mathbb{E}[\log p(X, T; \theta)|X = x; \bar{\theta}] \stackrel{\theta}{=} -\frac{1}{2}q(x, v(x, \bar{\theta}), w(x, \bar{\theta}), \theta),$$

since $v(x, \theta) = \mathbb{E}[T|X = x]$ and $w(x, \theta) = \mathbb{E}[T^2|X = x]$ by definition.

Lemma 3. Let X_0 and X_1 be two continuous random variables. Let X be a mixture random variable defined as $X = ZX_1 + (1 - Z)X_0$, where $Z \sim \text{Bernoulli}(\alpha)$ for $\alpha \in [0, 1]$. Then For any random variable T , not necessarily independent of X and for any real function h , the following statements are true

$$\begin{aligned} \mathbb{E}[Zh(T)|X] &= p(Z = 1|X) \cdot \mathbb{E}[h(T)|X, Z = 1] \\ \mathbb{E}[(1 - Z)h(T)|X] &= p(Z = 0|X) \cdot \mathbb{E}[h(T)|X, Z = 0] \end{aligned}$$

Proof.

By the law of iterated expectation,

$$\begin{aligned} \mathbb{E}[Zh(T)|X] &= \mathbb{E}[\mathbb{E}[Zh(T)|X, Z]|X] \\ &= \mathbb{E}[Z\mathbb{E}[h(T)|X, Z]|X] \\ &= p(Z = 1|X) \cdot 1 \cdot \mathbb{E}[h(T)|X, Z = 1] + p(Z = 0|X) \cdot 0 \cdot \mathbb{E}[h(T)|X, Z = 0] \\ &= p(Z = 1|X)\mathbb{E}[h(T)|X, Z = 1] \end{aligned}$$

To prove the expression for $\mathbb{E}[(1 - Z)h(T)|X]$, use the same argument with $1 - Z$, instead of Z .

Lemma 4. Let S_1 be the random variable corresponding to the top score under the 1SMix model; i.e., $S_1 = ZX_c + (1 - Z)X_1$, where $X_c \sim \text{SN}(\theta_c)$, $X_1 \sim \text{SN}(\theta_1)$ and $Z \sim \text{Bernoulli}(\alpha)$. Let $T_c, T_1 \sim \text{HN}(0, 1)$ and $U_c, U_1 \sim \text{N}(0, 1)$ be related to X_c and X_1 respectively by the probabilistic representation of SN random variable as $X_c \stackrel{d}{=} \mu_c + \Delta_c T_c + \Gamma_c^{1/2} U_c$ and $X_1 \stackrel{d}{=} \mu_1 + \Delta_1 T_1 + \Gamma_1^{1/2} U_1$. Further, let $T = ZT_c + (1 - Z)T_1$. Then for the joint density of S_1, T, Z given by $p(s_1, t, z; \zeta)$,

$$\begin{aligned} & \mathbb{E}[\log p(S_1, T, Z; \zeta) | S_1 = s_1; \bar{\zeta}] \\ & \stackrel{\zeta}{=} \bar{p}_c(s_1) \left(\log \alpha - \frac{1}{2} q(s_1, v(s_1, \bar{\theta}_c), w(s_1, \bar{\theta}_c), \theta_c) \right) \\ & \quad + \bar{p}_1(s_1) \left(\log(1 - \alpha) - \frac{1}{2} q(s_1, v(s_1, \bar{\theta}_1), w(s_1, \bar{\theta}_1), \theta_1) \right) \end{aligned}$$

where ζ serves as placeholder for $(\alpha, \theta_c, \theta_1)$; the expectation is taken w.r.t. the distribution of Z, T given $S_1 = s_1$ with $\bar{\zeta}$ used as the parameter for the conditional distribution;

$$\begin{aligned} \bar{p}_c(s_1) &= p(Z = 1 | S_1 = s_1, \bar{\zeta}) \\ \bar{p}_1(s_1) &= p(Z = 0 | S_1 = s_1, \bar{\zeta}). \end{aligned}$$

Proof. The joint distribution of S_1, T, Z can be written as

$$\begin{aligned} p(s_1, t, z) &= p(s_1, t, Z = 1)^z p(s_1, t, Z = 0)^{1-z} \\ &= [p(s_1, t | Z = 1)p(Z = 1)]^z [p(s_1, t | Z = 0)p(Z = 0)]^{1-z} \\ &= [p(X_c = s_1, T_c = t)\alpha]^z [p(X_1 = s_1, T_1 = t)(1 - \alpha)]^{1-z} \end{aligned}$$

Now treating $p(s_1, t, z; \zeta)$ as function of random variables S_1, T and Z and taking expectation w.r.t. the distribution of T and Z given $S_1 = s_1$ and $\bar{\zeta}$.

$$\begin{aligned} & \mathbb{E}[\log p(S_1, T, Z; \zeta) | S_1 = s_1; \bar{\zeta}] \\ &= \mathbb{E}[Z(\log \alpha + \log p(X_c = s_1, T_c = t; \theta_c)) | S_1 = s_1; \bar{\zeta}] + \mathbb{E}[(1 - Z)(\log(1 - \alpha) + \log p(X_1 = s_1, T_1 = t; \theta_1)) | S_1 = s_1; \bar{\zeta}] \\ &\stackrel{\zeta}{=} \bar{p}_c(s_1) \left(\log \alpha - \frac{1}{2} q(s_1, v(s_1, \bar{\theta}_c), w(s_1, \bar{\theta}_c), \theta_c) \right) + \bar{p}_1(s_1) \left(\log(1 - \alpha) - \frac{1}{2} q(s_1, v(s_1, \bar{\theta}_1), w(s_1, \bar{\theta}_1), \theta_1) \right), \end{aligned}$$

where the last line is obtained by first applying Lemma 3 and then Lemma 2.

Lemma 5. Let S_2 be the random variable corresponding to the second score distributed as $S_2 \stackrel{d}{=} Y_1 X_1 + Y_2 X_2 + Y_c X_c$, where $X_c \sim \text{SN}(\theta_c)$, $X_1 \sim \text{SN}(\theta_1)$, $X_2 \sim \text{SN}(\theta_2)$, $Y = [Y_1, Y_2, Y_c] \sim \text{Categorical}([\alpha, 1 - \alpha - \beta, \beta])$, $\alpha, \beta \in [0, 1]$ and $\alpha + \beta < 1$. Let $T_c, T_1, T_2 \sim \text{HN}(0, 1)$ and $U_c, U_1, U_2 \sim \text{N}(0, 1)$ be related to X_c , X_1 and X_2 respectively by the probabilistic representation of SN random variable as $X_c \stackrel{d}{=} \mu_c + \Delta_c T_c + \Gamma_c^{1/2} U_c$, $X_1 \stackrel{d}{=} \mu_1 + \Delta_1 T_1 + \Gamma_1^{1/2} U_1$ and $X_2 \stackrel{d}{=} \mu_2 + \Delta_2 T_2 + \Gamma_2^{1/2} U_2$. Further, let $T \stackrel{d}{=} Y_1 T_1 + Y_2 T_2 + Y_c T_c$. Then for the joint density of S_2, T, Y given by $p(s_2, t, y; \zeta)$,

$$\begin{aligned} \mathbb{E}[\log p(S_2, T, Y; \zeta) | S_2 = s_2; \bar{\zeta}] &\stackrel{\zeta}{=} \bar{r}_1(s_2) \left(\log \alpha - \frac{1}{2} q(s_2, v(s_2, \bar{\theta}_1), w(s_2, \bar{\theta}_1), \theta_1) \right) \\ &\quad + \bar{r}_2(s_2) \left(\log(1 - \alpha - \beta) - \frac{1}{2} q(s_2, v(s_2, \bar{\theta}_2), w(s_2, \bar{\theta}_2), \theta_2) \right) \\ &\quad + \bar{r}_c(s_2) \left(\log \beta - \frac{1}{2} q(s_2, v(s_2, \bar{\theta}_c), w(s_2, \bar{\theta}_c), \theta_c) \right), \end{aligned}$$

where ζ serves as placeholder for $(\alpha, \beta, \theta_c, \theta_1, \theta_2)$; the expectation is taken w.r.t. the distribution of T, Y given $S_2 = s_2$ with $\bar{\zeta}$ used as the parameter for the conditional distribution;

$$\begin{aligned} \bar{r}_1(s_2) &= p(Y_1 = 1 | S_2 = s_2, \bar{\zeta}) \\ \bar{r}_2(s_2) &= p(Y_2 = 1 | S_2 = s_2, \bar{\zeta}) \\ \bar{r}_c(s_2) &= p(Y_c = 1 | S_2 = s_2, \bar{\zeta}). \end{aligned}$$

Proof. The joint distribution of S_2, T, Y can be written as

$$\begin{aligned} p(s_2, t, y) &= p(s_2, t, Y_1 = 1)^{y_1} p(s_2, t, Y_2 = 1)^{y_2} p(s_2, t, Y_c = 1)^{y_c} \\ &= [p(s_2, t|Y_1 = 1)p(Y_1 = 1)]^{y_1} [p(s_2, t|Y_2 = 1)p(Y_2 = 1)]^{y_2} [p(s_2, t|Y_c = 1)p(Y_c = 1)]^{y_c} \\ &= [p(X_1 = s_2, T_1 = t)\alpha]^{y_1} [p(X_2 = s_2, T_2 = t)(1 - \alpha - \beta)]^{y_2} [p(X_c = s_2, T_c = t)\beta]^{y_c} \end{aligned}$$

Now treating $p(s_2, t, y; \zeta)$ as function of random variables s_2, T and Y and taking expectation w.r.t. the distribution of T, Y given $S_2 = s_2$ and $\bar{\zeta}$.

$$\begin{aligned} \mathbb{E}[\log p(S_2, T, Y; \zeta)|S_2 = s_2; \bar{\zeta}] &= \mathbb{E}[Y_1(\log \alpha + \log p(X_1 = s_2, T_1 = t; \theta_1))|S_2 = s_2; \bar{\zeta}] \\ &\quad + \mathbb{E}[Y_2(\log(1 - \alpha - \beta) + \log p(X_2 = s_2, T_2 = t; \theta_2))|S_2 = s_2; \bar{\zeta}] \\ &\quad + \mathbb{E}[Y_c(\log \beta + \log p(X_c = s_2, T_c = t; \theta_c))|S_2 = s_2; \bar{\zeta}] \end{aligned} \tag{7}$$

Now, S_2 can be expressed in the form required by Lemma 3 with Y_i ($i = 1, 2, c$) replacing Z (in Lemma 3) as follows

$$\begin{aligned} S_2 &\stackrel{d}{=} Y_1 X_1 + (1 - Y_1)(Y_2 X_2 + Y_c X_c), \\ S_2 &\stackrel{d}{=} Y_2 X_2 + (1 - Y_2)(Y_1 X_1 + Y_c X_c), \\ S_2 &\stackrel{d}{=} Y_c X_c + (1 - Y_c)(Y_1 X_1 + Y_2 X_2). \end{aligned}$$

First applying Lemma 3 to Equation 7 and then applying Lemma 2 gives

$$\begin{aligned} \mathbb{E}[\log p(S_2, T, Y; \zeta)|S_2 = s_2; \bar{\zeta}] &\stackrel{\zeta}{=} \bar{r}_1(s_2) \left(\log \alpha - \frac{1}{2} q(s_2, v(s_2, \bar{\theta}_1), w(s_2, \bar{\theta}_1), \theta_1) \right) \\ &\quad + \bar{r}_2(s_2) \left(\log(1 - \alpha - \beta) - \frac{1}{2} q(s_2, v(s_2, \bar{\theta}_2), w(s_2, \bar{\theta}_2), \theta_2) \right) \\ &\quad + \bar{r}_c(s_2) \left(\log \beta - \frac{1}{2} q(s_2, v(s_2, \bar{\theta}_c), w(s_2, \bar{\theta}_c), \theta_c) \right). \end{aligned}$$

2 Derivation of the parameter update equations

2.1 Partial derivatives of the q -function

$$\begin{aligned} \frac{\partial}{\partial \mu} q(x, t, \tau, \theta) &= \frac{2\mu - 2(x - \Delta t)}{\Gamma} \\ \frac{\partial}{\partial \Delta} q(x, t, \tau, \theta) &= \frac{2\Delta\tau - 2(x - \mu)t}{\Gamma} \\ \frac{\partial}{\partial \Gamma} q(x, t, \tau, \theta) &= \frac{1}{\Gamma} - \frac{(x - \mu)^2 - 2\Delta(x - \mu)t + \Delta^2\tau}{\Gamma^2} \end{aligned}$$

2.2 1SMix

The pdf of S_1 (top score) under the 1SMix model is given by

$$f_1(x) = \alpha f_{SN}(x; \theta_c) + (1 - \alpha) f_{SN}(x; \theta_1).$$

The log-likelihood of \mathbb{S}_1 (containing all top scores) is given by

$$\mathcal{L}(\mathbb{S}_1; \zeta) = \frac{1}{|\mathbb{S}_1|} \sum_{s_1 \in \mathbb{S}_1} \log f_1(s_1).$$

We obtain maximum likelihood estimate of ζ by indirectly maximizing the log-likelihood using the Expectation Maximization (EM) approach.

To derive the EM algorithm, notice that the 1SMix model for S_1 can be alternatively expressed as that in Lemma 4. Let $\mathbb{S}_1^{\mathcal{E}}$ be a partially observed sample containing triples of the form (s_1, t, z) for all $s_1 \in \mathbb{S}_1$, where t and z are the unobserved values for T and Z corresponding to S_1 . We will use

$\mathbb{S}_1^{\mathcal{C}}$ as the complete data in the Expectation Maximization (EM) framework. The so called “Q-function” can be derived as

$$\begin{aligned}
Q(\zeta|\bar{\zeta}) &= \mathbb{E} \left[\log p(\mathbb{S}_1^{\mathcal{C}}; \zeta) \middle| \mathbb{S}_1, \bar{\zeta} \right] \\
&= \mathbb{E} \left[\log \prod_{(s_1, t, z) \in \mathbb{S}_1^{\mathcal{C}}} p(s_1, t, z; \zeta) \middle| \mathbb{S}_1, \bar{\zeta} \right] \\
&= \mathbb{E} \left[\sum_{(s_1, t, z) \in \mathbb{S}_1^{\mathcal{C}}} \log p(s_1, t, z; \zeta) \middle| \mathbb{S}_1, \bar{\zeta} \right] \\
&= \sum_{s_1 \in \mathbb{S}_1} \mathbb{E} [\log p(s_1, t, z; \zeta) | S_1 = s_1, \bar{\zeta}] \\
&\stackrel{c}{=} \sum_{s_1 \in \mathbb{S}_1} \bar{p}_c(s_1) \left(\log \alpha - \frac{1}{2} q(s_1, v(s_1, \bar{\theta}_c), w(s_1, \bar{\theta}_c), \theta_c) \right) + \sum_{s_1 \in \mathbb{S}_1} \bar{p}_1(s_1) \left(\log(1 - \alpha) - \frac{1}{2} q(s_1, v(s_1, \bar{\theta}_1), w(s_1, \bar{\theta}_1), \theta_1) \right)
\end{aligned}$$

The update equations are obtained by taking the partial derivatives of $Q(\zeta|\bar{\zeta})$ w.r.t. $\alpha, \mu_c, \mu_1, \Delta_c, \Delta_1, \Gamma_c$ and Γ_1 and equating them to 0.

2.3 2SMix

The pdfs of S_1 (top score) and S_2 (second score) under the 2SMix model are given by

$$\begin{aligned}
f_1(x) &= \alpha f_{SN}(x; \theta_c) + (1 - \alpha) f_{SN}(x; \theta_1), \\
f_2(x) &= \alpha f_{SN}(x; \theta_1) + (1 - \alpha - \beta) f_{SN}(x; \theta_2) + \beta f_{SN}(x; \theta_c),
\end{aligned}$$

respectively. The combined log-likelihood of \mathbb{S}_1 (containing all top scores) and \mathbb{S}_2 (containing all second scores) is given by

$$\mathcal{L}(\mathbb{S}_1, \mathbb{S}_2; \zeta) = \frac{1}{|\mathbb{S}_1|} \sum_{s_1 \in \mathbb{S}_1} \log f_1(s_1) + \frac{1}{|\mathbb{S}_2|} \sum_{s_2 \in \mathbb{S}_2} \log f_2(s_2).$$

We obtain maximum likelihood estimate of ζ by indirectly maximizing the log-likelihood using the Expectation Maximization (EM) approach. To derive the EM algorithm, notice that the 2SMix model for S_1 and S_2 can be alternatively expressed as that in Lemma 4 and Lemma 5, respectively. The parameters $\alpha, \theta_c, \theta_1$ are shared between the models for S_1 and S_2 . Let $\mathbb{S}_1^{\mathcal{C}}$ be a partially observed sample containing triples of the form (s_1, t, z) for all $s_1 \in \mathbb{S}_1$, where t and z are the unobserved values for T and Z corresponding to S_1 according to Lemma 4. Let $\mathbb{S}_2^{\mathcal{C}}$ be a partially observed sample containing triples of the form (s_2, t, y) for all $s_2 \in \mathbb{S}_2$, where t and y are the unobserved values for T and Y corresponding to S_2 according to Lemma 5. We will use $\mathbb{S}_1^{\mathcal{C}}$ and $\mathbb{S}_2^{\mathcal{C}}$ as the complete data in the Expectation Maximization (EM) framework. The so called “Q-function” can be derived as

$$\begin{aligned}
Q(\zeta|\bar{\zeta}) &= \mathbb{E} \left[\log p(\mathbb{S}_1^{\mathcal{C}}, \mathbb{S}_2^{\mathcal{C}}; \zeta) \middle| \mathbb{S}_1, \mathbb{S}_2, \bar{\zeta} \right] \\
&= \mathbb{E} \left[\log p(\mathbb{S}_1^{\mathcal{C}}; \zeta) \middle| \mathbb{S}_1, \bar{\zeta} \right] + \mathbb{E} \left[\log p(\mathbb{S}_2^{\mathcal{C}}; \zeta) \middle| \mathbb{S}_2, \bar{\zeta} \right] \\
&= \mathbb{E} \left[\log \prod_{(s_1, t, z) \in \mathbb{S}_1^{\mathcal{C}}} p(s_1, t, z; \zeta) \middle| \mathbb{S}_1, \bar{\zeta} \right] + \mathbb{E} \left[\log \prod_{(s_2, t, y) \in \mathbb{S}_2^{\mathcal{C}}} p(s_2, t, y; \zeta) \middle| \mathbb{S}_2, \bar{\zeta} \right] \\
&= \mathbb{E} \left[\sum_{(s_1, t, z) \in \mathbb{S}_1^{\mathcal{C}}} \log p(s_1, t, z; \zeta) \middle| \mathbb{S}_1, \bar{\zeta} \right] + \mathbb{E} \left[\sum_{(s_2, t, y) \in \mathbb{S}_2^{\mathcal{C}}} \log p(s_2, t, y; \zeta) \middle| \mathbb{S}_2, \bar{\zeta} \right] \\
&= \sum_{s_1 \in \mathbb{S}_1} \mathbb{E} [\log p(s_1, t, z; \zeta) | S_1 = s_1, \bar{\zeta}] + \sum_{s_2 \in \mathbb{S}_2} \mathbb{E} [\log p(s_2, t, y; \zeta) | S_2 = s_2, \bar{\zeta}] \\
&\stackrel{c}{=} \sum_{s_1 \in \mathbb{S}_1} \bar{p}_c(s_1) \left(\log \alpha - \frac{1}{2} q(s_1, v(s_1, \bar{\theta}_c), w(s_1, \bar{\theta}_c), \theta_c) \right) + \sum_{s_1 \in \mathbb{S}_1} \bar{p}_1(s_1) \left(\log(1 - \alpha) - \frac{1}{2} q(s_1, v(s_1, \bar{\theta}_1), w(s_1, \bar{\theta}_1), \theta_1) \right) \\
&\quad + \sum_{s_2 \in \mathbb{S}_2} \bar{r}_1(s_2) \left(\log \alpha - \frac{1}{2} q(s_2, v(s_2, \bar{\theta}_1), w(s_2, \bar{\theta}_1), \theta_1) \right) + \sum_{s_2 \in \mathbb{S}_2} \bar{r}_2(s_2) \left(\log(1 - \alpha - \beta) - \frac{1}{2} q(s_2, v(s_2, \bar{\theta}_2), w(s_2, \bar{\theta}_2), \theta_2) \right) \\
&\quad + \sum_{s_2 \in \mathbb{S}_2} \bar{r}_c(s_2) \left(\log \beta - \frac{1}{2} q(s_2, v(s_2, \bar{\theta}_c), w(s_2, \bar{\theta}_c), \theta_c) \right),
\end{aligned}$$

The update equations are obtained by taking the partial derivatives of $Q(\zeta|\bar{\zeta})$ w.r.t. $\alpha, \beta, \mu_c, \mu_1, \mu_2, \Delta_c, \Delta_1, \Delta_2, \Gamma_c, \Gamma_1$ and Γ_2 and equating them to 0. To obtain a simpler update equation for α , we use the reparametrization: $\beta = (1 - \alpha)c_1$ and $1 - \alpha - \beta = (1 - \alpha)c_2$, where $c_1, c_2 \in [0, 1]$ and $c_1 + c_2 = 1$. Similarly, to obtain a simpler update equation for β we use the reparametrization: $\alpha = (1 - \beta)c_3$ and $1 - \alpha - \beta = (1 - \beta)c_4$, where $c_3, c_4 \in [0, 1]$ and $c_3 + c_4 = 1$.

3 Method of moments estimator for SN parameters

Given a sample \mathbb{S} from $\text{SN}(\mu, \omega, \lambda)$, the method of moment estimator is given by the following equations.

SN Method of Moments (Arnold *et al.*, 1993):

$$\hat{\delta} = \text{sgn}(m_3) \left(a_1^2 + m_2(b_1/|m_3|)^{2/3} \right)^{-1/2},$$

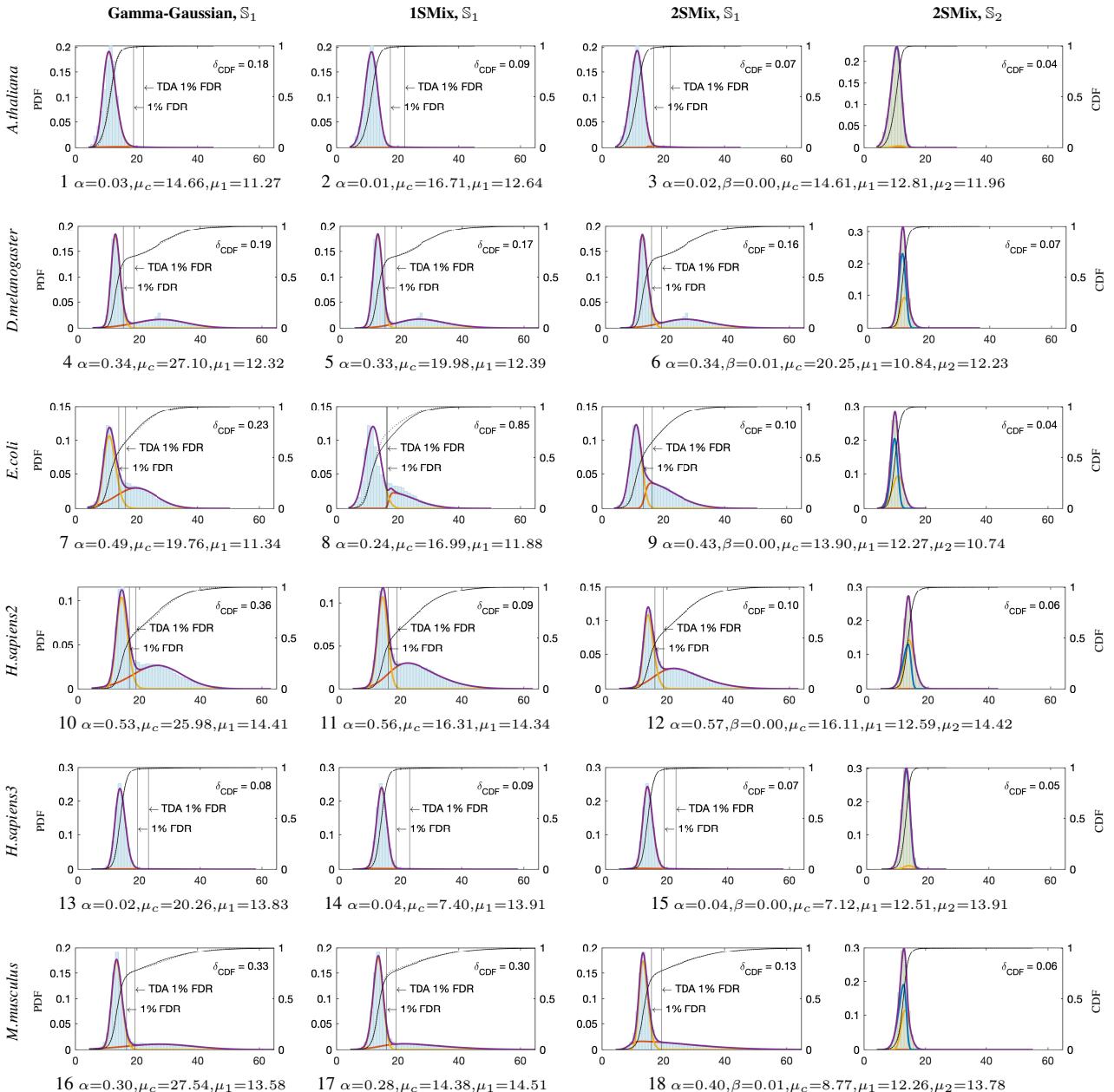
$$\hat{\omega}^2 = \frac{m_2}{1 - a_1^2 \hat{\delta}^2},$$

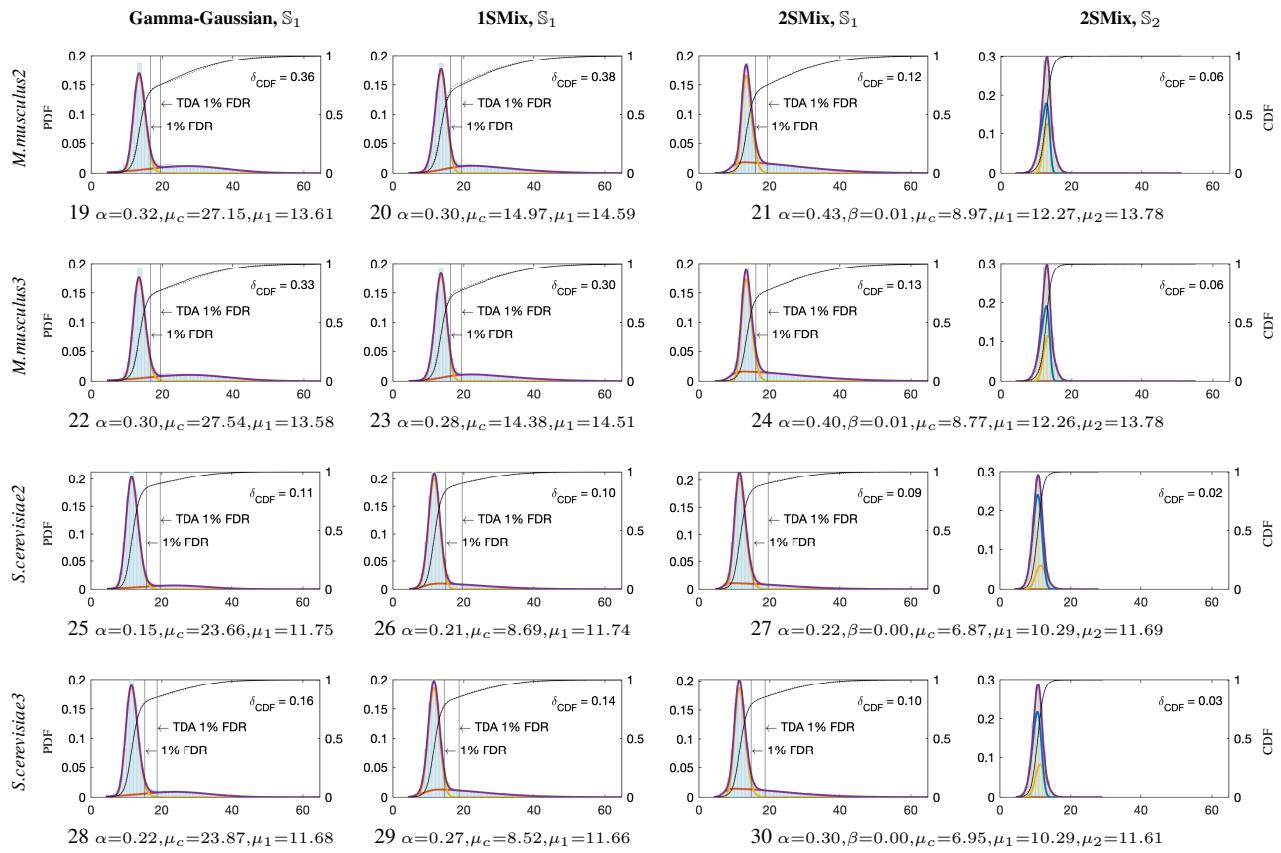
$$\hat{\mu} = m_1 - a_1 \hat{\delta} \hat{\omega},$$

where $a_1 = \sqrt{2/\pi}$, $b_1 = (4/\pi - 1)a_1$, $m_1 = 1/|\mathbb{S}| \sum_{s \in \mathbb{S}} s$, $m_2 = 1/(|\mathbb{S}| - 1) \sum_{s \in \mathbb{S}} (s - m_1)^2$ and $m_3 = 1/(|\mathbb{S}| - 1) \sum_{s \in \mathbb{S}} (s - m_1)^3$.

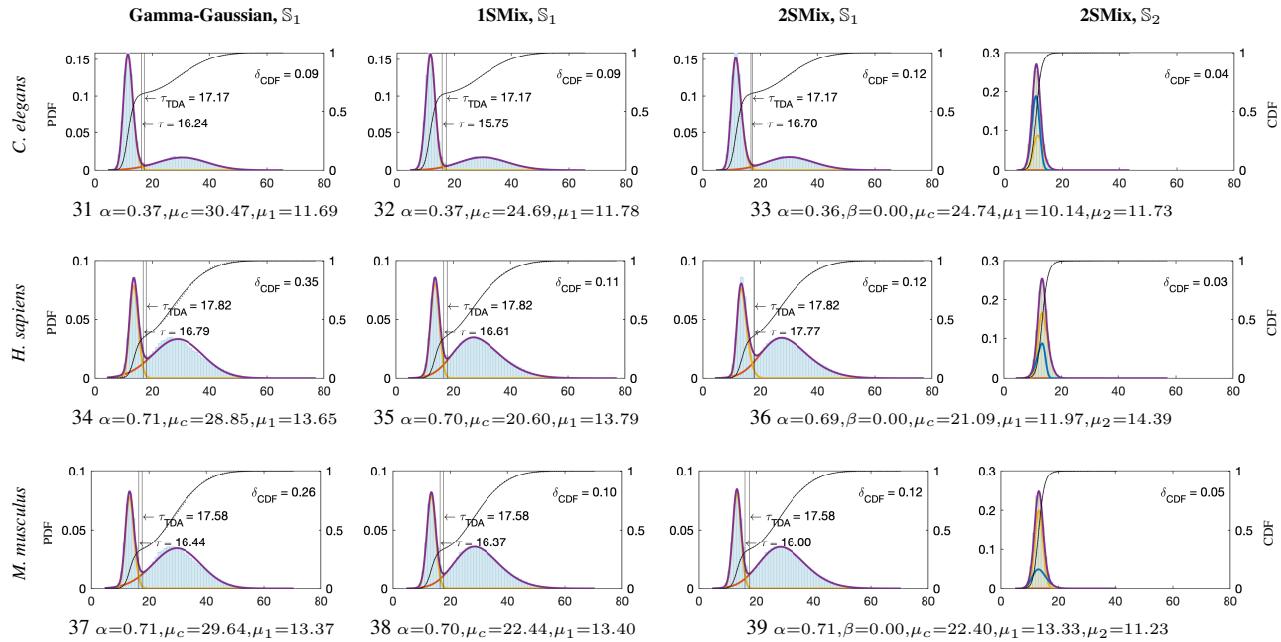
4 Supporting Results

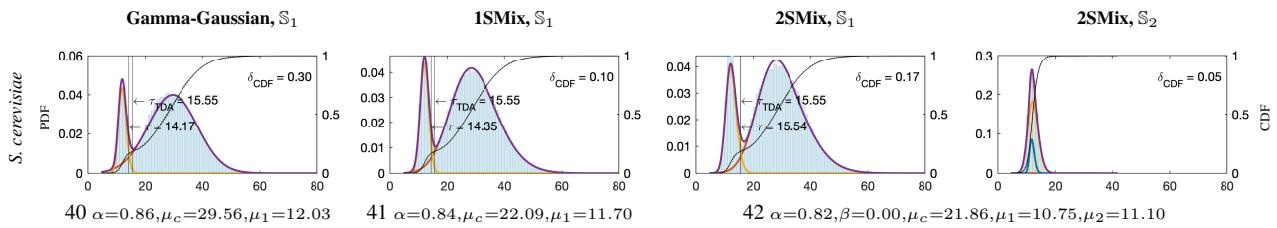
4.1 PRIDE Datasets



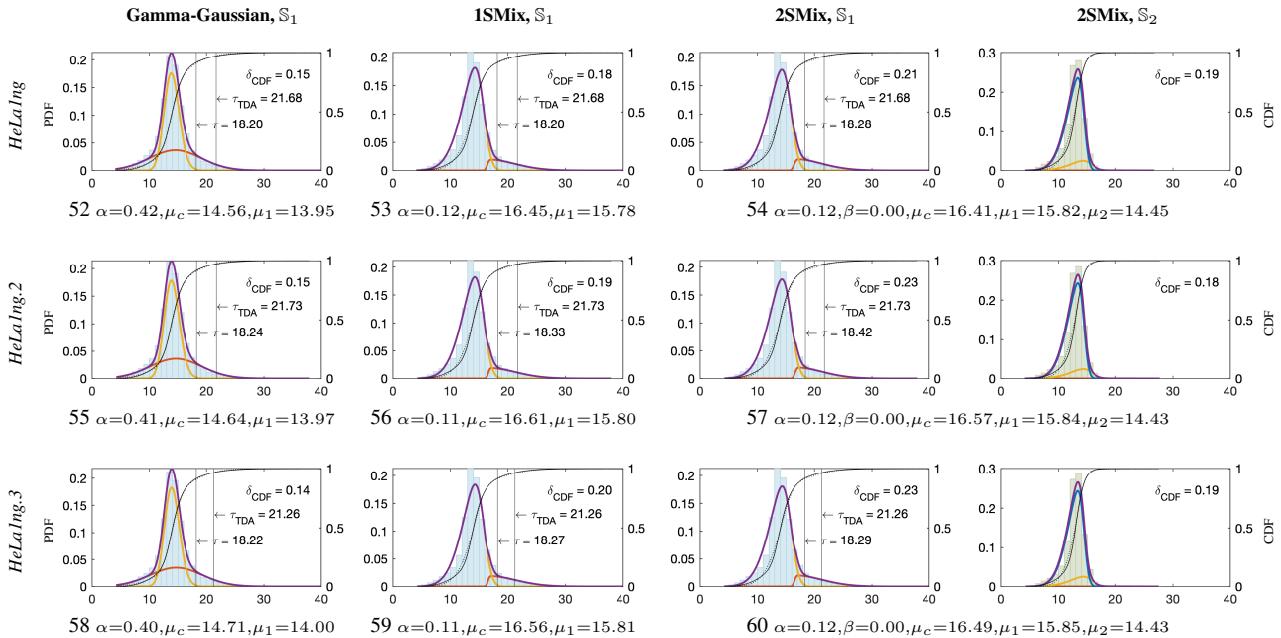
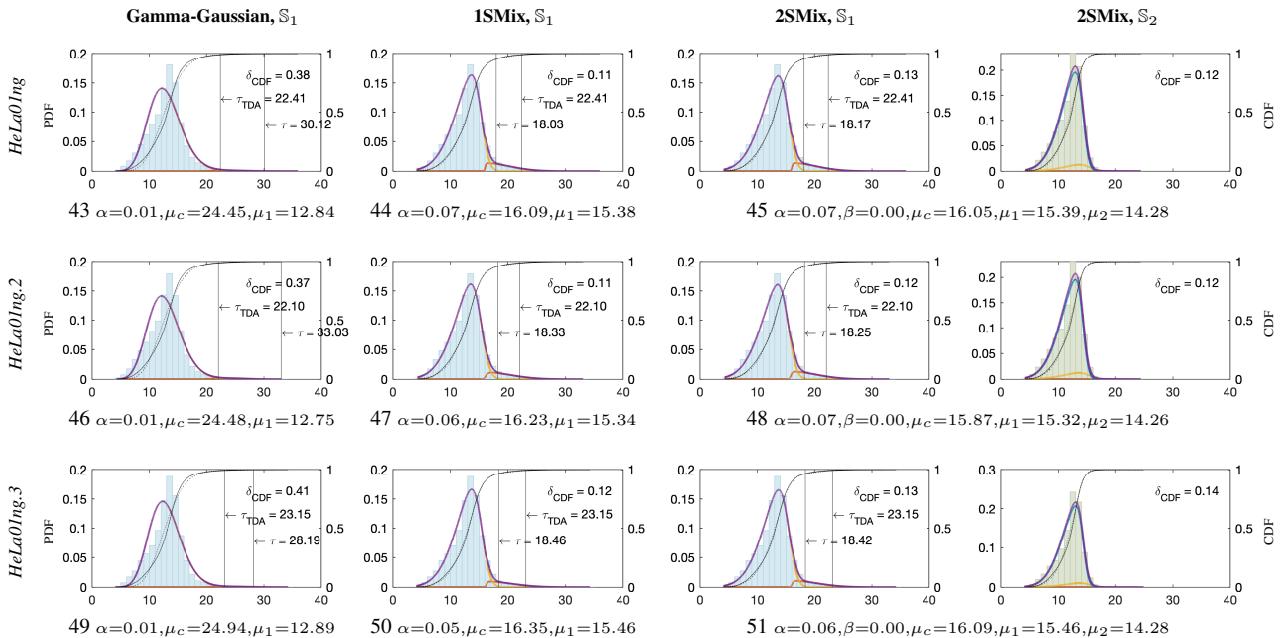


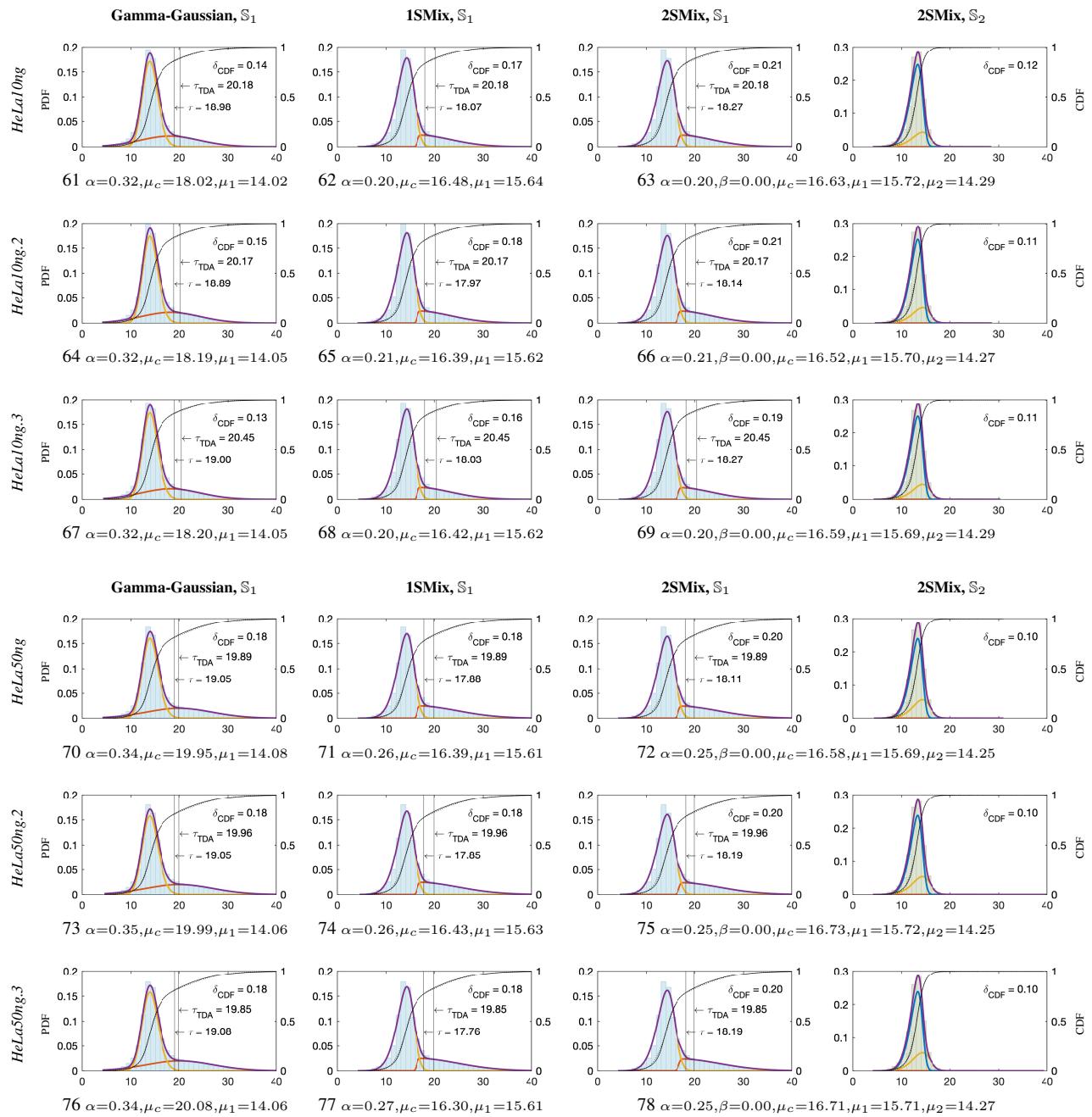
4.2 NIST Datasets





4.3 HeLa Datasets





4.4 All Results

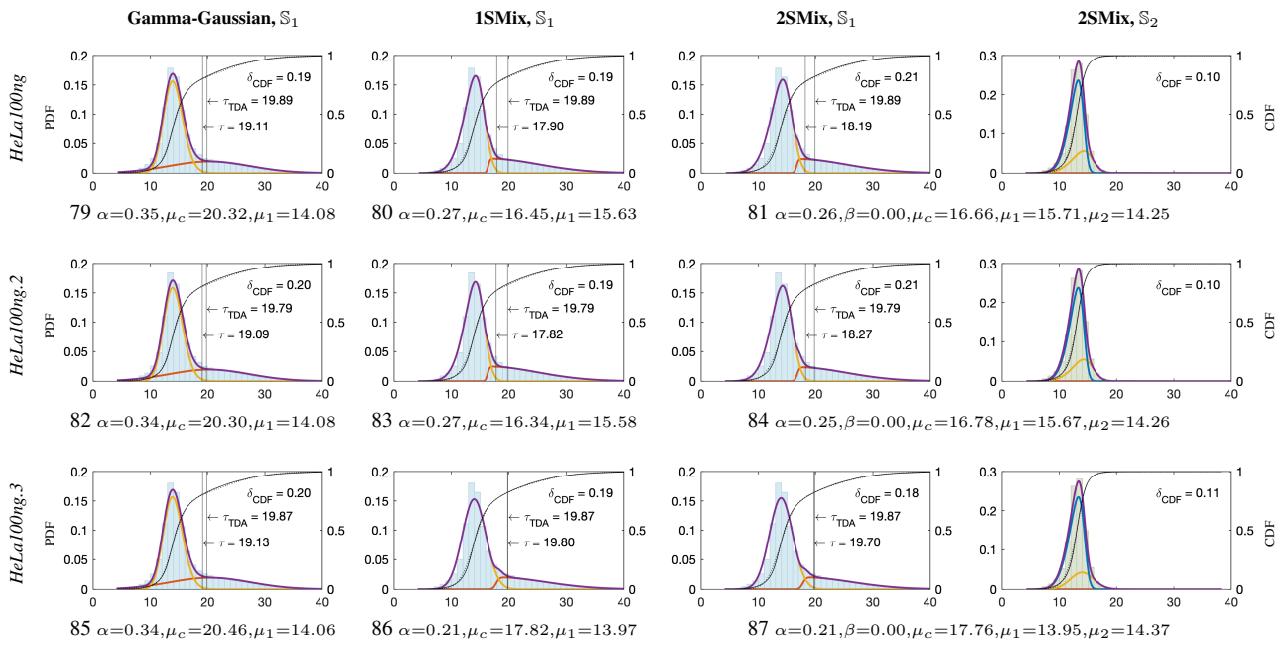


Table 1. Results for all experiments. # hits is the number of identified peptides above 1% FDR threshold; τ is the 1% FDR threshold; δ_{CDF} is the distance between empirical and estimated cdf; $\ell\ell_1$ is the log-likelihood on \mathbb{S}_1 ; $\ell\ell_{1,2}$ is the log-likelihood on \mathbb{S}_1 and \mathbb{S}_2 together.

Dataset	TDA		Gamma-Gaussian DFA				1SMix DFA				2SMix DFA					
	# hits	τ	# hits	τ	δ_{CDF}	$\ell\ell_1$	# hits	τ	δ_{CDF}	$\ell\ell_1$	# hits	τ	δ_{CDF}	$\ell\ell_1$	$\ell\ell_{1,2}$	
PRIDE	<i>A. thaliana</i>	354	22.13	433	21.17	0.182	-2.2452	723	18.79	0.089	-2.2256	857	18.07	0.066	-2.2239	-4.1987
	<i>D. melanogaster</i>	21425	18.23	22591	16.42	0.193	-2.9175	22958	16.09	0.171	-2.9172	22181	16.90	0.162	-2.9146	-4.7033
	<i>E. coli</i>	15001	16.33	14407	16.72	0.226	-3.0582	11070	18.95	0.850	-3.0806	16713	15.31	0.100	-3.0479	-4.9211
	<i>H. sapiens</i> 1	21486	18.80	21109	19.05	0.358	-3.2573	22918	17.89	0.085	-3.2461	22094	18.41	0.099	-3.2461	-5.1892
	<i>H. sapiens</i> 2	247	22.89	313	20.77	0.081	-2.0241	401	19.68	0.087	-2.0220	328	20.49	0.071	-2.0188	-3.7564
	<i>M. musculus</i> 1	12578	19.36	13283	18.45	0.335	-2.9247	13964	17.67	0.298	-2.9112	13757	17.88	0.131	-2.9072	-4.7852
	<i>M. musculus</i> 2	16419	19.40	17431	18.40	0.361	-2.9822	18192	17.70	0.375	-2.9689	18115	17.76	0.124	-2.9636	-4.8401
	<i>M. musculus</i> 3	12578	19.36	13283	18.45	0.335	-2.9247	13964	17.67	0.298	-2.9112	13757	17.88	0.131	-2.9072	-4.7852
	<i>S. cerevisiae</i> 1	3867	19.50	4502	17.47	0.112	-2.5084	5016	16.29	0.103	-2.5034	4648	17.08	0.086	-2.5006	-4.3121
	<i>S. cerevisiae</i> 2	5402	18.65	6010	17.11	0.158	-2.6846	6653	15.94	0.140	-2.6756	6271	16.53	0.104	-2.6736	-4.5201
NIST	<i>C. elegans</i>	23224	17.17	23764	16.24	0.085	-3.1058	24155	15.75	0.091	-3.1070	23467	16.70	0.122	-3.1061	-5.0126
	<i>H. sapiens</i>	218487	17.82	225599	16.79	0.353	-3.5219	227023	16.61	0.106	-3.5154	218963	17.77	0.125	-3.5167	-5.4989
	<i>M. musculus</i>	99358	17.58	102065	16.44	0.263	-3.5261	102241	16.37	0.096	-3.5221	103481	16.00	0.116	-3.5234	-5.5007
	<i>S. cerevisiae</i>	75529	15.55	77385	14.17	0.301	-3.6286	77067	14.35	0.100	-3.6214	75606	15.54	0.172	-3.6244	-5.5864
HeLa Cell Digest	HeLa01ng	218	22.41	9	30.12	0.378	-2.5319	1058	18.03	0.115	-2.4826	997	18.17	0.126	-2.4827	-4.6226
	HeLa01ng.2	255	22.10	1	33.03	0.370	-2.5310	895	18.33	0.110	-2.4848	920	18.25	0.120	-2.4849	-4.6320
	HeLa01ng.3	152	23.15	18	28.19	0.412	-2.5027	757	18.46	0.124	-2.4416	771	18.42	0.135	-2.4418	-4.5397
	HeLa1ng	1314	21.68	3798	18.20	0.152	-2.4335	3802	18.20	0.185	-2.4433	3702	18.28	0.211	-2.4436	-4.3975
	HeLa1ng.2	1305	21.73	3837	18.24	0.147	-2.4249	3713	18.33	0.193	-2.4356	3612	18.42	0.226	-2.4360	-4.3732
	HeLa1ng.3	1500	21.26	3812	18.22	0.140	-2.4173	3748	18.27	0.203	-2.4315	3733	18.29	0.228	-2.4318	-4.3607
	HeLa10ng	7480	20.18	9375	18.98	0.142	-2.5938	11229	18.07	0.172	-2.5961	10759	18.27	0.211	-2.5967	-4.4853
	HeLa10ng.2	7560	20.17	9602	18.89	0.147	-2.5898	11538	17.97	0.176	-2.5924	11140	18.14	0.212	-2.5929	-4.4620
	HeLa10ng.3	7153	20.45	9299	19.00	0.133	-2.5828	11284	18.03	0.165	-2.5847	10760	18.27	0.194	-2.5851	-4.4594
	HeLa50ng	12312	19.89	13925	19.05	0.175	-2.7327	16591	17.88	0.178	-2.7305	16018	18.11	0.202	-2.7311	-4.6222
	HeLa50ng.2	12581	19.96	14307	19.05	0.175	-2.7538	17071	17.85	0.176	-2.7514	16223	18.19	0.201	-2.7521	-4.6550
	HeLa50ng.3	12721	19.85	14165	19.08	0.179	-2.7509	17208	17.76	0.176	-2.7474	16160	18.19	0.197	-2.7483	-4.6488
	HeLa100ng	14009	19.89	15516	19.11	0.191	-2.7702	18477	17.90	0.191	-2.7670	17718	18.19	0.214	-2.7678	-4.6735
	HeLa100ng.2	13833	19.79	15230	19.09	0.196	-2.7519	18495	17.82	0.189	-2.7484	17195	18.27	0.208	-2.7492	-4.6437
	HeLa100ng.3	13941	19.87	15449	19.13	0.198	-2.7637	14090	19.80	0.188	-2.7677	14279	19.70	0.177	-2.7678	-4.6690