Accumulation Curve Probabilistic Framework

Jeremy Ash

May 31, 2018

Part I

As stated in the prompt, S is the score from a ranking algorithm, where larger values of S are indicative of the $+$ class and smaller values are indicative of the $-$ class.

$S|+$ has cumulative distribution function (cdf), $F_+(t)$ and probability distribution function (pdf), $f_+(t)$. $S|-$ has cdf, $F_-(t)$, and pdf, $f_-(t)$.

$X = I(active)$, where $I(\cdot)$ is the indicator function. That is, $X = 1$ when an item belongs to the $+$ class and $X = 0$ when an item belongs to the $-$ class.

S is a random variable drawn from the mariginal distribution of $F_S(t)$. Since the class that S belongs to is unknown, we treat X as a random variable, where $P(X = 1) = \pi$.

Binormal

Under the binormal model, the conditional distributions of $S|+$ and $S|-$ are given by:

$$S|+ \sim N(\mu_+, \sigma_+^2)$$
$$S|- \sim N(\mu_-, \sigma_-^2)$$

Where $Normal(\mu, \sigma^2)$ is the Normal distribution with mean μ and variance σ^2.

First, I derive the marginal pdf of S:

$$
\begin{aligned}
F_S(t) = P(S < t) &= P((S < t \cap +) \cup (S < t \cap -)) && \text{+ and - partition the sample space} \\
&= P((S < t \cap +) + P(S < t \cap -)) && \text{+ and - are mutually exclusive} \\
&= P(+)P(S < t|+) + P(-)P(S < t|-)) && \text{law of conditional probability} \\
&= \pi F_+(t) + (1 - \pi)F_-(t)
\end{aligned}
$$

Taking the derivative wrt t provides the following mixture as the mariginal pdf of S:

$$f_S(t) = \pi f_+(t) + (1 - \pi)f_-(t)$$

For the binormal model, this is a mixture of Gaussians:

$$S \sim \pi N(\mu_+, \sigma_+^2) + (1 - \pi)N(\mu_-, \sigma_-^2)$$

Next, I derive the conditional distribution of $X|S = t$. To do so, I note that $X|S = 1$ with probability $\theta = P(+|S = t)$ and $X|S = 0$ with probability $1 - \theta$, so $X|S \sim Bern(\theta)$. $P(+|S = t)$ is known since S is known. It can be computed as follows:
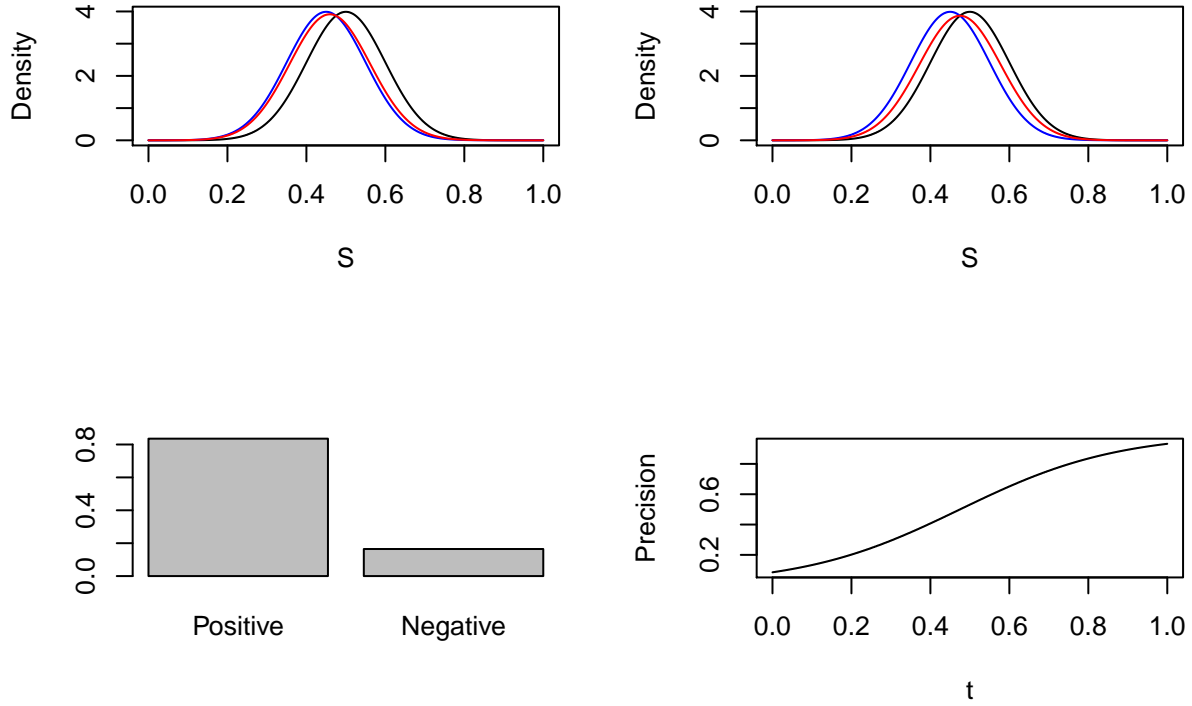
1

Figure 1: Densities of S and X|S for the binormal model. (top left) The pdf of $S|+$ (black), $S|-$ (blue), and S (red) when $u_+ = .6$, $u_- = .4$, $\sigma_+ = \sigma_- = .1$, and $\pi = .2$. (top right) pdfs when $S|+$ and $S|-$ have the same parameters and $\pi = .5$. (bottom left) The pdf of $X|S = t$, when $t = .8$ and $\pi = .5$ (bottom right) Precision, or $P(+|S = t)$ for a range of values of $t \in [0, 1]$ when $\pi = .5$

$$P(+|S = t) = \frac{P(+)P(S = t|+)}{P(S = t)} \quad \text{(Bayes rule)}$$

$$= \frac{\pi f_+(S = t)}{f(S = t)}$$

$$= \frac{\pi}{\pi + (1 - \pi)\frac{f_-(S=t)}{f_+(S=t)}}$$

Another expression for θ is precision, which is plotted by the population precision recall curve as a function of $F_+(\cdot)$, or true positive rate, *tpr*.

In **Figure 1**, I illustrate the distribution of S and $X|S$ under a binormal model for S where $u_+ > u_-$. This means the ranking algorithm has generated scores such that large scores are indicative of the $+$ class and small scores are indicative of the $-$ class. Two mixing probabilities are considered, .2 and .5. $\pi = .2$ corresponds to the case when the negative class has four times more observations than the positive class. This is a scenario that occurs frequently in real world data sets such as case control studies or drug discovery settings where the chemicals with a desired activity are rare.
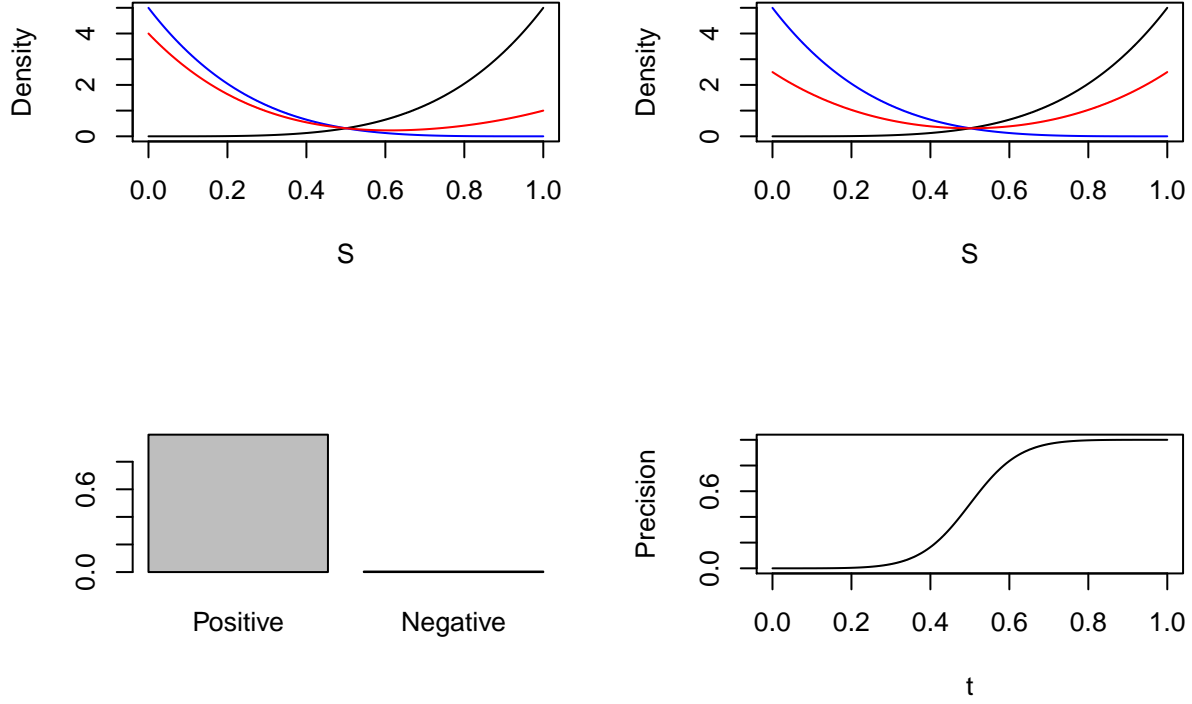
Figure 2: Densities of S and X|S for the bibeta model. (top left) The pdf of $S|+$ (black), $S|-$ (blue), and S (red) when $\alpha_+ = 5$, $\beta_+ = 1$, $\alpha_- = 1$, $\beta_- = 5$ and $\pi = .2$. (top right) pdfs when $S|+$ and $S|-$ have the same parameters and $\pi = .5$. (bottom left) The pdf of $X|S = t$, when $t = .8$ and $\pi = .5$ (bottom right) Precision, or $P(+|S = t)$ for a range of values of $t \in [0, 1]$ when $\pi = .5$

Bibeta

Under the bibeta model, conditional distributions of $S|+$ and $S|-$ are given by:

$$S|+ \sim Beta(\alpha_+, \beta_+)$$
$$S|- \sim Beta(\alpha_-, \beta_-)$$

Where $Beta(\alpha, \beta)$ is the Beta distribution with shape parameters α and β.

The same arguements above can be used to prove that, under the bibeta model, the marginal pdf of S is mixture of Beta distributions:

$$S \sim \pi Beta(\alpha_+, \beta_+) + (1 - \pi)Beta(\alpha_-, \beta_-)$$

Again, $X|S \sim Bern(\theta)$, where $\theta = \frac{\pi}{\pi + (1-\pi)\frac{f_-(S=t)}{f_+(S=t)}}$ can be determined by computing the corresponding Beta densities at $S = t$.

In **Figure 2**, I illustrate the distribution of S and X|S under a bibeta model where $\alpha_+ = 5$, $\beta_+ = 1$, $\alpha_- = 1$, $\beta_- = 5$. $S|+$ has a heavy left skew, while $S|-$ has a heavy right skew. Again, the ranking algorithm has

generated scores such that large scores are indicative of the $+$ class and small scores are indicative of the $-$ class. This scoring algorithm is performing better than the binormal model, because the probability that an item is in the positive class given that the score is high (e.g. above .5) is higher under this model. **Figures 1 and 2** show an instance of this when $S = .8$. The probability that an item is positive given $S = .8$ is much larger under the bibeta model.

Part II

The accumulation curve plots the cumulative fraction of actives, y_{acc}, identified as a function of the relative ranks of items, x_{acc}, according to some scoring algorithm. The prompt states that the cumulative fraction of actives identified should be used since we are considering an infinite population. I argue that a *percentile rank* should be used for x_{acc}, where the $(100i)$th percentile ranked item, $i \in (0, 1]$, has $100i$ percent of scores greater than or equal to it. As on the y-axis of the accumulation curve, the percentile is simply a rescaling of the rank of the compound. At percentile rank, x_{acc}, all compounds in the precentile are "tested" and the cumulative fraction of actives determined, y_{acc}. Another way of determining this percentile would be to chose a threshold t such that the fraction of items with $S > t$ is x_{acc}.

Borrowing notation from Hughes-Oliver (Hughes-oliver 2000). We consider an independent random sample of scores $\{S_i; i = 1, ..., n\}$ from the marginal distribution of S. Let $\{S_i^+; i = 1, ..., n^+\}$ be the scores that were independently sampled from the $+$ class mixture component, $F_+(\cdot)$, and $\{S_i^-; i = 1, ..., n^-\}$ be the scores that were independently sampled from the $-$ class mixture component, $F_-(\cdot)$. We define, $\hat{F}(\cdot)$, $\hat{F}_+(\cdot)$, $\hat{F}_-(\cdot)$ to be the empirical cumlative distribution functions (cdf) for all, $+$, and $-$ scores:

$$\hat{F}(t) = \frac{1}{n} \sum_{i=1}^{n} I(S_i \leq t)$$

$$\hat{F}_+(t) = \frac{1}{n^+} \sum_{i=1}^{n^+} I(S_i^+ \leq t)$$

$$\hat{F}_-(t) = \frac{1}{n^-} \sum_{i=1}^{n^-} I(S_i^- \leq t)$$

We then chose a threshold t such that $x_{acc} = 1 - \hat{F}(t)$ is the precentile selected for testing. The fraction of the active compounds that were correctly predicted to be active (i.e. $S_i^+ > t$) is $y_{acc} = 1 - \hat{F}_+(t)$. This is called the empirical true positive fraction.

Since we are considering a population of infinite size, we consider the convergence of empirical cdf $1 - \hat{F}^n(t)$ as $n \to \infty$ and $1 - \hat{F}_+^{n^+}(t)$ as $n^+ \to \infty$. We assume that the mixing porportion is $0 < \pi < 1$ so that $n^+ \to \infty$ as $n \to \infty$.

By the Glivenko-Cantelli theorem (Shuster 1972), as the number of independently and identically distributed samples goes to infinity, $\hat{F}_n(t)$ converges to $F(t)$ almost surely. That is, for every $\epsilon > 0$ and all $t \in [0, 1]$.

$$P(\lim_{n \to \infty} |\hat{F}_n(t) - F(t)| \leq \epsilon) = 1$$

Therefore we conclude that

$$x_{acc} = 1 - \hat{F}^n(t) \stackrel{a.s.}{\to} 1 - F(t) = P(S > t)$$

$$y_{acc} = 1 - \hat{F}_+^{n^+}(t) \stackrel{a.s.}{\to} 1 - F_+(t) = P(S > t|+)$$

For a population of infinite size, the population accumulation curve plots the pairs $\{P(S > t), P(S > t|+)\}$. If $\pi = 0$ or $\pi = 1$, then the accumulation curve actually plots $\{P(S > t), 0\}$ and $\{P(S > t), P(S > t)\}$, respectively.

Binormal

To plot the population accumulation curve under the binormal model for a given t, we compute:

$$P(S > t) = \pi(1 - \Phi(\frac{t - \mu_+}{\sigma_+^2})) + (1 - \pi)(1 - \Phi(\frac{t - \mu_-}{\sigma_-^2}))$$

$$P(S > t|+) = 1 - \Phi(\frac{t - \mu_+}{\sigma_+^2})$$

Where $\Phi(t)$ is the standard normal cdf.

Bibeta

To plot the population accumulation curve under the binormal model for a given t, we compute:

$$P(S > t) = \pi(1 - I_t(\alpha_+, \beta_+)) + (1 - \pi)(1 - I_t(\alpha_-, \beta_-))$$
$$P(S > t|+) = 1 - I_t(\alpha_+, \beta_+)$$

Where $I_t(\alpha, \beta)$ is the regularized incomplete beta function:

$$I_t(\alpha, \beta) = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha + \beta)} \int_0^t x^{\alpha-1}(1 - x)^{\beta-1}$$

Functional form of the accumulation curve

We can write $t = F_S^{-1}(x_{acc})$, were $F^{-1}(\cdot)$ is the generalized inverse of $F(\cdot)$. This allows us to write y_{acc} as a function of x_{acc}:

$$y_{ACC}(x_{ACC}) = 1 - F_+(F_S^{-1}(1 - x_{ACC}))$$

I will refer to this as the population accumulation curve $ACC(x)$. This curve can be estimated using the empirical cdfs specified above:

$$\widehat{ACC}(x) = 1 - \hat{F}_+(\hat{F}_S^{-1}(1 - x))$$

Hsieh and Turnbull (Hsieh and Turnbull 1996) have derived uniform consistency and strong approximation properties of the empirical ROC curve, $1 - \hat{F}_+(\hat{F}_-^{-1}(1 - x))$. Since their proofs are for general distributions F_- and F_+ that satisfy conditions that the corresponding distributions, F_+ and F_S, in the ACC curve also satisfy. I assume that uniform consistency of the functional form of the ACC curve holds, that is, for some $\epsilon > 0$.

$$\sup_{x \in [\epsilon, 1-\epsilon]} |\widehat{ACC}(x) - ACC(x)| \to 0 \quad \text{a.s. as } n \to 0$$

And that:

$$\sqrt{n}(\widehat{ACC}(x) - ACC(x)) \xrightarrow{d} N(0, \sigma^2(x))$$

I attempted to derive the form of $\sigma^2(x)$ the asymptotic variance of empirial ACC curve, but I am not confident in this. Instead, I will state how I expect the variance to change as a number of variables change. I expect that the variance will have two components like the empirical ROC asymptotic variance (see (Pepe 2003)):

$$var(\widehat{ROC}(t)) = (\frac{f_+(F_+^{-1}(1-x))}{f_-(F_+^{-1}(1-x))})^2 \frac{x_{ROC}(1-x_{ROC})}{n^+} + \frac{y_{ROC}(1-y_{ROC})}{n^-}$$

The first term is the variability due to selection of the threshold, which in the ACC curve should depend on n and not only n^+. I expect that the slope term in the ACC curve will resemble $\frac{f_S(F_S^{-1}(1-x))}{f_+(F_S^{-1}(1-x))}$. The second term is the variance due to uncertainity in the false positive rate (fpr) on the y-axis, which in the ACC curve is the true positive rate (tpr), so in the ACC curve this should depend on $n+$. From what I have seen in simulations, I expect the sampling variance of $\widehat{ACC}(x)$ to increase when the skewness increases, so this term should be included in the expression.

I expect the sampling variance of $\widehat{ACC}(x)$ to increase when the skewness increases, to increase as y_{ACC} or x_{ACC} approaches .5, and to decrease as n gets larger. I also expect variance to be larger for the beta distribution than the normal distribution because the slope term in the variance tends to be larger for the beta distribution. These are the conditions I will checking for during my simulations.

Part III

$\{S_i; i = 1, ..., n\}$ were independently sampled from the marginal distributions of S derived in **Part 1** under both the binormal and bibeta models with $N = 1000, 5000, 10000$ and $\pi = 1/101, 1/11, 1/2$. We define skew as $\frac{1-\pi}{\pi}$. Then, the values of π used in this simulation correspond to skew = 100, 10, 1. This was when the sample size of the negative class is 100, 10, and 1 fold greater than the positive class. These sample sizes and skew values were selected because the accumulation curve is frequently used to assess the performance of a classifier in situations where there is large sample sizes and extreme class imbalance. For example, in drug discovery, data sets will typically contain tens or hundreds of thousands of compounds, while only small fraction will be active. Extreme skewness was incorporated into the simulation study to assess how reasonable the empirical estimator of the ACC curve is in these situations.

The scores were sampled from the mixture distribution of S as follows: one of the mixture components $(+, -)$ were selected at random where $P(+) = \pi$ and $P(-) = 1 - \pi$. Then, a score was sampled from the distribution corresponding to the sampled class, $f_+(\cdot)$ or $f_-(\cdot)$.

The sampled scores were then ranked from highest to lowest and at each rank, or number of tests, the fraction of the total actives identified was plotted. Let T be the set of thresholds that that identify the percentile of each ranked compound. That is, for each $i = 1, ..., n$, we choose a t that is below the score of the ith ranked compound and above or equal to the score of the $(i + 1)$th ranked compound. This provides an empirical estimator of the population accumulation curve:

$$\widehat{ACC}(t) = \{\text{proportion of } S \text{ values} > \text{t, proportion of } S^+ \text{ values} > \text{t}; t \in T\}$$
$$= \{1 - \hat{F}_S(t), \ 1 - \hat{F}_+(t); t \in T\}$$

Binormal

We assume the binormal model with, $\mu_+ = .6$, $\mu_- = .4$, $\sigma_+ = \sigma_- = 1$.

As we have shown in Part II, the empirical ACC curve converges almost surely to the population ACC curve $\forall t \in [0, 1]$ as $n \to \infty$. Therefore, I expect that by plotting the empirical ACC curve along with the population binormal accumulation curve derived in **Part 2**, I will be able to visually see pointwise convergence as n becomes larger. **Figure 3** shows the estimation of the population ACC curve for single simulation replicate when n is moderate but n^+ is small ($n = 1000$, $\pi = 1/101$). We can consider specifying a point on the ACC curve in a number of different ways. One way would be to specify a threshold and then compute $\{P(S > t), P(S > t|+)\}$. At a specific threshold value, t, the empirical estimate of $P(S > t)$ is very good and has clearly converged pointwise. This is because n is sufficiently large. However, the empirical estimate

of $P(S > t|+)$ is poor because the estimate of this relies on $n^+ \approx 10$. The empirical cdf for $S|+$ is a step function that has clearly not converged to the smooth population cdf. This results in considerable sampling variability in $\widehat{ACC}(x)$ and a poor approximation of the population curve. Interestingly, the very early on in the $\widehat{ACC}(x)$ is a good estimate of the population accumulation curve, which is usually the region of greatest interest to researchers, as there will only be resources to test the very top percentile. Since both x_{acc} and y_{acc} are close to zero in this region, the sample variance of $\widehat{ACC}(x)$ is small.

Figure 4 shows that either decreasing the skew to 1 or increasing the sample size by 10 fold will dramatically improve the estimation of the population curve. However, when $N = 10,000 and \pi = \frac{1}{101}$, $\widehat{ACC}(x)$ undershoots the population curve by a good margin. An even larger sample size would be necessary to accurately capture the population curve when there is this much class imbalance.

Bibeta

We assume the bibeta model as illustrated previously, with $\alpha_+ = 5$, $\beta_+ = 1$, $\alpha_- = 1$, $\beta_- = 5$. Again, the mixing proportions used were $\pi = 1/101, 1/11, 1/2$.

The empirical ACC curve was visually inspected for pointwise convergence as N became larger. **Figure 5** shows the results when $N = 1000$ $\pi = 1/11$. The population accumulation curve is much more optimal in this scenario. $P(S > t|+)$ grows to near 1 much earlier in this curve. This means all of the actives are likely to be identified early on in the curve, when only a small fraction of the data set is tested. This is due to a better separation of the scores sampled from the positive and negative classes. $\widehat{ACC}(x)$ is a much better estimate of the population accumulation curve when $N = 1000$ and $\pi = 1/101$.

Figure 6 shows that $\widehat{ACC}(x)$ is an excellent estimator of the population accumulation curve when N is increased to 10000 or the skew is decreased to 1.

Larger scale simulation

Under both models, for $\pi = 1/101, 1/2$ and $N = 1000, 5000, 10000$, $\widehat{ACC}(x)$ was estimated at x = .01, .1, .5, .9. The top 1 and top 10 percent of the data set was considered, as researchers are often most interested in testing a small top ranked fraction of the data. Therefore, the quality of $\widehat{ACC}(x)$ early on in the curve was of great interest. 1000 simulation replicates were performed. Convergence rate to the normal distribution of the $\widehat{ACC}(x)$ estimates was assessed and MSE was computed for the $\widehat{ACC}(x)$ estimator of the $ACC(x)$. To compute the population $ACC(x)$, the inverse of the mixture distribution $F_S(\cdot)$ needed to be computed. In order to do so, I took many samples (N = 1000000) from $F_S(\cdot)$ and then estimated the desired quantile using the sample distribution.

Binormal

Figure 7 shows the histograms of $\widehat{ACC}(x)$ estimates under the binormal model with large skew, $\pi = 1/101$. Non-parameteric kernel densities were estimated for the empirical $\widehat{ACC}(x)$ using a gaussian kernel with Silverman's rule of thumb as the bandwidth parameter (Silverman, n.d.). Kernel density estimates demonstrated the quality of the normal approximation. When $\pi = 1/101$, the normal approximation is good at small values of x, even n is small. However, the approximation is very poor when x = .9. As n increases the normal approximation improves. **Figure 8** shows that when $\pi = 1/2$, the normal approximation when x = .01 is poor for all n considered.

Table 1 shows the MSE values for the binormal model. As the skewness decreases to 1, the MSEs decrease substantially, suggesting a decrease in the sampling variance of the $\widehat{ACC}(x)$ curve. When $\pi = 1/2$, x = .5 results in the highest MSE, while the highest MSE was often observed for x = .01 when $\pi = 1/101$. Increasing N decreases the MSE.

7
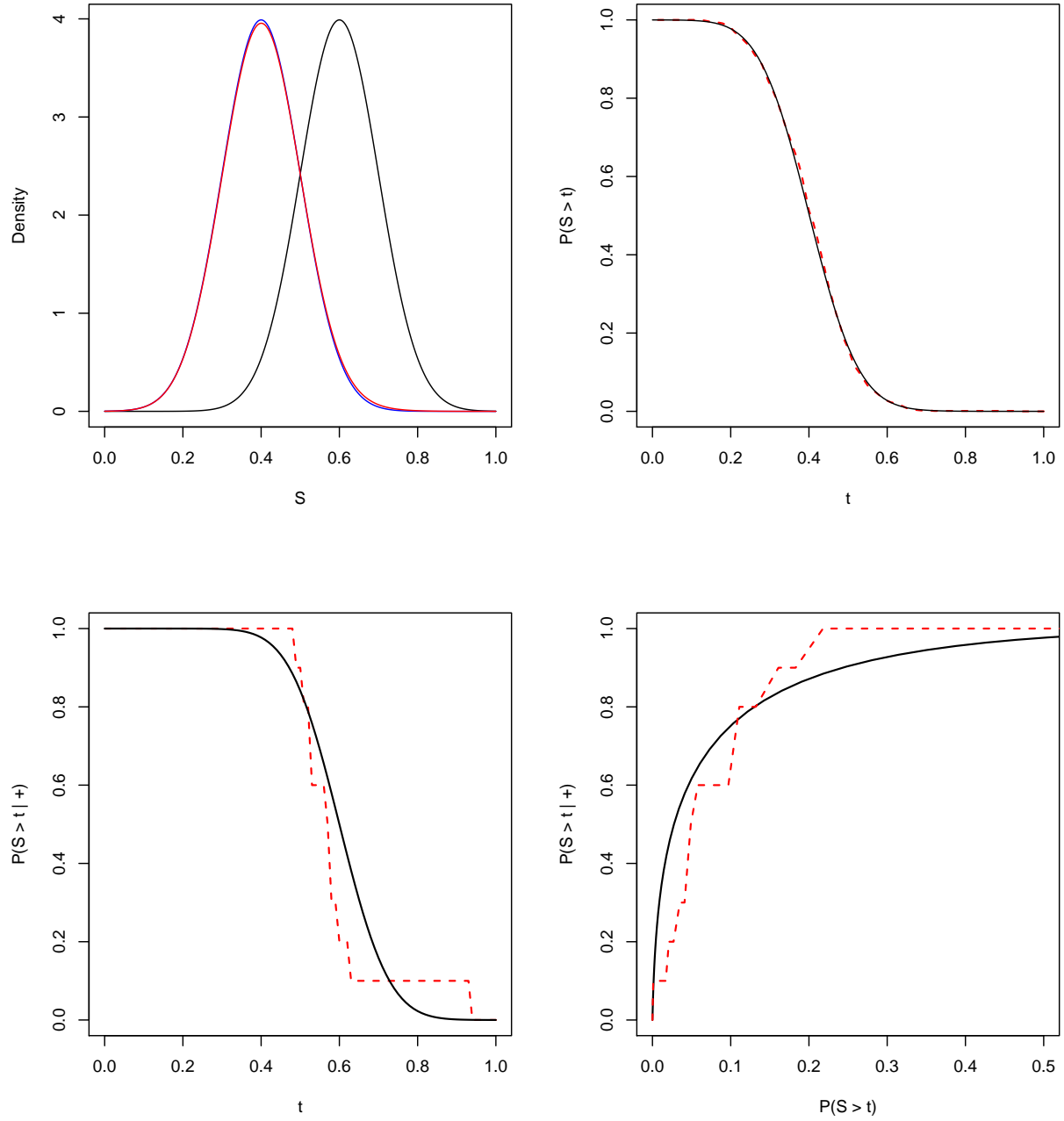
Figure 3: Estimation of the population ACC curve under the binormal model with small sample size (N = 1,000) and large skew ($\pi = 1/101$). $u_+ = .6$, $u_- = .4$, $\sigma_+ = \sigma_- = .1$. (top left) The pdf of S (red), $S|+$ (black) and $S|-$ (blue). (top right) Empirical cdf used to estimate $P(S > t)$ (red) and the true population curve (black) as t ranges from 0 to 1.(bottom left) Empirical cdf used to estimate $P(S > t \mid +)$ (red) and the true population curve (black). (bottom right) The empirical estimate of the ACC curve (red) and the true population curve (black).

8
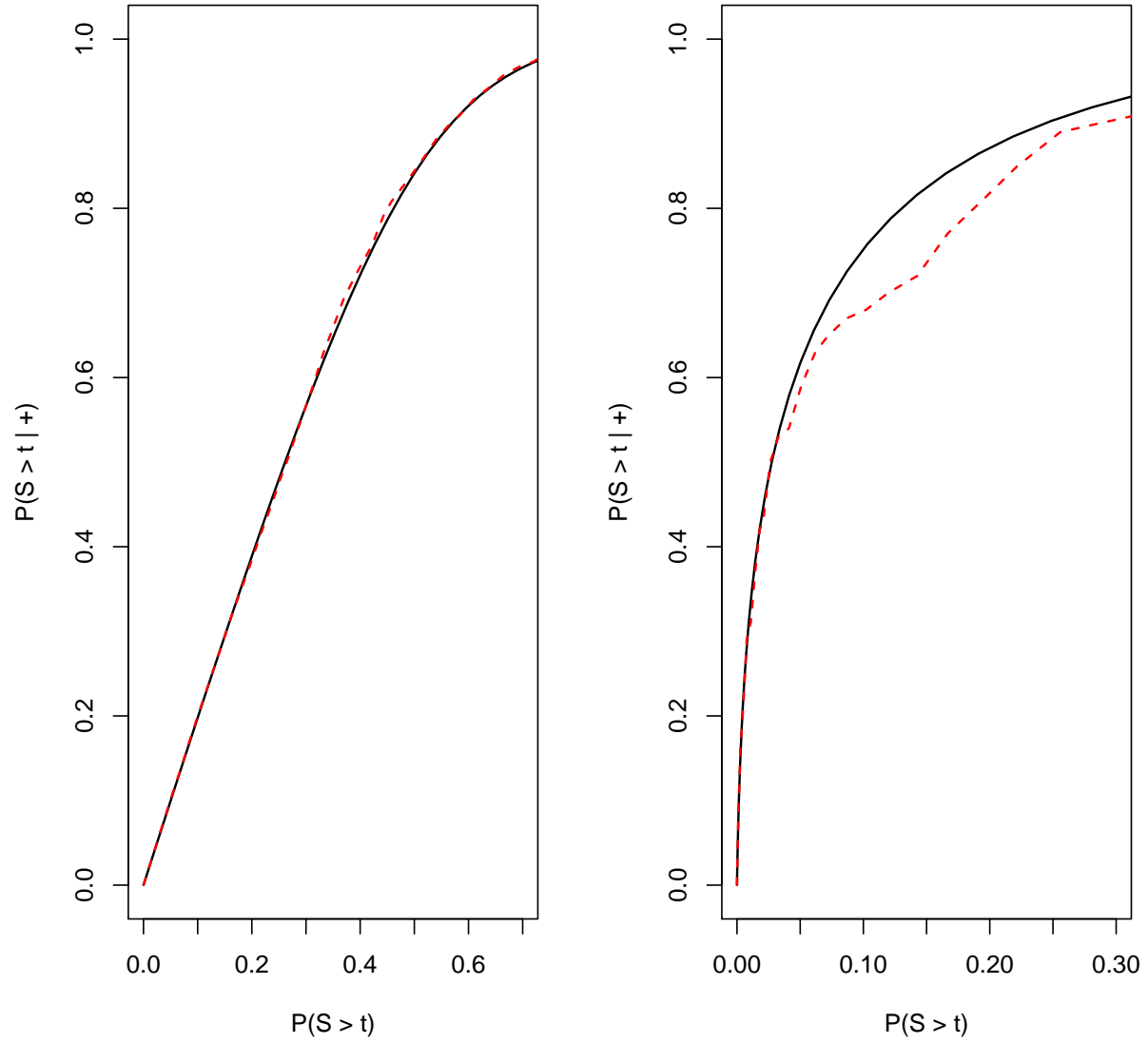
Figure 4: Estimation of the population ACC curve under the same binormal model as Figure 1 with different combinations of sample size and π. (left) Empirical ACC curve estimate when N = 1,000 and $\pi = 1/2$. (right) Empirical ACC curve estimate when N = 10,000 and $\pi = 1/101$.
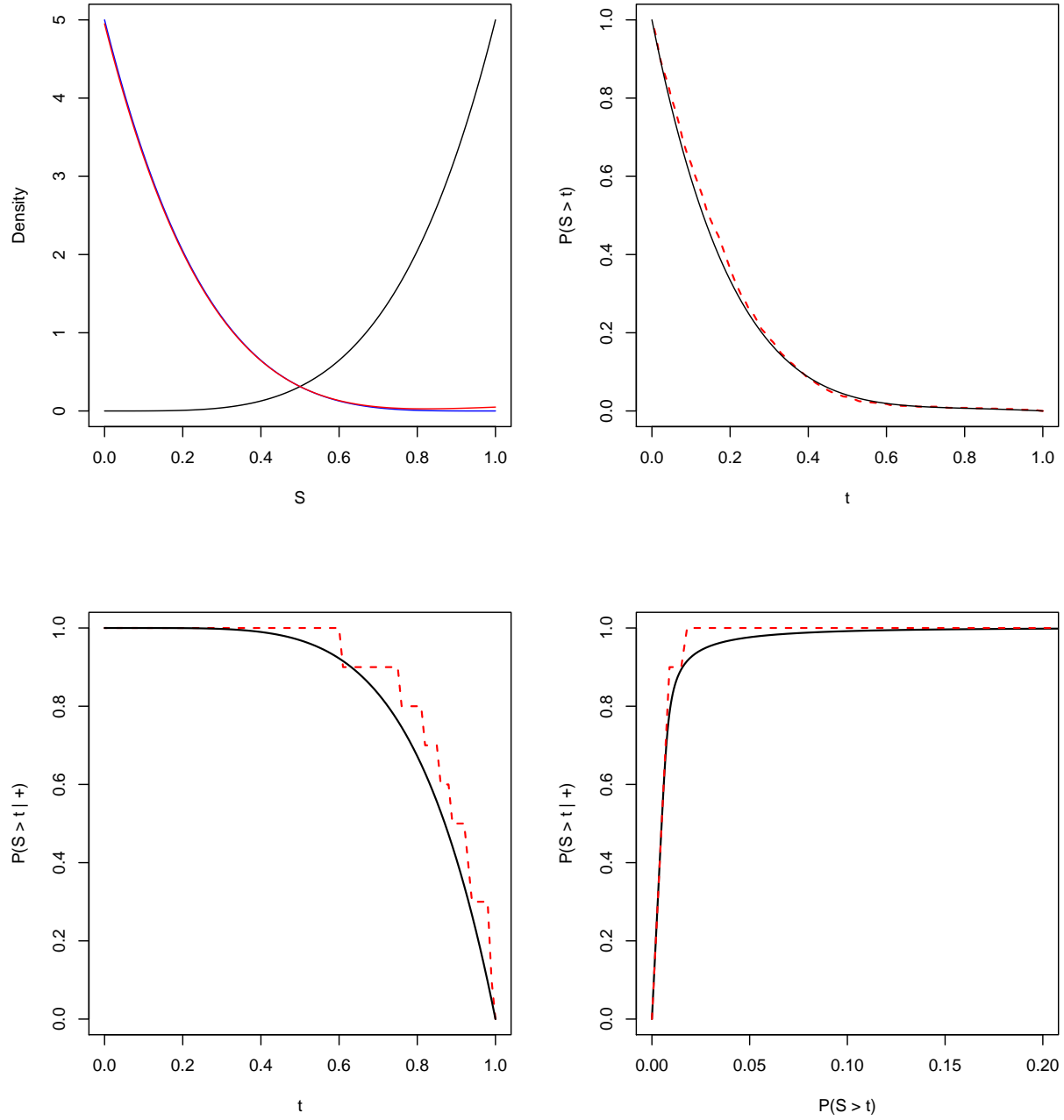
Figure 5: Estimation of the population ACC curve under the bibeta model with small sample size (N = 1,000) and large skew ($\pi = 1/101$). $\alpha_+ = 5$, $\beta_+ = 1$, $\alpha_- = 1$, $\beta_- = 5$. (top left) The pdf of S (red), $S|+$ (black) and $S|-$ (blue). (top right) Empirical cdf used to estimate $P(S > t)$ (red) and the true population curve (black) as t ranges from 0 to 1.(bottom left) Empirical cdf used to estimate $P(S > t \mid +)$ (red) and the true population curve (black). (bottom right) The empirical estimate of the ACC curve (red) and the true population curve (black).
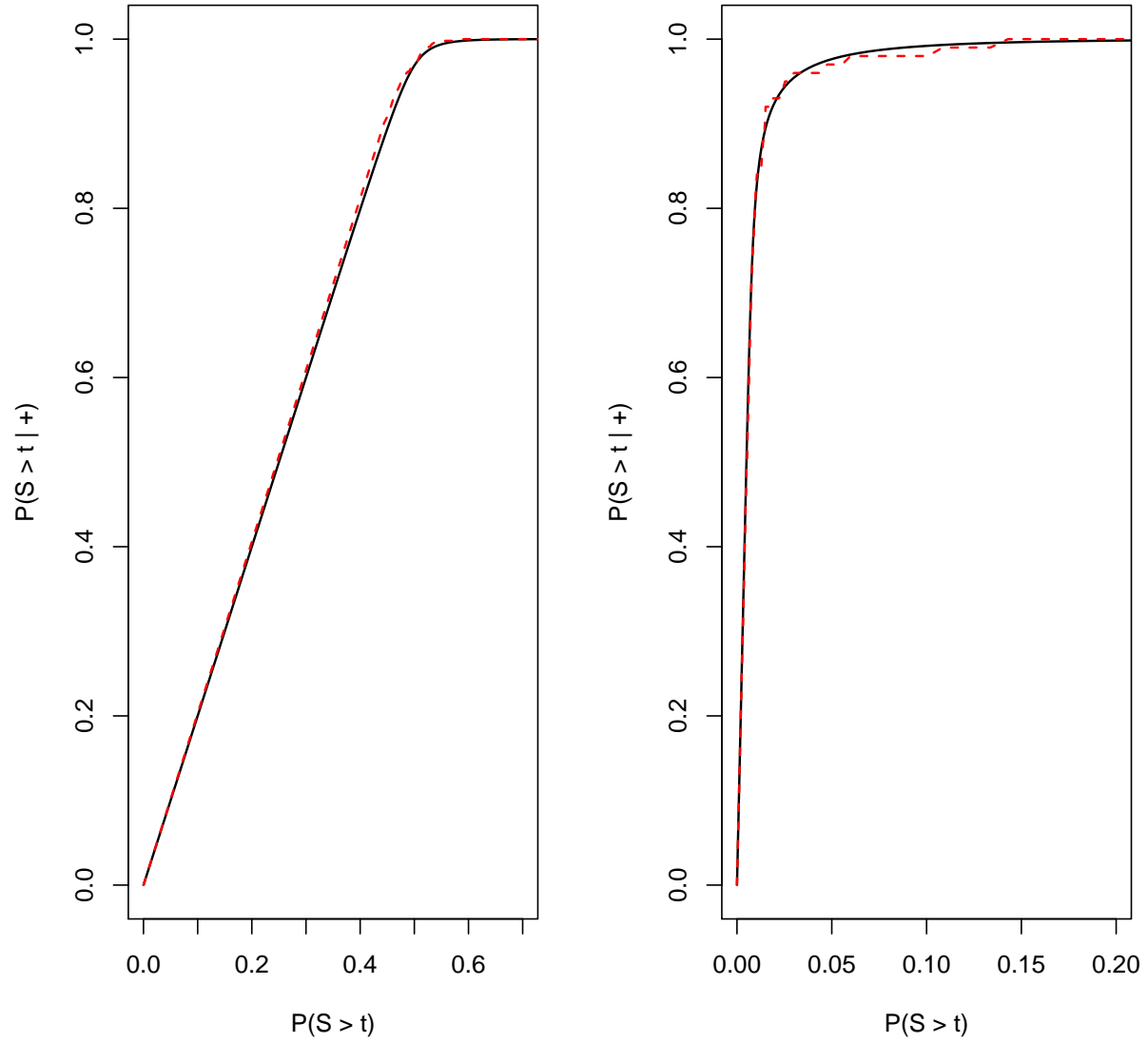
10

Figure 6: Estimation of the population ACC curve under the same binormal model as Figure 1 with different combinations of sample size and π. (left) Empirical ACC curve estimate when N = 1,000 and $\pi = 1/2$. (right) Empirical ACC curve estimate when N = 10,000 and $\pi = 1/101$.
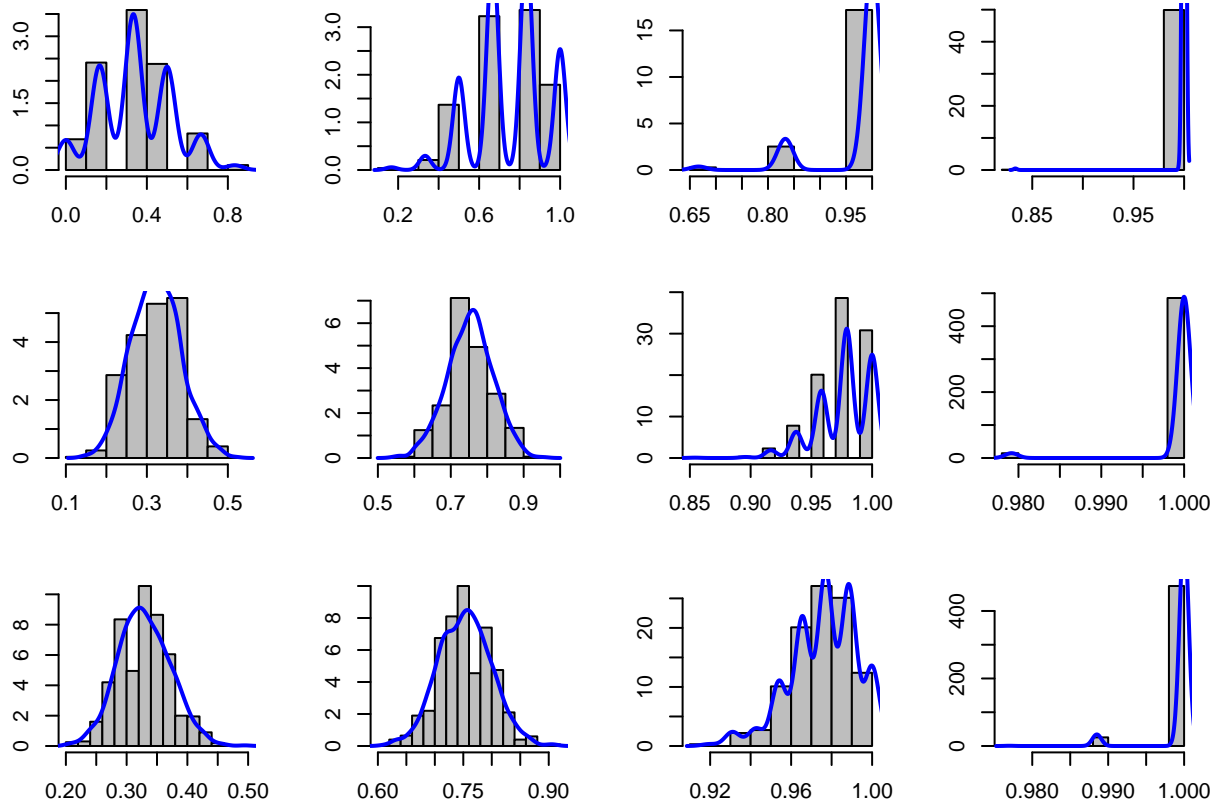
Figure 7: Histograms of observed $\widehat{ACC}(x)$ for binormal model with large skew ($\pi = 1/101$). 1000 simulation replicates were performed. Kernel density estimates are overlayed to demonstrate cases in which there was approximate normality. A gaussian kernel (rows) N = 1000, 5000, 10000. (columns) x = .01, .1, .5, .9.

Bibeta

Many of the same trends were observed under the bibeta model. Interestingly, the normal approximation of $\widehat{ACC}(x)$ seems to perform much better for the bibeta model, but the MSEs are on average substantially higher for all x.
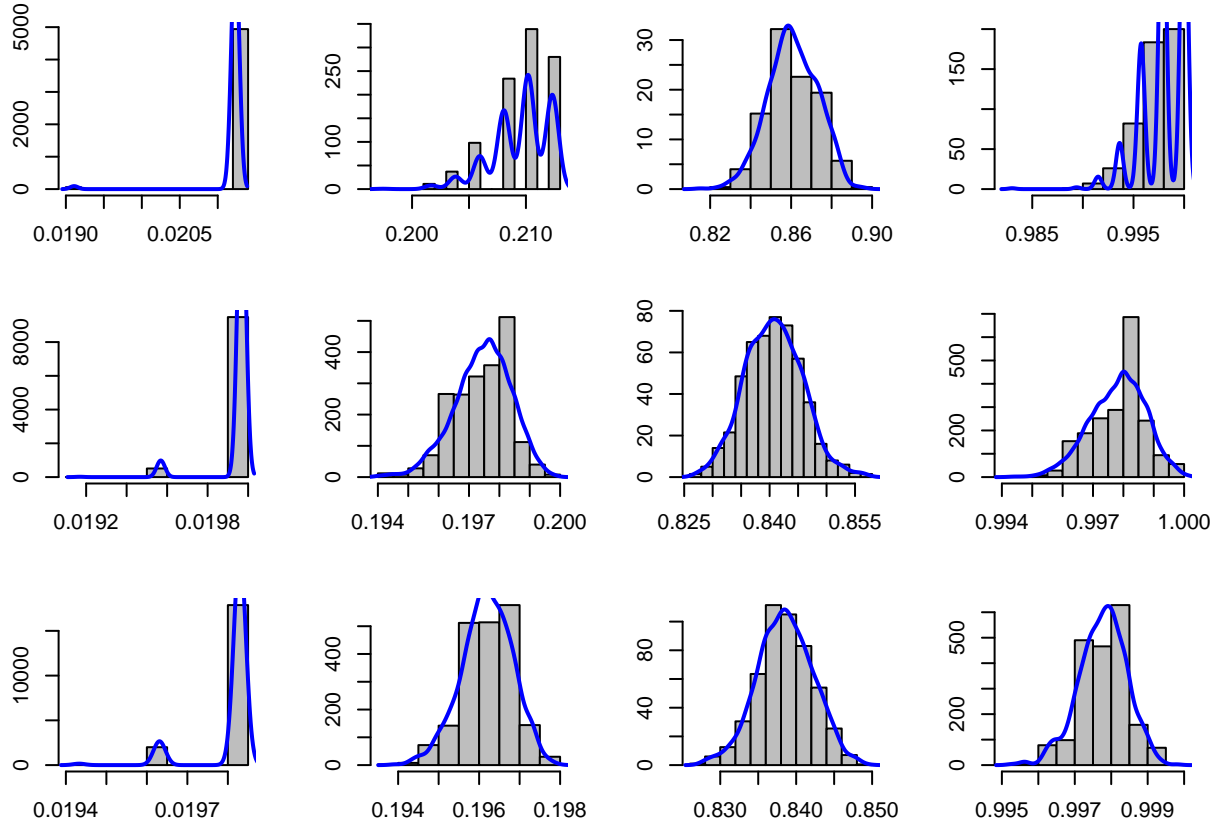
Figure 8: Histograms of observed $\widehat{ACC}(x)$ for binormal model with skew $= 1$ ($\pi = 1/2$). 1000 simulation replicates were performed. Kernel density estimates are overlayed to demonstrate cases in which there was approximate normality. (rows) N = 1000, 5000, 10000. (columns) x = .01, .1, .5, .9.

Table 1: MSE for the $\widehat{ACC}(x)$ estimator for the binormal model. On the columns are the x_{acc} values considered. Both cases of large skew ($\pi = 1/101$) and skew $= 1$ are included.

	pi = 1/101				pi = 1/2			
	.01	.1	.5	.9	.01	.1	.5	.9
N = 1000	3.31e-02	2.93e-02	4.40e-03	5.55e-05	1.19e-06	1.40e-04	5.33e-04	4.44e-06
N = 5000	3.64e-03	3.93e-03	4.75e-04	1.22e-05	4.48e-08	1.13e-06	2.54e-05	8.18e-07
N = 10000	1.89e-03	2.13e-03	2.67e-04	6.90e-06	1.14e-07	3.56e-06	2.02e-05	4.33e-07

Table 2: MSE for the $\widehat{ACC}(x)$ estimator for the bibeta model. On the columns are the x_{acc} values considered. Both cases of large skew ($\pi = 1/101$) and skew $= 1$ are included.

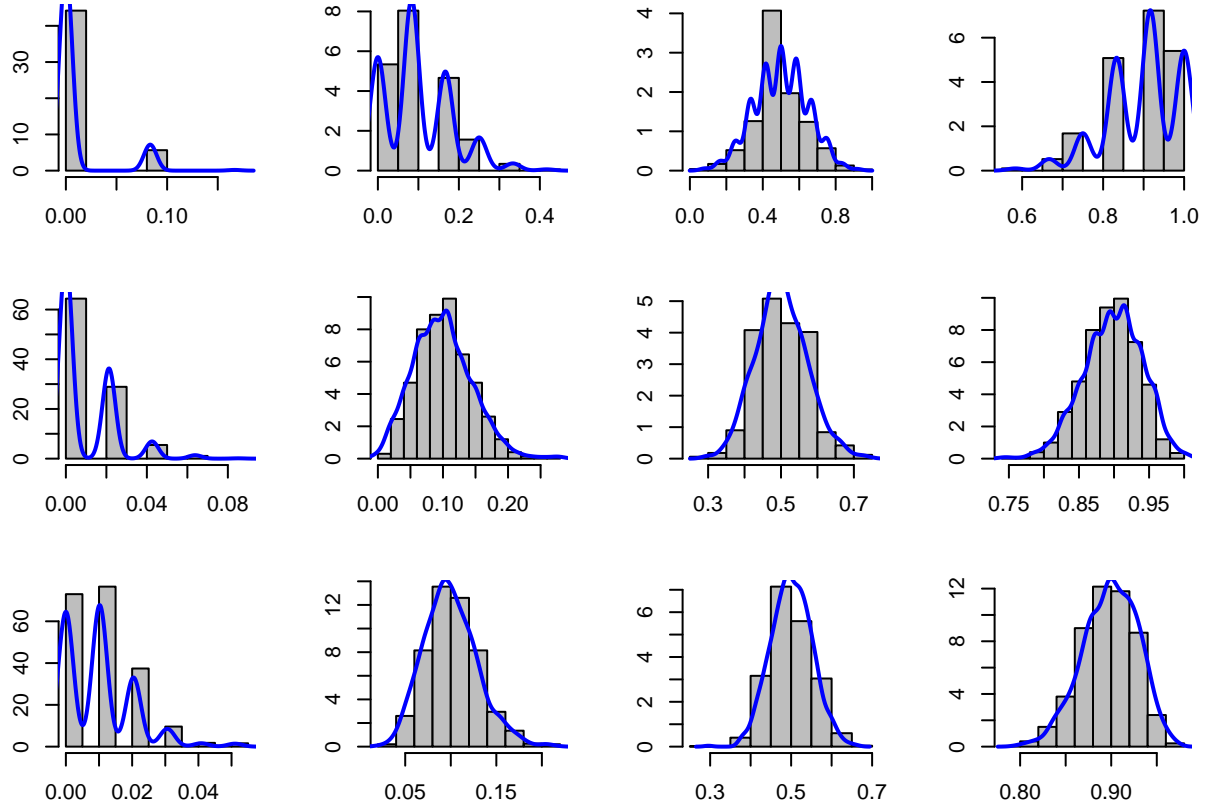|	pi = 1/101				pi = 1/2			
	.01	.1	.5	.9	.01	.1	.5	.9
N = 1000	6.44e-01	8.05e-01	2.71e-01	1.86e-02	1.00e-04	1.01e-02	2.22e-01	1.02e-02
N = 5000	6.44e-01	7.99e-01	2.54e-01	1.20e-02	9.35e-05	9.94e-03	2.21e-01	1.00e-02
N = 10000	6.43e-01	7.97e-01	2.51e-01	1.11e-02	9.21e-05	9.93e-03	2.20e-01	1.00e-02

Figure 9: Histograms of observed $\widehat{ACC}(x)$ for bibeta model with large skew ($\pi = 1/101$). 1000 simulation replicates were performed. Kernel density estimates are overlayed to demonstrate cases in which there was approximate normality. (rows) N = 1000, 5000, 10000. (columns) x = .01, .1, .5, .9.
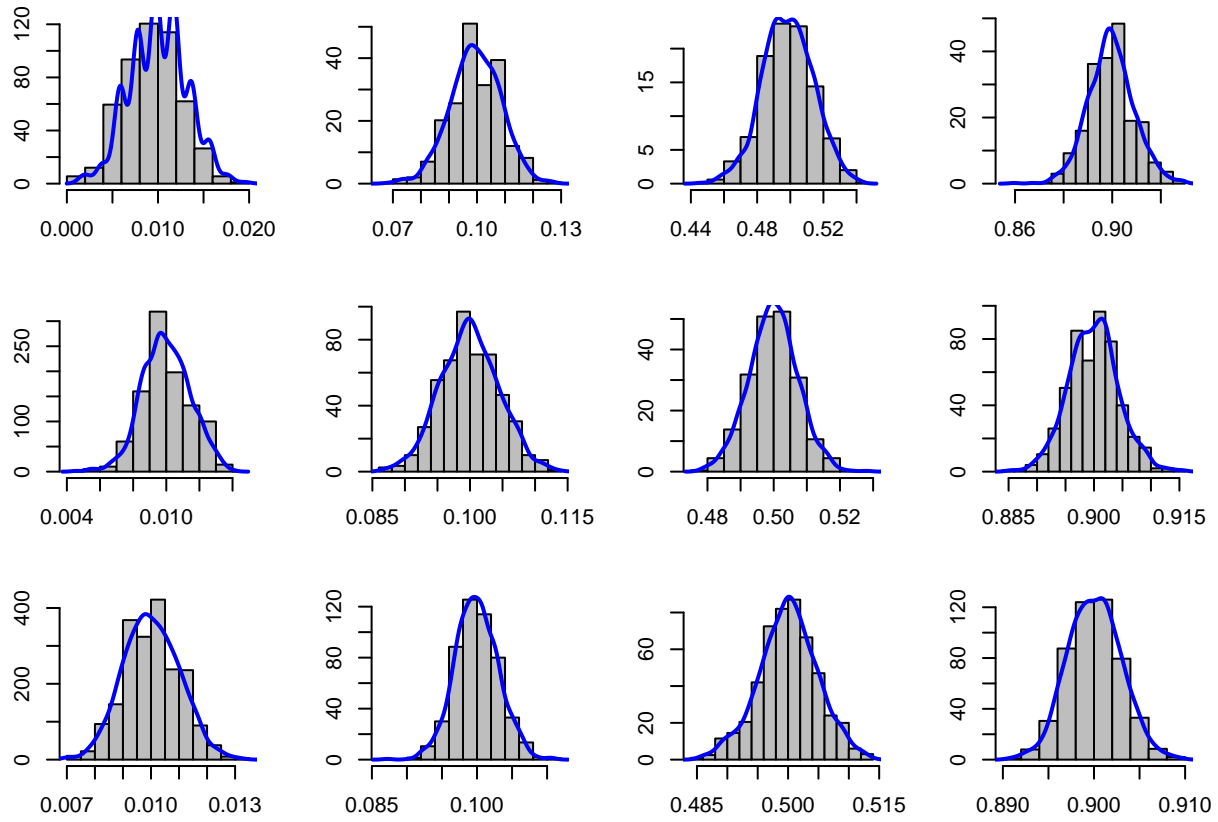
Figure 10: Histograms of observed $\widehat{ACC}(x)$ for bibeta model with skew $= 1$ ($\pi = 1/101$). 1000 simulation replicates were performed. Kernel density estimates are overlayed to demonstrate cases in which there was approximate normality. (rows) N = 1000, 5000, 10000. (columns) x = .01, .1, .5, .9.

References

Hsieh, Fushing, and Bruce W Turnbull. 1996. "NONPARAMETRIC AND SEMIPARAMETRIC ESTI-MATION OF THE RECEIVER OPERATING CHARACTERISTIC CURVE." *The Annals of Statistics* 24 (1):25–40. https://projecteuclid.org/download/pdf{_}1/euclid.aos/1033066197.

Hughes-oliver, Jacqueline M. 2000. "Population and Empirical PR Curves for Assessment of Ranking Algorithms" 1.

Pepe, Margaret Sullivan. 2003. *The statistical evaluation of medical tests for classification and prediction.* https://www.amazon.ca/dp/0198565828/ref=rdr{_}ext{_}tmb{\%}5Cnhttp://www.mendeley.com/research/statistical-evaluation-medical-tests-biomarkers-classification/.

Shuster, Jonathan. 1972. "A Generalization of the Weak Version of the Glivenko-Cantelli Theorem." *Ann. Math. Statist.* 43 (3). The Institute of Mathematical Statistics:1010–1. https://doi.org/10.1214/aoms/1177692565.

Silverman, B W. n.d. "DENSITY ESTIMATION FOR STATISTICS AND DATA ANALYSIS." https://ned.ipac.caltech.edu/level5/March02/Silverman/paper.pdf.