An Old Tale of Two Computational Approaches to Regression, Updated for the 21st Century

Lawrence Hubert

Department of Psychology
The University of Illinois

Psychometric Society Meeting, Beijing, China, July, 2015

A more expansive version of this talk with greater detail and more extensive quotes and formulas:

`cda.psych.uiuc.edu/psychometric_paper.pdf`

The slides you are seeing now:

`cda.psych.uiuc.edu/psychometric_talk_beijing.pdf`

Over the last several years, I have given talks on two individuals who have some importance to the history of the Psychometric Society — Truman Lee Kelley (an invited talk at the 2013 Psychometric Society meeting in Arnheim), and Henry A. Wallace (an invited talk on Best Practices in Statistics in 2014 on the occasion of Willem Heiser's retirement at Leiden).

What I hope to do today is to connect these two individuals, or more precisely, discuss the differing approaches they each took to multiple regression.

I will then point out that the type of least-squares iterative computational strategy developed by Truman Lee Kelley for regression is alive and well in my own work over the last several decades on the structural representation of proximity matrices through ultrametrics, additive trees, city-block scalings, and similar structures.

Truman Lee Kelley

Truman Lee Kelley (1884–1961)

Kelley was one of the most prominent psychometricians of the 20th century; Professor at Stanford (1920–1930) and then at Harvard until his retirement (1931–1950)

`cda.psych.uiuc.edu/kelley_handout.pdf`

`kelley_beamer_talk_psychometric_society.pdf`

Henry A. Wallace

Henry A. Wallace (1888–1965)

Wallace was one of the most prominent politicians of the 20th century; USDA Secretary of Agriculture under Franklin Roosevelt and the New Deal (1933–1940); U.S. Vice President under Roosevelt (1941–1945)

`cda.psych.uiuc.edu/wallace_handout.pdf`

`wallace_beamer_talk.pdf`

The Multiple Regression Connection

Both careers intersect in the early 1920s around the issue of multiple regression, and in particular, its computational aspects.

For Truman Lee Kelley, it was in the publication of his *Statistical Method* (1923), and specifically, in his approach to multiple regression and how it might be done depending on the number of variables involved.

Hint: for many variables, think alternating least-squares (with the setting of a convergence criterion, and no fixed number of operations)

For Henry A. Wallace, it was in the 1925 publication, *Correlation and Machine Calculation*, with George Snedecor (Snedecor's first publication, by the way), and how to organize the computational steps in solving the multiple regression normal equations.

Hint: think Gaussian elimination or the Doolittle method (an algorithmic approach with a fixed number of operations)

This Wallace and Snedecor reference (and a second edition in 1931) was the main source cited well into the 1940s when computational aspects of multiple regression were of concern; see, for example, Guilford's 1936 *Psychometric Methods*, the year after the Psychometric Society was formed and the same year the first issue of *Psychometrika* appeared.

The Normal Equations

Suppose I have $p+1$ variables, $Z_0, Z_1, \ldots, Z_p$, each standardized to have mean 0.0 and variance 1.0

I would like to find the set of weights, $b_1, \ldots, b_p$, such that $b_1 Z_1 + \cdots + b_p Z_p \equiv (\hat{Z}_0)$ defines a least-squares fit to the values on the dependent variable, Z_0

Let $\mathbf{R}_{p \times p} = \{r_{ij}\}$ be the intercorrelation matrix among the p independent variables, $Z_1, \ldots, Z_p$; $\mathbf{r}_{p \times 1} = \{r_{0i}\}$ contains the correlations between Z_0 and $Z_1, \ldots, Z_p$

To find the weights, $[b_1, \ldots, b_p] = \mathbf{b}^{'}$, the normal equations

$$\mathbf{R}_{p \times p} \mathbf{b}_{p \times 1} = \mathbf{r}_{p \times 1}$$

must be solved for $\mathbf{b}$ (there are p equations in p unknowns)

Direct Solution Methods

Direct methods based on a *fixed* number of steps: Gaussian elimination (named as such, confusedly, in the 1950s; it was known to Chinese mathematics from the second century); the Doolittle method (from the late 1800s); Cramer's Rule using determinants (from Gabriel Cramer, 1750)

In Gaussian elimination and the Doolittle method we can represent what is being done by what is called an **LU** decomposition for $\mathbf{R}$: $\mathbf{L}$ is a unit lower-triangular matrix (thus, with ones along its main diagonal) and $\mathbf{U}$ an upper-triangular matrix.

First, find $\mathbf{y}$ so that $\mathbf{Ly} = \mathbf{r}$, and then find $\mathbf{b}$ so $\mathbf{Ub} = \mathbf{y}$. The **LU** decomposition was introduced by Alan Turing in 1948 ("Rounding-Off Errors in Matrix Processes")

Indirect Solution Methods

Direct methods are algorithmic in the sense that the process will terminate by itself after a finite number of operations. Indirect methods terminate when we conclude that the accuracy of the result thus far achieved is sufficient for our present purposes.

Indirect methods based on iteration: suppose the least-squares fit equation, $b_1 Z_1 + \cdots + b_p Z_p$, is relaxed to $w_1 Z_1 + (b_2 Z_2 + \cdots + b_p Z_p)$, where w_1 replaces b_1; b_1 can be retrieved by a three variable problem predicting Z_0 from Z_1 and $(b_2 Z_2 + \cdots + b_p Z_p)$

So, one iterative scheme for multiple regression starts with arbitrary weights, $w_1, \ldots, w_p$, and then improves the weights with three variable systems, one-at-a-time, until convergence.

So, where are our two people (Wallace and Kelley) with respect to computational matters and the solution of the normal equations?

Wallace: the monograph with Snedecor discusses Gaussian elimination (the Doolittle method) for the solution; this approach was taken up by Mordecai Ezekiel (1899–1974), a close colleague of Wallace at the Department of Agriculture, who compared it to Truman Lee Kelley's suggestion of an iterative method for "big data" (that is, for many variables)

Mordecai Ezekiel was the economic advisor for Wallace during his time as U.S. Secretary of Agriculture (during the depression years of 1933–1940); and as far as I can tell, he operated as a thug for Wallace

Kelley: his 1923 text discussed a number of methods for solving the normal equations: simple Gauss elimination for a small number of variables; determinantal solution for an intermediate number of variables (Cramer's Rule); a Kelley iterative method for a large number of variables.

Papers From JASA in the 1920s

I want to mention briefly four papers from the *Journal of the American Statistical Association* in the 1920s.

a) H. R. Tolley and M.J.B. Ezekiel (1923). A method of handling multiple correlation problems.

This emphasized only the Doolittle method (from 1878) for solving the normal equations (as developed in Yule's classic text). However, there is the following sentence: " … Kelley in his first paper on partial correlation [in 1916] has a very suggestive table of gross correlation coefficients, with a method outline for obtaining the final regression equation by a series of approximations."

Spoiler alert: Kelley published a working alternating least-squares method in 1916 for solving the multiple regression problem; this was foreshadowed in Kelley's 1914 thesis under E.L. Thorndike (*Educational Guidance*). By the way, Thorndike provided the Introduction to this Kelley 1916 publication: *Tables: to facilitate the calculation of partial coefficients of correlation and regression equations*

b) Truman L. Kelley and Frank S. Salisbury (1926). An iteration method for determining multiple correlation constants.

c) H.R. Tolley and Mordecai Ezekiel (1927). The Doolittle method for solving multiple correlation equations versus the Kelley-Salisbury "iteration" method.

Quote from Tolley and Ezekiel

... Since, however, so eminent a statistician as Dr. Kelley has apparently not become acquainted with it [the Doolittle method], it would seem worth while taking space in this JOURNAL to present this technique, which gives exact results to five decimal places and an automatic check on the accuracy of all the arithmetic with approximately one-half the work required by the iteration method to get results accurate only to the second decimal place.

... It would certainly be desirable to lighten the arithmetic of multiple correlation problems, which is heavy enough even with the Doolittle method; but until further evidence is forthcoming it would seem that the iteration method has not done so.

Kelley and McNemar Reaction

d) Truman L. Kelley and Quinn McNemar (1929). Doolittle versus the Kelley-Salisbury iteration method for computing multiple regression coefficients.

... It is believed that for a problem containing a small number of variables the Doolittle method is the best available, but it seems evident that for a large number of variables the iteration method is much the shorter. If great accuracy is demanded the Doolittle method tends to become the more expeditious, but if computation accuracy of the order of .1 of the probable error of the regression coefficient involved (assuming samples of less than 20,000) is sufficient, and the authors do so deem it to be, then for practical purposes there is no real point in choosing the longer Doolittle method.

… The long delay in presenting this present reply to Tolley and Ezekiel's 1927 criticism of the iteration method is due to the fact that the present authors have felt it incumbent upon them actually to test out the two methods in question upon a critical problem before rushing into print, and they have not earlier had the time in which to do this.

Data Representation Uses of an Iterative Projection Strategy for Solving Linear Systems of Equations

A common problem in linear algebra:

Given $\mathbf{A} = \{a_{ij}\}$ of order $m \times n$, $\mathbf{x}' = \{x_1, \ldots, x_n\}$, $\mathbf{b}' = \{b_1, \ldots, b_m\}$, and assuming the linear system $\mathbf{Ax} = \mathbf{b}$ is consistent, find $\mathbf{x}$.

Direct methods having a fixed number of steps (such as **LU** matrix factorization) may be the most well-known strategies for solving such linear systems of equations, but another method, typically attributed to Kaczmarz (1937) and based on an iterative projection strategy, could also be used.

The Kaczmarz Strategy

Define the set $C_i = \{\mathbf{x} \mid \sum_{j=1}^{n} a_{ij}x_j = b_i\}$, for $1 \leq i \leq m$.

The projection of any $n \times 1$ vector $\mathbf{y}$ onto C_i is simply $\mathbf{y} - (\mathbf{a}_i'\mathbf{y} - b_i)\mathbf{a}_i(\mathbf{a}_i'\mathbf{a}_i)^{-1}$, where $\mathbf{a}_i' = \{a_{i1}, \ldots, a_{in}\}$.

Begin with a vector $\mathbf{x}_0$, and successively project $\mathbf{x}_0$ onto C_1, and that result onto C_2, and so on, and cyclically and repeatedly reconsidering projections onto the sets $C_1, \ldots, C_m$.

At convergence we have a vector $\mathbf{x}_0^*$ closest (in a least-squares sense) to $\mathbf{x}_0$ satisfying $\mathbf{A}\mathbf{x}_0^* = \mathbf{b}$.

So, if we start with the data to be fit as x_0 (such as a given proximity measure), x_0^* is a least-squares fitted structure to the proximities that satisfy the equality constraints.

The Dykstra Strategy

A close relative to the Kaczmarz strategy (for solving linear **equality** constrained least-squares tasks) is known as Dykstra's method for solving linear **inequality** constrained weighted least-squares tasks (JASA, 1983).

The Dykstra extension to the use of a weight vector and weighted least-squares allows for an Iteratively Reweighted Least-Squares method, and for a general strategy of replacing an L_2 (least squares) loss criterion with L_1 (least absolute residuals).

Applied Statistics/Psychometrics Use of the Dykstra Method

The Dykstra method is currently serving as the major computational tool for a variety of newer data representation devices in applied statistics/psychometrics (AS/P).

For an arbitrary symmetric proximity matrix $\mathbf{P} = \{p_{ij}\}$ (of order $q \times q$ and with diagonal entries typically set to zero), a number of applications of Dykstra's method have been discussed for approximating $\mathbf{P}$ in a least-squares sense by $\mathbf{P}_1 + \cdots + \mathbf{P}_K$, where K is typically small (such as 2 or 3).

each $\mathbf{P}_k$ is patterned in a particularly informative way that can be characterized by a set of linear inequality constraints that its entries should satisfy.

We will note three exemplar classes of patterns that $\mathbf{P}_k$ might have, and all with a substantial history in the AS/P literature.

In each instance, Dykstra's method can be used to fit the additive structures satisfying the inequality constraints once they are identified,

possibly through an initial combinatorial optimization task seeking an optimal reordering of a given (residual) data matrix,

or in some instances in a heuristic form to identify the constraints to impose in the first place.

Linear and Circular Unidimensional Scales:

The entries in $\mathbf{P}_k$ should be represented by a linear unidimensional scale (suppressing throughout an additional subscript k for clarity):

$p_{ij} = |x_j - x_i|$ for some set of coordinates $x_1, \ldots, x_q$;

(or $|x_j - x_i| - c$, for an additional additive constant c)

or a circular unidimensional scale:

$p_{ij} = \min\{|x_j - x_i|,\ x_0 - |x_j - x_i|\}$ for some set of coordinates $x_1, \ldots, x_q$ and x_0 representing the circumference of the circular structure

(or $\min\{|x_j - x_i|,\ x_0 - |x_j - x_i|\} - c$, for an additional additive constant c)

Ultrametric and additive trees:

The entries in $\mathbf{P}_k$ should be represented by an ultrametric:

for all i, j, and h, $p_{ij} \leq \max\{p_{ih}, p_{jh}\}$;

(or equivalently, among p_{ij}, p_{ih}, and p_{jh}, the largest two values are equal)

or an additive tree:
for all i, j, h, and l, $p_{ij} + p_{hl} \leq \max\{p_{ih} + p_{jl}, p_{il} + p_{jh}\}$.

(or equivalently, among $p_{ij} + p_{hl}$, $p_{ih} + p_{jl}$, and $p_{il} + p_{jh}$, the largest two values are equal)

Order constraints:

The entries in $\mathbf{P}_k = \{p_{ij}\}$ should satisfy the anti-Robinson constraints:

there exists a permutation on the first q integers $\rho(\cdot)$ such that $p_{\rho(i)\rho(j)} \leq p_{\rho(i)\rho(j')}$ for $1 \leq i < j < j' \leq q$,

and $p_{\rho(i)\rho(j)} \leq p_{\rho(i')\rho(j)}$ for $1 \leq i < i' < j' \leq q$.

Resources

Hubert, L. J., Arabie, P., & Meulman, J. (2006). *The structural representation of proximity matrices with MATLAB*. Philadelphia: SIAM. (214 pp.) [ASA-SIAM Series on Statistics and Applied Probability] `srpm_mfiles`

Hubert, L., Köhn, H.-F., & Steinley, D. (2009). Cluster analysis: A Toolbox for MATLAB. In R.E. Millsap & A. Maydeu-Olivares (Eds.), *The SAGE handbook of quantitative methods in psychology* (pp. 444–511). Los Angeles: SAGE. `clusteranalysis_mfiles`

Hubert, L., Köhn, H.-F., & Steinley, D. (2010). Order-constrained proximity matrix representations: Ultrametric generalizations and constructions with MATLAB. In S. Kolenikov, D. Steinley, & L. Thombs (Eds.), *Current methodological developments of statistics in the social sciences* (pp. 81–112). New York: Wiley. `ordered_reps_mfiles`; `cda_miscellany_mfiles`