Randomization and Data Partioning.

From Gentle Chpt 3.

Although subsampling, resampling & o.w. rearranging a given dataset cannot increase its information content it can be helpful in extracting information.

3.1 Randomization Methods

Basic idea: Compare an observed configuration of outcomes w/all possible configurations.

These types of randomization tests have been used for a long time (Fisher 1935 - "lady tasting tea" experiment) on small data sets, & only recently have been more widespread due to extensive computations.

Consider the problem of testing if the means of two data ~~sets~~ generating processes are equal. The decision will be based on observations of the two samples

$$x_1, x_2, \ldots, x_{n_1}$$
$$y_1, y_2, \ldots, y_{n_2}$$

resulting from the two different treatments. There are several statistical test for the null hypotheses H_0 = the means are equal however we will use the unscaled test stat.

$$t_0 = \bar{x} - \bar{y}.$$

To apply a randomization test to this problem we will estimate the significance of the test stat. by comparing it to the same test statistic computed for all possible config of the observations. (that is for all possible arrangements of the observations) and then ranking the observed value t_0, within the set of all computed values. The if it has a low prob under the null hyp. but a rel higher prob under the alternate hyp. we reject it. or v/v.

More specifically, for our example consider a diff configuration of the same set of observations.

$$y_1, x_2, \ldots, x_{n_1}$$
$$x_1, y_2, \ldots, y_{n_2}$$

not all single pair switches.

where y_1 & x_1 have been switched. We now compute $t_1 = \bar{X}_{new} - \bar{Y}_{new}$ And so we continue - switching a ~~single pair~~ values in the data sets, eventually obtaining $\binom{n_1+n_2}{n_2}$ different configurations. and hence $\binom{n_1+n_2}{n_2}$ test statistics.

For this ex. if all $\binom{n_1+n_2}{n_2}$ compared & t_0 is the kth largest then we reject Ho with sig level k/M.

Then w/o making any assumptions about the distribution of the r.v. corresp. to the test statistic, we can consider the set of computed values to be a realization of a random sample from the null hypotheses and compute the empirical significance of t_0 compared to the other values. (rank).

See 2A

These ideas can be expanded & used in a variety of ways. There is an example in the handout where they divide a rectangle into quadrants & subquand & the randomization comes in the form of rearranging the grid. Also - if the # of rearrangements is too high sometimes a random sampling is used instead.

Your book briefly discusses Randomization Methods in Sec 9.7 "Permutation Tests".

Ex/9.9 pg 272.

Consider a medical experiments in which rats are randomly assigned to treatment & control groups.

- For rat i the outcome X_i is measured.
- null hypothesis: outcome does not depend on whether the rat was in treatment or control.
- alternative hypo: outcomes larger for rats labeled treatment.
- A test statistic T is used to measure the diff of the two outcomes.

  Ex: $T = \text{mean (treatment)} - \text{mean (control)}$

  and has value t_1 for the obs. data.
- under null hypo labels mean nothing $\therefore$ any shuffling of the labels should not change the joint null dist of the data.
- So calc $t_2, \ldots, t_M$ for all M permutations and compare to t_1.
- Back to 2.

3.2 Cross Validation for Smoothing & Fitting.

Assume some observed data set from $S \subseteq X \times Y$.

Consider the problem of fitting Y using X, ie the problem of determining a function $g_{XY}(X)$ such that $Y \sim g_{XY}(X)$. Examples: regression, classification, & density estimation.

For a given point (x_0, y_0) we can ask how well does $g_{xy}(x_0)$ match y_0? This answer probably depends on whether (x_0, y_0) was used in determining g_{xy}, if so, you expect a closer fit than if not. So the question really is
"How well will our fitted model g_{XY} perform at new points – how useful is it as a predictor?"

First we need a measure of the error between the observed value y and the predicted value g.
- $R(y, g)$
(Example: $(y-g)^2$)

To answer our question we need the expected value of this error w.r.t to cond dist of Y given X, $P(Y|X)$

or approx of ^ This is the value that we would like to minimize with the fit.

- $E_{P_{Y|X}}\left(R(Y_0, g_{XY}(x_0))\right)$

However, we don't know $P_{Y|X}$. so we use our fitted function $g_{XY}(x)$ as an estimate of $P_{Y|X}$ in order to estimate the expected value.

- $E_{\hat{P}_{Y|X}}\left(R(Y_0, g_{XY}(x_0))\right) = \frac{1}{n}\sum_{i=1}^{n} R(y_i, g_{XY}(x_i))$.
For observed $(y_i, x_i) \in S$

This estimate is called the "apparent error" and is often smaller than the true error for a specific x_0 $\rightarrow$ (since fit aims to minimize).

Q: Can we get a better estimate of the true error?

Consider instead, that we partion our data set into two parts S_1 & S_2.

S_1 - training or estimating set and will be used to get the fit g_{1XY}.

S_2 - validation or test set and can be used to estimate the expected error

$$E_{\hat{P}_1 Y|X}\left(R(y_i, g_{xy}(x_i)) \right) = \frac{1}{\#(S_2)} \sum_{i \in S_2} R(y_i, g_{1XY}(x_i))$$

This quantity is likely to be larger & closer to the true error.

Similarly we can switch the roles of S_1 & S_2 and combine the estimates to get:

$$E_{\hat{P} Y|X}\left(R(Y_0, g_{XY}(x_0)) \right) = \frac{1}{n} \left[\sum_{i \in S_2} R(y_i, g_{1xy}(x_i)) + \sum_{i \in S_1} R(y_i, g_{2xy}(x_i)) \right]$$

This idea is an old one and is called balanced half-sampling, but it illustrates the basic idea behind cross validation.

More generally, CV is forming multiple partial data sets with overlap (by leaving out one or more observations) and then comparing the fitted values w/ the observed values.

a common example: K-fold cross validation

- Divide the sample into K approx equal sized subsets
- 1 by 1 hold each subset back & fit gen ~~[illegible]~~ a w/ the remain K-1 subsets
- measure the prediction error by using the subset held back.
- This gives K estimates of the pred error which can be averaged to find an overall estimate.

In addition Cross validation can be useful in model building to help avoid overfitting by choosing a smaller subset ~~S~~ that provides a good fit.

The book gives a good example of CV in least squares analysis.

Sec 3.3 Jackknife Methods.

- Methods that systematically partition the data set to estimate properties of an estimator computed from the full sample. (1st used to estimate the bias of an estimator).

Suppose we have a random sample $Y_1, \ldots, Y_n$ which we use to compute a statistic T as an estimator of some parameter θ for* the population from which the Y_i's were drawn. In the jack knife method we partition the Y_i into r groups of size k. (to ease our discussion assume $n = kr$).

*go to (7)

* Recall that a functional Θ is linear if for any two functions f & g in the domain of Θ & $\forall a \in \mathbb{R}$
$\Theta(af+g) = a\,\Theta(f) + \Theta(g)$.

Now, consider removing the jth group from the sample & computing the new estimator; $T_{(-j)}$ from the remaining $r-1$ groups. This new estimator $T_{(-j)}$* will have properties similar to T.

* from sample size $n-k$.

For ex: if T is unbiased - so is $T_{(-j)}$
if T is biased - so is $T_{(-j)}$ although prob. diff

Let $\bar{T}_{(\cdot)} = \frac{1}{r}\sum_{j=1}^{r} T_{(-j)}$ be the mean of the $T_{(-j)}$

the $\bar{T}_{(\cdot)}$ can also be used as an estimate for Θ.

Why should we do this?
Often the $T_{(-j)}$ can be used to obtain more information about T. (Examples to come).

First note: If T is a linear functional of the ECDF ( ) then $\bar{T}_{(\cdot)} = T$ and this will not help us. *

So how do we gain access to this additional info?
— Introducing the Jackknife.

Consider the weighted differences in the estimate from the full sample & the reduced samples

$$T_j^* = rT - (r-1)T_{(-j)}$$

We call the T_j^* "pseudo values" and their mean

$$J(T) = \frac{1}{r}\sum_{j=1}^{r} T_j^* = \bar{T}^*$$

is the jackknifed T. and this value is of interest.

(H.W.?)

Things to note about $J(T)$.

1. $J(T) = T + (n-1)(T - \bar{T}_{(\cdot)})$
2. $J(T) = nT - (n-1)\bar{T}_{(\cdot)}$
3. In most applications, $K=1 \Rightarrow r=n$. and under certain assumptions it can be shown that this is optimal.

from 5 more earlier

Jackknife ~~[illegible]~~ Estimates. (or why bother).

We would like to know the variance (bias, etc.) of the estimator T, of θ. (ie. we would like to know char. of the distribution of T). How do we get this information?

If we had enough time & resources, we could generate S samples $\underline{Y}^{(1)}, \dots, \underline{Y}^{(S)}$ from and compute $T^{(i)}$ for each.

Then estimate

- $E(T)$ by $\hat{E}(T) = \bar{T} = \frac{1}{S}\sum_{s=1}^{n} T^{(s)}$
- $\text{Bias}(T)$ by $\widehat{\text{Bias}}(T) = \bar{T} - \theta$
- variance $V(T)$ by $\hat{V}(T) = \frac{1}{S-1}\sum_{s=1}^{S}(T^{(s)} - \bar{T})^2$.

However, we don't know θ & samples are hard to get. So, we turn to clever ways of utilizing our single sample.

back to 5.

uses of the Jackknife

Jackknife Variance Estimate.

The basic idea: Although the pseudovalues T_j^* are not independent, we treat them as if they were and use $\text{Var}(J(T))$ to estimate $\text{var}(T)$!

Intuition: small variation in pseudovalues
$\Rightarrow$ small variation in the estimator.
(ie. removal of k of data doesn't cause major change).

So, in jackknife variance estimation we use as our estimator for $V(T)$, the sample var. of the mean of the T_j^*.

$$\widehat{V(T)_J} = \sum_{j=1}^{r} \frac{(T_j^* - J(T))^2}{r(r-1)} \qquad (*)$$

Often $(*)$ is taken to be the est. of the variance of the Jackknife $J(T)$.

Comments:

1. If T is the mean & $k=1$ then $(*)$ is the standard variance estimator.
2. There are other ways to est variance, and MC studies show that $\widehat{V(T)_J}$ is often a conservative est. - often larger than the true value.
3. A variant of $(*)$ using the original estimator T is sometimes used
$$\sum_{j=1}^{r} (T_j^* - T)^2 \Big/ r(r-1) \qquad (**)$$
4. $(**) \geq (*)$

omit til later.

5. Sim, we have the jackknife est. of bias.
$$\widehat{\text{Bias}(T)_J} = (r-1)(J(T) - T).$$

Jackknife Bias correction

Another common use of the jackknife is to reduce the bias of an estimator.
(From now on we assume: $k=1, \Rightarrow r=n.$).

Suppose that we can represent the bias of T as a powerseries in $1/n$.
i.e.

$$\text{Bias}(T) = E(T) - \theta = \sum_{q=1}^{\infty} \frac{a_q}{n^q}$$

where the a_q do not involve n.

If all $a_q = 0$, then T is unbiased.
If $a_1 \neq 0$, then the order of the bias is $1/n$.
and so on.

Now consider the bias of the Jackknife estimator.

$$\text{Bias}(J(T)) = E(J(T)) - \theta$$

From $J(T) = nT - (n-1)\bar{T}_{(\cdot)}$

$$= n(E(T) - \theta) - \frac{(n-1)}{n} \sum_{j=1}^{n} E(T_{(-j)}) - \theta$$

$$= n\left(\sum_{q=1}^{\infty} \frac{a_q}{n^q}\right) - (n-1)\left(\sum_{q=1}^{\infty} \frac{a_q}{(n-1)^q}\right)$$

$$= a_1 + n\left(\sum_{q=2}^{\infty} \frac{a_q}{n^q}\right) - a_1 - (n-1)\left(\sum_{q=2}^{\infty} \frac{a_q}{(n-1)^q}\right)$$

$$= a_2\left(\frac{1}{n} - \frac{1}{(n-1)}\right) + a_3\left(\frac{1}{n^2} - \frac{1}{(n-1)^2}\right) + \cdots$$

$$= -a_2\left(\frac{1}{n(n-1)}\right) + a_3\left(\frac{1}{n^2} - \frac{1}{(n-1)^2}\right) + \cdots$$

and so the bias for the jackknife is at most of order $1/n^2$. In fact if $a_q = 0 \ \forall q \geq 2$ the Jackknife is unbiased, true even if the original estimator T has bias of order $1/n$.

This reduction in bias is the major reason for using the jackknife.

Downside: depends on representation of bias as a power series in $1/n$.

$J(T) = T + (n-1)(T - \bar{T}_{(\cdot)})$. Also from

$$E(J(T)) - \theta = E(T) - \theta + (n-1)\left(E(T) - \frac{1}{n}\sum_{j=1}^{n} E(T_{(-j)})\right)$$

we obtain the jackknife est of the bias of T

$$B_j = (\bar{T}^* - T)(n-1).$$

and hence the jackknife bias-corrected est of θ is

$$T_j = nT - (n-1)\bar{T}^*.$$

Higher Order Bias Correction

Suppose we want to pursue bias correction to higher orders. One possibility is by using a second application of the jackknife. Here the pseudovalues are

$$T_j^{**} = nJ(T) - (n-1)J(T_{(-j)})$$

jackknife applied to $J(T)$.

Now, assuming the same series rep for the bias. the new second order jackknife estimator.

(HW?)

$$J^2(T) = \frac{n^2 J(T) - (n-1)^2 \sum_{j=1}^{n} J(T_{(-j)})/n}{n^2 - (n-1)^2}$$

is unbiased to order $O(n^{-3})$.

However

① w/ $J(T)$, it differs from T by at most $O(1/n)$ ∴ if T has var $O(1/n)$ the var of $J(T)$ is asymptotically the same. Not true of $J^2(T)$
∴ tradeoff ↓ bias ↑ variance.

② w/ $J(T)$ if $q \geq 2$ has $a_q = 0$ then $J(T)$ is unbias.
w/ $J^2(T)$ $q \geq 3$ has $a_q = 0$ may still be biased because of the term $\frac{a_2}{(n-1)(n-2)(2n-1)}$. is $O(1/n^3)$.

These ideas can be further generalized to systematically reduce bias by combining higher order jackknifes (pgs. 80-82).

The Delete-k- Jackknife

- Usually deleting one observation at a time is optimal and leads to reliable estimators J(T).

- However, this does not always work.

- Consider the jackknife estimator of the variance of the sample median., If we leave out only one observation at a time the median of the reduced samples will always be one of two values. ∴ The jackknife method cannot lead to a good est. of the variance. (no matter how big n is).

- Now what? Instead, delete k-observations at a time*. How big should k be?

  For median

  $n^{1/2}/k \to 0$ and $n-k \to \infty$

for all subsets of size k.

- However, like in randomization, this yields $\binom{n}{k}$ pseudovalues, which can be large.

- if necessary use a random sampling.