Supplementary Material for "Robust Analysis of High Throughput Screening (HTS) Assay Data" by Changwon Lim, Pranab K. Sen, Shyamal D. Peddada

This supplementary material text is divided into three sections as listed in the following table of contents. In Section 1, we provide the regularity conditions which are needed for proving the main results discussed in the paper. In Section 2, we provide some additional simulation results. Several issues about the condition number of the information matrix of the estimators when estimating parameters in the Hill model are discussed in Section 3.

Table of contents

1. *Regualrity conditions for the main results*

For the asymptotic normality of the WME (Theorem 1, Lim et al. 2012a) which are needed for proving Theorem 1 and 2 in the main paper, we require the following sets of regularity conditions.

[S1]:

(i) $\lim_{n\to\infty} n^{-1}\Gamma_{1n}(\theta,\tau) = \Gamma_1(\theta,\tau)$, where

$$\Gamma_{1n}(\theta,\tau) = \gamma_2 \sum_{i=1}^{n} k^2(z_i,\tau) f_\theta(x_i,\theta) f_\theta^{\mathrm{T}}(x_i,\theta).$$

(ii) $\lim_{n\to\infty} n^{-1}\Gamma_{31n}(\theta,\tau) = \Gamma_{31}(\theta,\tau)$, where

$$\Gamma_{31n}(\theta,\tau) = \sigma_{\psi 1}^2 \sum_{i=1}^{n} k^2(z_i,\tau) f_\theta(x_i,\theta) f_\theta^{\mathrm{T}}(x_i,\theta),$$

and $\Gamma_{31}(\theta,\tau)$ is a positive definite matrix.

(iii) $\max_i \left\{ k^2(z_i,\tau) f_\theta^{\mathrm{T}}(x_i,\theta) \Gamma_{31n}^{-1}(\theta,\tau) f_\theta(x_i,\theta) \right\} \longrightarrow 0$, as $n \to \infty$

[S2]:

(i) $\lim_{n\to\infty} n^{-1}\Gamma_{2n}(\theta,\tau) = \Gamma_2(\theta,\tau)$, where

$$\Gamma_{2n}(\theta,\tau) = \sum_{i=1}^{n} \left\{ \frac{2\gamma_1 + \gamma_3 - 1}{\sigma^2(z_i,\tau)} \sigma_\tau(z_i,\tau) \sigma_\tau^{\mathrm{T}}(z_i,\tau) + \frac{1-\gamma_1}{\sigma(z_i,\tau)} \Sigma_\tau(z_i,\tau) \right\},$$

and $\Sigma_\tau(z_i,\tau) = (\partial^2/\partial\tau\partial\tau^{\mathrm{T}})\sigma(z_i,\tau)$.

(ii) $\lim_{n\to\infty} n^{-1}\Gamma_{32n}(\theta,\tau) = \Gamma_{32}(\theta,\tau)$, where

$$\Gamma_{32n}(\theta,\tau) = \sigma_{\psi 2}^2 \sum_{i=1}^{n} k^2(z_i,\tau) \sigma_\tau(z_i,\tau) \sigma_\tau^{\mathrm{T}}(z_i,\tau),$$

and $\Gamma_{32}(\theta,\tau)$ is a positive definite matrix.

(iii) $\max_i \left\{ k^2(z_i,\tau) \sigma_\tau^{\mathrm{T}}(z_i,\tau) \Gamma_{32n}^{-1}(\theta,\tau) \sigma_\tau(z_i,\tau) \right\} \longrightarrow 0$, as $n \to \infty$

For the asymptotic linearity of WME (Lim et al. 2012b) which are needed for proving Thoerem 1 and 2 in the main paper, we require the following regularity conditions.

[S3]:

ψ is a nonconstant, odd function which is absolutely continuous and differentiable with respect to θ.

[S4]: Let $\epsilon = \{y - f(x,\theta)\}/\sigma(z,\tau)$,

(i) $E\psi(\epsilon) = 0$; $E\psi^2(\epsilon) = \sigma_{\psi 1}^2 < \infty$; $E\{\psi(\epsilon)\epsilon\} = \gamma_1 (\neq 0)$;

$\mathrm{var}\{\psi(\epsilon)\epsilon\} = \sigma_{\psi 2}^2 < \infty$

(ii) $E|\psi'(\epsilon)|^{1+\delta} < \infty$, $E|\psi'(\epsilon)\epsilon|^{1+\delta} < \infty$, $E|\psi'(\epsilon)\epsilon^2|^{1+\delta} < \infty$ for some $0 < \delta \leq 1$, and

$E\psi'(\epsilon) = \gamma_2 (\neq 0)$; $E\{\psi'(\epsilon)\epsilon\} = 0$; $E\{\psi'(\epsilon)\epsilon^2\} = \gamma_3 (\neq 0)$;

$E\psi'(\sigma(z,\tau)\epsilon) = \gamma_4 (\neq 0)$;

[S5]: Let $\epsilon(\theta,\tau) = \{y - f(x,\theta)\}/\sigma(z,\tau)$,

(i) $$\lim_{\delta_1 \to 0} \lim_{\delta_2 \to 0} E \left\{ \sup_{\|\Delta_1\| \leq \delta_1, \|\Delta_2\| \leq \delta_2} \left| \psi\left(\epsilon(\theta + \Delta_1, \tau + \Delta_2)\right) - \psi\left(\epsilon(\theta,\tau)\right) \right| \right\} = 0$$

(ii) $$\lim_{\delta_1 \to 0} \lim_{\delta_2 \to 0} E \left\{ \sup_{\|\Delta_1\| \leq \delta_1, \|\Delta_2\| \leq \delta_2} \left| \psi\left(\epsilon(\theta + \Delta_1, \tau + \Delta_2)\right) \epsilon(\theta + \Delta_1, \tau + \Delta_2) - \psi\left(\epsilon(\theta,\tau)\right) \epsilon(\theta,\tau) \right| \right\} = 0$$

(iii) $$\lim_{\delta_1 \to 0} \lim_{\delta_2 \to 0} E \left\{ \sup_{\|\Delta_1\| \leq \delta_1, \|\Delta_2\| \leq \delta_2} \left| \psi'\left(\epsilon(\theta + \Delta_1, \tau + \Delta_2)\right) - \psi'\left(\epsilon(\theta,\tau)\right) \right| \right\} = 0$$

(iv) $$\lim_{\delta_1 \to 0} \lim_{\delta_2 \to 0} E \left\{ \sup_{\|\Delta_1\| \leq \delta_1, \|\Delta_2\| \leq \delta_2} \left| \psi'\left(\epsilon(\theta + \Delta_1, \tau + \Delta_2)\right) \epsilon(\theta + \Delta_1, \tau + \Delta_2) - \psi'\left(\epsilon(\theta,\tau)\right) \epsilon(\theta,\tau) \right| \right\} = 0$$

(v) $$\lim_{\delta_1 \to 0} \lim_{\delta_2 \to 0} E \left\{ \sup_{\|\Delta_1\| \leq \delta_1, \|\Delta_2\| \leq \delta_2} \left| \psi'\left(\epsilon(\theta + \Delta_1, \tau + \Delta_2)\right) \epsilon^2(\theta + \Delta_1, \tau + \Delta_2) - \psi'\left(\epsilon(\theta,\tau)\right) \epsilon^2(\theta,\tau) \right| \right\} = 0$$

[S6]:

$f(x, \theta)$ is continuous and twice differentiable with respect to $\theta \in \Re^p$.

[S7]: For $j, l = 1, \ldots, p$

(i) $\lim_{\delta \to 0} \sup_{\|\Delta\| \leq \delta} \Big| (\partial/\partial\theta_j) f(x, \theta + \Delta)(\partial/\partial\theta_l) f(x, \theta + \Delta)$

$$- (\partial/\partial\theta_j) f(x, \theta)(\partial/\partial\theta_l) f(x, \theta) \Big| = 0$$

(ii) $\lim_{\delta \to 0} \sup_{\|\Delta\| \leq \delta} \Big| (\partial^2/\partial\theta_j \partial\theta_l) f(x, \theta + \Delta) - (\partial^2/\partial\theta_j \partial\theta_l) f(x, \theta) \Big| = 0$

[S8]:

$\sigma(z, \tau)$ is continuous and twice differentiable with respect to $\tau \in \Re^q$.

[S9]: For $j, l = 1, \ldots, q$

(i) $\lim_{\delta \to 0} \sup_{\|\Delta\| \leq \delta} \Big| (\partial/\partial\tau_j) \sigma(z, \tau + \Delta)(\partial/\partial\tau_l) \sigma(z, \tau + \Delta)$

$$- (\partial/\partial\tau_j) \sigma(z, \tau)(\partial/\partial\tau_l) \sigma(z, \tau) \Big| = 0$$

(ii) $\lim_{\delta \to 0} \sup_{\|\Delta\| \leq \delta} \Big| (\partial^2/\partial\tau_j \partial\tau_l) \sigma(z, \tau + \Delta) - (\partial^2/\partial\tau_j \partial\tau_l) \sigma(z, \tau) \Big| = 0$

For the asymptotic results regarding PTE (Theorem 2, Lim et al. 2012a) which are needed for proving Theorem 1(b) and 2 in the main paper, we require the following sets of regularity conditions.

[S10]: Let $\epsilon = \{y - f(x, \theta)\}/\sigma(z, \tau)$. Then, $E\psi'(\sigma(z, \tau)\epsilon) = \gamma_4 \, (\neq 0)$, $E\psi^2(\sigma(z, \tau)\epsilon) = \sigma^2_{\psi 3} w_1(x) < \infty$ and $E\{\psi(\epsilon)\psi(\sigma(z, \tau)\epsilon)\} = \sigma^2_{\psi 4} w_2(x) < \infty$.

[S11]:

(i) $\lim_{n \to \infty} n^{-1} \Gamma_{4n}(\theta) = \Gamma_4(\theta)$, where

$$\Gamma_{4n}(\theta) = \gamma_4 \sum_{i=1}^{k} n_i f_\theta(x_i, \theta) f_\theta^{\mathrm{T}}(x_i, \theta).$$

(ii) $\lim_{n \to \infty} n^{-1} \Gamma_{33n}(\theta) = \Gamma_{33}(\theta)$, where

$$\Gamma_{33n}(\theta) = \sigma^2_{\psi 3} \sum_{i=1}^{k} n_i w_1(x_i) f_\theta(x_i, \theta) f_\theta^{\mathrm{T}}(x_i, \theta),$$

and $\Gamma_{33}(\theta)$ is a positive definite matrix.

(iii) $\lim_{n\to\infty} n^{-1}\Gamma_{34n}(\theta,\tau) = \Gamma_{34}(\theta,\tau)$, where

$$\Gamma_{34n}(\theta,\tau) = \sigma_{\psi 4}^2 \sum_{i=1}^{k} \frac{n_i w_2(x_i)}{\sigma_i} f_\theta(x_i,\theta) f_\theta^{\mathrm{T}}(x_i,\theta).$$

(iv) $\lim_{n\to\infty} n^{-1}G_{2n}(\theta,\tau) = G_2(\theta,\tau)$, where

$$G_{2n}(\theta,\tau) = \begin{pmatrix} \Gamma_{31n}(\theta,\tau) & \Gamma_{34n}(\theta,\tau) & 0 \\ \Gamma_{34n}(\theta,\tau) & \Gamma_{33n}(\theta) & 0 \\ 0 & 0 & 2n^2 \sum_{i=1}^{k} n_i w_{i2}^2 \end{pmatrix},$$

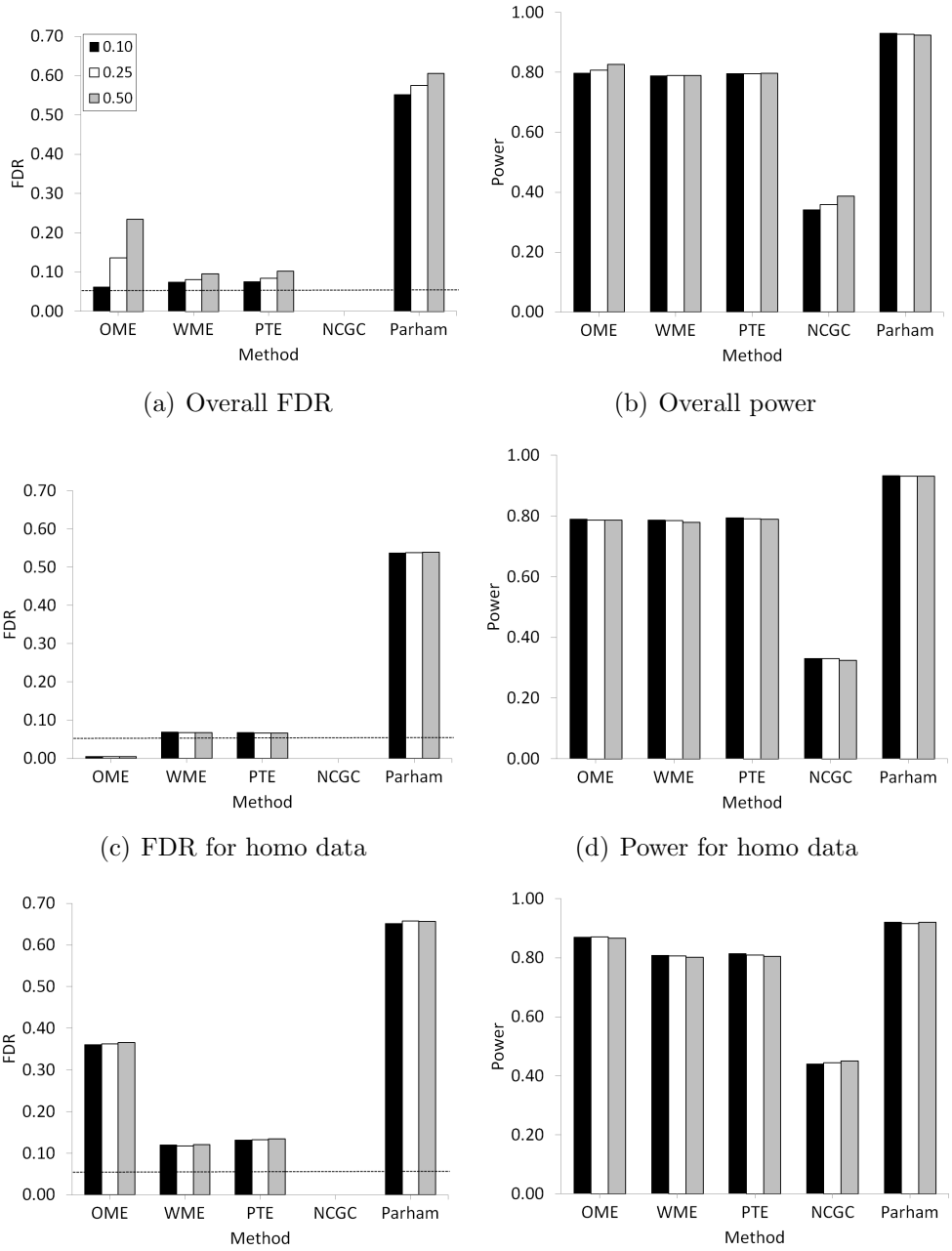
w_{i2} is the second element of $w_i = (Z^{\mathrm{T}}Z)^{-1}z_i$, and $G_2(\theta,\tau)$ is a positive definite matrix.

(v) $\max_i ch_1\Big\{G_1(x_i,\theta,\tau)\big(G_{2n}(\theta,\tau)\big)^{-1}\Big\} \longrightarrow 0$, as $n \to \infty$, where

$$G_1(x_i,\theta,\tau) = \begin{pmatrix} \sigma_{\psi 1}^2 \sigma_i^{-2} H_i & \sigma_{\psi 4}^2 w_2(x_i)\sigma_i^{-1} H_i & 0 \\ \sigma_{\psi 4}^2 w_2(x_i)\sigma_i^{-1} H_i & \sigma_{\psi 3}^2 w_1(x_i) H_i & 0 \\ 0 & 0 & 2n^2 w_{i2}^2 \end{pmatrix},$$

and $H_i = f_\theta(x_i,\theta) f_\theta^{\mathrm{T}}(x_i,\theta)$.

2. *Some additional simulation results*



(a) Overall FDR

(b) Overall power

(c) FDR for homo data

(d) Power for homo data

(e) FDR for hetero data

(f) Power for hetero data

Figure S1: Estimated FDR and power for OME (dotted line), WME (dashed line), PTE (solid line), NCGC (dash dotted line) and Parham (long dashed line) methods when the proportion of heteroscedastic data is 0.10, 0.25 and 0.50. Here $\gamma = 0.05$ and $\alpha = 0.05/10,000$.

Table S1: Estimated proportion of inconclusive (marginal) data among non-null and null data based on NCGC and Parham methods with $\alpha = 0.05/10,000$.

γ	Hetero.	Method	Overall		Homoscedastic		Heteroscedastic	
			Non-null	Null	Non-null	Null	Non-null	Null
0.05	0.10	NCGC	0.634	0.951	0.645	0.952	0.536	0.943
		Parham	0.068	0.285	0.067	0.294	0.074	0.199
	0.25	NCGC	0.615	0.950	0.647	0.952	0.516	0.943
		Parham	0.071	0.271	0.069	0.295	0.078	0.201
	0.50	NCGC	0.585	0.947	0.647	0.952	0.524	0.943
		Parham	0.071	0.248	0.065	0.295	0.077	0.201

3. *Issues about the condition number of the information matrix*

Figure S2 represents a simulated data generated from a Hill model with true parameters, $(\theta_0, \theta_1, \theta_2, \theta_3) = (-46, 39, 1.2, 31.01)$. The fitted curve is reasonable based on the data and quite close to the true curve. However, the OME of θ_1 and θ_3 are 126.154 and 484.308, which are substantially different (very large) from the true values (39 and 31.01), respectively. The standard errors of these estimates are also very large, 361.50 and 3203.22, respectively. The underlying problem is that the information matrix of OME is almost singular, with a condition number (ratio of the largest eigenvalue to the smallest eigenvalue) of the order 10^9. Unfortunately, this is a common phenomenon when fitting Hill models. Consequently, many toxicologists discount data with either large slopes or large ED_{50} values since they can't trust the fit (Parham et al. 2009). This problem arises not because of model mis-specification, since in this example we generated the data using the true Hill model, but possibly because of the dose spacing and/or range of the doses. As seen in Figure S2, a visually good fit does not always imply sensible estimates for parameters.

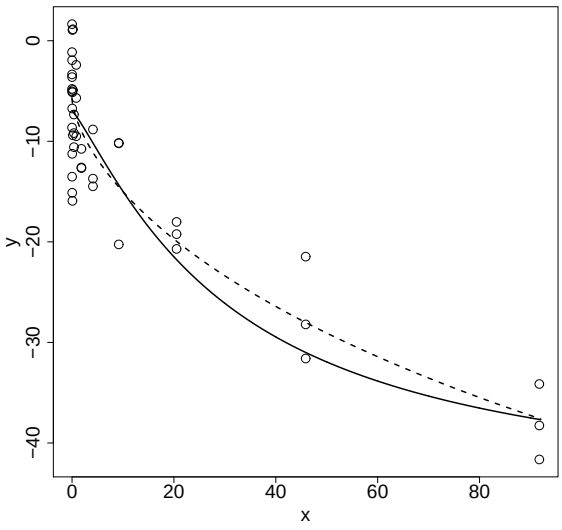
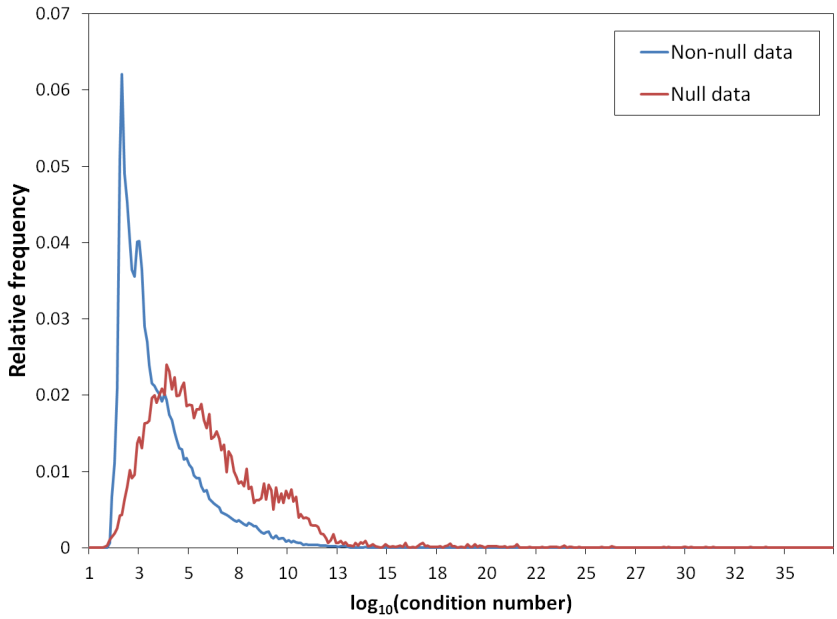


Figure S2: Example of data generated from Hill model, true curve (dotted line), and fitted curve using OME (solid line).

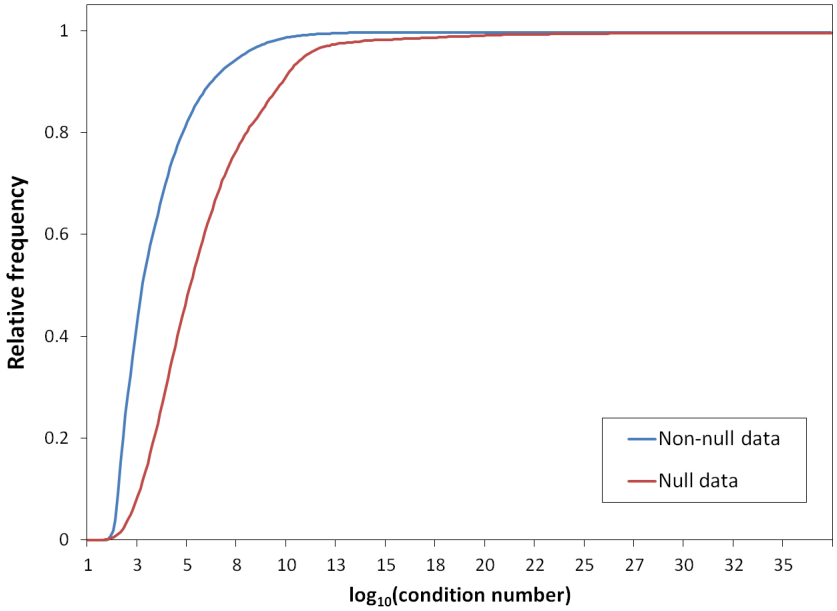
Table S2: Empirical cumulative distribution of condition number for null and non-null data.

$\log_{10}$(condition number)	Percentile	
	Null data	Non-null data
2	0.01	0.06
3	0.08	0.43
4	0.23	0.64
5	0.40	0.78
6	0.56	0.86
7	0.69	0.92

To better understand the distribution of condition numbers, using the data from the simulation experiment described in Section 3 of this paper, we estimated the distribution functions of condition numbers based on OME under the null hypothesis and under the alternative hypothesis. We used 9,000 and 90,000 samples to obtain the empirical distribution functions of the condition numbers under the null and the alternative hypotheses, respectively. The resulting empirical distribution functions are plotted in Figure S3. Whether the data are homoscedastic or heteroscedastic, the condition numbers for the non-null data appear to be stochastically smaller than those of the null data. For the simulated non-null data, the median of the distribution is $10^{3.2}$; and the first and the third quartiles are $10^{2.4}$ and $10^{4.7}$, respectively. On the other hand, for the simulated null data, the median is $10^{5.6}$; and the first and the third quartiles are $10^{4.1}$ and $10^{7.6}$, respectively. See Table S2 for some examples of the empirical CDF under the null and alternative hypothesis. From our simulation study it is clear that the condition number provides valuable information when assessing the goodness of fit of a Hill model.



(a) Empirical probability density function



(b) Empirical cumulative distribution function

Figure S3: Distributions of the maximum of the condition numbers of the information matrix of OME and WME for the simulated non-null data (red line) and the simulated null data (black line).

References

[1] Lim, C., Sen, P. K., and Peddada, S. D. (2012a), "Accounting for uncertainty in heteroscedasticity in nonlinear regression", *Journal of Statistical Planning and Inference*, 142, 1047–1062.

[2] ——(2012b), "Robust nonlinear regression methods in applications", *Journal of Indian Society of Agricultural Statistics*, submitted.