How Jerry ACEd Regression

Andreas Buja

Department of Statistics, The Wharton School
University of Pennsylvania

Stanford — 2019/05/15
Getting to know Jerry in the late 1970s, early 1980s:

- 1976 Jerry visits ETHZ:
  - Talk about “Recursive Partitioning” (later CART)
  - Showed the PRIM-9 movie (Fisher, Keller, JHF, Tukey, 1974)

- Projection Pursuit regression with WXS:
  PP regression, PP density estimation, PP classification, ...

- ORION workstation at SLAC: 'statistical graphics’ with WXS and JAM, inspired by the PRIM-9 system

- WXS takes a sabbatical 1981/82, has AB substitute for him:
  50% SLAC with JHF + 50% Stats Dept, teaching 233A/B
Which One is Jerry?

Andreas Buja (Wharton, UPenn)

How Jerry ACEd Regression

2019/05/15
Late 1970s: JHF and WXS revamp 233A/B, 2 of 3 quarters of the applied stats sequence for 1st year PhD students.

1980/81 WXS teaches: Trevor Hastie, ...

1981/82 AB (substitute) teaches: Art Owen, Rob Tibshirani, Mark Segal, Jan Pedersen, Tim Hesterberg, ...

Course content, in part:
- Smoothing (local averaging, local lines)
- Cross-Validation (Wahba’s leave-one-out formula)
- Permutation testing
- Bootstrap (2 years after Brad’s 1979 paper)
- Tree-Based Regression (before the CART book)
- Numerical linear algebra
- ...

Andreas Buja (Wharton, UPenn)
Early ACE in 1981/82

- Jerry tells us about work with Leo Breiman:

**How to “learn” data transformations:**

$$g(Y) \approx f_1(X_1) + f_2(X_2) + \ldots + f_p(X_p)$$

- Algorithmic idea:
  - Computational building block: A Very Fast Smoother $S(Y|X)$ to estimate Conditional Expectations $E(Y|X)$.
  - Initialize somehow, e.g., $g(Y) = Y$ and $f_j(X_j) = \beta_j X_j$ from lin. regr.
  - Alternate updating the transformations:
    $$S(\ g(Y) - \sum_{k(\neq j)} f_k(X_k) \ | \ X_j ) \ \rightarrow \ f_j(X_j) \quad (j = 1, 2, \ldots, p)$$
    $$S(\sum_k f_k(X_k) \ | \ Y) \ \rightarrow \ g(Y) \quad (g(Y)/\|g(Y)\| \rightarrow g(Y))$$
The “backfitting loop” over $X_1, \ldots, X_p$ was easy for Jerry because he and WXS had already used backfitting in PP regression.

Surprising was the idea of smoothing $\sum_j f_j(X_j)$ on $Y$ to update $g(Y)$.  
$\Rightarrow$ ACE is a canonical correlation problem rather than a regression:
$$\max_{g,f_1,\ldots,f_p} \text{Cor}(g(Y), \sum_j f_j(X_j))$$

As an eigen value/vector problem, ACE has a hierarchy of solutions.
$\Rightarrow$ Solutions are quasi-linear (monotone), quasi-quadratic, quasi-cubic, ...
$\Rightarrow$ In noisy problems, quasi-quadratic non-sense may dominate meaningful quasi-linear solutions. $\Rightarrow$ Null finding!

JHF’s smoother implementation accommodates continuous, categorical, and circular/periodic variables.

ACE was published in 1985, with discussions.
At the time, ...

- ... data analysts were timidly tinkering with Box-Cox transformations, one variable at a time;
- ... it was methodologically inconceivable how transformations could be estimated simultaneously;
- ... statisticians were mostly not computationally interested;
- ... the idea was computationally out of reach;
- ... the idea of computationally fast building blocks had been developed in adjacent engineering fields (Cooley-Tukey FFT), but its potential was not yet recognized in statistics;
- ... Tukey, JHF, WXS were among the very few with a vision of the potential of algorithmic thinking.
Subsequent Developments (Close to Home)

- Tibshirani on AVAS (1988): asymptotically variance stabilizing $g(Y)$
- Hastie & Tibs on GAMs (1990): generalizing GLMs
  \[
  \min_{f_1, \ldots, f_p} E \left[ b \left( \sum_j f_j(X_j) \right) - Y \sum_j f_j(X_j) \right], \quad b() \text{ convex}
  \]
- Donnell, AB, WXS (1994) on APCs:
  \[
  \min_{f_1, \ldots, f_p} E \left[ \left( \sum_j f_j(X_j) \right)^2 \right] \quad \text{subject to} \quad \sum_j E[f_j(X_j)^2] = 1
  \]
  - Additive implicit rather than explicit equations: $\sum_j f_j(X_j) \approx 0$
  - Symmetric analysis w/o singling out a response
  - “Concurvity” analysis for additive regressions
- AB, Brown, Kuchibhotla, George, Zhao (2019): Reweighting Diagnostic
  - Fit a linear model and pick a coefficient of special interest.
  - Up-weight the data as a function of each regressor in turn.
  - How does the coefficient change?
  \[ \Rightarrow \] Detect interactions w/o modeling them.
LA Homeless Data — Linear model

LA Homeless Data: \( Y = \text{StreetTotal} \) (# homeless in 505 LA census tracts)

Source: Richard Berk (Criminology & Stats, UPenn)

<table>
<thead>
<tr>
<th></th>
<th>( \hat{\beta}_j )</th>
<th>( \text{SE}_{\text{lin}} )</th>
<th>( \text{SE}_{\text{boot}} )</th>
<th>( \frac{\text{SE}<em>{\text{boot}}}{\text{SE}</em>{\text{lin}}} )</th>
<th>( t_{\text{lin}} )</th>
<th>( t_{\text{boot}} )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>0.760</td>
<td>22.767</td>
<td>16.505</td>
<td>0.726</td>
<td>0.033</td>
<td>0.046</td>
</tr>
<tr>
<td>MedianIncome ($K)</td>
<td>-0.183</td>
<td>0.187</td>
<td>0.114</td>
<td>0.610</td>
<td>-0.977</td>
<td>-1.601</td>
</tr>
<tr>
<td>PercVacant</td>
<td>4.629</td>
<td>0.901</td>
<td>1.385</td>
<td>1.531</td>
<td>5.140</td>
<td>3.341</td>
</tr>
<tr>
<td>PercMinority</td>
<td>0.123</td>
<td>0.176</td>
<td>0.165</td>
<td>0.937</td>
<td>0.701</td>
<td>0.748</td>
</tr>
<tr>
<td>PercResidential</td>
<td>-0.050</td>
<td>0.171</td>
<td>0.112</td>
<td>0.653</td>
<td>-0.292</td>
<td>-0.446</td>
</tr>
<tr>
<td>PercCommercial</td>
<td>0.737</td>
<td>0.273</td>
<td>0.390</td>
<td>1.438</td>
<td>2.700</td>
<td>1.892</td>
</tr>
<tr>
<td>PercIndustrial</td>
<td>0.905</td>
<td>0.321</td>
<td>0.577</td>
<td>1.801</td>
<td>2.818</td>
<td>1.570</td>
</tr>
</tbody>
</table>

\( R^2 = 0.12 \)
LA Homeless Data — ACE

\[ R^2 = 0.27 \]
LA Homeless Data — GAM

![Graphs showing the relationship between various factors and homelessness.]

Andreas Buja (Wharton, UPenn)
How Jerry ACEd Regression
2019/05/15 12/14
LA Homeless Data — Reweighting for PercVacant

Andreas Buja (Wharton, UPenn)
How Jerry ACEd Regression
2019/05/15 13 / 14
CHEERS, JERRY!

Andreas Buja (Wharton, UPenn)

How Jerry ACEd Regression

2019/05/15 14 / 14