Resampling Methods

CAPT David Ruth, USN

Mathematics Department, United States Naval Academy

Conference on Applied Statistics in Defense

27 October 2016

Outline

- Overview of resampling methods
- Bootstrapping
- Cross-validation
- Permutation tests

Overview of resampling methods

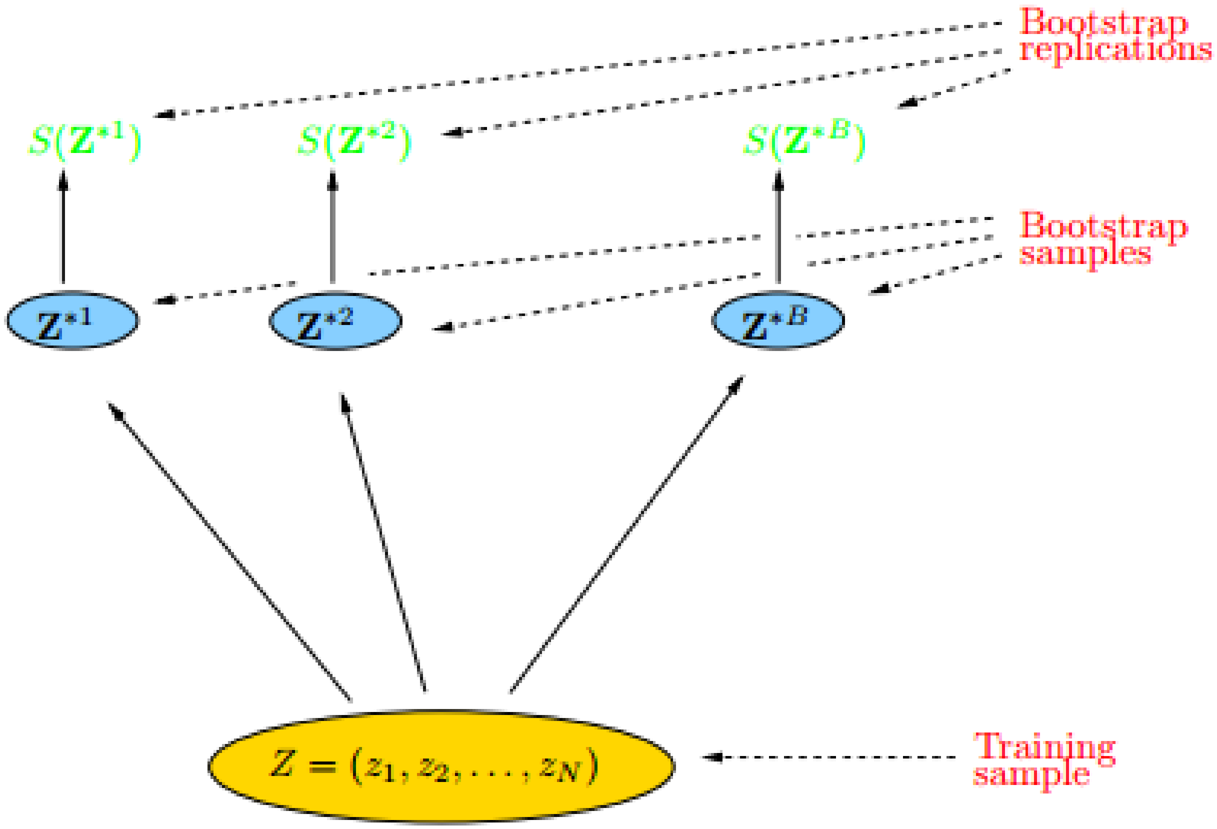
- "Re-"sampling methods => methods applied to an existing sample.
- Gist:
  - Sample from existing sample to obtain new sample(s).
  - *Bootstrapping*: Sample WITH replacement, treating original sample as a proxy for the population of interest.
  - *Cross-validation and permutation tests*: Sample WITHOUT replacement, using exchangeability assumptions for inference.

Bootstrapping

Bootstrapping

- The bootstrap is a tool for **assessing statistical accuracy**.
- Goal: Estimate any aspect of the distribution of $S(\boldsymbol{Z})$, where S is a statistic of interest and $\boldsymbol{Z} = (Z_1, \dots, Z_n)$.
- Idea:
  - Approximate the (unknown) distribution function, F, for the $Z_i's$ with the empirical distribution function, $\hat{F}$.
  - Draw "bootstrap" samples from $\hat{F}$ to estimate quantity of interest.

Bootstrapping



For example, we can estimate the variance of $S(Z)$ by

$$\widehat{\mathrm{Var}}[S(\boldsymbol{Z})] = \frac{1}{B-1} \sum_{i=1}^{B} \left(S\left(Z^{*i}\right) - \bar{S}^{*}\right)$$

Bootstrapping

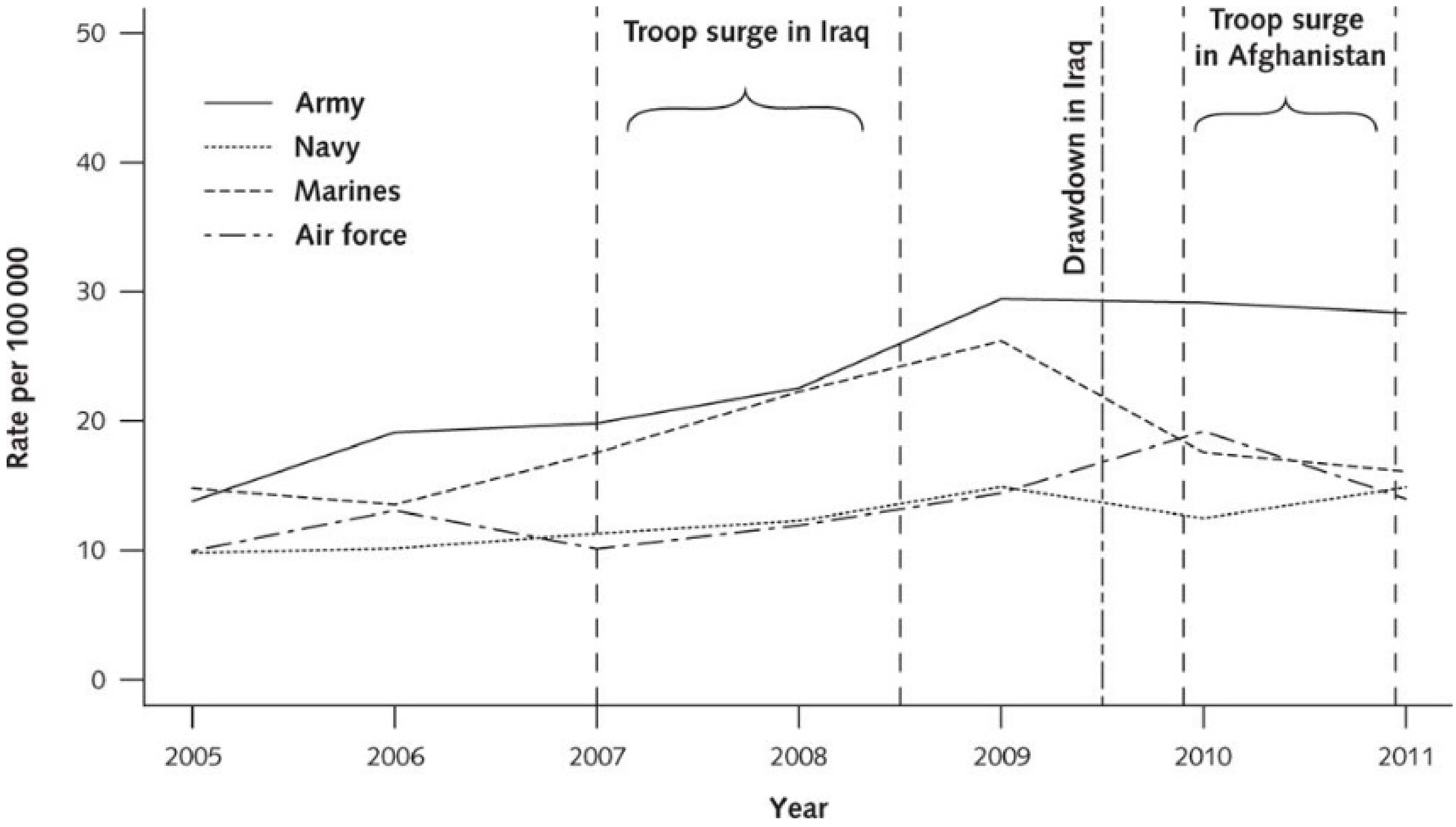
Algorithm: Given original sample $\boldsymbol{Z} = (Z_1, \dots, Z_N)$,

- Choose some large B as the number of bootstrap samples.
- for (i in 1:B){

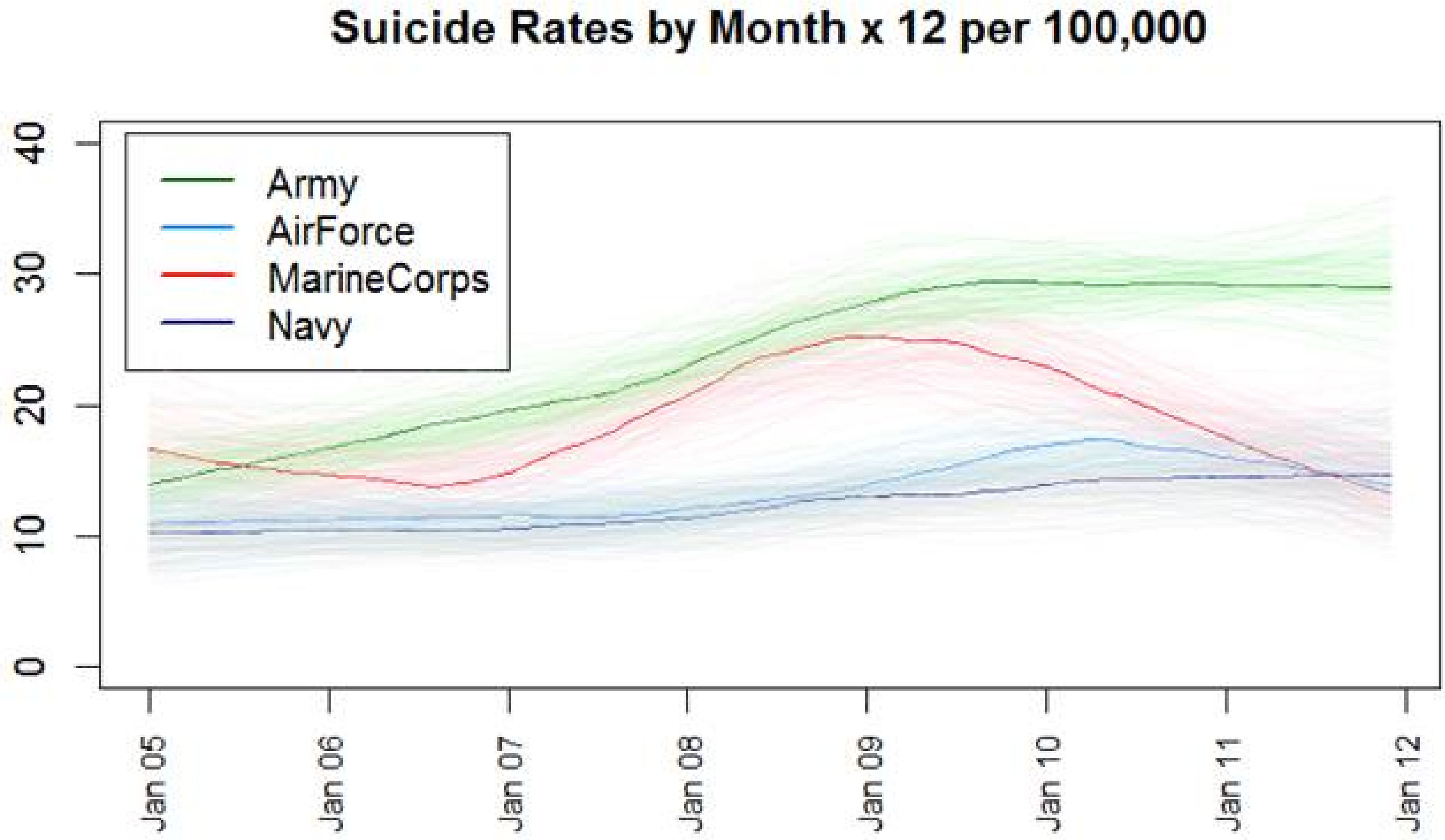
```
    # Sample $\boldsymbol{Z}$ WITH REPLACEMENT to obtain
    bootstrap sample $\boldsymbol{Z}^{*i} = (Z_{i_1}, \dots, Z_{i_N})$
    # Compute statistic $S(\boldsymbol{Z}^{*i})$
    }
```

- Use the B bootstrap replications of the $S(\boldsymbol{Z}^{*i})$'s to estimate quantity of interest.

Bootstrapping



Bootstrapping



Bootstrapping

R

Cross-validation

Cross-validation

- One of the simplest and most widely used methods for **estimating prediction error**.
- *If we had enough data*, we might set aside a "test set" or "validation set" to assess a fitted prediction model's performance.
- Why not just use training data as test data?

OVERFITTING!

Cross-validation

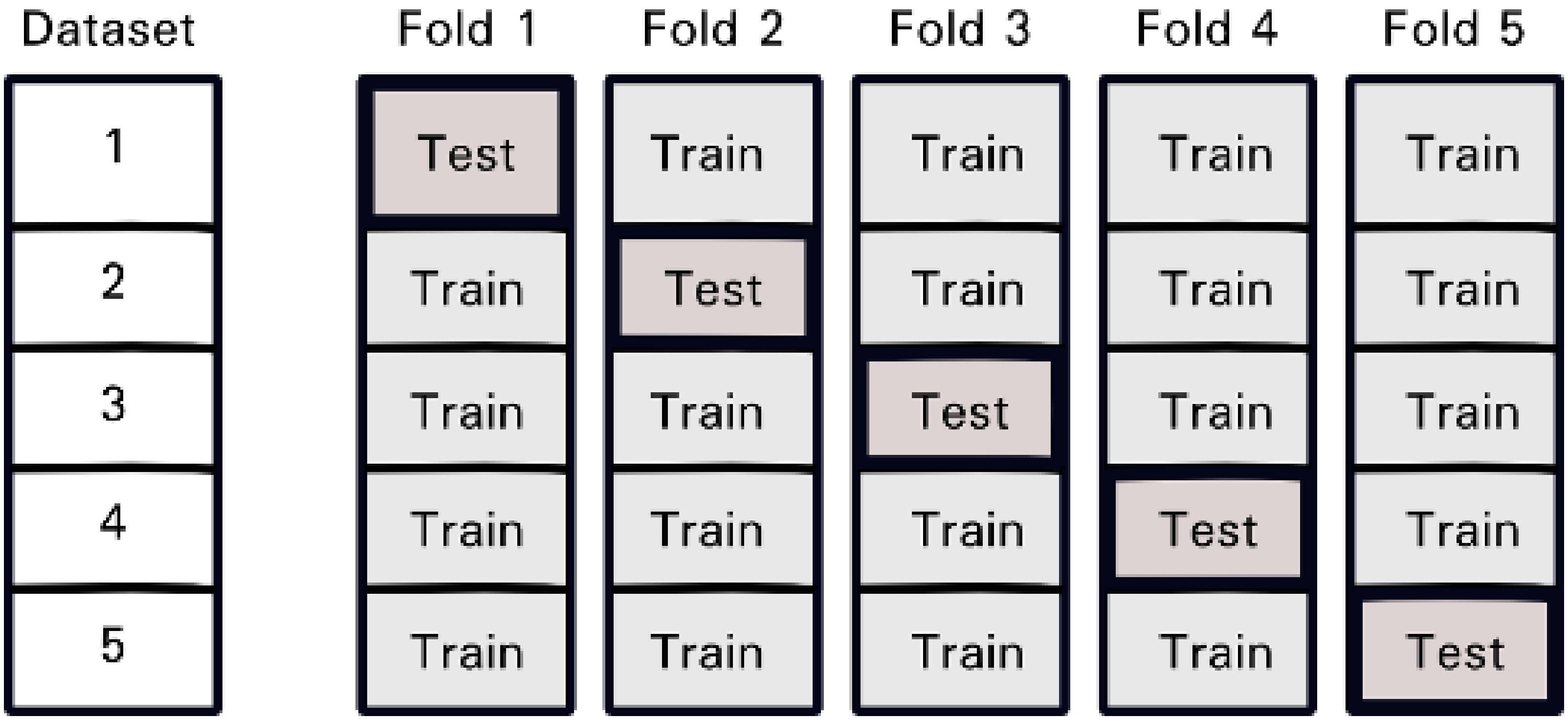
- Goal: Estimate expected extra-sample error $E\left[L\left(Y,\hat{f}(X)\right)\right]$ when X and Y are drawn from a from an independent test sample (e.g., $E\left[\left(Y - X\hat{\beta}\right)^2\right]$).
- Idea:
  - Partition original data set into training and test sets.
  - Fit model to training set; validate analysis on test set.
  - Perform multiple rounds of partitioning and average prediction error over all rounds.

Cross-validation

- K-fold cross validation is widely used with $K = 5$ or $K = 10$ recommended as a good compromise.
- Algorithm:
  - Split data into K roughly equal-sized parts.

  -
    ```
    for (k in 1:K){
    # Leave out set k, and fit model to
    remaining parts.
    # Compute prediction error for fitted model
    on set k.}
    ```
  - Average the k prediction errors.

Cross-validation



Dataset	Fold 1	Fold 2	Fold 3	Fold 4	Fold 5
1	Test	Train	Train	Train	Train
2	Train	Test	Train	Train	Train
3	Train	Train	Test	Train	Train
4	Train	Train	Train	Test	Train
5	Train	Train	Train	Train	Test

Real-world example: Heart data

- Data: heart disease diagnosis for 303 patients at the Cleveland Clinic Foundation, plus 75 other attributes
- We consider 5 quantitative and 8 categorical explanatory variables with separate binary response (presence of heart disease; evident in 139 of 303 patients):

```
  age sex cp trestbps chol fbs restecg thalach exang oldpeak slope  ca thal
1 63   1  1      145  233   1       2     150     0     2.3    3   0.0  6.0
2 67   1  4      160  286   0       2     108     1     1.5    2   3.0  3.0
3 67   1  4      120  229   0       2     129     1     2.6    2   2.0  7.0
4 37   1  3      130  250   0       0     187     0     3.5    3   0.0  3.0
```

- Data available from the UCI Machine Learning Repository at http://archive.ics.uci.edu/ml/datasets/Heart+Disease

Cross-validation

R

Permutation tests

Permutation tests

- Permutation tests may be used to test for a **distributional difference** among groups.
- Under the assumption that observations are *exchangeable*, the distribution of a test statistic may be approximated by its empirical distribution under all possible label permutations.
- "All possible" may be rather large, so…

RESAMPLE!

Permutation tests

- Simple example: difference in resting blood pressure from heart data (*univariate*)
- Noisy example: difference in resting blood pressure from heart data with tainted labels (*univariate*)
- MCC example: difference in resting blood pressure from heart data with tainted labels (*multivariate*)

Permutation tests

R

Mean Cross Count Test

Problem statement

- **Given:**

Two sets of independent multivariate observations (with categorical or quantitative attributes, or both). Formally:

$N = m + n$ independent p-variate observations

$$\mathcal{A} = \{\mathbf{X}_1, \ldots, \mathbf{X}_m\}, \qquad \mathcal{B} = \{\mathbf{X}_{m+1}, \cdots, \mathbf{X}_N\},$$

$$\mathbf{X}_i \sim F \;\; \forall i: 1 \le i \le m,$$

$$\boldsymbol{X}_j \sim G \;\; \forall j: m + 1 \le j \le N.$$

- **Goal:**

Develop a robust two-sample test against

$$H_0: \;\; F = G \, .$$

Gower dissimilarity for mixed data

- Given p-variate observations $X_1, X_2, \ldots, X_N$, the dissimilarity $d_{ij,k}$ between observations X_i and X_j on covariate k is given by

$$d_{ij,k} = \begin{cases} 0 & \text{if covariate } k \text{ is categorical and } x_{ik} = x_{jk}\,, \\ 1 & \text{if covariate } k \text{ is categorical and } x_{ik} \neq x_{jk}\,, \\ \dfrac{|x_{ik} - x_{jk}|}{R_k} & \text{if covariate } k \text{ is quantitative,} \end{cases}$$

where x_{ik} and x_{jk} are the i and j entries, respectively, in the column associated with covariate k, and R_k is the range of covariate k.

- Gower's dissimilarity measure is a weighted average:

$$d_{\text{Gower}}(X_i, X_j) = \frac{\sum_{k=1}^{p} \partial_{ij,k} d_{ij,k}}{\sum_{k=1}^{p} \partial_{ij,k}}$$

Tree-based dissimilarity

- In the CART setting, consider two observations to be "alike" if they fall into the same leaf of a tree.
- Create p trees, **using each predictor in turn as a response variable** (may prune to avoid overfitting). Let

$$\boxed{I_k(i,j) = I\{X_i \text{ and } X_j \text{ are in different leaves of tree } k\}.}$$

- Define "treeClust" dissimilarity as

$$\boxed{d_{\text{treeClust}}(X_i, X_j) = \frac{1}{K}\sum_{k=1}^{K} w_k I_k(i,j)}$$

(or variation on this theme).

- The R package "`treeClust`" finds these pairwise dissimilarities.

Mean Cross Count test

- Given some choice of informative distance measure for mixed data, we use a graph-theoretic approach to determine whether a difference exists between groups.
- Let each observation be a vertex in a graph, $\mathcal{G}$, and each pair of observations be an (undirected) edge.
- Assign interpoint dissimilarities as edge weights (from treeClust, for example).

Mean Cross Count test

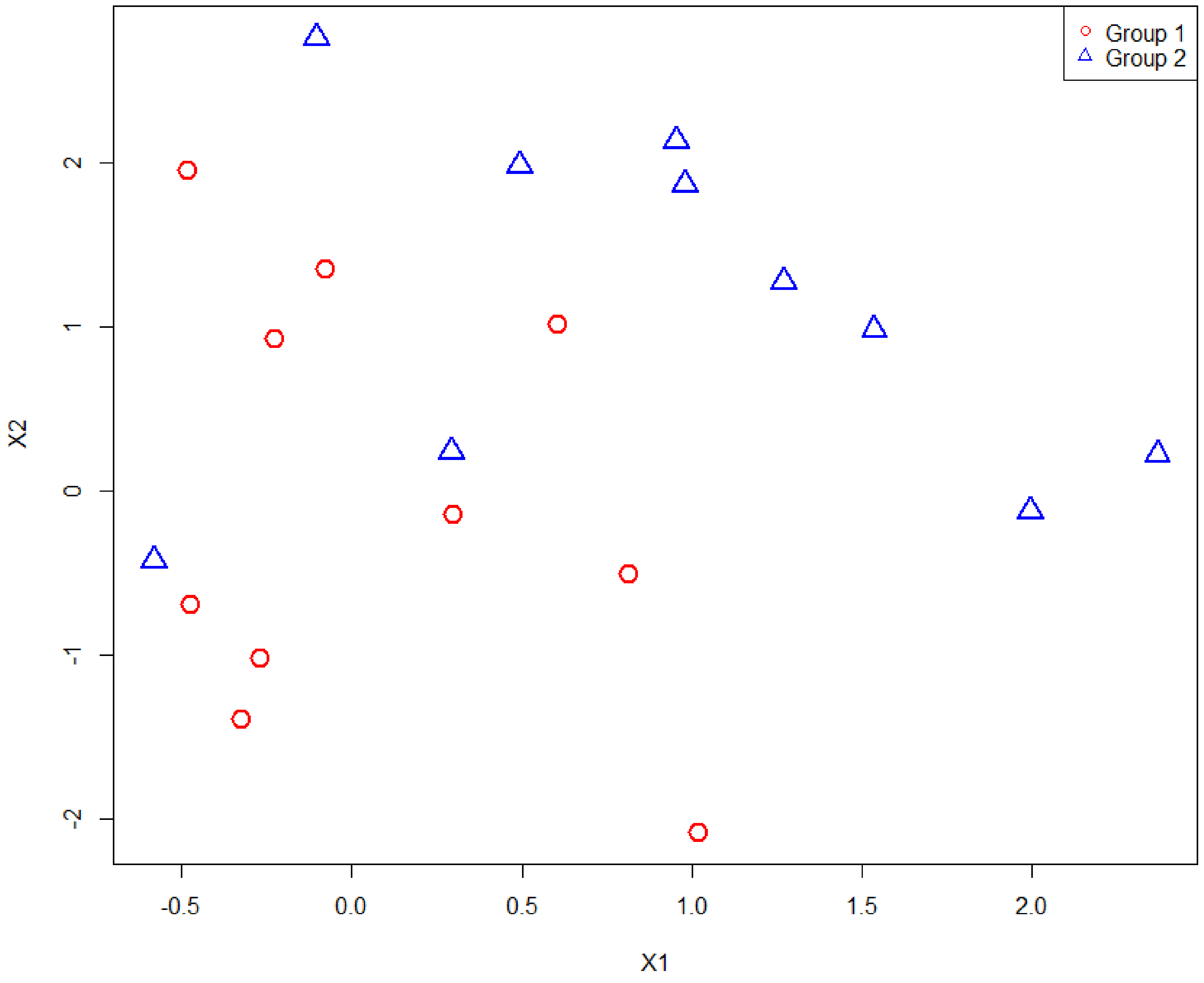
- With respect to the weighted graph, $\mathcal{G}$, find a *minimum-weight r-regular spanning subgraph*, $\mathcal{G}_r^*$.

  Note: $\mathcal{G}_r^*$ does **not** depend on group labels.

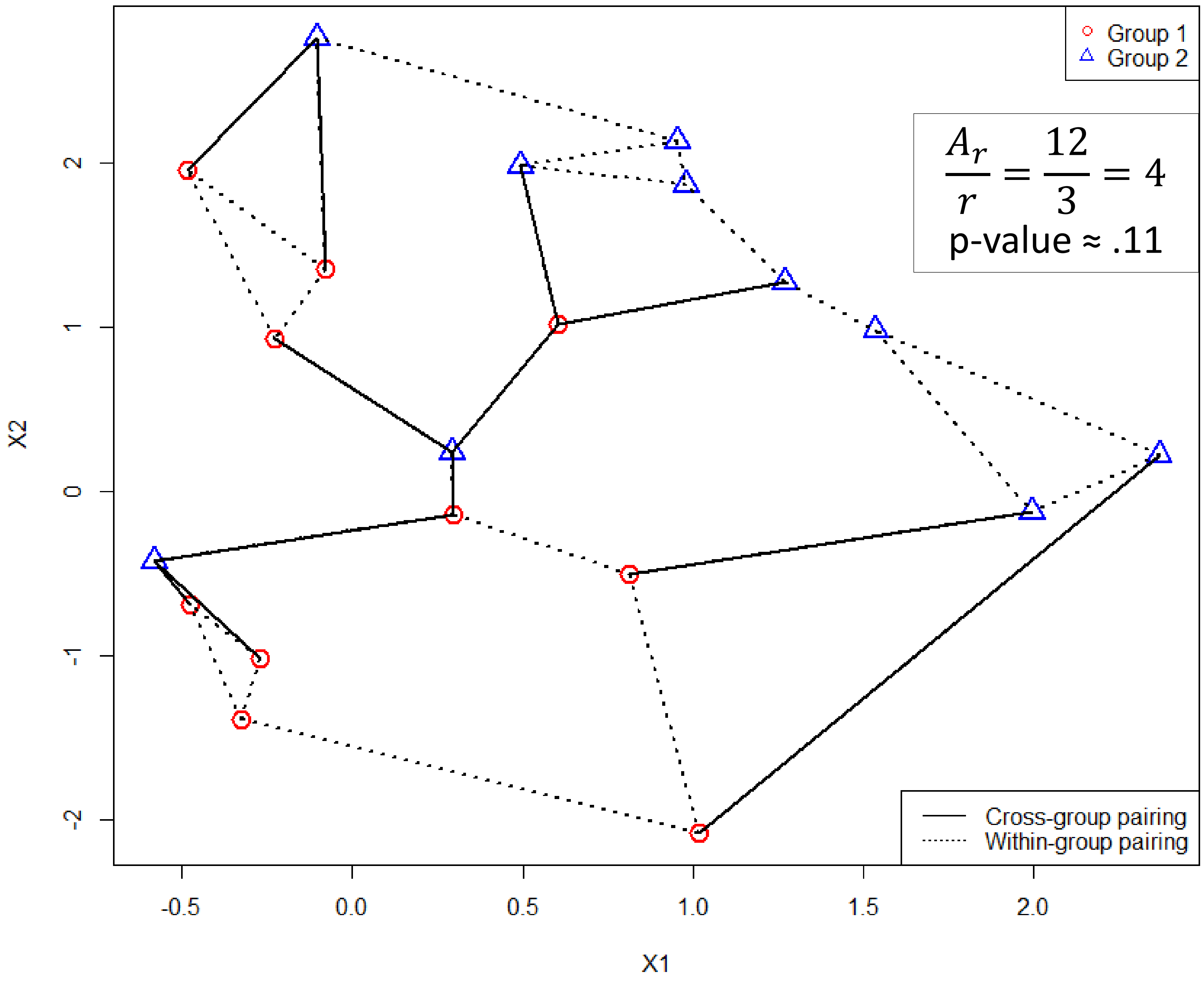
- Let A_r be the number of edges in $\mathcal{G}_r^*$ which have one vertex in the first group and the other in the second group.

- Call $T_r = \frac{A_r}{r}$ the **Mean Cross-Count** (**MCC**) and use this statistic for our two-sample test.

Bivariate data with m = n = 10



Bivariate data with m = n = 10



$$\frac{A_r}{r} = \frac{12}{3} = 4$$

p-value ≈ .11

Supporting R packages

- `treeClust` (***treeClust***ering)
  - produces tree-based dissimilarities
  - allows a number of user options
- `AcrossTic` (***A*** ***c***ost-minimal ***r***egular ***s***panning ***s***ubgraph with ***T***ree ***c***lustering)
  - finds minimum weight regular spanning subgraphs and associated test statistic for two-sample problem
  - `ptest` performs permutation test for p-values
  - `plot` allows 2D view of spanning subgraph and MCC

Mean Cross Count test

R

References

- Anglemyer, A., Miller, M., Buttrey, S., Whitaker, L. (2015), "Suicide Rates and Methods in Active Duty Military Personnel, 2005 to 2011; A Cohort Study, " *Annals of Internal Medicine,* 165(3).
- Blake, C. and Merz, C. (1998), UCI repository of machine learning databases, University of California, Irvine, Department of Information and Computer Sciences.
- Buttrey, S. and Whitaker, L. (2015), "treeClust: An R Package for Tree-Based Clustering Dissimilarities," *R Journal,* 7(2).
- Friedman, J., Hastie, T., Tibshirani, R., *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, Springer Series in Statistics, 2009.
- Ruth, D. (2014), "A New Multivariate Two-Sample Test Using Regular Minimum-Weight Spanning Subgraphs," *Journal of Statistical Distributions and Applications*, 1(1).