STAT 750: Bootstrap for Large p Clustering Problems

We saw in class that for fixed p and large n there are Bootstrap theoretical results supporting a general resampling approach to calculating the distribution of centered and standardized statistic values from the corresponding values in repeated nonparametric bootstrap samples of the same size. For the main examples of these results, the distribution in question must satisfy a Central Limit Theorem, and the statistics should be of the form

$$T(\mathbf{X}) \;=\; g\big(\frac{1}{n}\sum_{i=1}^{n} h(\underline{X}_i)\big)$$

for a continuously differentiable scalar function g, where $\mathbf{X}$ is a multivariate dataset with n independent p-vector rows $\underline{X}_i \sim (\mu, \Sigma)$, and h is a fairly general vector-valued function such that $h(\underline{X}_i)$ has second moment. Here $\mathbf{X}^*$ is a nonparametric-bootstrap resampled data matrix with rows $\underline{X}^*_i \equiv \underline{X}_{r(i)}$, where $r(i)$ are random indices in $\{1, \ldots, n\}$ that are selected independently and equiprobably.

This bootstrap theorem, which you can read about in variance references (especially those by Wassermann and Das Gupta) given on the STAT 818D Course web-page `http://www.math.umd.edu/~evs/s818D`, says that

$$\sqrt{n}\,\Big(T(\mathbf{X}) - g(E(h(\underline{X}_1)))\Big)/\sigma_T \;\;\overset{\mathcal{D}}{\approx}\;\; \sqrt{n}\,\Big(T(\mathbf{X}^*) - T(\mathbf{X})\Big)/\hat{\sigma}_T \qquad (*)$$

where the distribution on the right-hand side is understood conditionally given $\mathbf{X}$, and $\hat{\sigma}_T$ is a consistent estimator of σ_T based upon $\mathbf{X}$. This consistent estimator could be taken via the Delta Method as

$$\hat{\sigma}_T \;\;=\;\; \big[\nabla g(n^{-1}\sum_{i=1}^{n} h(\underline{X}_i))\big]'\; Var(\{h(\underline{X}_i)\}_{i=1}^{n})\; \big[\nabla g(n^{-1}\sum_{i=1}^{n} h(\underline{X}_i))\big]$$

and Var denotes the sample covariance of the sequence of n vectors $h(\underline{X}_i)$. The main value of the Theorem is that the distribution of T can be closely approximated from auxiliary randomization and a single matrix of data, and the two sides of (*) differ $o_P(1/\sqrt{n})$, less than the amount $O_P(1/\sqrt{n})$ that they differ from $\mathcal{N}(0,1)$.

In the context of clustering, the functions h of main interest are indicators related to $\underline{X}_i$ or $\underline{X}_j$ falling in specified clusters. This covers the behavior of confusion matrices for true and assigned clusters. Other questions about clusters – the choice of an optimal number of clusters, the sizes of smallest and largest clusters – may not be accessible to this theory based on sums of functions of independent random vectors.

Approach of van der Laan and Bryan 2001 when p is large

In the paper of van der Laan and Bryan (2001, Biostatistics), a **parametric** Bootstrap approach is given to problems with moderate n (say of order 50 to 100) and very large p (of order at least 10^4). The main idea is to concentrate on clustering methods (now for **variables** not subjects) that are functions of $\big(\hat{\mu}\,\hat{\Sigma}\big)$, i.e., only of the estimated means and covariances among the p variables.

The (method-of-moments) estimates $\hat{\mu}_k,\ \hat{\Sigma}_{jk}$ are expressed simply in terms of averages

$$n^{-1}\sum_{i=1}^{n}\ \big(X_{i,j},\ X_{i,k},\ X_{i,j}^2,\ X_{i,k}^2,\ X_{i,j}X_{i,k}\big)$$

Assuming that the $X_{i,k}$ variables are uniformly bounded, these averages converge with rate governed by an *exponential inequality* of Bernstein: for *iid* variables $Z_i \in [-W, W]$,

$$P\big(\frac{1}{n}\big|\sum_{i=1}^{n}(Z_i-\mu_Z)\big|\ \geq b\big)\ \leq 2\cdot\exp\big(\ -n\,\frac{b^2}{2W(1+b/3)}\big)$$

So if $n \gg \log(p)$, it follows that the right-hand side decays for large n faster than any power of p, and if the same bound W applies to all variables $X_{i,k}$, we conclude

$$P\big(\max_{1\leq k\leq p}\frac{1}{n}\big|\sum_{i=1}^{n}(X_{i,k}-\mu_k)\big|\ \geq b\big)\ \leq 2p\cdot e^{-cn}\ \ \to\ \ 0$$

for $c = b^2/(2W(1+b/3)) > 0$, and similarly (with probability bound $p(p+1)e^{-cn}$) for the $p(p+1)/2$ covariance estimates. Assuming variances are bounded away from 0, a similar statement holds for estimated correlations. These small upper-bounds on probabilities are summable over p, hence by the Borel-Cantelli Lemma with probability 1, as n, p get large, $|\hat{\mu}_k - \mu_k| > b$ for at most finitely many values of p.

The rules considered by van der Laan and Bryan for determining clusters depend smoothly on thresholds involving only estimates of μ_k, Σ_{kk} and correlations ρ_{jk}. These rules first involve determining a large subset of relatively useless variables, providing a large preliminary cluster whose complement is of much smaller size. The additional thresholding rules are used in clustering. The idea is then to use joint Central Limit Theorems for $\hat{\mu}_k,\ \hat{\Sigma}_{jk}$ to justify a parametric bootstrap. (The parametric bootstrap would be based essentially on the normal distribition for $\hat{\mu}, \Sigma$, which one could simulate approximately by treating the $X_{i,j}$ variables as though they were normal with μ, Σ parameters.) This idea worked well enough in simulations reported in van der Laan & Bryan (2001)..

Bootstrapping Residuals from Linear Models as in Kerr & Churchill 2001

In standard (univariate or multivariate) regression models $Y_i = \underline{X}_i' B + A e_i$, $1 \le i \le n$ where e_i are iid and sometimes assumed to have simple structure (indep e_i coordinates). One can bootstrap either fully nonparametrically or, with greater accuracy (as in (*) for the bootstrappeing of studentized deviates) by NP bootstrapping of residuals: with $\mathbf{X}$ fixed, and $\hat{B}, \hat{A}$ estimated from Y via least squares,

$$\hat{\epsilon}_i \equiv \hat{A}^{-1}(Y_i - \underline{X}_i' \hat{B}) \ , \qquad e_i^* = \hat{\epsilon}_{r(i)} \ , \qquad Y_i^* = \underline{X}_i' \hat{B} + \hat{A} e_i^* \qquad (\dagger)$$

This idea is discussed theoretically and practically in many bootstrap references: see especially the Das Gupta (2008) chapter 29 referenced on the STAT 818D web-page. A similar idea can be used in large-p ANOVA model settings.

In the genomics setting of Kerr & Churchill, the measurements were taken in a series of "microarrays", which gene-level activity for genes $1, \ldots, p$, are collected through pairs of "arrays" consisting of measurements contrasted with controls in "dyed" subject tissue samples (organized by "variety" and "array" (replicate or time-point). "Dyes" represent two alternative chemical preparations applied to all the "arrays", one of which is the control source. So the subject indices are $i = (v, d, a)$ and the outcome variables g. The models are

$$Y_{vdag} = \mu + A_v + D_d + (AD)_{vd} + G_g + (AG)_{ag} + (VG)_{vg} + (DG)_{dg} + \epsilon_{vdag}$$

where many of the parwise interactions were special to the genomics technology prevalent at the time of the Kerr & Churchill paper. Interesting gene-activity contrasts over varieties are estimated as $\widehat{(VG)_{vg}} - \widehat{(VG)_{0g}}$ where $v = 0$ corresponds to "control".

The key point for us is that in this application, $n = |\{(v, a, d)\}|$ is moderate and $p > 5000$ is large, and the science apparently dictates that there is no independence across genes. However: this model attempts to account for apparent joint activity among genes through the various main and interaction (fixed-effect) parameters involving g, and the **residuals** are assumed *iid* across all indices. (Later bioinformatics authors treat some of the parameter terms as random effects.) Various clustering procedures can be used, and the bootstrapping is done as in (†) above. The number of parameters is a constant time p. If cluster rules are simple and depend on these parameters (including the common variance only), then the accuracy of bootstrap can be justified as in van der Laan and Bryan. Otherwise there are no master theorems guaranteeing theoretical validity, but for some kinds of hierarchical clustering ('phylogenies'), other papers (of Felsenstein 1985 and Efron et al. 1996) provide justification.

Bootstrapping 'Phylogenies', Efron et al. 1996 (PNAS)
with relevance to hierarchical clustering

In hierarchical clustering of outcomes $k = 1, \ldots, p$ based on distances between observations $\underline{X}_i$, Efron, Halloran and Holmes (1996) would have us imagine the universe of all possible observations $\underline{x}$ (expressed in continuous-data problems through the density $f(\underline{x})$), leading to the distance matrix $D_{jk} = \left[\int_{\mathbb{R}^p} (x_j - x_k)^2 \, f(\underline{x}) d\underline{x} \right]^{1/2}$.

For this ideal universe of data, a hierarchical clustering algorithm maps to an ideal tree $\mathcal{T}(D)$. (Think: *'dendrogram'*.) With an actual $n \times p$ data matrix $\mathbf{X}$, we have instead an empirical distance matrix and tree

$$\hat{D}_{jk} \;=\; \left[n^{-1} \sum_{i=1}^{n} (X_{i,j} - X_{i,k})^2 \right]^{1/2} \quad , \qquad \widehat{\mathcal{T}} = \mathcal{T}(\hat{D})$$

Now, although the hierarchical-clustering algorithms are not smooth mappings of the distance matrix, it still makes sense to view the trees as elements in a metric space; to bootstrap the data matrix and distance matrix and tree

$$\underline{X}^*_i \;\equiv\; \underline{X}_{r(i)} \quad , \qquad D^*_{j,k} \;=\; \left[n^{-1} \sum_{i=1}^{n} (X^*_{i,j} - X^*_{i,k})^2 \right]^{1/2} \quad , \qquad \mathcal{T}^* \;=\; \mathcal{T}(D^*)$$

The central bootstrap idea, restated here, is to relate $\mathcal{T}^*$ to $\widehat{\mathcal{T}}$ in the same way that $\widehat{\mathcal{T}}$ relates to $\mathcal{T}$. For example, to estimate as a parameter the frequency θ with which $\{k_1, k_2, k_3, k_4\}$ appear together in a final four-element cluster (*'clade'*), we argue that $\hat{\theta} - \theta$ over many Monte Carlo repetitions of the data $\mathbf{X}$ is distributed nearly the same as the conditional distribution of $\theta^* - \hat{\theta}$ conditionally given $\mathbf{X}$, estimated as a relative frequency through many bootstrap repetitions.

The co-occurrence probabilities, Sensitivity and PPV we considered before, and that van der Laan and Bryan (2001) also considered, are parameters θ of just this sort and can be estimated via Nonparametric Bootstrap in this way.

The specific bioinformatics problem considered by Felsenstein (1985) and Efron et al. (1996) was discrete, related to DNA sequence matching. I am not sure whether (apart from special clustering methods like those of van der Laan depending only on μ_k, Σ_{jk} parameters) the theory establishes the validity for very large p, but the method is widely used and works well in simulations.