Simulation set up

A single optimal cut-off point was identified using the Youden method within a Bayesian classification framework, applied to a bivariate mixture model of CSF and plasma. For each sample size, the average cut-off point, the maximum of sensitivity + specificity, accuracy, area under the ROC curve (AUC), and prevalence were calculated across 1,000 datasets. However, it remains unclear why the maximum of sensitivity + specificity, accuracy, and AUC values decrease as the sample size increases.

$$f\begin{pmatrix} y^{\text{csf}} \\ y^{\text{plasma}} \end{pmatrix} = \sum_{k=1}^{K} \pi_k \cdot N \begin{pmatrix} \mu_k^{\text{CSF}} \\ \mu_k^{\text{Plasma}} \end{pmatrix}, \Sigma \end{pmatrix}$$

$$\text{where } \mu_k^{\text{csf}} = (0.05, 0.1), \mu_k^{\text{plasma}} = (0.08, 0.1), \Sigma = \begin{pmatrix} 0.0001082 & 0.0000375 \\ 0.0000375 & 0.0001030 \end{pmatrix}$$

Findings

A single optimal cut-off point was identified using the Youden method within a Bayesian framework classification, applied to a bivariate mixture model of CSF and plasma. For each sample size, the average cut-off point, maximum of sensitivity + specificity, accuracy, area under the ROC curve (AUC), and prevalence were calculated across 1,000 datasets. However, I have yet to resolve why the maximum of sensitivity + specificity, accuracy, and AUC values decrease as the sample size increases.

|	optimal_cutpoint	sum_sens_spec	acc	sensitivity	specificity	AUC	prevalence
real data with sample size 130	0.0915	1.6569	0.8231	0.8182	0.8387	0.8912	0.7615
simulated data with sample size 100	0.0904	1.7473	0.8659	0.8543	0.8930	0.9265	0.7012
simulated data with sample size 200	0.0898	1.7213	0.8590	0.8566	0.8647	0.9230	0.7010
simulated data with sample size 500	0.0899	1.6974	0.8483	0.8478	0.8497	0.9194	0.7005
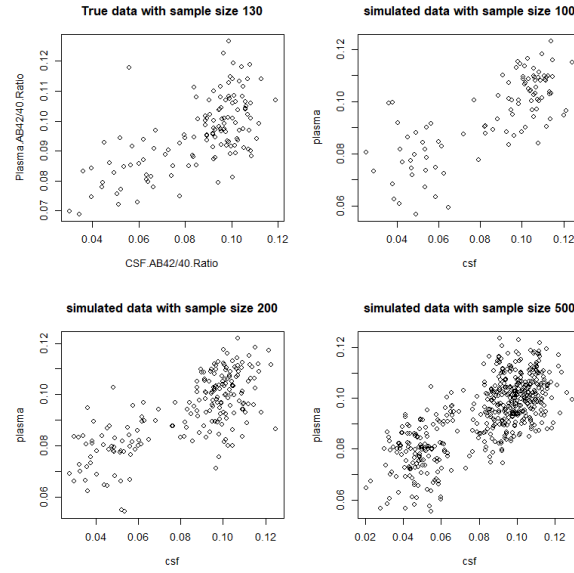
Figure 1: scatter plot of real data and one sample simuation data of each sample size
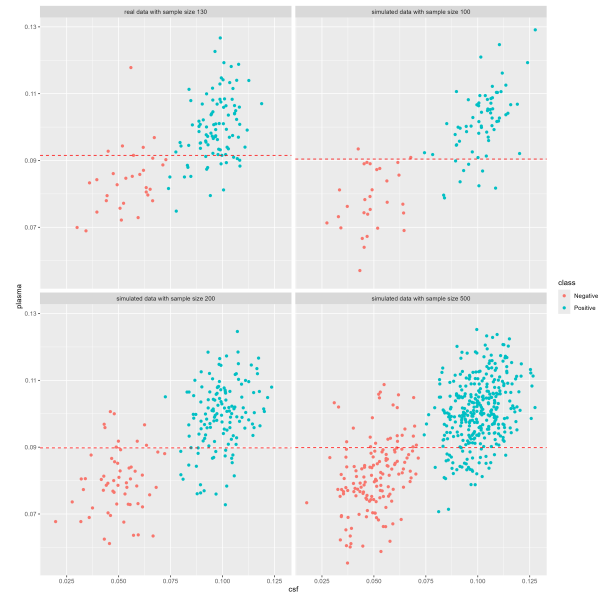

Figure 2: scatter plot of real data and one sample simuation data of each sample size with classification