Using differential variability to increase the power of the homogeneity test in a two-sample problem

Guanfu Liu[1], Yuejiao Fu[2], Pengfei Li[3], and Xiaolong Pu[1]

[1] *East China Normal University,* [2] *York University,* [3] *University of Waterloo*

Supplementary Material

This is a supplementary document to the corresponding paper submitted to the *Statistica Sinica*. It contains the definition of the complete data penalized log-likelihood, the monotonicity property of EM-algorithm, some additional simulation results, R code for sample size calculation, further analysis of the second real data, regularity conditions, and proofs of Theorems 1–3.

1. Complete data penalized log-likelihood and monotonicity property of EM-algorithm

Recall that based on the two sample data $\{x_{11},\ldots,x_{1n_1};x_{21},\ldots,x_{22}\}$, the log-likelihood function for the unknown parameters $(\lambda,\mu_1,\mu_2,\sigma_1,\sigma_2)$ is

$$
\begin{aligned}
& l_n(\lambda,\mu_1,\mu_2,\sigma_1,\sigma_2) \\
= & \sum_{i=1}^{n_1}\log f(x_{1i};\mu_1,\sigma_1)+\sum_{i=1}^{n_2}\log\{(1-\lambda)f(x_{2i};\mu_1,\sigma_1)+\lambda f(x_{2i};\mu_2,\sigma_2)\}.
\end{aligned}
$$

The penalized log-likelihood function is defined as

$$pl_n(\lambda, \mu_1, \mu_2, \sigma_1, \sigma_2) = l_n(\lambda, \mu_1, \mu_2, \sigma_1, \sigma_2) + p(\lambda) + p_n(\sigma_2). \qquad (1)$$

To explicitly establish the EM-test, we define the complete data penalized log-likelihood function as follows. Let $z_i = 1$ if x_{2i} comes from $f(x; \mu_2, \sigma_2)$, and $z_i = 0$ if x_{2i} comes from $f(x; \mu_1, \sigma_1)$, $i = 1, \ldots, n_2$. Hence, the complete data consists of $\{x_{1i}, i = 1, \ldots, n_1\}$ and $\{(z_i, x_{2i}), 1, \ldots, n_2\}$. The complete data log-likelihood function is

$$\begin{aligned} l_n^c(\lambda, \mu_1, \mu_2, \sigma_1, \sigma_2) = & \sum_{i=1}^{n_1} \log f(x_{1i}; \mu_1, \sigma_1) + \sum_{i=1}^{n_2} \left[(1 - z_i) \log \{(1 - \lambda) f(x_{2i}; \mu_1, \sigma_1)\}\right] \\ & + \sum_{i=1}^{n_2} \left[z_i \log \{\lambda f(x_{2i}; \mu_2, \sigma_2)\}\right]. \end{aligned}$$

Further, the complete data penalized log-likelihood function is defined as

$$pl_n^c(\lambda, \mu_1, \mu_2, \sigma_1, \sigma_2) = l_n^c(\lambda, \mu_1, \mu_2, \sigma_1, \sigma_2) + p(\lambda) + p_n(\sigma_2).$$

Built upon $pl_n(\lambda, \mu_1, \mu_2, \sigma_1, \sigma_2)$ and $pl_n^c(\lambda, \mu_1, \mu_2, \sigma_1, \sigma_2)$, the EM-test statistics are defined through the following iteration process.

Step 1. Let $k = 1$. For each $j = 1, 2, \ldots, J$, let $\lambda_j^{(1)} = \lambda_j$ and compute

$$(\mu_{1j}^{(1)}, \mu_{2j}^{(1)}, \sigma_{1j}^{(1)}, \sigma_{2j}^{(1)}) = \arg \max_{\{\mu_1, \mu_2, \sigma_1, \sigma_2\}} pl_n(\lambda_j^{(1)}, \mu_1, \mu_2, \sigma_1, \sigma_2).$$

Step 2. For $i = 1, \ldots, n_2$ and the current k, we use an E-step to compute

$$w_{ij}^{(k)} = \frac{\lambda_j^{(k)} f(x_{2i}; \mu_{2j}^{(k)}, \sigma_{2j}^{(k)})}{(1-\lambda_j^{(k)}) f(x_{2i}; \mu_{1j}^{(k)}, \sigma_{1j}^{(k)}) + \lambda_j^{(k)} f(x_{2i}; \mu_{2j}^{(k)}, \sigma_{2j}^{(k)})}.$$

Update λ and other parameters by a M-step such that

$$\lambda_j^{(k+1)} = \arg\max_{\lambda} \left\{ \left(n_2 - \sum_{i=1}^{n_2} w_{ij}^{(k)} \right) \log(1-\lambda) + \sum_{i=1}^{n_2} w_{ij}^{(k)} \log \lambda + p(\lambda) \right\}$$

and

$$(\mu_{1j}^{(k+1)}, \sigma_{1j}^{(k+1)}) = \arg\max_{\{\mu_1,\sigma_1\}} \left\{ \sum_{i=1}^{n_1} \log f_1(x_{1i}) + \sum_{i=1}^{n_2} (1-w_{ij}^{(k)}) \log f_1(x_{2i}) \right\},$$

$$(\mu_{2j}^{(k+1)}, \sigma_{2j}^{(k+1)}) = \arg\max_{\{\mu_2,\sigma_2\}} \left\{ \sum_{i=1}^{n_2} w_{ij}^{(k)} \log f_2(x_{2i}) + p_n(\sigma_2) \right\}.$$

The E-step and M-step are iterated $K-1$ times.

Step 3. For each k and j, we define

$$M_n^{(k)}(\lambda_j) = 2\{pl_n(\lambda_j^{(k)}, \mu_{1j}^{(k)}, \mu_{2j}^{(k)}, \sigma_{1j}^{(k)}, \sigma_{2j}^{(k)}) - pl_n(1, \hat{\mu}_0, \hat{\mu}_0, \hat{\sigma}_0, \hat{\sigma}_0)\},$$

where $(\hat{\mu}_0, \hat{\sigma}_0) = \arg\max_{\{\mu,\sigma\}} pl_n(1, \mu, \mu, \sigma, \sigma)$. The EM-test statistic is then defined as

$$EM_n^{(K)} = \max\{M_n^{(K)}(\lambda_j) : j = 1, \ldots, J\}.$$

In the study of the asymptotic properties of the EM-test statistics, the monotonicity property of the EM algorithm plays an important role. We summarize it in the following proposition.

Proposition 1. *For $k \geq 1$,*

$$pl_n\left(\lambda_j^{(k+1)}, \mu_{1j}^{(k+1)}, \mu_{2j}^{(k+1)}, \sigma_{1j}^{(k+1)}, \sigma_{2j}^{(k+1)}\right) \geq pl_n\left(\lambda_j^{(k)}, \mu_{1j}^{(k)}, \mu_{2j}^{(k)}, \sigma_{1j}^{(k)}, \sigma_{2j}^{(k)}\right). \tag{2}$$

Proof. Let

$$l_{n1}(\mu_1, \sigma_1) = \sum_{i=1}^{n_1} \log f(x_{1i}; \mu_1, \sigma_1)$$

and

$$l_{n2}(\lambda, \mu_1, \mu_2, \sigma_1, \sigma_2) = \sum_{i=1}^{n_2} \log\{(1-\lambda) f(x_{2i}; \mu_1, \sigma_1) + \lambda f(x_{2i}; \mu_2, \sigma_2)\}.$$

Define the Q-function as

$$\begin{aligned} Q(\lambda, \mu_1, \mu_2, \sigma_1, \sigma_2) = & \left(n_2 - \sum_{i=1}^{n_2} w_{ij}^{(k)}\right) \log(1-\lambda) + \sum_{i=1}^{n_2} w_{ij}^{(k)} \log \lambda + p(\lambda) \\ & + \sum_{i=1}^{n_1} \log f(x_{1i}; \mu_1, \sigma_1) + \sum_{i=1}^{n_2} (1 - w_{ij}^{(k)}) \log f(x_{2i}; \mu_1, \sigma_1) \\ & + \sum_{i=1}^{n_2} w_{ij}^{(k)} \log f(x_{2i}; \mu_2, \sigma_2) + p_n(\sigma_2). \end{aligned}$$

It can be easily verified that

$$\left(\lambda_j^{(k+1)}, \mu_{1j}^{(k+1)}, \mu_{2j}^{(k+1)}, \sigma_{1j}^{(k+1)}, \sigma_{2j}^{(k+1)}\right) = \arg \max_{\lambda, \mu_1, \mu_2, \sigma_1, \sigma_2} Q(\lambda, \mu_1, \mu_2, \sigma_1, \sigma_2). \tag{3}$$

Next, we argue that

$$\begin{aligned}
& pl_n(\lambda, \mu_1, \mu_2, \sigma_1, \sigma_2) - pl_n\left(\lambda_j^{(k)}, \mu_{1j}^{(k)}, \mu_{2j}^{(k)}, \sigma_{1j}^{(k)}, \sigma_{2j}^{(k)}\right) \\
\geq & Q(\lambda, \mu_1, \mu_2, \sigma_1, \sigma_2) - Q\left(\lambda_j^{(k)}, \mu_{1j}^{(k)}, \mu_{2j}^{(k)}, \sigma_{1j}^{(k)}, \sigma_{2j}^{(k)}\right),
\end{aligned} \tag{4}$$

which together with (3) implies (2).

Note that

$$\begin{aligned}
& l_{n2}(\lambda, \mu_1, \mu_2, \sigma_1, \sigma_2) - l_{n2}\left(\lambda_j^{(k)}, \mu_{1j}^{(k)}, \mu_{2j}^{(k)}, \sigma_{1j}^{(k)}, \sigma_{2j}^{(k)}\right) \\
= & \sum_{i=1}^{n_2} \log\left\{\frac{(1-\lambda) f(x_{2i}; \mu_1, \sigma_1) + \lambda f(x_{2i}; \mu_2, \sigma_2)}{\left(1-\lambda_j^{(k)}\right) f\left(x_{2i}; \mu_{1j}^{(k)}, \sigma_{1j}^{(k)}\right) + \lambda_j^{(k)} f\left(x_{2i}; \mu_{2j}^{(k)}, \sigma_{2j}^{(k)}\right)}\right\} \\
= & \sum_{i=1}^{n_2} \log\left\{\left(1-w_{ij}^{(k)}\right)\frac{(1-\lambda) f(x_{2i}; \mu_1, \sigma_1)}{\left(1-\lambda_j^{(k)}\right) f\left(x_{2i}; \mu_{1j}^{(k)}, \sigma_{1j}^{(k)}\right)} + w_{ij}^{(k)} \frac{\lambda f(x_{2i}; \mu_2, \sigma_2)}{\lambda_j^{(k)} f\left(x_{2i}; \mu_{2j}^{(k)}, \sigma_{2j}^{(k)}\right)}\right\} \\
\geq & \sum_{i=1}^{n_2} \left\{\left(1-w_{ij}^{(k)}\right)\log\frac{(1-\lambda) f(x_{2i}; \mu_1, \sigma_1)}{\left(1-\lambda_j^{(k)}\right) f\left(x_{2i}; \mu_{1j}^{(k)}, \sigma_{1j}^{(k)}\right)} + w_{ij}^{(k)} \log\frac{\lambda f(x_{2i}; \mu_2, \sigma_2)}{\lambda_j^{(k)} f\left(x_{2i}; \mu_{2j}^{(k)}, \sigma_{2j}^{(k)}\right)}\right\},
\end{aligned} \tag{5}$$

where in the last step, we have applied the Jensen's inequality.

Note that

$$pl_n(\lambda, \mu_1, \mu_2, \sigma_1, \sigma_2) = l_{n1}(\mu_1, \sigma_1) + l_{n2}(\lambda, \mu_1, \mu_2, \sigma_1, \sigma_2) + p(\lambda) + p_n(\sigma_2). \tag{6}$$

After some algebra work, it can be verified that (5) and (6) together imply (4). This finishes the proof. □.

2. Additional simulation results

Table 1 and Tables 2–4 respectively show the type I error and the power comparison of the MLRT and EM-test under the logistic kernel. The conclusions are similar to those for the normal kernel case.

Table 1: Type I error comparison with $f_1 = f_2 = Logistic(0, 1)$.

	$\alpha = 0.05$			$\alpha = 0.01$		
	MLRT	$EM_n^{(3)}$	$EM_n^{(3)}$	MLRT	$EM_n^{(3)}$	$EM_n^{(3)}$
$n_1 = 50, n_2 = 50$	0.0548	0.0538	0.0546	0.0115	0.0113	0.0115
$n_1 = 50, n_2 = 100$	0.0495	0.0556	0.0564	0.0099	0.0118	0.0121
$n_1 = 100, n_2 = 50$	0.0514	0.0540	0.0541	0.0107	0.0098	0.0100
$n_1 = 100, n_2 = 100$	0.0534	0.0534	0.0540	0.0112	0.0106	0.0107

Results in column (3,6) used $\{\lambda_1, \ldots, \lambda_J\} = \{0.1, 0.4, 0.7, 1.0\}$. Results in column (4, 7) used $\{\lambda_1, \ldots, \lambda_J\} = \{0.1, 0.2, \ldots, 1.0\}$.

In Figures 1 and 2, we further present the quantile-quantile plots of $EM_n^{(3)}$ with $\{\lambda_1, \ldots, \lambda_J\} = \{0.1, 0.4, 0.7, 1.0\}$ under the normal kernel and logistic kernel cases, respectively. It shows that the limiting distribution of EM-test provides an accurate approximation to its finite sample distribution in cases.

Next, we present the power comparison of the t-test with unequal variance (denoted as T), Wilcoxon rank sum test (denoted as Wilc) and the EM-test (denoted as $EM_n^{(3)}$) under the normal kernel in Table 5, the logistic kernel in Table 6 and the extreme value type I kernel in Table 7.

Table 2: Power comparison with $f_1 = Logistic(0,1),\ f_2 = Logistic(1,1)$.

	$\alpha = 0.05$			$\alpha = 0.01$		
$\lambda = 0.9$	MLRT	$EM_n^{(3)}$	$EM_n^{(3)}$	MLRT	$EM_n^{(3)}$	$EM_n^{(3)}$
$n_1 = 50, n_2 = 50$	0.708	0.604	0.602	0.464	0.384	0.379
$n_1 = 50, n_2 = 100$	0.861	0.776	0.775	0.673	0.559	0.554
$n_1 = 100, n_2 = 50$	0.825	0.745	0.742	0.631	0.536	0.536
$n_1 = 100, n_2 = 100$	0.944	0.901	0.899	0.846	0.772	0.774
	$\alpha = 0.05$			$\alpha = 0.01$		
$\lambda = 0.7$	MLRT	$EM_n^{(3)}$	$EM_n^{(3)}$	MLRT	$EM_n^{(3)}$	$EM_n^{(3)}$
$n_1 = 50, n_2 = 50$	0.501	0.404	0.399	0.259	0.199	0.199
$n_1 = 50, n_2 = 100$	0.637	0.551	0.552	0.414	0.315	0.314
$n_1 = 100, n_2 = 50$	0.627	0.511	0.510	0.377	0.301	0.301
$n_1 = 100, n_2 = 100$	0.776	0.697	0.696	0.575	0.468	0.470

Results in column (3,6) used $\{\lambda_1, \ldots, \lambda_J\} = \{0.1, 0.4, 0.7, 1.0\}$. Results in column (4, 7) used $\{\lambda_1, \ldots, \lambda_J\} = \{0.1, 0.2, \ldots, 1.0\}$.

In Tables 5–6, it can be seen that if the two samples only differ in location, T and Wilc are better than the EM-test. In contrast, when f_1 and f_2 have different scales, the EM-test always possesses more power than T and Wilc. Especially for the case that f_1 and f_2 have the same location but different scales, the power of the EM-test is much bigger than that of T and Wilc, and T and Wilc have almost no power.

For the extreme value distribution case, we choose $a_1 = 1$ and $a_2 = 2$

Table 3: Power comparison with $f_1 = Logistic(0,1)$, $f_2 = Logistic(0.5,1.5)$.

	$\alpha = 0.05$			$\alpha = 0.01$		
$\lambda = 0.9$	MLRT	$EM_n^{(3)}$	$EM_n^{(3)}$	MLRT	$EM_n^{(3)}$	$EM_n^{(3)}$
$n_1 = 50, n_2 = 50$	0.253	0.592	0.592	0.114	0.335	0.330
$n_1 = 50, n_2 = 100$	0.287	0.735	0.732	0.128	0.494	0.492
$n_1 = 100, n_2 = 50$	0.383	0.719	0.716	0.211	0.522	0.522
$n_1 = 100, n_2 = 100$	0.442	0.889	0.890	0.272	0.724	0.727
	$\alpha = 0.05$			$\alpha = 0.01$		
$\lambda = 0.7$	MLRT	$EM_n^{(3)}$	$EM_n^{(3)}$	MLRT	$EM_n^{(3)}$	$EM_n^{(3)}$
$n_1 = 50, n_2 = 50$	0.166	0.422	0.422	0.076	0.228	0.228
$n_1 = 50, n_2 = 100$	0.153	0.516	0.516	0.056	0.279	0.276
$n_1 = 100, n_2 = 50$	0.243	0.552	0.555	0.122	0.345	0.344
$n_1 = 100, n_2 = 100$	0.272	0.708	0.709	0.138	0.489	0.489

Results in column (3,6) used $\{\lambda_1, \ldots, \lambda_J\} = \{0.1, 0.4, 0.7, 1.0\}$. Results in column (4, 7) used $\{\lambda_1, \ldots, \lambda_J\} = \{0.1, 0.2, \ldots, 1.0\}$.

for EM-test in order to obtain accurate type I errors and reasonable power. The simulation results are provided in Table 7. It can be seen that when f_1 and f_1 have different locations and same scale, the EM-test has greater power than T and Wilc. In contrast, when f_1 and f_2 have the same location and different scales, the performance of the EM-test is much better than T, and slightly better than Wilc. When f_1 and f_2 differ in both location and scale, the EM-test is again more powerful than T and Wilc. In summary,

Table 4: Power comparison with $f_1 = Logistic(0,1)$, $f_2 = Logistic(0,1.5)$.

	$\alpha = 0.05$			$\alpha = 0.01$		
$\lambda = 0.9$	MLRT	$EM_n^{(3)}$	$EM_n^{(3)}$	MLRT	$EM_n^{(3)}$	$EM_n^{(3)}$
$n_1 = 50, n_2 = 50$	0.102	0.481	0.480	0.038	0.262	0.262
$n_1 = 50, n_2 = 100$	0.077	0.615	0.617	0.020	0.363	0.363
$n_1 = 100, n_2 = 50$	0.157	0.646	0.643	0.067	0.426	0.425
$n_1 = 100, n_2 = 100$	0.142	0.807	0.806	0.063	0.614	0.615
	$\alpha = 0.05$			$\alpha = 0.01$		
$\lambda = 0.7$	MLRT	$EM_n^{(3)}$	$EM_n^{(3)}$	MLRT	$EM_n^{(3)}$	$EM_n^{(3)}$
$n_1 = 50, n_2 = 50$	0.076	0.367	0.366	0.017	0.169	0.166
$n_1 = 50, n_2 = 100$	0.050	0.439	0.442	0.011	0.218	0.219
$n_1 = 100, n_2 = 50$	0.111	0.448	0.450	0.044	0.262	0.261
$n_1 = 100, n_2 = 100$	0.108	0.605	0.604	0.042	0.385	0.386

Results in column (3,6) used $\{\lambda_1, \ldots, \lambda_J\} = \{0.1, 0.4, 0.7, 1.0\}$. Results in column (4, 7) used $\{\lambda_1, \ldots, \lambda_J\} = \{0.1, 0.2, \ldots, 1.0\}$.

under the asymmetric distribution case, when there is no mixture in the second sample, the performance of EM-test is always superior to T and Wilc.

Table 5: Power comparison with $\lambda = 1$ under the normal kernel.

	$\alpha = 0.05$				$\alpha = 0.01$			
	T	Wilc	$EM_n^{(3)}$	$EM_n^{(3)}$	T	Wilc	$EM_n^{(3)}$	$EM_n^{(3)}$
$f_1 = N(0,1), f_2 = N(0.5,1)$								
$n_1 = 50, n_2 = 50$	0.680	0.661	0.579	0.574	0.443	0.404	0.346	0.348
$n_1 = 50, n_2 = 100$	0.814	0.804	0.731	0.730	0.608	0.567	0.472	0.473
$n_1 = 100, n_2 = 50$	0.808	0.792	0.731	0.731	0.598	0.571	0.485	0.485
$n_1 = 100, n_2 = 100$	0.931	0.924	0.881	0.881	0.818	0.799	0.723	0.725
$f_1 = N(0,1), f_2 = N(0,1.5^2)$								
$n_1 = 50, n_2 = 50$	0.047	0.049	0.724	0.716	0.010	0.006	0.485	0.485
$n_1 = 50, n_2 = 100$	0.049	0.044	0.815	0.813	0.007	0.007	0.609	0.608
$n_1 = 100, n_2 = 50$	0.052	0.066	0.852	0.852	0.009	0.016	0.671	0.671
$n_1 = 100, n_2 = 100$	0.052	0.049	0.959	0.959	0.009	0.013	0.877	0.875
$f_1 = N(0,1), f_2 = N(0.5,1.5)$								
$n_1 = 50, n_2 = 50$	0.459	0.450	0.879	0.877	0.257	0.217	0.697	0.697
$n_1 = 50, n_2 = 100$	0.631	0.576	0.957	0.955	0.410	0.339	0.829	0.829
$n_1 = 100, n_2 = 50$	0.586	0.572	0.944	0.944	0.321	0.346	0.851	0.851
$n_1 = 100, n_2 = 100$	0.813	0.794	0.993	0.993	0.576	0.554	0.982	0.982

Results in column (4,8) used $\{\lambda_1, \ldots, \lambda_J\} = \{0.1, 0.4, 0.7, 1.0\}$. Results in column (5,9) used $\{\lambda_1, \ldots, \lambda_J\} = \{0.1, 0.2, \ldots, 1.0\}$.

Table 6: Power comparison with $\lambda = 1$ under the logistic kernel.

$f_1 = Logistic(0,1), f_2 = Logistic(1,1)$								
	$\alpha = 0.05$				$\alpha = 0.01$			
	T	Wilc	$EM_n^{(3)}$	$EM_n^{(3)}$	T	Wilc	$EM_n^{(3)}$	$EM_n^{(3)}$
$n_1 = 50, n_2 = 50$	0.779	0.811	0.733	0.732	0.571	0.587	0.496	0.494
$n_1 = 50, n_2 = 100$	0.869	0.898	0.851	0.852	0.702	0.730	0.652	0.649
$n_1 = 100, n_2 = 50$	0.889	0.909	0.846	0.843	0.744	0.760	0.699	0.699
$n_1 = 100, n_2 = 100$	0.971	0.979	0.954	0.955	0.903	0.927	0.880	0.880

$f_1 = Logistic(0,1), f_2 = Logistic(0,1.5)$								
	$\alpha = 0.05$				$\alpha = 0.01$			
	T	Wilc	$EM_n^{(3)}$	$EM_n^{(3)}$	T	Wilc	$EM_n^{(3)}$	$EM_n^{(3)}$
$n_1 = 50, n_2 = 50$	0.052	0.048	0.586	0.586	0.012	0.007	0.337	0.334
$n_1 = 50, n_2 = 100$	0.051	0.036	0.692	0.691	0.008	0.010	0.446	0.444
$n_1 = 100, n_2 = 50$	0.049	0.066	0.671	0.669	0.008	0.021	0.447	0.445
$n_1 = 100, n_2 = 100$	0.047	0.059	0.867	0.866	0.010	0.024	0.699	0.700

$f_1 = Logistic(0,1), f_2 = Logistic(0.5,1.5)$								
	$\alpha = 0.05$				$\alpha = 0.01$			
	T	Wilc	$EM_n^{(3)}$	$EM_n^{(3)}$	T	Wilc	$EM_n^{(3)}$	$EM_n^{(3)}$
$n_1 = 50, n_2 = 50$	0.185	0.205	0.641	0.640	0.067	0.075	0.403	0.399
$n_1 = 50, n_2 = 100$	0.261	0.260	0.800	0.800	0.111	0.090	0.544	0.543
$n_1 = 100, n_2 = 50$	0.226	0.261	0.799	0.798	0.063	0.105	0.589	0.586
$n_1 = 100, n_2 = 100$	0.339	0.366	0.927	0.927	0.141	0.159	0.805	0.805

Results in column (3,6) used $\{\lambda_1, \ldots, \lambda_J\} = \{0.1, 0.4, 0.7, 1.0\}$. Results in column (4, 7) used $\{\lambda_1, \ldots, \lambda_J\} = \{0.1, 0.2, \ldots, 1.0\}$.

Table 7: Power comparison with $\lambda = 1$ under the extreme value type I kernel.

$f_1 = Extreme(0,1), f_2 = Extreme(\log(2),1)$								
	$\alpha = 0.05$				$\alpha = 0.01$			
	T	Wilc	$EM_n^{(3)}$	$EM_n^{(3)}$	T	Wilc	$EM_n^{(3)}$	$EM_n^{(3)}$
$n_1 = 50, n_2 = 50$	0.288	0.135	0.990	0.989	0.096	0.035	0.943	0.942
$n_1 = 50, n_2 = 100$	0.250	0.170	0.999	0.999	0.097	0.065	0.993	0.994
$n_1 = 100, n_2 = 50$	0.431	0.135	0.999	0.999	0.222	0.040	0.987	0.987
$n_1 = 100, n_2 = 100$	0.518	0.223	1.000	1.000	0.277	0.083	1.000	1.000
$f_1 = Extreme(0,1), f_2 = Extreme(0,1/1.75)$								
	$\alpha = 0.05$				$\alpha = 0.01$			
	T	Wilc	$EM_n^{(3)}$	$EM_n^{(3)}$	T	Wilc	$EM_n^{(3)}$	$EM_n^{(3)}$
$n_1 = 50, n_2 = 50$	0.597	0.685	0.695	0.694	0.366	0.433	0.447	0.447
$n_1 = 50, n_2 = 100$	0.689	0.800	0.806	0.807	0.458	0.571	0.580	0.579
$n_1 = 100, n_2 = 50$	0.701	0.797	0.824	0.824	0.494	0.559	0.652	0.652
$n_1 = 100, n_2 = 100$	0.866	0.929	0.943	0.943	0.702	0.794	0.822	0.821
$f_1 = Extreme(0,1), f_2 = Extreme(\log(1.5),1/1.25)$								
	$\alpha = 0.05$				$\alpha = 0.01$			
	T	Wilc	$EM_n^{(3)}$	$EM_n^{(3)}$	T	Wilc	$EM_n^{(3)}$	$EM_n^{(3)}$
$n_1 = 50, n_2 = 50$	0.475	0.360	0.583	0.584	0.237	0.165	0.352	0.352
$n_1 = 50, n_2 = 100$	0.541	0.456	0.792	0.791	0.266	0.236	0.574	0.572
$n_1 = 100, n_2 = 50$	0.660	0.451	0.709	0.709	0.392	0.219	0.487	0.488
$n_1 = 100, n_2 = 100$	0.740	0.619	0.906	0.906	0.532	0.378	0.749	0.749

Results in column (4,8) used $\{\lambda_1, \ldots, \lambda_J\} = \{0.1, 0.4, 0.7, 1.0\}$. Results in column (5,9) used $\{\lambda_1, \ldots, \lambda_J\} = \{0.1, 0.2, \ldots, 1.0\}$.

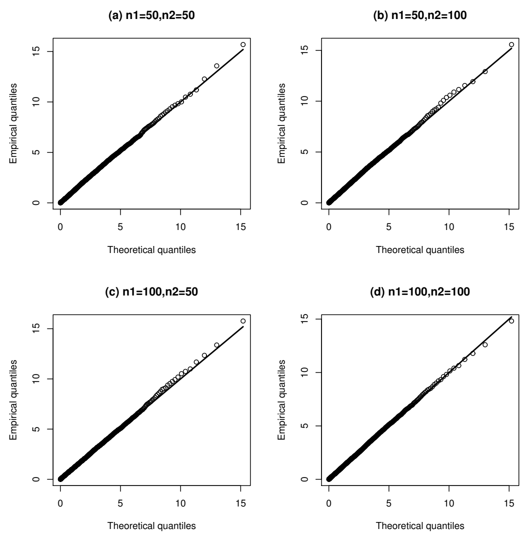

Figure 1: Quantile-quantile plots of $EM_n^{(3)}$ with $\{\lambda_1, \ldots, \lambda_J\} = \{0.1, 0.4, 0.7, 1.0\}$ under the normal kernel

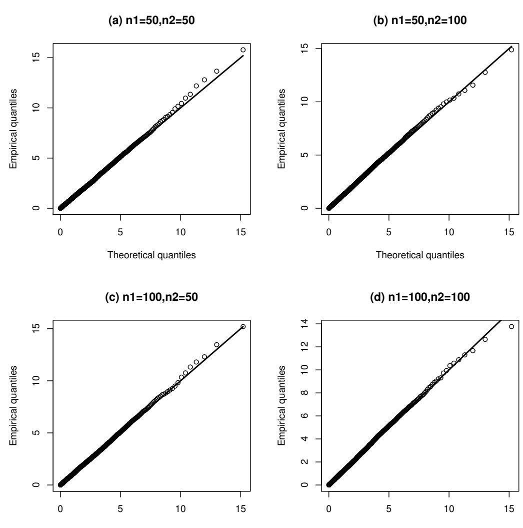

Figure 2: Quantile-quantile plots of $EM_n^{(3)}$ with $\{\lambda_1, \ldots, \lambda_J\} = \{0.1, 0.4, 0.7, 1.0\}$ under the logistic kernel

3. R code for sample size calculation

Given the density function $f(x;\mu_1,\sigma_1)$ for the first sample, the mixture density $(1-\lambda_0)f(x;\mu_1,\sigma_1)+\lambda_0 f(x;\mu_2,\sigma_2)$ for the second sample, and ρ_1, the following R functions *sizenorm()* and *sizelogis()* calculate the required sample sizes (n_1,n_2) to reject the null hypothesis with the target power $1-\beta$ at the significance level α for the normal kernel and logistic kernel, respectively.

For example, suppose $\lambda_0 = 0.5$, $(\mu_1,\sigma_1) = (0,1)$, $(\mu_2,\sigma_2) = (1,1.5)$, and $\rho_1 = 1/3$. If the target power is 80% at the 5% significance level, the required sample sizes are found to be $(n_1,n_2) = (39,78)$ under the normal kernel and $(n_1,n_2) = (84,168)$ under the logistic kernel by using R functions *sizenorm()* and *sizelogis()*.

```
sizenorm=function(lam0,rho1,mu1,sigma1,mu2,sigma2,alpha,target_power)
{
n2=2
power0=target_power; diff_power=1
while(diff_power>0.001){
Delta1=sqrt(n2)*(mu2-mu1); Delta2=sqrt(n2)*(sigma2-sigma1)
c0_squ=lam0^2*rho1/(sigma1^2)*(Delta1^2+2*Delta2^2)
power1=pchisq(qchisq(1-alpha,2),2,ncp=c0_squ,lower.tail = F)
```

```
diff_power=power0-power1

n2=n2+1

}

n1=round(rho1*n2/(1-rho1),0)

data.frame(n1=n1,n2=n2,row.names="sample size")

}


sizelogis=function(lam0,rho1,mu1,sigma1,mu2,sigma2,alpha,target_power)

{

  n2=2

  power0=target_power; diff_power=1

  while(diff_power>0.001){

    Delta1=sqrt(n2)*(mu2-mu1); Delta2=sqrt(n2)*(sigma2-sigma1)

    c0_squ=lam0^2*rho1/(sigma1^2)*(Delta1^2/3+Delta2^2*(3+pi^2)/9)

    power1=pchisq(qchisq(1-alpha,2),2,ncp=c0_squ,lower.tail = F)

    diff_power=power0-power1

    n2=n2+1

  }

n1=round(rho1*n2/(1-rho1),0)

data.frame(n1=n1,n2=n2,row.names="sample size")
```

```
}

> sizenorm(0.5,1/3,0,1,1,1.5,0.05,0.8)
            n1 n2
sample size 39 78

> sizelogis(0.5,1/3,0,1,1,1.5,0.05,0.8)
            n1  n2
sample size 84 168
```

4. Further analysis of the second real data

For the second real data of the main paper, 16 CpG sites are chosen as further analysis. The box plots of the control groups and case groups in each site are given in Figure 3. The box plots show that the control group and case group in each CpG site may have different variances, while some of them may have the same mean. The K-S normality testing p-values for the 16 control groups in 16 CpG sites are displayed in Table 8, which shows that the control groups should be taken as normal data. The p-values of three tests are given in Table 9. We can see that the p-values of EM-test are much smaller than the t-test and MLRT since only EM-test takes into the consideration of differential variabilities in addition to the location shift in order to detect heterogeneity in the two samples.

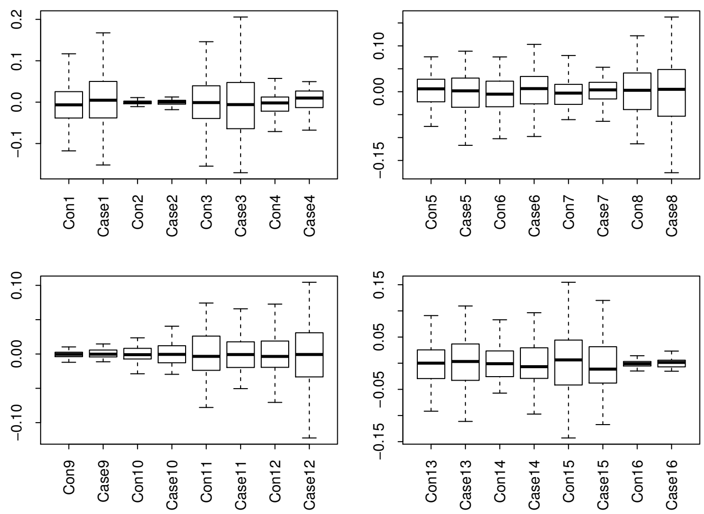

Figure 3: The parallel box plots for control groups and case groups of CpG sites 1-16. Con: Control group; Case: Case group.

Table 8: K-S normality testing p-values for the 16 control groups in 16 CpG sites.

CpG1	CpG2	CpG3	CpG4	CpG5	CpG6	CpG7	CpG8
0.835	0.633	0.704	0.993	0.831	0.651	0.641	0.677
CpG9	CpG10	CpG11	CpG12	CpG13	CpG14	CpG15	CpG16
0.896	0.715	0.645	0.587	0.669	0.784	0.735	0.933

5. Proofs of Theorems 1–3 in the main paper

Recall that the asymptotic properties of the EM-test rely on some regularity conditions on the kernel function $f(x; \mu, \sigma)$, and penalty func-

Table 9: The p-values of each test method for 16 CpG sites.

Methods	CpG1	CpG2	CpG3	CpG4	CpG5	CpG6	CpG7	CpG8
t-test	0.183	0.497	0.499	0.189	0.180	0.145	0.285	0.369
MLRT	0.177	0.492	0.494	0.184	0.174	0.142	0.286	0.304
$EM_n^{(3)}$	0.035	0.039	0.006	0.040	0.040	0.033	0.024	0.021
Methods	CpG9	CpG10	CpG11	CpG12	CpG13	CpG14	CpG15	CpG16
t-test	0.110	0.670	0.955	0.998	0.574	0.950	0.254	0.423
MLRT	0.108	0.678	0.956	1.000	0.567	0.952	0.253	0.421
$EM_n^{(3)}$	0.028	0.044	0.044	0.005	0.028	0.004	0.031	0.029

Results for the $EM_n^{(3)}$ used $\{\lambda_1, \ldots, \lambda_J\} = \{0.1, 0.4, 0.7, 1.0\}$.

tions $p(\lambda)$ and $p_n(\sigma)$. For completeness, we repeat the regularity conditions here. We start listing the following mild regularity conditions on $f(x; \mu, \sigma)$ in which the expectations are taken under the true null distribution $f(x; \mu_0, \sigma_0)$.

B1. (Wald's integrability conditions) (i) $E\{|\log f(x; \mu_0, \sigma_0)|\} < \infty$; (ii) for sufficiently small ρ and sufficiently large r, $E[\log\{1+f(x; \mu, \sigma, \rho)\}] < \infty$ for $(\mu, \sigma) \in \Theta$ and $E[\log\{1 + \phi(x; r)\}] < \infty$, where Θ is parameter space of (μ, σ) , $f(x; \mu, \sigma, \rho) = \sup_{|\mu'-\mu|^2+|\sigma'-\sigma|^2\leq\rho} f(x; \mu', \sigma')$, and $\phi(x; r) = \sup_{\mu^2+\sigma^2\geq r} f(x; \mu, \sigma)$; (iii) $f(x; \mu, \sigma) \to 0$ in probability as $\mu^2 + \sigma^2 \to \infty$.

B2. (Smoothness) The kernel $f(x;\mu,\sigma)$ has common support and is three times continuously differentiable with respect to μ and σ.

B3. (Identifiability) For any two mixing distribution functions Ψ_1 and Ψ_2 with two supporting points such that $\int f(x;\mu,\sigma)\,d\Psi_1(\mu,\sigma) = \int f(x;\mu,\sigma)\,d\Psi_2(\mu,\sigma)$ for all x, we must have $\Psi_1 = \Psi_2$.

B4. (Uniform boundedness) There exists a function g with finite expectation such that

$$\left|\frac{\partial^{(h+l)} f(x;\mu_0,\sigma_0)/\partial\mu^h\partial\sigma^l}{f(x;\mu_0,\sigma_0)}\right|^3 \leq g(x), for\ h+l \leq 2,$$

where h and l are two nonnegative integers. Moreover, there exists a positive ϵ such that

$$\sup_{|\mu-\mu_0|^2+|\sigma-\sigma_0|^2\leq\epsilon}\left|\frac{\partial^{(h+l)} f(x;\mu,\sigma)/\partial\mu^h\partial\sigma^l}{f(x;\mu_0,\sigma_0)}\right|^3 \leq g(x), for\ h+l = 3.$$

B5. (Positive definiteness) The covariance matrix of (U,V) is positive definite, where

$$U = \frac{\partial f(x_{11};\mu_0,\sigma_0)/\partial\mu}{f(x_{11};\mu_0,\sigma_0)} \quad\text{and}\quad V = \frac{\partial f(x_{11};\mu_0,\sigma_0)/\partial\sigma}{f(x_{11};\mu_0,\sigma_0)}.$$

B6. (Tail condition) There exists positive constants v_0, v_1 and β_0 with $\beta_0 > 1$ such that $f(x;0,1) \leq \min\{v_0, v_1|x|^{-\beta_0}\}$.

B7. (Upper bound function) There exist a nonnegative function $s(x;\mu,\sigma)$ which satisfies Condition B1 and is continuous in (μ,σ), a positive number a with $1/\beta_0 < a < 1$, a positive number b, and a positive number ϵ^* with $\epsilon^* < 1$ such that for $\sigma \in (0, \epsilon^*\sigma_0)$, $s(x; \mu, \sigma)$ is uniformly bounded, $\int s(x; \mu, \sigma)dx < 1$, and

$$f(x; \mu, \sigma) \leq \begin{cases} \sigma^{-1} s(x; \mu, \sigma), & \text{if } |x - \mu| \leq \sigma^{1-a} \\ \sigma^b s(x; \mu, \sigma), & \text{if } |x - \mu| > \sigma^{1-a} \end{cases}.$$

Next, we list regularity conditions on $p(\lambda)$ and $p_n(\sigma)$.

C1. $p(\lambda)$ is a continuous function such that it is maximized at $\lambda = 1$ and goes to negative infinity as $\lambda \to 0$.

C2. $\sup_{\sigma>0} \max\{p_n(\sigma), 0\} = o(n)$ and $p_n(\sigma) = o(n)$ for any σ.

C3. $p_n'(\sigma) = o_p(n^{1/2})$, for all $\sigma > 0$, where $p_n'(\sigma)$ is the derivative function with respect to σ.

C4. $p_n(\sigma) \leq 4(\log n_2)^2 \log(\sigma)$, when $0 < \sigma \leq (n_2 M_0)^{-1}$ and n_2 is large. Here $M_0 = \max\{\sup_x f(x; \mu_0, \sigma_0), 8\}$.

C5. $p_n(b_1\sigma; b_1X_1 + b_0, ..., b_1X_n + b_0) = p_n(\sigma; X_1, ..., X_n)$.

Some useful lemmas

Before proving Theorems 1–3 provided in the main paper, we present three lemmas, which summarize some useful properties of the point estimators. Lemma 1 indicates that any estimator with λ bounded away from 0 and with a large likelihood value, is consistent for μ_h and σ_h, $h = 1, 2$, under the null model. Lemma 2 strengthens Lemma 1 by providing specific convergence rates. Lemma 3 makes Lemmas 1 and 2 applicable to $(\lambda_j^{(k)}, \mu_{1j}^{(k)}, \mu_{2j}^{(k)}, \sigma_{1j}^{(k)}, \sigma_{2j}^{(k)})$ by showing that the EM-iteration keeps $\lambda_j^{(k)}$ in a small neighbourhood of λ_j , and therefore away from 0. Theorems 1–3 in the main paper then follow easily.

Since the EM-test is location-scale invariant, we assume $(\mu_0, \sigma_0) = (0, 1)$ for the convenience of presentation.

Lemma 1. *(Consistency with non-zero mixing proportion) Assume the same conditions in Theorem 1. Let $(\bar{\lambda}, \bar{\mu}_1, \bar{\mu}_2, \bar{\sigma}_1, \bar{\sigma}_2)$ be any estimator of $(\lambda, \mu_1, \mu_2, \sigma_1, \sigma_2)$ such that*

$$pl_n(\bar{\lambda}, \bar{\mu}_1, \bar{\mu}_2, \bar{\sigma}_1, \bar{\sigma}_2) - pl_n(1, 0, 0, 1, 1) > c > -\infty$$

and $\bar{\lambda} \in [\delta, 1]$ for some $\delta \in (0, 1)$. Then under the null model $f(x; 0, 1)$, $\bar{\mu}_1 = o_p(1)$, $\bar{\mu}_2 = o_p(1)$, $\bar{\sigma}_1 - 1 = o_p(1)$ and $\bar{\sigma}_2 - 1 = o_p(1)$.

A critical step of the proof for Lemma 1 is to show that there exists a positive constant τ_0 such that

$$\lim_{n\to\infty} P(\bar{\sigma}_1 > \tau_0, \bar{\sigma}_2 > \tau_0) = 1. \tag{7}$$

First, we establish a technical lemma, labelled by Lemma A.1, which gives the uniform upper bound for the number of observations in the σ^{1-a}

neighbourhood of μ. Let $\delta_n(\sigma) = M_0\sigma^{1-a} + 1/n$, where a and M_0 are respectively the constants defined in Conditions B7 and C4. Throughout the proof, all the expectations are taken under $f(x; 0, 1)$, unless stated otherwise.

Lemma A.1. *Suppose $t_1, \ldots, t_n$ are iid randoms sample from the probability density function $f(x; 0, 1)$ and $f(x; 0, 1)$ is a continuous function of x. Except for a zero-probability event not depending on σ, as $n \to \infty$ and almost surely, we have:*

1. for each given σ between $\exp(-2)$ *and* $8/(nM_0)$,

$$\sup_{\mu} \sum_{i=1}^{n} I(|t_i - \mu| < \sigma^{1-a}) \leq 8n\delta_n(\sigma);$$

2. uniformly for σ between 0 *and* $8/(nM_0)$,

$$\sup_{\mu} \sum_{i=1}^{n} I(|t_i - \mu| < \sigma^{1-a}) \leq 4(\log n)^2;$$

Lemma A.1 can be proved in similar fashion to the proof of Lemmas 1 and 2 in Chen, Tan and Zhang (2008). Hence we omit it.

Next, we move back to the proof of (7). In the following discussion we concentrate on the case that $\sigma_1 \leq \sigma_2$. The proof for the case that $\sigma_1 > \sigma_2$ is similar. Let

$$\Gamma_1 = \{(\lambda, \mu_1, \mu_2, \sigma_1, \sigma_2) : \sigma_1 \leq \sigma_2 \leq \epsilon_0\}$$

and

$$\Gamma_2 = \{(\lambda, \mu_1, \mu_2, \sigma_1, \sigma_2) : \sigma_1 \leq \tau_0, \sigma_2 \geq \epsilon_0\}.$$

The choice of ϵ_0 will be discussed later and the choice of τ_0 will be given in Lemma A.3. We prove in Lemmas A.2 and A.3 that

$$\lim_{n\to\infty} P\Big((\bar{\lambda}, \bar{\mu}_1, \bar{\mu}_2, \bar{\sigma}_1, \bar{\sigma}_2) \in \Gamma_1\Big) = 0$$

and

$$\lim_{n\to\infty} P\Big((\bar{\lambda}, \bar{\mu}_1, \bar{\mu}_2, \bar{\sigma}_1, \bar{\sigma}_2) \in \Gamma_2\Big) = 0,$$

respectively. Then (7) follows.

We now define the choice of ϵ_0. Let $K_0 = E\{\log f(X; 0, 1)\}$. Condition B1 ensures that $|K_0| < \infty$. We require that the positive constant ϵ_0 satisfies

1. $\epsilon_0 < \min\Big\{\epsilon^*, e^{-2}, (32M_0)^{-1/(1-a)}, v_0, (v_0/v_1)^{1/(a\beta_0)}, v_1^{-1/(a\beta_0-1)}\Big\}$, where ϵ^* is given in Condition B7;

2. $-8\rho_1 M_0 \epsilon_0^{1-a} \log \epsilon_0 + 16\rho_2 M_0 \epsilon_0^{1-a}(\log v_0 - \log \epsilon_0) \leq \rho_2$;

3. $\log v_1 + \log \epsilon_0^{\beta_0 a - 1} \leq 2K_0 - 4$.

Lemma A.2. *Assume the same conditions in Lemma 1. Then*

$$\sup_{(\lambda,\mu_1,\mu_2,\sigma_1,\sigma_2)\in\Gamma_1} pl_n(\lambda, \mu_1, \mu_2, \sigma_1, \sigma_2) - pl_n(1, 0, 0, 1, 1) \to -\infty$$

almost surely when $n \to \infty$.

Proof. Let $A_{h1} = \{i : |x_{hi} - \mu_1| \le \sigma_1^{1-a}, i = 1, \ldots, n_h\}$ and $A_{h2} = \{j : |x_{hi} - \mu_2| \le \sigma_2^{1-a}, i = 1, \ldots, n_h\}$, $h = 1, 2$. Further, let $l_{n1}(\mu_1, \sigma_1)$ and $l_{n2}(\lambda, \mu_1, \mu_2, \sigma_1, \sigma_2)$ be the log-likelihood functions based on the first and second samples, respectively.

For any index set, say S, define

$$l_{n1}(\mu_1, \sigma_1; S) = \sum_{i \in S} f(x_{1i}; \mu_1, \sigma_1) \quad \text{and}$$

$$l_{n2}(\lambda, \mu_1, \mu_2, \sigma_1, \sigma_2; S) = \sum_{i \in S} \log \left\{ (1-\lambda) f\left(x_{2i}; \mu_1, \sigma_1\right) + \lambda f\left(x_{2i}; \mu_2, \sigma_2\right) \right\}.$$

Let $n_h(S) = \#\{x_{hi} : i \in S\}$ be the number of observations in the hth sample such that their indexes belongs to S, $h = 1, 2$.

We first find an upper bound for $l_{n1}(\mu_1, \sigma_1) - l_{n1}(0, 1)$. Using Condition B7, we have that for $\sigma_1 \le \epsilon_0$,

$$\begin{aligned} l_{n1}(\mu_1, \sigma_1; A_{11}^c) - l_{n1}(0, 1; A_{11}^c) &= \sum_{i \in A_{11}^c} \log f\left(x_{1i}; \mu_1, \sigma_1\right) - \sum_{i \in A_{11}^c} \log f(x_{1i}; 0, 1) \\ &\le \sum_{i \in A_{11}^c} \log\{\sigma_1^b s(x_{1i}; \mu_1, \sigma_1)\} - \sum_{i \in A_{11}^c} \log f(x_{1i}; 0, 1) \\ &\le b n_1(A_{11}^c) \log \sigma_1 + \sum_{i \in A_{11}^c} \log \frac{s(x_{1i}; \mu_1, \sigma_1)}{f(x_{1i}; 0, 1)} \end{aligned} \tag{8}$$

and

$$\begin{aligned} l_{n1}(\mu_1,\sigma_1;A_{11}) - l_{n1}(0,1;A_{11}) &= \sum_{i\in A_{11}} \log f\,(x_{1i};\mu_1,\sigma_1) - \sum_{i\in A_{11}} \log f(x_{1i};0,1) \\ &\le - n_1(A_{11})\log\sigma_1 + \sum_{i\in A_{11}} \log\frac{s(x_{1i};\mu_1,\sigma_1)}{f(x_{1i};0,1)}. \end{aligned} \tag{9}$$

Combining (8) and (9), we have

$$l_{n1}(\mu_1,\sigma_1) - l_{n1}(0,1) \le \{bn_1(A_{11}^c) - n_1(A_{11})\}\log\sigma_1 + \sum_{i=1}^{n_1}\log\frac{s(x_{1i};\mu_1,\sigma_1)}{f(x_{1i};0,1)}. \tag{10}$$

Lemma A.1 implies that

$$-n_1(A_{11})\log\sigma_1 \le \begin{cases} -4(\log n_1)^2\log\sigma_1, & \sigma_1\in(0,\frac{8}{n_1M_0}], \\ -8\log\sigma_1 - 8n_1M_0\sigma_1^{1-a}\log\sigma_1, & \sigma_1\in(\frac{8}{n_1M_0},\epsilon_0) \end{cases}$$

and $n_1(A_{11}^c)\ge(3/4)n_1$ for large enough n_1. Hence when $\sigma_1\le\epsilon_0$ and n_1 is large enough,

$$\frac{1}{2}\frac{3bn_1}{4}\log\sigma_1 - n_1(A_{11})\log\sigma_1 \le -8n_1M_0\epsilon_0^{1-a}\log\epsilon_0 + 9\log n_1. \tag{11}$$

Combining (10)-(11) and the fact that $n_1(A_{11}^c)\ge(3/4)n_1$ for large enough n_1 gives

$$\begin{aligned} l_{n1}(\mu_1,\sigma_1) - l_{n1}(0,1) \le& \frac{1}{2}\frac{3bn_1}{4}\log\sigma_1 - 8n_1M_0\epsilon_0^{1-a}\log\epsilon_0 + 9\log n_1 \\ &+ \sum_{i=1}^{n_1}\log\frac{s(y_{1i};\mu_1,\sigma_1)}{f(y_{1i};0,1)}. \end{aligned} \tag{12}$$

By the strong law of large numbers, Jensen Inequality, and Condition B7, for any given (μ_1, σ_1) with $\sigma_1 \leq \epsilon_0$,

$$\frac{1}{n_1}\sum_{i=1}^{n_1} \log \frac{s(x_{1i};\mu_1,\sigma_1)}{f(x_{1i};0,1)} \to E\left\{\log \frac{s(x_{11};\mu_1,\sigma_1)}{f(x_{11};0,1)}\right\} \leq \log E\left\{\frac{s(x_{11};\mu_1,\sigma_1)}{f(x_{11};0,1)}\right\} < 0$$

almost surely when $n_1 \to \infty$. Consequently, we can easily show, as in Wald (1949), that

$$\sup_{\{(\mu_1,\sigma_1):\ \sigma_1 \leq \epsilon_0\}} \frac{1}{n_1}\sum_{i=1}^{n_1} \log \frac{s(x_{1i};\mu_1,\sigma_1)}{f(x_{1i};0,1)} < 0 \tag{13}$$

almost surely when n_1 is large enough. Combining (12) and (13), we get that for $\sigma_1 \leq \epsilon_0$ and large n_1,

$$l_{n1}(\mu_1,\sigma_1) - l_{n1}(0,1) \leq \frac{1}{2}\frac{3bn_1}{4}\log\sigma_1 - 8n_1M_0\epsilon_0^{1-a}\log\epsilon_0 + 9\log n_1. \tag{14}$$

Next we find an upper bound for $l_{n2}(\lambda,\mu_1,\mu_2,\sigma_1,\sigma_2) - l_{n2}(1,0,0,1,1)$. From Condition B6, we get that for $(\lambda,\mu_1,\mu_2,\sigma_1,\sigma_2) \in \Gamma_1$,

$$\begin{aligned}
& l_{n2}(\lambda,\mu_1,\mu_2,\sigma_1,\sigma_2;A_{21}) \\
&= \sum_{i\in A_{21}} \log\left\{\frac{1-\lambda}{\sigma_1} f\left(\frac{x_{2i}-\mu_1}{\sigma_1};0,1\right) + \frac{\lambda}{\sigma_2} f\left(\frac{x_{2i}-\mu_2}{\sigma_2};0,1\right)\right\} \\
&\leq n_2(A_{21})\log\left(\frac{1-\lambda}{\sigma_1}v_0 + \frac{\lambda}{\sigma_2}v_0\right) \\
&\leq n_2(A_{21})(\log v_0 - \log\sigma_1),
\end{aligned} \tag{15}$$

and

$$
\begin{aligned}
& l_{n2}(\lambda, \mu_1, \mu_2, \sigma_1, \sigma_2; A_{21}^c A_{22}) \\
= & \sum_{i \in A_{21}^c A_{22}} \log \left\{ \frac{1-\lambda}{\sigma_1} f\left(\frac{x_{2i} - \mu_1}{\sigma_1}; 0, 1 \right) + \frac{\lambda}{\sigma_2} f\left(\frac{x_{2i} - \mu_2}{\sigma_2}; 0, 1 \right) \right\} \\
\leq & \sum_{i \in A_{21}^c A_{22}} \log \left\{ (1-\lambda) v_1 \sigma_1^{\beta_0 a - 1} + \frac{\lambda v_0}{\sigma_2} \right\} \\
\leq & \, n_2(A_{22})(\log v_0 - \log \sigma_2).
\end{aligned}
$$

Using Lemma A.1 and (15), we get that

$$
l_{n2}(\lambda, \mu_1, \mu_2, \sigma_1, \sigma_2; A_{21}) \leq \begin{cases} 4(\log n_2)^2 \log \frac{v_0}{\sigma_1}, \ \sigma_1 \in (0, \frac{8}{n_2 M_0}], \\ 8 \log \frac{v_0}{\sigma_1} + 8 n_2 M_0 \sigma_1^{1-a} \log \frac{v_0}{\sigma_1}, \ \sigma_1 \in (\frac{8}{n_2 M_0}, \epsilon_0), \end{cases}
$$

which implies that for large n

$$
\frac{1}{2} \frac{3bn_1}{4} \log \sigma_1 + l_{n2}(\lambda, \mu_1, \mu_2, \sigma_1, \sigma_2; A_{21}) \leq 8 n_2 M_0 \epsilon_0^{1-a} \log \frac{v_0}{\epsilon_0} + 9 \log n_2. \quad (16)
$$

By Condition C4, we similarly have that for large enough n_2,

$$
l_{n2}(\lambda, \mu_1, \mu_2, \sigma_1, \sigma_2; A_{21}^c A_{22}) + p_n(\sigma_2) \leq 8 n_2 M_0 \epsilon_0^{1-a} \log \frac{v_0}{\epsilon_0} + 9 \log n_2 + o(n). \quad (17)
$$

For $l_{n2}(\lambda, \mu_1, \mu_2, \sigma_1, \sigma_2; A_{21}^c A_{22}^c)$, by using Condition B6, we get

$$
\begin{aligned}
l_{n2}(\lambda, \mu_1, \mu_2, \sigma_1, \sigma_2; A_{21}^c A_{22}^c) \leq & \sum_{i \in A_{21}^c A_{22}^c} \log\{(1-\lambda) v_1 \sigma_1^{\beta_0 a - 1} + \lambda v_1 \sigma_2^{\beta_0 a - 1}\} \\
\leq & \, n_2(A_{21}^c A_{22}^c) \log(v_1 \epsilon_0^{\beta_0 a - 1}).
\end{aligned}
$$

From Lemma A.1, it can be checked that when n_2 is large enough, $n_2(A_{21}^c A_{22}^c) \geq n_2 - \{n_2(A_{21}) + n_2(A_{22})\} \geq n_2/2$. Hence we get

$$l_{n2}(\lambda, \mu_1, \mu_2, \sigma_1, \sigma_2; A_{21}^c A_{22}^c) \leq (n_2/2)(\log v_1 + \log \epsilon_0^{\beta_0 a - 1}). \quad (18)$$

Combining (16)–(18), it follows that for large n,

$$\begin{aligned}
&\frac{1}{2}\frac{3bn_1}{4}\log\sigma_1 + l_{n2}(\lambda, \mu_1, \mu_2, \sigma_1, \sigma_2) + p_n(\sigma_2) \\
\leq & 16 n_2 M_0 \epsilon_0^{1-a} \log\frac{v_0}{\epsilon_0} + 18\log n_2 + (n_2/2)(\log v_1 + \log\epsilon_0^{\beta_0 a-1}) + o(n). \quad (19)
\end{aligned}$$

By strong law of large numbers, we have that $l_{n2}(1, 0, 0, 1, 1) = n_2\{K_0 + o(1)\}$. Combining (1), (14), (19) and using Conditions C1–C2, we get

$$\begin{aligned}
& pl_n(\lambda, \mu_1, \mu_2, \sigma_1, \sigma_2) - pl_n(1, 0, 0, 1, 1) \\
\leq & -8n_1 M_0 \epsilon_0^{1-a}\log\epsilon_0 + 16 n_2 M_0 \epsilon_0^{1-a}\log\frac{v_0}{\epsilon_0} \\
& + \frac{n_2}{2}(\log v_1 + \log\epsilon_0^{\beta_0 a - 1}) + 27\log n - l_{n2}(1,0,0,1,1) - p_n(1) + o(n) \\
\leq & n\rho_2 + \frac{n\rho_2}{2}(2K_0 - 4) - n\rho_2 K_0 + o(n) \\
= & -n\rho_2 + o(n).
\end{aligned}$$

Therefore

$$\sup_{(\lambda,\mu_1,\mu_2,\sigma_1,\sigma_2)\in\Gamma_1} pl_n(\lambda, \mu_1, \mu_2, \sigma_1, \sigma_2) - pl_n(1,0,0,1,1) \to -\infty$$

almost surely when $n \to \infty$. This finishes the proof. □

To establish a similar result in Γ_2, we define

$$g(x; \lambda, \mu_1, \mu_2, \sigma_1, \sigma_2) = a_1(1-\lambda)s(x; \mu_1, \sigma_1) + a_2\lambda f(x; \mu_2, \sigma_2),$$

where $a_1 = I(\sigma_1 \neq 0, \mu_1 \neq \pm\infty)$, $a_2 = I(\mu_2 \neq \pm\infty)$. It is easy to show that the function $g(x; \lambda, \mu_1, \mu_2, \sigma_1, \sigma_2)$ has the following properties:

1. $g(x; \lambda, \mu_1, \mu_2, \sigma_1, \sigma_2)$ is continuous in $(\lambda, \mu_1, \mu_2, \sigma_1, \sigma_2)$ almost surely with respect to $f(x; 0, 1)$;

2. $E[\log\{g(X; \lambda, \mu_1, \mu_2, \sigma_1, \sigma_2)/f(X; 0, 1)\}] < 0$ for any $(\lambda, \mu_1, \mu_2, \sigma_1, \sigma_2)$ in Γ_2 (Jensen Inequality);

3. $g(x; \lambda, \mu_1, \mu_2, \sigma_1, \sigma_2)$ is uniformly bounded for $(\lambda, \mu_1, \mu_2, \sigma_1, \sigma_2) \in \Gamma_2$.

We choose $\tau_0 < \epsilon_0$, which is smaller than 1. Hence σ_1 is smaller than the true value σ_0. Consequently, we can easily show, as in Wald (1949), that

$$\sup_{(\lambda,\mu_1,\mu_2,\sigma_1,\sigma_2)\in\Gamma_2} \frac{1}{n_2} \sum_{i=1}^{n_2} \log \frac{g(x_{2i}; \lambda, \mu_1, \mu_2, \sigma_1, \sigma_2)}{f(x_{2i}; 0, 1)} \to -\delta(\tau_0) < 0$$

almost surely as $n_2 \to \infty$. Note that $\delta(\tau_0)$ is a decreasing function of τ_0. Then $\delta(\epsilon_0) \leq \delta(\tau_0)$.

Lemma A.3. *Assume the same conditions as in Lemma A.2. Then*

$$\sup_{(\lambda,\mu_1,\mu_2,\sigma_1,\sigma_2)\in\Gamma_2} pl_n(\lambda, \mu_1, \mu_2, \sigma_1, \sigma_2) - pl_n(1, 0, 0, 1, 1) \to -\infty$$

almost surely when $n \to \infty$.

Proof. Using Condition B7, we have that for $(\lambda, \mu_1, \mu_2, \sigma_1, \sigma_2) \in \Gamma_2$.

$$\begin{aligned}
l_{n2}(\lambda, \mu_1, \mu_2, \sigma_1, \sigma_2) =& l_{n2}(\lambda, \mu_1, \mu_2, \sigma_1, \sigma_2; A_{21}) + l_{n2}(\lambda, \mu_1, \mu_2, \sigma_1, \sigma_2; A_{21}^c) \\
=& \sum_{i \in A_{21}} \log \left[(1-\lambda) f\left(x_{2i}; \mu_1, \sigma_1\right) + \lambda f\left(x_{2i}; \mu_2, \sigma_2\right)\right] \\
&+ \sum_{i \in A_{21}^c} \log \left[(1-\lambda) f\left(x_{2i}; \mu_1, \sigma_1\right) + \lambda f\left(x_{2i}; \mu_2, \sigma_2\right)\right] \\
\leq& - n_2(A_{21}) \log \sigma_1 + \sum_{i \in A_{21}} \log g(x_{2i}; \lambda, \mu_1, \mu_2, \sigma_1, \sigma_2) \\
&+ \sum_{i \in A_{21}^c} \log g(x_{2i}; \lambda, \mu_1, \mu_2, \sigma_1, \sigma_2) \\
=& - n_2(A_{21}) \log \sigma_1 + \sum_{i=1}^{n_2} \log g(x_{2i}; \lambda, \mu_1, \mu_2, \sigma_1, \sigma_2).
\end{aligned}$$

By Conditions C1–C3, we further have

$$\begin{aligned}
& l_{n2}(\lambda, \mu_1, \mu_2, \sigma_1, \sigma_2) - l_{n2}(1, 0, 0, 1, 1) + p(\lambda) - p(1) + p_n(\sigma_2) - p_n(1) \\
\leq& - n_2(A_{21}) \log \sigma_1 + \sum_{i=1}^{n_2} \log \frac{g(x_{2i}; \lambda, \mu_1, \mu_2, \sigma_1, \sigma_2)}{f(x_{2i}; 0, 1)} + p_n(\sigma_2) - p_n(1) \\
\leq& - n_2(A_{21}) \log \sigma_1 - n_2 \delta(\tau_0) + o(n).
\end{aligned} \tag{20}$$

Similar to (14), we get

$$l_{n1}(\mu_1, \sigma_1) - l_{n1}(0, 1) \leq \frac{1}{2} \frac{3 b n_1}{4} \log \sigma_1 - 8 n_1 M_0 \tau_0^{1-a} \log \tau_0 + 9 \log n_1. \tag{21}$$

Note that when n_1 and n_2 are large enough and $\sigma_1 \leq \tau_0$, it can be verified that

$$\frac{1}{2} \frac{3 b n_1}{4} \log \sigma_1 - n_2(A_{21}) \log \sigma_1 \leq -8 n_2 M_0 \tau_0^{1-a} \log \tau_0 + 9 \log n_2.$$

Hence, from (1), (20) and (21), we have

$$pl_n(\lambda, \mu_1, \mu_2, \sigma_1, \sigma_2) - pl_n(1, 0, 0, 1, 1) \leq -16nM_0\tau_0^{1-a} \log \tau_0 - n\rho_2\delta(\tau_0) + 18 \log n + o(n).$$

We select τ_0 such that $-16M_0\tau_0^{1-a} \log \tau_0 \leq 0.5\rho_2\delta(\epsilon_0) \leq 0.5\rho_2\delta(\tau_0)$. Then we get

$$pl_n(\lambda, \mu_1, \mu_2, \sigma_1, \sigma_2) - pl_n(1, 0, 0, 1, 1) \leq -0.5n\rho_2\delta(\tau_0) + 18 \log n + o(n).$$

Therefore

$$\sup_{(\lambda,\mu_1,\mu_2,\sigma_1,\sigma_2)\in\Gamma_2} pl_n(\lambda, \mu_1, \mu_2, \sigma_1, \sigma_2) - pl_n(1, 0, 0, 1, 1) \to -\infty$$

almost surely when $n \to \infty$, as required. □

Combining the condition $pl_n(\bar{\lambda}, \bar{\mu}_1, \bar{\mu}_2, \bar{\sigma}_1, \bar{\sigma}_2) - pl_n(1, 0, 0, 1, 1) > c > -\infty$ in Lemma 1, and Lemmas A.2–A.3, we know that there exists a constant $\tau_0 > 0$ such that

$$\lim_{n\to\infty} P(\bar{\sigma}_1 > \tau_0, \bar{\sigma}_2 > \tau_0) = 1.$$

Further, by Conditions B1 and B3, the consistency of $(\bar{\mu}_h, \bar{\sigma}_h)$ for $h = 1, 2$ is covered by Kiefer and Wolfowitz (1956). More arguments can be found in the proof of Theorem 3 in Chen, Tan and Zhang (2008). This finishes the proof of Lemma 1.

In the next lemma, we strengthen the conclusion of Lemma 1 by providing an order assessment.

Lemma 2. *(Convergence rate with non-zero mixing proportion) Assume the same conditions as in Lemma 1. If* $\bar{\lambda} - \lambda_0 = o_p(1)$ *for some* $\lambda_0 \in (0, 1]$*, then* $\bar{\mu}_h = O_p(n^{-1/2})$ *and* $\bar{\sigma}_h - 1 = O_p(n^{-1/2})$*,* $h = 1, 2$*.*

Proof. Recall that

$$l_{n1}(\mu_1, \sigma_1) = \sum_{i=1}^{n_1} \log f(x_{1i}; \mu_1, \sigma_1)$$

and

$$l_{n2}(\lambda, \mu_1, \mu_2, \sigma_1, \sigma_2) = \sum_{i=1}^{n_2} \log\{(1 - \lambda) f(x_{2i}; \mu_1, \sigma_1) + \lambda f(x_{2i}; \mu_2, \sigma_2)\}.$$

Further let

$$R_{n1}(\mu_1, \sigma_1) = l_{n1}(\mu_1, \sigma_1) - l_{n1}(0, 1)$$

and

$$R_{n2}(\lambda, \mu_1, \mu_2, \sigma_1, \sigma_2) = l_{n2}(\lambda, \mu_1, \mu_2, \sigma_1, \sigma_2) - l_{n2}(1, 0, 0, 1, 1).$$

Then,

$$\begin{aligned} pl_n(\bar{\lambda}, \bar{\mu}_1, \bar{\mu}_2, \bar{\sigma}_1, \bar{\sigma}_2) - pl_n(1, 0, 0, 1, 1) =& R_{n1}(\bar{\mu}_1, \bar{\sigma}_1) + R_{n2}(\bar{\lambda}, \bar{\mu}_1, \bar{\mu}_2, \bar{\sigma}_1, \bar{\sigma}_2) \\ &+ p_n(\bar{\sigma}_2) - p_n(1) + p(\bar{\lambda}) - p(1). \end{aligned}$$

Next we derive an upper bound for $pl_n(\bar{\lambda}, \bar{\mu}_1, \bar{\mu}_2, \bar{\sigma}_1, \bar{\sigma}_2) - pl_n(1, 0, 0, 1, 1)$. Together with the lower bound c, we get the order assessment of $\bar{\mu}_h$ and $\bar{\sigma}_h$, $h = 1, 2$.

We first find an approximation for $R_{n1}(\bar\mu_1, \bar\sigma_1)$. From Lemma 1, we have the consistency results $\bar\mu_h = o_p(1)$ and $\bar\sigma_h - 1 = o_p(1)$, $h = 1, 2$. Applying the second Taylor expansion to $R_{n1}(\bar\mu_1, \bar\sigma_1)$ around $(0, 1)$, and the weak law of large numbers with Conditions B2 and B4, we get that

$$\begin{aligned}
R_{n1}(\bar\mu_1, \bar\sigma_1) =& l_{n1}(\bar\mu_1, \bar\sigma_1) - l_{n1}(0, 1) \\
=& \frac{\partial l_{n1}(0,1)}{\partial \mu_1}\bar\mu_1 + \frac{\partial l_{n1}(0,1)}{\partial \sigma_1}(\bar\sigma_1 - 1) \\
&+ \frac{1}{2}\frac{\partial^2 l_{n1}(0,1)}{\partial \mu_1^2}\bar\mu_1^2 + \frac{\partial^2 l_{n1}(0,1)}{\partial \mu_1 \partial \sigma_1}\bar\mu(\bar\sigma_1 - 1) + \frac{1}{2}\frac{\partial^2 l_{n1}(0,1)}{\partial \sigma_1^2}(\bar\sigma_1 - 1)^2 \\
&+ o_p(n)\{\bar\mu_1^2 + (\bar\sigma_1 - 1)^2\}. \qquad (22)
\end{aligned}$$

Let

$$Y_{hi} = \frac{\partial f(x_{hi}; 0, 1)/\partial \mu}{f(x_{hi}, 0, 1)} \quad \text{and} \quad Z_{hi} = \frac{\partial f(x_{hi}; 0, 1)/\partial \sigma}{f(x_{hi}, 0, 1)}, \;\; h = 1, 2, \;\; i = 1, \ldots, n_h.$$

Then we can verify that

$$\frac{\partial l_{n1}(0,1)}{\partial \mu_1}\bar\mu_1 + \frac{\partial l_{n1}(0,1)}{\partial \sigma_1}(\bar\sigma_1 - 1) = \sum_{i=1}^{n_1}\left\{\bar\mu_1 Y_{1i} + (\bar\sigma_1 - 1)Z_{1i}\right\}. \qquad (23)$$

Further,

$$\begin{aligned}
\frac{\partial^2 l_{n1}(0,1)}{\partial \mu_1^2} =& \sum_{i=1}^{n_1}\left[\frac{\partial^2 f(x_{1i}; 0, 1)/\partial \mu_1^2}{f(x_{1i}; 0, 1)} - \left\{\frac{\partial f(x_{1i}; 0, 1)/\partial \mu_1}{f(x_{1i}; 0, 1)}\right\}^2\right] \\
=& -\sum_{i=1}^{n_1} Y_{1i}^2 + o_p(n), \qquad (24)
\end{aligned}$$

where in the last step, we have used the definition of Y_{1i} and the weak law of large numbers. Similarly to (24), we have that

$$\frac{\partial^2 l_{n1}(0,1)}{\partial\mu_1\partial\sigma_1} = -\sum_{i=1}^{n_1} Y_{1i}Z_{1i} + o_p(n) \text{ and } \frac{\partial^2 l_{n1}(0,1)}{\partial\sigma_1^2} = -\sum_{i=1}^{n_1} Z_{1i}^2 + o_p(n). \quad (25)$$

By Condition B5 and the weak law of large numbers, we get that

$$\left[\sum_{i=1}^{n_1}\{\bar\mu_1 Y_{1i} + (\bar\sigma_1 - 1)Z_{1i}\}^2\right] o_p(1) = o_p(n)\{\bar\mu_1^2 + \bar\mu_1(\bar\sigma_1 - 1) + (\bar\sigma_1 - 1)^2\}. \quad (26)$$

Combining (22)–(26), we obtain that

$$\begin{aligned} R_{n1}(\bar\mu_1, \bar\sigma_1) = &\sum_{i=1}^{n_1}\{\bar\mu_1 Y_{1i} + (\bar\sigma_1 - 1)Z_{1i}\} \\ &- \frac{1}{2}\left[\sum_{i=1}^{n_1}\{\bar\mu_1 Y_{1i} + (\bar\sigma_1 - 1)Z_{1i}\}^2\right]\{1 + o_p(1)\}. \end{aligned} \quad (27)$$

We next study $p_n(\bar\sigma_2) - p_n(1)$. By Condition C3, we have

$$p_n(\bar\sigma_2) - p_n(1) = o_p(n^{1/2})(\bar\sigma_2 - 1) \le o_p(1) + o_p(n)(\bar\sigma_2 - 1)^2. \quad (28)$$

We now find an upper bound for $R_{n2}(\bar\lambda, \bar\mu_1, \bar\mu_2, \bar\sigma_1, \bar\sigma_2)$. Write $R_{n2}(\bar\lambda, \bar\mu_1, \bar\mu_2, \bar\sigma_1, \bar\sigma_2) = \sum_{i=1}^{n_2} \log(1 + \delta_i)$ with

$$\delta_i = \frac{(1-\bar\lambda)\{f(x_{2i}; \bar\mu_1, \bar\sigma_1) - f(x_{2i}; 0, 1)\} + \bar\lambda\{f(x_{2i}; \bar\mu_2, \bar\sigma_2) - f(x_{2i}; 0, 1)\}}{f(x_{2i}; 0, 1)}.$$

By the inequality $\log(1 + x) \le x - x^2/2 + x^3/3$, we have

$$R_{n2}(\bar\lambda, \bar\mu_1, \bar\mu_2, \bar\sigma_1, \bar\sigma_2) \le \sum_{i=1}^{n_2}\delta_i - \sum_{i=1}^{n_2}\delta_i^2/2 - \sum_{i=1}^{n_2}\delta_i^3/3.$$

Let $\bar m_1 = (1 - \bar\lambda)\bar\mu_1 + \bar\lambda\bar\mu_2$, $\bar m_2 = (1 - \bar\lambda)(\bar\sigma_1 - 1) + \bar\lambda(\bar\sigma_2 - 1)$. By the consistency $\bar{\mu}_h = o_p(1)$ and $\bar{\sigma}_h - 1 = o_p(1)$, $h = 1, 2$, we have that

$$\bar{m}_1 = o_p(1) \text{ and } \bar{m}_2 = o_p(1).$$

Applying the first order Taylor expansion to $f(x_{2i}; \bar{\mu}_h, \bar{\sigma}_h)$, we find that

$$\delta_i = \bar{m}_1 Y_{2i} + \bar{m}_2 Z_{2i} + \varepsilon_{ni}$$

and the remainder term $\varepsilon_n = \sum_{i=1}^{n_2} \varepsilon_{ni}$ satisfies

$$\varepsilon_n = O_p(n_2^{1/2}) \sum_{h=1}^{2} \left\{\bar{\mu}_h^2 + (\bar{\sigma}_h - 1)^2\right\} = O_p(n^{1/2}) \sum_{h=1}^{2} \left\{\bar{\mu}_h^2 + (\bar{\sigma}_h - 1)^2\right\}.$$

From Condition B4 and the weak law of large numbers, after some straightforward algebra, we get

$$\begin{aligned} R_{n2}(\bar{\lambda}, \bar{\mu}_1, \bar{\mu}_2, \bar{\sigma}_1, \bar{\sigma}_2) \;\; \leq \;\; & \sum_{i=1}^{n_2} \{\bar{m}_1 Y_{2i} + \bar{m}_2 Z_{2i}\} \\ & -(1/2) \sum_{i=1}^{n_2} \{\bar{m}_1 Y_{2i} + \bar{m}_2 Z_{2i}\}^2 \{1 + o_p(1)\} \\ & +(1/3) \sum_{i=1}^{n_2} \{\bar{m}_1 Y_{2i} + \bar{m}_2 Z_{2i}\}^3 + O_p(\varepsilon_n). \end{aligned}$$

For the cubic term in the upper bound of $R_{n2}(\bar{\lambda}, \bar{\mu}_1, \bar{\mu}_2, \bar{\sigma}_1, \bar{\sigma}_2)$, we have that

$$\begin{aligned} \sum_{i=1}^{n_2} (\bar{m}_1 Y_{2i} + \bar{m}_2 Z_{2i})^3 = & \sum_{i=1}^{n_2} \{\bar{m}_1^3 Y_{2i}^3 + 3\bar{m}_1^2 \bar{m}_2 Y_{2i}^2 Z_{2i} + 3\bar{m}_1 \bar{m}_2^2 Y_{2i} Z_{2i}^2 + \bar{m}_2^3 Z_{2i}^3\} \\ = & o_p(1) \left[\sum_{i=1}^{n_2} \left\{\bar{m}_1^2 Y_{2i}^3 + 3\bar{m}_1^2 Y_{2i}^2 Z_{2i} + 3\bar{m}_2^2 Y_{2i} Z_{2i}^2 + \bar{m}_2^2 Z_{2i}^3\right\} \right] \\ = & o_p(n)(\bar{m}_1^2 + \bar{m}_2^2), \end{aligned}$$

where in the second step, we have used the fact that $\bar{m}_1 = o_p(1)$ and $\bar{m}_2 = o_p(1)$, and in the third step we have used Condition B4.

Hence,

$$\begin{aligned} R_{n2}(\bar{\lambda}, \bar{\mu}_1, \bar{\mu}_2, \bar{\sigma}_1, \bar{\sigma}_2) \leq & \sum_{i=1}^{n_2} \bar{m}_1 Y_{2i} + \sum_{i=1}^{n_2} \bar{m}_2 Z_{2i} \\ & -1/2 \sum_{i=1}^{n_2} (\bar{m}_1 Y_{2i} + \bar{m}_2 Z_{2i})^2 \{1 + o_p(1)\} \\ & + O_p(\varepsilon_n). \end{aligned} \tag{29}$$

Combining (27)–(29) and Condition C1, we get

$$\begin{aligned} & pl_n(\bar{\lambda}, \bar{\mu}_1, \bar{\mu}_2, \bar{\sigma}_1, \bar{\sigma}_2) - pl_n(1, 0, 0, 1, 1) \\ \leq & \sum_{i=1}^{n_1} \{\bar{\mu}_1 Y_{1i} + (\bar{\sigma}_1 - 1) Z_{1i}\} + \sum_{i=1}^{n_2} (\bar{m}_1 Y_{2i} + \bar{m}_2 Z_{2i}) \\ & -1/2 \left[\sum_{i=1}^{n_1} \{\bar{\mu}_1 Y_{1i} + (\bar{\sigma}_1 - 1) Z_{1i}\}^2 + \sum_{i=1}^{n_2} (\bar{m}_1 Y_{2i} + \bar{m}_2 Z_{2i})^2 \right] \{1 + o_p(1)\} \\ & + O_p(\varepsilon_n) + o_p(1). \end{aligned}$$

Condition $\bar{\lambda} - \lambda_0 = o_p(1)$ with $\lambda_0 \in (0, 1]$ implies that

$$O_p(\varepsilon_n) = O_p(n^{1/2}) \left\{ \bar{\mu}_1^2 + (\bar{\sigma}_1 - 1)^2 + \bar{m}_1^2 + \bar{m}_2^2 \right\} = o_p(n) \left\{ \bar{\mu}_1^2 + (\bar{\sigma}_1 - 1)^2 + \bar{m}_1^2 + \bar{m}_2^2 \right\}.$$

Hence, by the weak law of large numbers with Condition B5,

$$\begin{aligned} c \leq\ & pl_n(\bar{\lambda}, \bar{\mu}_1, \bar{\mu}_2, \bar{\sigma}_1, \bar{\sigma}_2) - pl_n(1, 0, 0, 1, 1) \\ \leq\ & \sum_{i=1}^{n_1} \{\bar{\mu}_1 Y_{1i} + (\bar{\sigma}_1 - 1) Z_{1i}\} + \sum_{i=1}^{n_2} (\bar{m}_1 Y_{2i} + \bar{m}_2 Z_{2i}) \\ & - 1/2 \left[\sum_{i=1}^{n_1} \{\bar{\mu}_1 Y_{1i} + (\bar{\sigma}_1 - 1) Z_{1i}\}^2 + \sum_{i=1}^{n_2} (\bar{m}_1 Y_{2i} + \bar{m}_2 Z_{2i})^2 \right] \{1 + o_p(1)\} + o_p(1) \\ \leq\ & 1/2 (\boldsymbol{U}_{\boldsymbol{1n}}^{\boldsymbol{\tau}} \boldsymbol{W}^{\boldsymbol{-1}} \boldsymbol{U}_{\boldsymbol{1n}} + \boldsymbol{U}_{\boldsymbol{2n}}^{\boldsymbol{\tau}} \boldsymbol{W}^{\boldsymbol{-1}} \boldsymbol{U}_{\boldsymbol{2n}}) + o_p(1), \end{aligned} \tag{30}$$

where "τ" denotes the transpose of a vector or matrix,

$$\boldsymbol{U_{hn}} = n_h^{-1/2} \left(\sum_{i=1}^{n_h} Y_{hi}, \sum_{i=1}^{n_h} Z_{hi} \right)^{\tau}, \quad \boldsymbol{W} = \begin{pmatrix} \sigma_Y^2 & \sigma_{Y,Z} \\ \sigma_{Y,Z} & \sigma_Z^2 \end{pmatrix}$$

with $\sigma_Y^2 = Var(Y_{11})$, $\sigma_Z^2 = Var(Z_{11})$ and $\sigma_{Y,Z} = Cov(Y_{11}, Z_{11})$. Therefore,

$$\bar{\mu}_1 = O_p(n^{-1/2}),\ \bar{\sigma}_1 - 1 = O_p(n^{-1/2}),\ \bar{m}_1 = O_p(n^{-1/2}),\ \bar{m}_2 = O_p(n^{-1/2}).$$

Any values of $(\bar{\mu}_1, \bar{\sigma}_1 - 1, \bar{m}_1, \bar{m}_2)$ out of this range will violate the inequality. With the condition that $\bar{\lambda} - \lambda_0 = o_p(1)$ for some $\lambda_0 \in (0, 1]$, we have

$$\bar{\mu}_1 = O_p(n^{-1/2}),\ \bar{\sigma}_1 - 1 = O_p(n^{-1/2}),\ \bar{\mu}_2 = O_p(n^{-1/2}),\ \bar{\sigma}_2 - 1 = O_p(n^{-1/2}).$$

□

Let $(\bar{\lambda}, \bar{\mu}_1, \bar{\mu}_2, \bar{\sigma}_1, \bar{\sigma}_2)$ be some estimators of $(\lambda, \mu_1, \mu_2, \sigma_1, \sigma_2)$ as before, and let

$$\bar{\omega}_i = \frac{\bar{\lambda} f(x_{2i}; \bar{\mu}_2, \bar{\sigma}_2)}{(1 - \bar{\lambda}) f(x_{2i}; \bar{\mu}_1, \bar{\sigma}_1) + \bar{\lambda} f(x_{2i}; \bar{\mu}_2, \bar{\sigma}_2)}.$$

We define

$$H_n(\lambda) = (n_2 - \sum_{i=1}^{n_2} \bar{\omega}_i) \log(1 - \lambda) + \sum_{i=1}^{n_2} \bar{\omega}_i \log(\lambda) + p(\lambda).$$

The EM-test updates λ by searching $\bar{\lambda}^* = \arg\max_{\lambda} H_n(\lambda)$.

Lemma 3. *(Consistency of mixing proportion after iteration) Under the same conditions as in Lemma A.1, and if* $\bar{\lambda} - \lambda_0 = o_p(1)$ *for some* $\lambda_0 \in (0, 1]$*, then* $\bar{\lambda}^* - \lambda_0 = o_p(1)$*.*

For the proof of Lemma 3 is similar to that of Lemma A3 of Li, Chen and Marriott (2009) and hence is omitted.

Proof of Theorem 1

Proof. For any $k \leq K$, due to the monotonicity property of the EM-algorithm that the penalized likelihood increases after each iteration (see Proposition 1), we have

$$pl_n(\lambda_j^{(k)}, \mu_{1j}^{(k)}, \mu_{2j}^{(k)}, \sigma_{1j}^{(k)}, \sigma_{2j}^{(k)}) \geq pl_n(\lambda_j^{(1)}, \mu_{1j}^{(1)}, \mu_{2j}^{(1)}, \sigma_{1j}^{(1)}, \sigma_{2j}^{(1)}) \geq pl_n(\lambda_j, 0, 0, 1, 1).$$

That is,

$$pl_n(\lambda_j^{(k)}, \mu_{1j}^{(k)}, \mu_{2j}^{(k)}, \sigma_{1j}^{(k)}, \sigma_{2j}^{(k)}) - pl_n(1, 0, 0, 1, 1) \geq p(\lambda_j) - p(1) > -\infty.$$

Then by Lemmas 1-3, Theorem 1 holds. □

Proof of Theorem 2

Proof. Under Conditions B2, B4 and B5, applying some of the classic results about regular models, we have (Serfling, 1980)

$$\begin{aligned} &\sup_{\mu,\sigma} pl_n(1, \mu, \mu, \sigma, \sigma) - pl_n(1, 0, 0, 1, 1) \\ =&1/2(\sqrt{\rho_1}\boldsymbol{U_{1n}^{\tau}} + \sqrt{\rho_2}\boldsymbol{U_{2n}^{\tau}})\boldsymbol{W^{-1}}(\sqrt{\rho_1}\boldsymbol{U_{1n}} + \sqrt{\rho_2}\boldsymbol{U_{2n}}) + o_p(1). \qquad (31) \end{aligned}$$

Due to the properties established in Theorem 1, $(\lambda_j^{(K)}, \mu_{1j}^{(K)}, \mu_{2j}^{(K)}, \sigma_{1j}^{(K)})$ satisfies the conditions of Lemma 2 and hence from (30), we have

$$\begin{aligned} & pl_n(\lambda_j^{(K)}, \mu_{1j}^{(K)}, \mu_{2j}^{(K)}, \sigma_{1j}^{(K)}, \sigma_{2j}^{(K)}) - pl_n(1,0,0,1,1) \\ \leq\ & 1/2(\boldsymbol{U_{1n}^{\tau} W^{-1} U_{1n}} + \boldsymbol{U_{2n}^{\tau} W^{-1} U_{2n}}) + o_p(1). \qquad (32) \end{aligned}$$

From (31) and (32), we find that

$$\begin{aligned} M_n^{(K)}(\lambda_j) =& 2\{pl_n(\lambda_j^{(K)}, \mu_{1j}^{(K)}, \mu_{2j}^{(K)}, \sigma_{1j}^{(K)}, \sigma_{2j}^{(K)}) - pl_n(1,0,0,1,1)\} \\ & - 2\{\sup_{\mu,\sigma} pl_n(1,\mu,\mu,\sigma,\sigma) - pl_n(1,0,0,1,1)\} \\ \leq & (\sqrt{\rho_2}\boldsymbol{U_{1n}^{\tau}} - \sqrt{\rho_1}\boldsymbol{U_{2n}^{\tau}})\boldsymbol{W^{-1}}(\sqrt{\rho_2}\boldsymbol{U_{1n}} - \sqrt{\rho_1}\boldsymbol{U_{2n}}) + o_p(1). \end{aligned}$$

Since the upper bound does not depend on λ_j, we further have

$$EM_n^{(K)} \leq (\sqrt{\rho_2}\boldsymbol{U_{1n}^{\tau}} - \sqrt{\rho_1}\boldsymbol{U_{2n}^{\tau}})\boldsymbol{W^{-1}}(\sqrt{\rho_2}\boldsymbol{U_{1n}} - \sqrt{\rho_1}\boldsymbol{U_{2n}}) + o_p(1). \quad (33)$$

Next, we show the upper bound for $EM_n^{(K)}$ is achievable. Recall that $\lambda_1 = 1$. Let

$$(\tilde{\mu}_h, \tilde{\sigma}_h - 1) = n_h^{-1/2}\boldsymbol{W}^{-1}\boldsymbol{U}_{hn},\ \ h = 1,2.$$

Since the EM-iteration always increases the penalized likelihood and $\lambda_1 = 1$, we have that

$$EM_n^{(K)} \geq M_n^{(K)}(\lambda_1) \geq M_n^{(1)}(\lambda_1) \geq 2\{pl_n(\lambda_1, \tilde{\mu}_1, \tilde{\mu}_2, \tilde{\sigma}_1, \tilde{\sigma}_2) - \sup_{\mu,\sigma} pl_n(1,\mu,\mu,\sigma,\sigma)\}. \quad (34)$$

Note that it is easy to verify that

$$\tilde{\mu}_h = O_p(n^{-1/2}) \ \text{ and } \ \tilde{\sigma}_h - 1 = O_p(n^{-1/2}), \quad h = 1,2.$$

With this order assessment and applying the second order Taylor expansion, we have that

$$\begin{aligned}&2\{pl_n(\lambda_1,\tilde{\mu}_1,\tilde{\mu}_2,\tilde{\sigma}_1,\tilde{\sigma}_2)-\sup_{\mu,\sigma}pl_n(1,\mu,\mu,\sigma,\sigma)\}\\=&(\sqrt{\rho_2}\boldsymbol{U}_{\boldsymbol{1n}}^{\boldsymbol{\tau}}-\sqrt{\rho_1}\boldsymbol{U}_{\boldsymbol{2n}}^{\boldsymbol{\tau}})\boldsymbol{W}^{\boldsymbol{-1}}(\sqrt{\rho_2}\boldsymbol{U}_{\boldsymbol{1n}}-\sqrt{\rho_1}\boldsymbol{U}_{\boldsymbol{2n}})+o_p(1).\end{aligned}\tag{35}$$

Combining (33)–(35), we get

$$EM_n^{(K)}=(\sqrt{\rho_2}\boldsymbol{U}_{\boldsymbol{1n}}^{\boldsymbol{\tau}}-\sqrt{\rho_1}\boldsymbol{U}_{\boldsymbol{2n}}^{\boldsymbol{\tau}})\boldsymbol{W}^{\boldsymbol{-1}}(\sqrt{\rho_2}\boldsymbol{U}_{\boldsymbol{1n}}-\sqrt{\rho_1}\boldsymbol{U}_{\boldsymbol{2n}})+o_p(1).$$

By central limit theorem, $\boldsymbol{W}^{\boldsymbol{-1/2}}\boldsymbol{U}_{\boldsymbol{hn}}$ for $h=1,2$ converges to $N(0,1)$ in distribution. Note that $\boldsymbol{U}_{\boldsymbol{1n}}$ and $\boldsymbol{U}_{\boldsymbol{2n}}$ are independent. Consequently, the null limiting distribution of $EM_n^{(K)}$ is χ_2^2. □

Proof of Theorem 3

Proof. Without loss of generality, we assume that the null model is $f(x;0,1)$. Then

$$E(U^2)=\sigma_0^{-2}\sigma_Y^2,\quad E(V^2)=\sigma_0^{-2}\sigma_Z^2,\quad E(UV)=\sigma_0^{-2}\sigma_{Y,Z}.\tag{36}$$

Further the local alternative H_a^n in (2.2) of the main paper becomes

$$H_a^n:\lambda=\lambda_0,\ (\mu_1,\sigma_1)=(0,1),\ (\mu_2,\sigma_2)=(n_2^{-1/2}\Delta_1/\sigma_0,1+n_2^{-1/2}\Delta_2/\sigma_0).$$

Let

$$\Lambda_n=\sum_{i=1}^{n_2}\log\frac{(1-\lambda_0)f(x_{2i};0,1)+\lambda_0 f(x_{2i};n_2^{-1/2}\Delta_1/\sigma_0,1+n_2^{-1/2}\Delta_2/\sigma_0)}{f(x_{2i};0,1)}.$$

Under Conditions B2 and B4, applying second order approximation, we can verify that under the null model,

$$\Lambda_n = \lambda_0\sigma_0^{-1}n_2^{-1/2}\sum_{i=1}^{n_2}(\Delta_1 Y_{2i} + \Delta_2 Z_{2i}) - 0.5\lambda_0^2\sigma_0^{-2}(\Delta_1, \Delta_2)\,\boldsymbol{W}\begin{pmatrix}\Delta_1\\ \Delta_2\end{pmatrix} + o_p(1).$$

Hence under the null model, $\Lambda_n \xrightarrow{d} N(-0.5c^2, c^2)$, where

$$c^2 = \lambda_0^2\sigma_0^{-2}(\Delta_1^2\sigma_Y^2 + 2\Delta_1\Delta_2\sigma_{Y,Z} + \Delta_2^2\sigma_Z^2).$$

Therefore, the local alternative H_a^n is contiguous to the null distribution (Le Cam and Yang, 1990 and Example 6.5 of van der Vaart, 2000). By Le Cam's contiguity theory, the limiting distribution of $EM_n^{(K)}$ under H_a^n is determined by the joint limiting distribution of $\boldsymbol{W^{-1/2}}(\sqrt{\rho_2}\boldsymbol{U_{1n}} - \sqrt{\rho_1}\boldsymbol{U_{2n}})$ and Λ_n under the null model.

By central limit theorem and Slutsky's theorem, the joint limiting distribution of $\boldsymbol{W^{-1/2}}(\sqrt{\rho_2}\boldsymbol{U_{1n}} - \sqrt{\rho_1}\boldsymbol{U_{2n}})$ and Λ_n under the null model is multivariate normal

$$\mathcal{N}_3\left(\begin{pmatrix}\boldsymbol{0}\\ -0.5c^2\end{pmatrix},\ \begin{pmatrix}\boldsymbol{I} & -\sqrt{\rho_1}\lambda_0\sigma_0^{-1}\boldsymbol{W^{1/2}}(\Delta_1,\Delta_2)^\tau\\ -\sqrt{\rho_1}\lambda_0\sigma_0^{-1}(\Delta_1,\Delta_2)\boldsymbol{W}^{1/2} & c^2\end{pmatrix}\right).$$

By Le Cam's third lemma (van der Vaart, 2000), we have under H_a^n,

$$\boldsymbol{W^{-1/2}}(\sqrt{\rho_2}\boldsymbol{U_{1n}} - \sqrt{\rho_1}\boldsymbol{U_{2n}}) \to \mathcal{N}_2\left(-\sqrt{\rho_1}\lambda_0\sigma_0^{-1}\boldsymbol{W}^{1/2}(\Delta_1,\Delta_2)^\tau,\ \ \boldsymbol{I}\right).$$

Since $EM_n^{(K)} = (\sqrt{\rho_2}\boldsymbol{U_{1n}^\tau} - \sqrt{\rho_1}\boldsymbol{U_{2n}^\tau})\boldsymbol{W^{-1}}(\sqrt{\rho_2}\boldsymbol{U_{1n}} - \sqrt{\rho_1}\boldsymbol{U_{2n}}) + o_p(1)$ holds under the null, by Le Cam's first lemma (van der Vaart, 2000), $EM_n^{(K)} = (\sqrt{\rho_2}\boldsymbol{U}_{\mathbf{1n}}^{\boldsymbol{\tau}} - \sqrt{\rho_1}\boldsymbol{U}_{\mathbf{2n}}^{\boldsymbol{\tau}})\boldsymbol{W}^{-\mathbf{1}}(\sqrt{\rho_2}\boldsymbol{U}_{\mathbf{1n}} - \sqrt{\rho_1}\boldsymbol{U}_{\mathbf{2n}}) + o_p(1)$ still holds under H_a^n. Therefore, the limiting distribution of $EM_n^{(K)}$ under the local alternative H_a^n is $\chi_2^2(c_0^2)$, where

$$c_0^2 = \rho_1 c^2 = \lambda_0^2 \rho_1 \sigma_0^{-2}(\Delta_1^2\sigma_Y^2 + 2\Delta_1\Delta_2\sigma_{Y,Z} + \Delta_2^2\sigma_Z^2).$$

With (36), we further have

$$c_0^2 = \lambda_0^2\rho_1\left\{\Delta_1^2 E(U^2) + 2\Delta_1\Delta_2 E(UV) + \Delta_2^2 E(V^2)\right\}.$$

This finishes the proof. □

References

Chen, J., Tan, X. and Zhang, R. (2008). Inference for normal mixtures in mean and variance. *Statist. Sinica* **18**, 443–465.

Kiefer, J. and Wolfowitz, J. (1956). Consistency of the maximum likelihood estimates in the presence of infinitely many incidental parameters. *Ann. Math. Statist.* **27**, 887–906.

Le Cam, L. and Yang, G. L. (1990). *Asymptotics in Statistics: Some Basic Concepts.* Springer-Verlag, New York.

Li, P., Chen, J. and Marriott, P. (2009). Non-finite Fisher information and homogeneity: The EM approach. *Biometrika* **96**, 411–426.

Serfling, R. J. (1980). *Approximation Theorem of Mathematical Statistics.* Wiley, New York.

van der Vaart, A.W. (2000). *Asymptotic Statistics.* Cambridge University Press, New York.

Wald, A. (1949). Note on the consistency of the maximum likelihood estimate. *Ann. Math. Statist.* **20**, 595–601.

School of Statistics, East China Normal University, Shanghai 200241, China.

E-mail: liuguanfu07@163.com

Department of Mathematics and Statistics, York University, Toronto, ON, M3J 1P3, Canada.

E-mail: yuejiao@mathstat.yorku.ca

Department of Statistics and Actuarial Science, University of Waterloo, Waterloo, ON, N2L 3G1, Canada

E-mail: pengfei.li@uwaterloo.ca

School of Statistics, East China Normal University, Shanghai 200241, China.

E-mail: xlpu@stat.ecnu.edu.cn