On Asymptotic Distribution of Expectation-Maximization estimators

Sihyung Park and Sohyeon Kim

1 Introduction

In the second chapter of the textbook (Boos and Stefanski 2013), as a method of solving likelihood equation, expectation-maximization was introduced. The rationale of the EM algorithm is to introduce a latent variable $\mathbf{Z}$ to maximize a complex likelihood $L(\theta; \mathbf{Y})$. By doing so, the complete data likelihood $L_C(\theta; \mathbf{Y}, \mathbf{Z})$ becomes more accessible form, sometimes even leads to analytic forms of estimators.

However, unlike other methods that are introduced in the same chapter, estimators from EM algorithm were deemed deterministic, and only the greedieness of the algorithm was briefly covered.

Surprisingly, there were few to no articles that covered asymptotic normality of EM estimators. Here, we would like to review an article (Wang et al. 2014) that covered the topic directly, for a modified version of the EM algorithm. (Wang et al. 2014) consists of three parts; the first part is about a modified version of EM algorithm that can be appplied to high-dimensional and sparse settings; the second part focuses on computational property; the third part focuses on defining test statistics based on EM estimates and deriving their asymptotic distribution under the null hypothesis.

In this post, we focused primarily on the third part of (Wang et al. 2014), i.e. the asymptotic distribution of EM estimators. Luckily, the modified EM used in the paper is a generalization of ordinary EM. This allows us to apply the result to derive asymptotic distribution of ordinary EM estimators. Upon reviewing the content from (Wang et al. 2014), we translated the notations into familiar form as in (Boos and Stefanski 2013). In addition, we set the settings that we covered in the class so that the results becomes clearer. Furthermore, application to and numerical results from specific latent variable model is proposed.

The rest of the post is organized as follows. In section 2, we review the basics of EM algorithm. We introduce the modified version as well and show that if no sparsity is assumed on parameter, it is equivalent to ordinary EM algorithm that we covered in the class. In section 3, we introduce score-type and Wald-type test statistic. We provide brief sketch of the proofs of asymptotic normality of them as well. We translated the result into familiar notation that was used in the class. In the last section 4, we apply the result to latent variable model and comment on the reasonability of conditions for asymptotic normality of test statistics. Numerical result supporting main theorems are provided in conjunction.

2 The Expectation-Maximization Algorithm

2.1 Ordinary EM

Suppose that $\mathbf{Y} = (Y_1, \cdots, Y_n)$ is an IID sample of size n from a density $f(y; \boldsymbol{\theta})$, $\boldsymbol{\theta} \in \mathbb{R}^d$. The log-likelihood becomes

$$\ell(\boldsymbol{\theta}) = \sum_{i=1}^{n} \log f(y_i; \boldsymbol{\theta}).$$

We assume that there exist latent variable $\mathbf{Z} = (Z_1, \cdots, Z_n)$ and the complete data $\{(X_1, Z_1), \cdots, (X_n, Z_n)\}$ is in fact randomly generated from a joint density $h(y, z; \boldsymbol{\theta})$. In this setting, the complete data log-likelihood is given as

$$\ell_C(\boldsymbol{\theta}) = \sum_{i=1}^{n} \log h(y_i, z_i; \boldsymbol{\theta}).$$

To derive a (possibly local) maximizer $\hat{\boldsymbol{\theta}}$ of $\ell(\boldsymbol{\theta})$, we use the fact that

$$\frac{1}{n}\left(\ell(\boldsymbol{\theta}) - \ell(\boldsymbol{\theta}')\right) \geq \mathbb{E}_{\boldsymbol{\theta}'}\left[\ell_C(\boldsymbol{\theta}) | \mathbf{Y} = \mathbf{y}\right] - c$$

for some constant c that depends only on $\mathbf{y}$ and $\boldsymbol{\theta}'$. This inequality can be acheived by applying Jensen's inequality. Thus, updating the value of $\boldsymbol{\theta}'$ to be that maximizes $Q_n(\boldsymbol{\theta}; \boldsymbol{\theta}') := E_{\boldsymbol{\theta}'}\left[\ell_C(\boldsymbol{\theta}) | \mathbf{Y} = \mathbf{y}\right]$, it will increase the value of the log-likelihood as a result.

The ordinary EM algorithm is defined as Algorithm 1. It is proved in (Boos and Stefanski 2013) that the algorithm is greedy. That is, $\ell(\boldsymbol{\theta}^{(t+1)}) \geq \ell(\boldsymbol{\theta}^{(t)})$ for all t.

Algorithm 1: Ordinary EM

```
Input: Maximum number of iteration $T$
for $t = 0$ to $T - 1$ do
    E-Step: Evaluate $Q_n(\boldsymbol{\theta}; \boldsymbol{\theta}^{(t)}) = E_{\boldsymbol{\theta}^{(t)}}\left[\ell_C(\boldsymbol{\theta}) | \mathbf{Y} = \mathbf{y}\right]$
    M-Step: $\boldsymbol{\theta}^{(t+1)} \leftarrow \arg\max_{\boldsymbol{\theta}} Q_n(\boldsymbol{\theta}; \boldsymbol{\theta}^{(t)})$
end
Output: EM estimate $\hat{\boldsymbol{\theta}} = \boldsymbol{\theta}^{(T)}$
```

2.2 High-dimensional sparse EM

(Wang et al. 2014) propose a generalized EM algorithm that can take sparse parameters into account for high-dimensional models. First, we define two functions to enforce sparsity of parameter.

$$\mathrm{supp}(\boldsymbol{\theta}, s) = \{j : \text{ index } j \text{ is of top-}s \text{ largest } |\theta_j|\text{'s}\}$$

$$[\mathrm{trunc}(\boldsymbol{\theta}, S)]_j = \begin{cases} \theta_j, \ j \in S \\ 0, \ j \notin S \end{cases}$$

The modified EM is provided as in Algorithm 2. Notice that if $\hat{s} = d$, then the algorithm is simplfied to ordinary EM. This allows us to use the statistical properties from Algorithm 2 directly to ordinary EM estimators.

Algorithm 2: High-dimensional sparse EM

```
Input: Sparsity parameter $\hat{s}$, maximum number of iterations T
Initialize: $\hat{S} \leftarrow \mathrm{supp}(\boldsymbol{\theta}^{\mathrm{init}}, \hat{s}), \boldsymbol{\theta}^{(0)} \leftarrow \mathrm{trunc}(\boldsymbol{\theta}^{\mathrm{init}}, \hat{S})$
for $t = 0$ to $T - 1$ do
    E-Step: Evaluate $Q_n(\boldsymbol{\theta}; \boldsymbol{\theta}^{(t)}) = E_{\boldsymbol{\theta}^{(t)}}\left[\ell_C(\boldsymbol{\theta}) | \mathbf{Y} = \mathbf{y}\right]$
    M-Step: $\boldsymbol{\theta}^{(t+0.5)} \leftarrow \arg\max_{\boldsymbol{\theta}} Q_n(\boldsymbol{\theta}; \boldsymbol{\theta}^{(t)})$
    T-Step: $\hat{S}^{(t+0.5)} \leftarrow \mathrm{supp}(\boldsymbol{\theta}^{(t+0.5)}, \hat{s}), \boldsymbol{\theta}^{(t+1)} \leftarrow \mathrm{trunc}(\boldsymbol{\theta}^{(t+0.5)}, \hat{S}^{(t+0.5)})$
end
Output: EM estimate $\hat{\boldsymbol{\theta}} = \boldsymbol{\theta}^{(T)}$
```

3 Test statistic based on EM estimators

The paper proposed a score-type statistic and a Wald-type statistic based on the estimator achieved from EM algorithm. For simplicity, denote T_S, T_W by score-type and Wald-type statistic, respectively. In addition, define $\nabla_1 Q_n(\boldsymbol{\theta}, \boldsymbol{\theta}')$ as the gradient with respect to the first argument, and similarly $\nabla_2 Q_n(\boldsymbol{\theta}, \boldsymbol{\theta}')$ as the gradient with respect to the second argument. Define $\nabla^2_{1,1} Q_n(\boldsymbol{\theta}, \boldsymbol{\theta}'), \nabla^2_{1,2} Q_n(\boldsymbol{\theta}, \boldsymbol{\theta}')$ as the second derivatives with respect to the first argument, and with respect to the first and then the second argument, respectively.

For simplicity, we let block components of any matrix A by

$$A = \begin{pmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{pmatrix},$$
$$A_{11} \in \mathbb{R}^{r \times r}, \ A_{12} \in \mathbb{R}^{r \times (d-r)}, \ A_{21} \in \mathbb{R}^{(d-r) \times r}, \ A_{22} \in \mathbb{R}^{(d-r) \times (d-r)},$$

and subvector of any vector $\mathbf{a}$ by

$$\mathbf{a} = \begin{bmatrix} \mathbf{a}_1 \\ \mathbf{a}_2 \end{bmatrix}, \ \mathbf{a}_1 \in \mathbb{R}^r, \ \mathbf{a}_2 \in \mathbb{R}^{(d-r)}.$$

3.1 Score-type test statistic

Suppose we are testing the null $H_0 : \boldsymbol{\theta}_1 = 0$ from $\boldsymbol{\theta} = (\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)^T \in \mathbb{R}^d, \boldsymbol{\theta}_1 \in \mathbb{R}^r$.

Now, define a score-type test statistic T_S as

$$T_S = \sqrt{n} S_n(\tilde{\boldsymbol{\theta}}, \lambda) \left\{ \left[\mathbb{I}_r, -W(\tilde{\boldsymbol{\theta}}, \lambda)^T \right] \hat{V}_n(\tilde{\boldsymbol{\theta}}) \left[\mathbb{I}_r, -W(\tilde{\boldsymbol{\theta}}, \lambda)^T \right]^T \right\}^{-1/2},$$

where $\mathbb{I}_r \in \mathbb{R}^{r \times r}$ is an identity, $\tilde{\boldsymbol{\theta}} = (0, \hat{\boldsymbol{\theta}}_2)^T$, $\hat{\boldsymbol{\theta}}_2$ is a part of the estimator from Algorithm 2,

$$\begin{aligned} S_n(\boldsymbol{\theta}, \lambda) &= [\nabla_1 Q_n(\boldsymbol{\theta}, \boldsymbol{\theta})]_1 - W(\boldsymbol{\theta}, \lambda)^T [\nabla_1 Q_n(\boldsymbol{\theta}, \boldsymbol{\theta})]_2, \\ W(\boldsymbol{\theta}, \lambda) &= \operatorname{argmin}_{W \in \mathbb{R}^{(d-r) \times r}} \|W\|_{1,1} \\ &\qquad \text{subject to} \left\| \left[\hat{V}_n(\boldsymbol{\theta}) \right]_{21} - \left[\hat{V}_n(\boldsymbol{\theta}) \right]_{22} W \right\|_\infty \leq \lambda, \end{aligned}$$

and

$$\hat{V}_n(\tilde{\boldsymbol{\theta}}) = -\nabla^2_{1,1} Q_n(\tilde{\boldsymbol{\theta}}, \tilde{\boldsymbol{\theta}}) - \nabla^2_{1,2} Q_n(\tilde{\boldsymbol{\theta}}, \tilde{\boldsymbol{\theta}}).$$

Equivalently, we can write that

$$\sqrt{n} S_n(\tilde{\boldsymbol{\theta}}, \lambda) \sim AN \left(0, \left[\mathbb{I}_r, -W(\tilde{\boldsymbol{\theta}}, \lambda)^T \right] \hat{V}_n(\tilde{\boldsymbol{\theta}}) \left[\mathbb{I}_r, -W(\tilde{\boldsymbol{\theta}}, \lambda)^T \right]^T \right).$$

Under regularity conditions, it is shown that $T_S \xrightarrow{d} N(0, 1)$ as $n \to \infty$.

3.2 Wald-type test statistic

The Wald-type test statistic is defined as

$$T_W = \sqrt{n} h(\hat{\boldsymbol{\theta}}, \lambda) \left\{ \left[\mathbb{I}_r, -W(\tilde{\boldsymbol{\theta}}, \lambda)^T \right] \hat{V}_n(\tilde{\boldsymbol{\theta}}) \left[\mathbb{I}_r, -W(\tilde{\boldsymbol{\theta}}, \lambda)^T \right]^T \right\}^{1/2},$$

where

$$h(\hat{\boldsymbol{\theta}}, \lambda) = \hat{\boldsymbol{\theta}}_1 - \left\{ \left[-\hat{V}_n(\hat{\boldsymbol{\theta}}) \right]_{11} + W(\hat{\boldsymbol{\theta}}, \lambda)^T \left[\hat{V}_n(\hat{\boldsymbol{\theta}}) \right]_{21} \right\}^{-1} S_n(\hat{\boldsymbol{\theta}}, \lambda).$$

Equivalently, we can write that

$$\sqrt{n} h(\hat{\boldsymbol{\theta}}, \lambda) \sim AN \left(0, \left[\mathbb{I}_r, -W(\hat{\boldsymbol{\theta}}, \lambda)^T \right] \hat{V}_n(\hat{\boldsymbol{\theta}}) \left[\mathbb{I}_r, -W(\hat{\boldsymbol{\theta}}, \lambda)^T \right]^T \right).$$

Under regularity conditions, T_W also converges in distribution to $N(0, 1)$ as n grows to infinity.

3.3 Simplest case: testing the full parameter

For a simplest example, consider testing the null hypothesis $H_0 : \boldsymbol{\theta} = \boldsymbol{\theta}_0$. In this case,

$$T_S = \sqrt{n} S_n(\boldsymbol{\theta}_0, \lambda) \left\{ \hat{V}_n(\boldsymbol{\theta}_0) \right\}^{-1/2},$$

where $S_n(\boldsymbol{\theta}_0, \lambda) = \nabla_1 Q_n(\boldsymbol{\theta}_0, \boldsymbol{\theta}_0)$ and $\hat{V}_n(\boldsymbol{\theta}_0) = -\nabla^2 Q_n(\boldsymbol{\theta}_0, \boldsymbol{\theta}_0)$. Similarly, the Wald-type statistic is derived as

$$T_W = \sqrt{n} \left(h(\hat{\boldsymbol{\theta}}, \lambda) - \boldsymbol{\theta}_0 \right) \left\{ \hat{V}_n(\hat{\boldsymbol{\theta}}) \right\}^{1/2},$$

where $h(\hat{\boldsymbol{\theta}}) = \hat{\boldsymbol{\theta}} - \left\{ -\hat{V}_n(\hat{\boldsymbol{\theta}}) \right\}^{-1} \nabla_1 Q_n(\hat{\boldsymbol{\theta}}, \hat{\boldsymbol{\theta}})$

By using lemma (3.3), we can further simplify T_S and T_W as the following familiar forms:

$$
\begin{aligned}
T_S &= \frac{1}{\sqrt{n}} S(\mathbf{Y};\boldsymbol{\theta}_0)\left\{\hat{I}(\boldsymbol{\theta}_0)\right\}^{-1/2}, && (3.1)\\
T_W &= \sqrt{n}\left(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0 + \frac{1}{n}\hat{I}^{-1}(\hat{\boldsymbol{\theta}})S(\mathbf{Y};\hat{\boldsymbol{\theta}})\right)\left\{\hat{I}(\hat{\boldsymbol{\theta}})\right\}^{1/2} && (3.2)\\
&\overset{\text{mle}}{=} \sqrt{n}\left(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0\right)\left\{\hat{I}(\hat{\boldsymbol{\theta}})\right\}^{1/2},
\end{aligned}
$$

where $S(\mathbf{Y};\boldsymbol{\theta}) = \nabla \ell_n(\boldsymbol{\theta})$ is the ordinary score function and $\hat{I}(\boldsymbol{\theta}) = \hat{V}_n(\boldsymbol{\theta})$ to emphasize its link with the information matrix $I(\boldsymbol{\theta})$. The last equality holds if $\hat{\boldsymbol{\theta}}$ is the global maximizer.

3.4 Sketch of the proof of asymptotic normality

The following lemma states that $\nabla_1 Q_n(\boldsymbol{\theta},\boldsymbol{\theta})$ is in fact equivalent to $\nabla \ell_n(\boldsymbol{\theta})/n$, and $\hat{V}_n(\boldsymbol{\theta}_0)$ is an unbiased estimator of the information matrix $I(\boldsymbol{\theta}_0)$.

Lemma 3.3.

$$
\text{For the true parameter } \boldsymbol{\theta}_0,\ \mathbb{E}_{\boldsymbol{\theta}_0}\hat{V}_n(\boldsymbol{\theta}_0) = I(\boldsymbol{\theta}_0) = -\mathbb{E}_{\boldsymbol{\theta}_0}\left[\nabla^2 \ell_n(\boldsymbol{\theta}_0)\right]/n. \text{ In addition, for any } \boldsymbol{\theta},\ \nabla_1 Q_n(\boldsymbol{\theta},\boldsymbol{\theta}) = \nabla \ell_n(\boldsymbol{\theta})/n. \quad (3.3)
$$

For the true value of $W(\boldsymbol{\theta},\lambda)$ under the null hypothesis, we define $W_0 = I_{22}(\boldsymbol{\theta}_0)^{-1}I_{21}(\boldsymbol{\theta}_0)$. We also define for simplicity that

$$
I^{-1}(\boldsymbol{\theta}) = \begin{bmatrix} I^{11}(\boldsymbol{\theta}) & I^{12}(\boldsymbol{\theta}) \\ I^{21}(\boldsymbol{\theta}) & I^{22}(\boldsymbol{\theta}) \end{bmatrix}.
$$

By defining W_0 and I^{ij}'s as the above, notice that the true value of the asymptotic variance of $\sqrt{n}S_n(\tilde{\boldsymbol{\theta}},\lambda)$ and $\sqrt{n}h(\hat{\boldsymbol{\theta}},\lambda)$ under the null hypothesis becomes

$$
\left[\,\mathbb{I}_r, -W_0^T\,\right] I(\boldsymbol{\theta}_0)\left[\,\mathbb{I}_r, -W_0^T\,\right]^T = I_{11}(\boldsymbol{\theta}_0) - I_{12}(\boldsymbol{\theta}_0)I_{22}^{-1}(\boldsymbol{\theta}_0)I_{21}(\boldsymbol{\theta}_0) = \left[I^{11}(\boldsymbol{\theta}_0)\right]^{-1}.
$$

and $I^{11}(\boldsymbol{\theta}_0)$, respectively.

3.4.1 Score-type statistic

Lemma (3.3) implies that it is sufficient to show that $\sqrt{n}S_n(\tilde{\boldsymbol{\theta}},\lambda)$ asymptotically follows $N\left(0, \left[I^{11}(\boldsymbol{\theta}_0)\right]^{-1}\right)$. Note again that

$$
\begin{aligned}
S_n(\tilde{\boldsymbol{\theta}},\lambda) &= \left[\nabla_1 Q_n(\tilde{\boldsymbol{\theta}},\tilde{\boldsymbol{\theta}})\right]_1 - W(\tilde{\boldsymbol{\theta}},\lambda)^T\left[\nabla_1 Q_n(\tilde{\boldsymbol{\theta}},\tilde{\boldsymbol{\theta}})\right]_2 \\
&= \left[\,\mathbb{I}_r, \quad -W(\tilde{\boldsymbol{\theta}},\lambda)^T\,\right]\nabla_1 Q_n(\tilde{\boldsymbol{\theta}},\tilde{\boldsymbol{\theta}}) \\
&= \left[\,\mathbb{I}_r, \quad -W_0^T\,\right]\nabla_1 Q_n(\boldsymbol{\theta}_0,\boldsymbol{\theta}_0) + o_{\mathbb{P}}(n^{-1/2}).
\end{aligned}
$$

The last equality is also from (Wang et al. 2014) and is used without proof. This leads to asymptotic variance

$$
\begin{aligned}
\operatorname{Var}\left(S_n(\tilde{\boldsymbol{\theta}},\lambda)\right) &\approx \left[\,\mathbb{I}_r, \quad -W_0^T\,\right]\operatorname{Var}\left(\nabla_1 Q_n(\boldsymbol{\theta}_0,\boldsymbol{\theta}_0)\right)\left[\,\mathbb{I}_r, \quad -W_0^T\,\right]^T \\
&= \left[\,\mathbb{I}_r, \quad -W_0^T\,\right]\operatorname{Var}\left(\nabla \ell_n(\boldsymbol{\theta}_0)/n\right)\left[\,\mathbb{I}_r, \quad -W_0^T\,\right]^T \\
&= \frac{1}{n}\left[\,\mathbb{I}_r, \quad -W_0^T\,\right] I(\boldsymbol{\theta}_0)\left[\,\mathbb{I}_r, \quad -W_0^T\,\right]^T \\
&= \frac{1}{n}\left[I^{11}(\boldsymbol{\theta}_0)\right]^{-1},
\end{aligned}
$$

which ends our sketch.

3.4.2 Wald-type statistic

Similar to the score-type statistic, it is sufficient to show that $\sqrt{n}\left(h(\hat{\boldsymbol{\theta}},\lambda) - \boldsymbol{\theta}_{10}\right)$ asymptotically follows $N\left(0, I^{11}(\boldsymbol{\theta}_0)\right)$. Again, under the true $\boldsymbol{\theta}_0$,

$$
\begin{aligned}
h(\hat{\boldsymbol{\theta}}, \lambda) - \boldsymbol{\theta}_{10} &= \hat{\boldsymbol{\theta}}_1 - \boldsymbol{\theta}_{10} - \left\{ \left[\hat{V}_n(\hat{\boldsymbol{\theta}})\right]_{11} - W(\hat{\boldsymbol{\theta}}, \lambda)^T \left[\hat{V}_n(\hat{\boldsymbol{\theta}})\right]_{21} \right\}^{-1} S_n(\hat{\boldsymbol{\theta}}, \lambda) \\
&= \hat{\boldsymbol{\theta}}_1 - \boldsymbol{\theta}_{10} - \left\{ \left[\mathbb{I}_r, \quad -W(\hat{\boldsymbol{\theta}}, \lambda)^T \right] \begin{bmatrix} \left[\hat{V}_n(\hat{\boldsymbol{\theta}})\right]_{11} \\ \left[\hat{V}_n(\hat{\boldsymbol{\theta}})\right]_{21} \end{bmatrix} \right\}^{-1} S_n(\hat{\boldsymbol{\theta}}, \lambda) \\
&= -I^{11}(\boldsymbol{\theta}_0) \left[\mathbb{I}_r, \quad -W_0^T \right] \nabla_1 Q_n(\boldsymbol{\theta}_0, \boldsymbol{\theta}_0) + o_{\mathbb{P}}(n^{-1/2}),
\end{aligned}
$$

where the last equality is from the fact that

$$
\begin{gathered}
\hat{\boldsymbol{\theta}}_1 - \boldsymbol{\theta}_{10} = o_{\mathbb{P}}(n^{-1/2}), \\
\left[\mathbb{I}_r, \quad -W(\hat{\boldsymbol{\theta}}, \lambda)^T \right] \begin{bmatrix} \left[\hat{V}_n(\hat{\boldsymbol{\theta}})\right]_{11} \\ \left[\hat{V}_n(\hat{\boldsymbol{\theta}})\right]_{21} \end{bmatrix} = \left[\mathbb{I}_r, \quad -W_0^T \right] \begin{bmatrix} I_{11}(\boldsymbol{\theta}_0) \\ I_{21}(\boldsymbol{\theta}_0) \end{bmatrix} + o_{\mathbb{P}}(n^{-1/2}).
\end{gathered}
$$

This results in

$$
\begin{aligned}
\operatorname{Var}\left(h(\tilde{\boldsymbol{\theta}}, \lambda) - \boldsymbol{\theta}_{10} \right) &\approx I^{11}(\boldsymbol{\theta}_0) \left[\mathbb{I}_r, \quad -W_0^T \right] \operatorname{Var}\left(\nabla_1 Q_n(\boldsymbol{\theta}_0, \boldsymbol{\theta}_0) \right) \left[\mathbb{I}_r, \quad -W_0^T \right]^T I^{11}(\boldsymbol{\theta}_0) \\
&= \frac{1}{n} I^{11}(\boldsymbol{\theta}_0)
\end{aligned}
$$

3.5 Regularity conditions

For clarity, we did not mention what conditions are needed for asymptotic normality of T_S and T_W to hold. The following conditions 1~4 and assumption 1 should be met.

Condition 1 (Parameter estimation).

$$
\|\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0\|_1 = O_{\mathbb{P}}(\zeta^{EM}),
$$

ζ^{EM} scales with s_0, d, n where s_0 is the true sparsity level.

Condition 2 (Gradient error). For the true $\boldsymbol{\theta}_0$,

$$
\|\nabla_1 Q_n(\boldsymbol{\theta}_0, \boldsymbol{\theta}_0) - \nabla_1 Q(\boldsymbol{\theta}_0, \boldsymbol{\theta}_0))|\|_\infty = O_{\mathbb{P}}(\zeta^G)
$$

where $Q(\boldsymbol{\theta}, \boldsymbol{\theta}') = \mathbb{E}_{\boldsymbol{\theta}_0} Q_n(\boldsymbol{\theta}, \boldsymbol{\theta}')$. ζ^G scales with s_0, d, n.

Condition 3 ($\hat{V}_n$ concentration). For the true $\boldsymbol{\theta}_0$,

$$
\|\hat{V}_n(\boldsymbol{\theta}_0) - \mathbb{E}_{\boldsymbol{\theta}_0} \hat{V}_n(\boldsymbol{\theta}_0)\|_{\infty,\infty} = O_{\mathbb{P}}(\zeta^T),
$$

ζ^T scales with d, n.

Condition 4 ($\hat{V}_n$ lipschitz). For the true $\boldsymbol{\theta}_0$ and any $\boldsymbol{\theta}$,

$$
\|\hat{V}_n(\boldsymbol{\theta}) - \hat{V}_n(\boldsymbol{\theta}_0)\|_{\infty,\infty} = O_{\mathbb{P}}(\zeta^L) \cdot \|\boldsymbol{\theta} - \boldsymbol{\theta}_0\|_1,
$$

ζ^T scales with d, n.

Assumption 1.

- $I(\boldsymbol{\theta}_0)$ is positive definite
- $\left[I^{11}(\boldsymbol{\theta}_0)\right]^{-1} = O(1), I^{11}(\boldsymbol{\theta}_0) = O(1)$
- $\lambda = C(\zeta^T + \zeta^L \zeta^{EM})(1 + \|W_0\|_{1,1})$ for a sufficiently large $C \geq 1$.
- n should be sufficiently large such that

$$
\begin{gathered}
\max\{\|W_0\|_{1,1}\} \cdot s_{W_0} \cdot \lambda = o(1), \ \zeta^{EM} = o(n^{-1/2}), \ \zeta^G \cdot s_{W_0} \cdot \lambda = o(n^{-1/2}), \\
\zeta^{EM} \cdot \lambda = o(n^{-1/2}), \ \max\{1, \|W_0\|_{1,1}\} \cdot \zeta^L \cdot (\zeta^{EM})^2 = o(n^{-1/2})
\end{gathered}
$$

where $s_{W_0} = \|W_0\|_0$.

We will check for specific model if these conditions are reasonable in practice. But before that, consider a low-dimensional case with no sparsity on parameters. It is a general fact proven in the paper that $\zeta^T \sim \sqrt{\log d/n}$. By the assumption 1, this implies $s_{W_0} = o(\sqrt{n/\log d})$ which means that $\left[I^{11}(\boldsymbol{\theta}_0)\right]_{-1}$ should be sparse with sparsity factor of $o(\sqrt{n/\log d})$. Since we assumed the low-dimensional case ($n >> d$), it allows s_{W_0} to be larger than d. This means our no-sparsity assumption is valid in low-dimensional case.

4 Application & Numerical Results

We check if regularity conditions are reasonable for the following model in practice. To restrict the problem to what we covered in class, low-dimensionality and no sparsity is assumed. Note that the assumption 1 is, in most cases, safe to be deemed true. Numerical verification of asymptotic normality follows.

4.1 Model description

Let $Y_1, \cdots, Y_n$ be IID random variables, Z be a Rademacher random variable. The model is

$$\mathbf{Y} = Z\boldsymbol{\beta}_0 + \mathbf{V}, \quad \mathbf{V} \sim N(\mathbf{0}, \sigma^2 \mathbb{I}_d).$$

We assume that σ^2 is known and $\boldsymbol{\beta}_0 \neq 0$.

It is proved that for some absolute constant $C > 0$, $C', C'' \geq 1$, if

$$\begin{aligned}
\zeta^{EM} &= \frac{\sqrt{d}\,\Delta^{GMM}}{1 - \exp(-Cr^2/2)}\sqrt{\frac{T\log d}{n}}, \\
\zeta^{G} &= (\|\boldsymbol{\beta}_0\|_\infty + \sigma)\sqrt{\frac{\log d}{n}}, \\
\zeta^{T} &= \left(\|\boldsymbol{\beta}_0\|_\infty^2 + \sigma^2\right)/\sigma^2\sqrt{\frac{\log d}{n}}, \\
\zeta^{L} &= \left(\|\boldsymbol{\beta}_0\|_\infty^2 + \sigma^2\right)^{3/2}/\sigma^4(\log d + \log n)^{3/2},
\end{aligned}$$

where $r > 0$ is a number that satisfies $\|\boldsymbol{\beta}_0\|_2/ger\sigma$, $R > 0$ is the diameter of initialization $\|\boldsymbol{\beta}_0 - \boldsymbol{\beta}^{\text{init}}\|_2 \leq R/2$, $T = \left\lceil \frac{\log(C'R) - \log(\Delta^{GMM}\sqrt{\log d/n})}{Cr^2/2} \right\rceil$, $\Delta^{GMM} = (1 + C'')\sqrt{d}\,(\|\boldsymbol{\beta}_0\|_\infty + \sigma)$, then all the conditions are met. To check the assumption 1, plugging in ζ's gives the requirement $d\log d = o(n/\log n)$ for our non-sparse case. This implies for low-dimensional case where $d\log d = o(n/\log n)$, we can assure asymptotic normality of test statistics.

4.2 Simulation

For simplicity, we focused on the full parameter case with $d = 10$. Furthermore, to view asymptotic multivariate normality, we redefined score and Wald-type statistics as the square of the corresponding ones introduced in previous section. That is,

$$T_S = nS_n(\boldsymbol{\theta}_0, \lambda)^T\left\{\hat{V}_n(\boldsymbol{\theta}_0)\right\}^{-1}S_n(\boldsymbol{\theta}_0, \lambda),$$

$$T_W = n\Big(h(\hat{\boldsymbol{\theta}}, \lambda) - \boldsymbol{\theta}_0\Big)^T\left\{\hat{V}_n(\hat{\boldsymbol{\theta}})\right\}\Big(h(\hat{\boldsymbol{\theta}}, \lambda) - \boldsymbol{\theta}_0\Big),$$

where $S_n(\boldsymbol{\theta}_0, \lambda) = \nabla_1 Q_n(\boldsymbol{\theta}_0, \boldsymbol{\theta}_0)$, $\hat{V}_n(\boldsymbol{\theta}_0) = -\nabla_{1,1}^2 Q_n(\boldsymbol{\theta}_0, \boldsymbol{\theta}_0) - \nabla_{1,2}^2 Q_n(\boldsymbol{\theta}_0, \boldsymbol{\theta}_0)$, and $h(\hat{\boldsymbol{\theta}}) = \hat{\boldsymbol{\theta}} + \left\{\hat{V}_n(\hat{\boldsymbol{\theta}})\right\}^{-1}\nabla_1 Q_n(\hat{\boldsymbol{\theta}}, \hat{\boldsymbol{\theta}})$.

We first set the case with a Gaussian mixture model with true parameter $\boldsymbol{\beta}_0 = (2, 2, \cdots, 2)^T$. We assume that the variance of random error σ^2 is known to be 1. In this case, it can be derived that

$$Q_n(\boldsymbol{\beta}', \boldsymbol{\beta}) = -\frac{1}{2n}\sum_{i=1}^{n}\omega_{\boldsymbol{\beta}}(\mathbf{y}_i)\|\mathbf{y}_i - \boldsymbol{\beta}'\|_2^2 + \left[1 - \omega_{\boldsymbol{\beta}}(\mathbf{y}_i)\right]\|\mathbf{y}_i - \boldsymbol{\beta}'\|_2^2.$$

Differentiate it by $\boldsymbol{\beta}'$ and set it to zero gives the maximizer as

$$\arg\max_{\boldsymbol{\beta}'} Q_n(\boldsymbol{\beta}', \boldsymbol{\beta}) = \frac{1}{2n}\sum_{i=1}^{n}\omega_{\boldsymbol{\beta}}(\mathbf{y}_i)\mathbf{y}_i - \frac{1}{n}\sum_{i=1}^{n}\mathbf{y}_i$$

Further computation gives

$$\hat{V}_n(\boldsymbol{\beta}) = \mathbb{I}_d - \frac{1}{n}\sum_{i=1}^{n} \nu_{\boldsymbol{\beta}}(\mathbf{y}_i)\mathbf{y}_i\mathbf{y}_i^T,$$

$$\nu_{\boldsymbol{\beta}}(\mathbf{y}_i) = \frac{4}{\sigma^2\left\{1+\exp(-2\boldsymbol{\beta}^T\mathbf{y}_i/\sigma^2)\right\}\left\{1+\exp(2\boldsymbol{\beta}^T\mathbf{y}_i/\sigma^2)\right\}}.$$

We ran 1000 simulation with each run computes test statistic with a random sample of size 50. Both score and Wald-type statistics are expected to follow Chi-squared distribution with degrees of freedom equal to $d = 10$. The following histograms empirically shows that our expectation is valid.

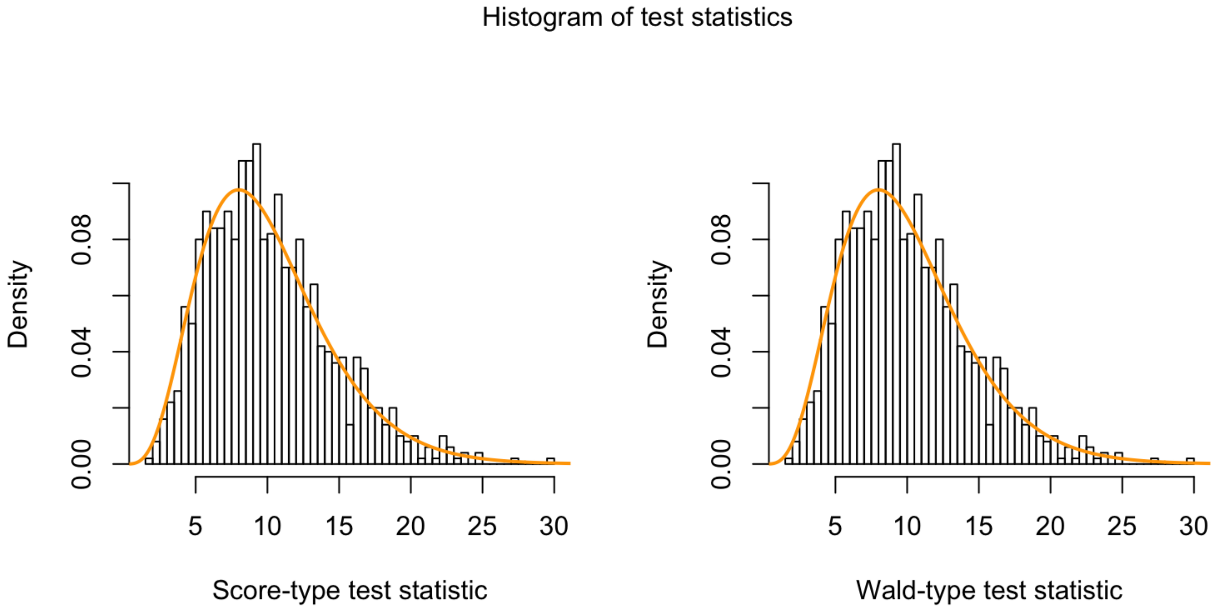

Figure 4.1: Black line is histograms of score and Wald-type test statistics. Orange line is the density of χ^2_{10}. This empirically shows that the test statistic follows asymptotic normality.

5 Discussions

We believe the result we reviewed is significant in the following manner.

- It is the first and the only attempt, as far as we know, to derive asymptotic normality of EM estimators. While previous research focused on computational convergence rate and convergence guarantee of the global maximizer, this result focused further on statistical properties of possibly local maximizers.
- It provides statistical properties of not only maximum likelihood estimators, but also local maximizer of likelihood function.
    - It is worth noting again that derivatives of Q-function successfully recovers score function and information matrix of incomplete data. As a result, the asymptotic distribution of EM estimator is surprisingly similar to that of MLE.
    - Rather than being consistent, local maximizers tend to be asymptotically biased as much as $\frac{1}{n}\hat{I}^{-1}(\hat{\boldsymbol{\theta}})S(\mathbf{Y};\hat{\boldsymbol{\theta}})$ as shown in (3.2).

However, it needs attention in practice that for some models, regularity conditions for asymptotic normality might not be easy to be met. Though we did not cover it, while the order requirement of the sample size n compared to dimension d is reasonable for the Gaussian mixture model, for a complex model such as a mixture of regression models it became exponentially large. Still, for the computational convergence guarantee, the required order of n could be much lower concerning d, which makes the algorithm applicable to most models.

References

Boos, D. D., and L. A. Stefanski. 2013. *Essential Statistical Inference: Theory and Methods*. Springer Texts in Statistics. Springer New York. https://books.google.com/books?id=8VNDAAAAQBAJ (https://books.google.com/books?id=8VNDAAAAQBAJ).

Wang, Zhaoran, Quanquan Gu, Yang Ning, and Han Liu. 2014. “High Dimensional Expectation-Maximization Algorithm: Statistical Optimization and Asymptotic Normality.” arXiv. https://doi.org/10.48550/ARXIV.1412.8729 (https://doi.org/10.48550/ARXIV.1412.8729).