BOOTSTRAPPING

NOVEMBER 11, 2010

1. Bootstrapping

Another way to analyze the behavior of $\hat{\boldsymbol{\pi}}$ is to resample the data via the bootstrap. Using either parametric or non-parametric approaches, we generate B bootstrap samples of $\left\{(x_i^*, y_i^*)\right\}_{i=1}^n$, which can be run through the expectation maximization algorithm to produce B bootstrapped estimates, $\hat{\boldsymbol{\pi}}_1^*, \ldots, \hat{\boldsymbol{\pi}}_B^*$.

1.1. **Parametric data generation.** Given a $\hat{\boldsymbol{\pi}}$ from the EM algorithm, we can construct B bootstrapped samples by plugging-in $\hat{\boldsymbol{\pi}}$ into our mixture model. Using uniform random numbers we first pick the π_j from which to sample from, and then, from that π_j, we use another random number generator to pick a point on the CDF of the model data for the π_j. We then add noise so as to uniformly randomly distribute the pick over the grid specified in the model data.

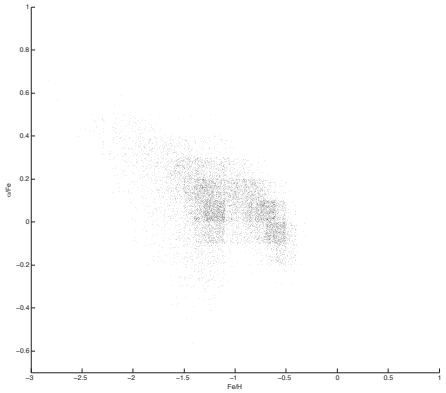

FIGURE 1. Generated parametrically

1.2. **Non-Parametric data generation.** An alternate approach is to resample the given dataset directly, without relying on a mixture model. We sample, with replacement, n data points from our original n data points, where each data point is equally likely to be picked.

1.3. **Confidence intervals for $\hat{\boldsymbol{\pi}}$.** Given B bootstrap estimates, $\hat{\boldsymbol{\pi}}_1^*, \ldots, \hat{\boldsymbol{\pi}}_B^*$, confidence intervals for each π_j can be constructed based on $\hat{\pi}_j^* \xrightarrow{d} \hat{\pi}_j$. Instead we use a centered and scaled approach based on the central limit theorem:

$$\sqrt{n}(\hat{\pi}_j - \pi_j) \xrightarrow{d} \mathcal{N}(0, \sigma_j)$$

Thus the $1 - \alpha$ confidence interval is bounded by $q_{\alpha/2}$ and $q_{1-\alpha/2}$ such that

$$P\big\{q_{\alpha/2} \leq \sqrt{n}(\hat{\pi}_j - \pi_j) \leq q_{1-\alpha/2}\big\} = 1 - \alpha$$

and since

$$\sqrt{n}(\hat{\pi}_j - \pi_j) \approx \sqrt{n}(\hat{\pi}_j^* - \hat{\pi}_j)$$

The $1 - \alpha$ bootstrapped confidence interval is bounded by $q^*_{\alpha/2}$ and $q^*_{1-\alpha/2}$ such that

$$P\big\{q^*_{\alpha/2} \leq \sqrt{n}(\hat{\pi}_j^* - \hat{\pi}_j) \leq q^*_{1-\alpha/2}\big\} \approx 1 - \alpha$$

Therefore confidence intervals can be computed as

$$\hat{\pi}_j - \frac{q^*_{1-\alpha/2}}{\sqrt{n}} \leq \pi_j \leq \hat{\pi}_j - \frac{q^*_{\alpha/2}}{\sqrt{n}} \tag{1}$$

1.4. **Covariance, correlation, and standard deviation.** The covariance matrix of the bootstrap estimates can be approximated by the sample covariance:

$$\mathrm{Covar}^*(\hat{\pi}_i, \hat{\pi}_j) = \frac{1}{B-1}\sum_{b=1}^{B}(\hat{\pi}^*_{bi} - \bar{\pi}_i)(\hat{\pi}^*_{bj} - \bar{\pi}_j) \tag{2}$$

where

$$\bar{\pi}_i = \frac{1}{B}\sum_{b=1}^{B}\hat{\pi}^*_i$$

With a standard deviation of

$$\hat{\sigma}^*_i = |\mathrm{Covar}^*(\hat{\pi}^*_i, \hat{\pi}^*_i)|$$

the sample correlation matrix is given by

$$\rho^*(\pi_i, \pi_j) = \frac{\mathrm{Covar}^*(\hat{\pi}_i, \hat{\pi}_j)}{\hat{\sigma}^*_i \hat{\sigma}^*_j} \tag{3}$$

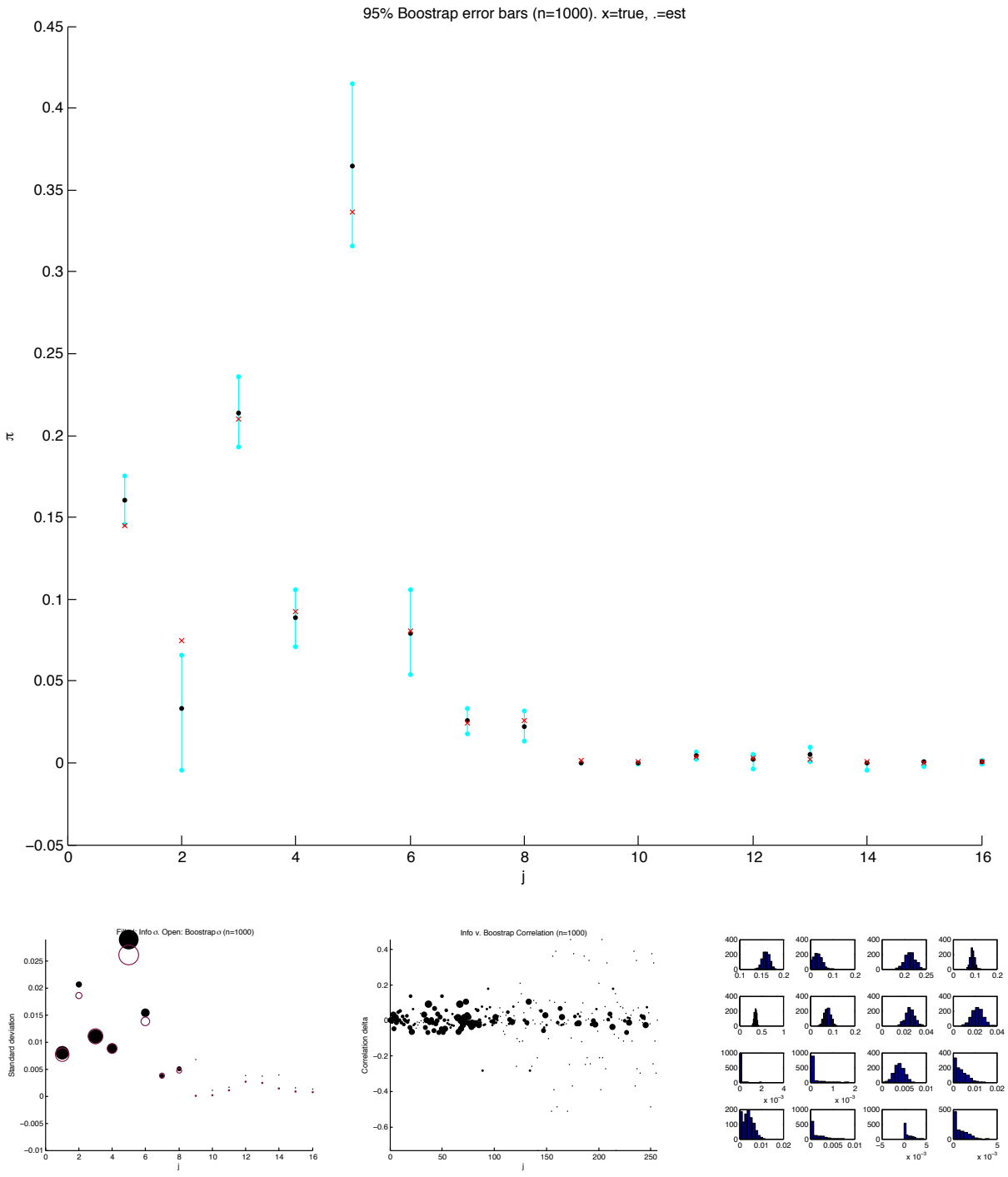

FIGURE 2. 95% bootstrapped confidence interval for n-out-of-n non-parametric resampling. Comparisons with information based confidence intervals below.

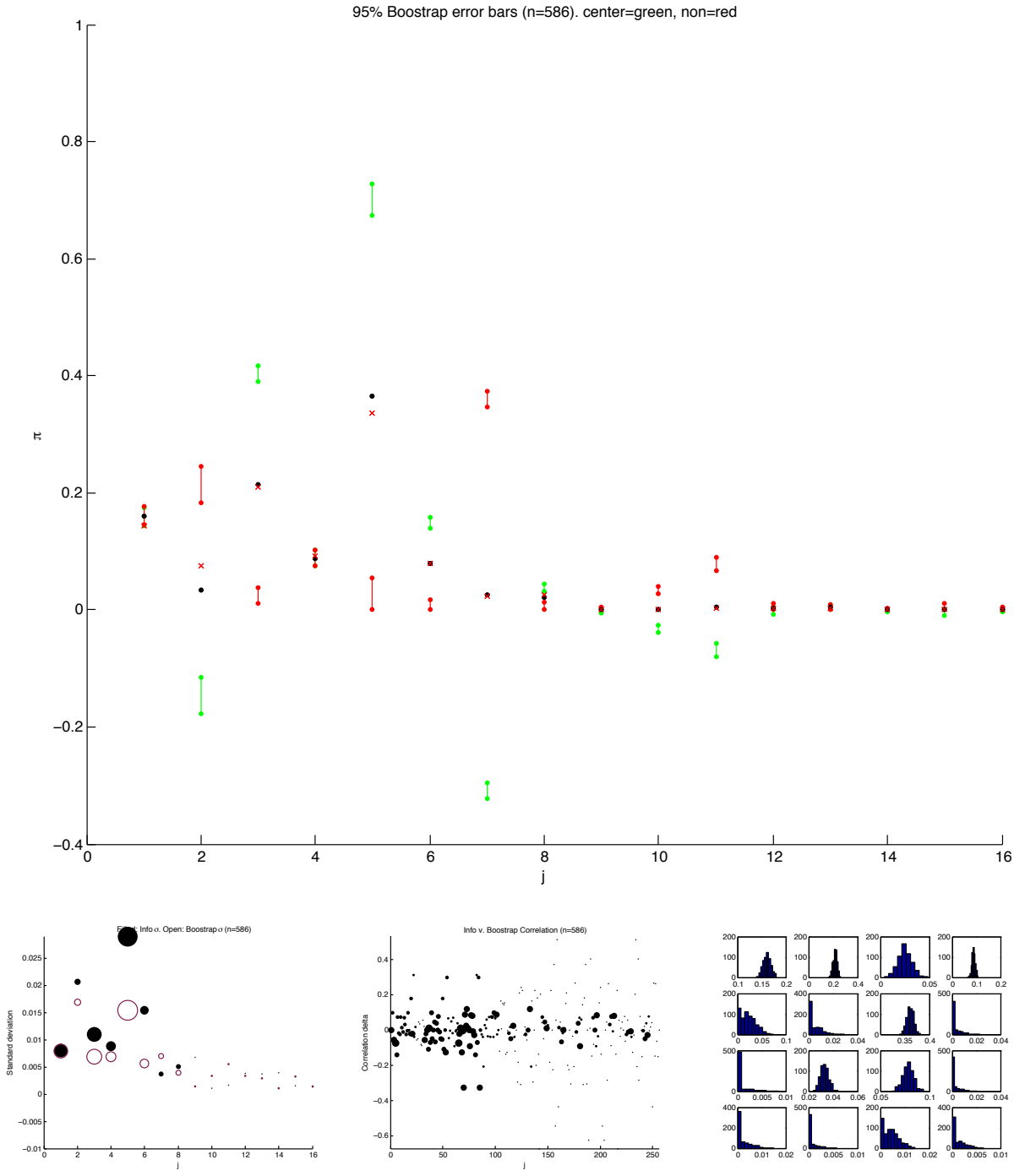

FIGURE 3. 95% bootstrapped confidence interval for generated parametric resampling. Green bars are centered, red bars are non-centered. Comparisons with information based confidence intervals below.

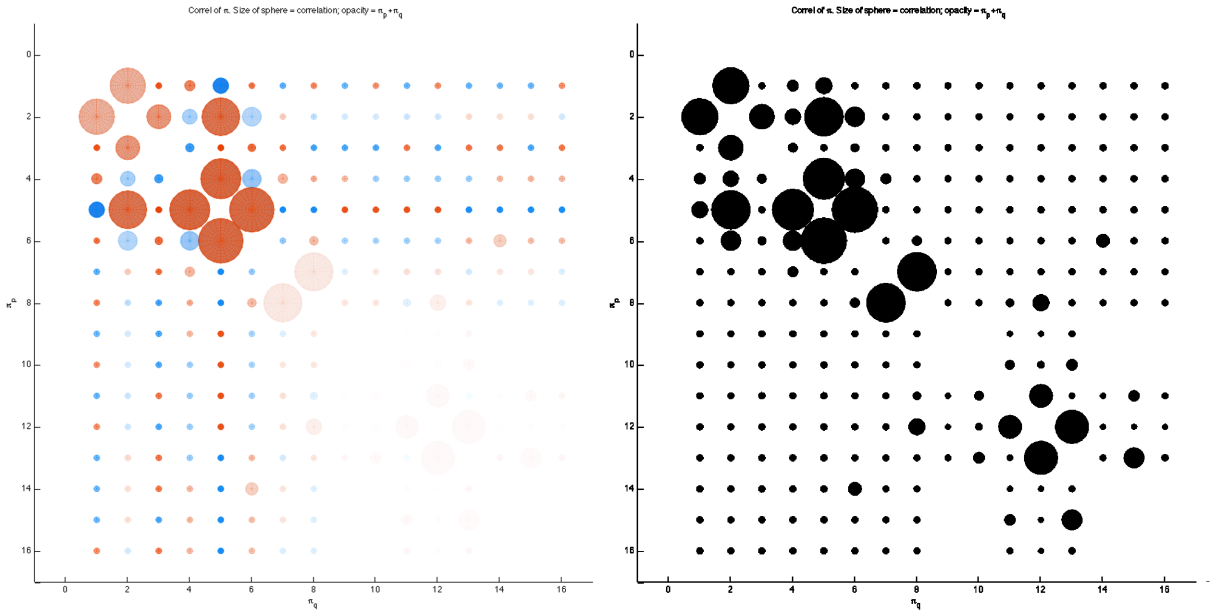

FIGURE 4. Left: correlation where size of sphere represents correlation value, and opacity of sphere represents $\pi_q + \pi_p$. Orange bubbles represent negative correlation, and blue positive. Right: Same graph, but without transparency.

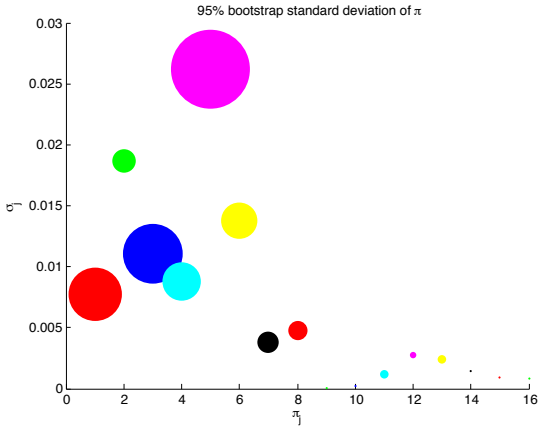

FIGURE 5. 95% n-out-of-n bootstrap standard deviation of each π_j.

TABLE 1. 95% bootstrap correlation matrix

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
1	1	-0.582	-0.058	-0.174	0.258	-0.087	0.01	-0.054	0.005	-0.058	0.009	-0.027	0.027	0.003	0	-0.029
2	-0.582	1	-0.396	0.248	-0.62	0.314	-0.004	0.013	0.018	0.007	0.028	0.013	-0.036	-0.029	-0.043	0.017
3	-0.058	-0.396	1	0.142	-0.101	-0.129	-0.03	0.004	0.007	0.061	-0.075	0.023	-0.005	-0.01	0.071	-0.029
4	-0.174	0.248	0.142	1	-0.658	0.31	-0.159	-0.022	-0.022	0.054	0.015	0.014	0.008	-0.028	-0.013	-0.013
5	0.258	-0.62	-0.101	-0.658	1	-0.731	0.038	0.018	-0.023	-0.03	-0.021	-0.004	0.016	0.087	0.015	0.025
6	-0.087	0.314	-0.129	0.31	-0.731	1	0.001	-0.147	0.019	0.019	0.011	0.006	-0.045	-0.207	-0.02	-0.064
7	0.01	-0.004	-0.03	-0.159	0.038	0.001	1	-0.627	0.057	-0.036	-0.051	-0.07	0.084	-0.032	-0.007	0.029
8	-0.054	0.013	0.004	-0.022	0.018	-0.147	-0.627	1	-0.032	-0.024	0.125	-0.261	-0.037	0.08	0.035	0.005
9	0.005	0.018	0.007	-0.022	-0.023	0.019	0.057	-0.032	1	-0.001	-0.053	0.014	-0.001	-0.021	0.038	-0.03
10	-0.058	0.007	0.061	0.054	-0.03	0.019	-0.036	-0.024	-0.001	1	-0.156	0.125	-0.176	-0.01	0.051	-0.055
11	0.009	0.028	-0.075	0.015	-0.021	0.011	-0.051	0.125	-0.053	-0.156	1	-0.373	0.046	0.056	-0.173	-0.048
12	-0.027	0.013	0.023	0.014	-0.004	0.006	-0.07	-0.261	0.014	0.125	-0.373	1	-0.54	-0.013	0.04	0.089
13	0.027	-0.036	-0.005	0.008	0.016	-0.045	0.084	-0.037	-0.001	-0.176	0.046	-0.54	1	-0.064	-0.327	-0.107
14	0.003	-0.029	-0.01	-0.028	0.087	-0.207	-0.032	0.08	-0.021	-0.01	0.056	-0.013	-0.064	1	-0.063	-0.132
15	0	-0.043	0.071	-0.013	0.015	-0.02	-0.007	0.035	0.038	0.051	-0.173	0.04	-0.327	-0.063	1	-0.087
16	-0.029	0.017	-0.029	-0.013	0.025	-0.064	0.029	0.005	-0.03	-0.055	-0.048	0.089	-0.107	-0.132	-0.087	1

TABLE 2. Means of generated $\hat{\boldsymbol{\pi}}$. B=1000 for parametric data, and B=586 for non-parametric data.

Parametric	Non-para	True
16.074	16.053	14.467
21.435	3.282	7.435
2.531	21.382	20.991
8.815	8.86	9.234
2.168	36.57	33.656
0.459	7.714	8.019
36.095	2.568	2.4
0.239	2.214	2.573
0.06	0	0.119
3.305	0.007	0.077
7.836	0.397	0.358
0.189	0.323	0.253
0.208	0.376	0.225
0.061	0.106	0.049
0.414	0.074	0.024
0.111	0.074	0.047