Journal de la Société Française de Statistique
Vol. 160 No. 2 (2019)

Cholesky and the Cholesky decomposition: a commemoration by an applied statistician [*]

Titre: Cholesky et la décomposition de Cholesky

Antoine de Falguerolles[1]

Abstract: Major André-Louis Cholesky was killed in action during the First World War on 31st August 1918. The centenary of his death in action is an opportunity to pay tribute to this outstanding scientist. Linear regression methods used in France at the time of his death are recalled. An early algorithm which Augustin-Louis Cauchy introduced to alleviate the computational burden in multiple linear regression is revisited. This algorithm iteratively builds an upper-triangular system of linear equations whose solution estimates the regression coefficients. It turns out that in the case of least-squares the upper-triangular system which is constructed is exactly that obtained by applying a closely related variant of the Cholesky decomposition to the normal equations.

Résumé : Le chef d'escadron André-Louis Cholesky a été tué sur le front durant la Première Guerre mondiale le 31 août 1918. Le centenaire de sa mort au combat est une occasion de rendre hommage à cet éminent scientifique. Les méthodes de régression linéaire utilisées en France au moment de son décès sont rappelées. Un algorithme anciennement introduit par Augustin-Louis Cauchy pour alléger le fardeau des calculs numériques à effectuer en régression linéaire multiple est revisité. Cet algorithme construit itérativement un système linéaire diagonal supérieur dont la solution estime les coefficients de régression. Il apparaît que dans le cas des moindres carrés ce système diagonal supérieur est exactement celui obtenu en appliquant une variante très proche de la décomposition de Cholesky aux équations normales.

Keywords: Cholesky decomposition, multiple linear regression, history of statistics
Mots-clés : décomposition de Cholesky, regression linéaire multiple, histoire de la statistique
AMS 2000 subject classifications: 35L05, 35L70

1. Introduction

Despite an eastern European sounding name, André-Louis Cholesky was a French citizen, a French officer and a French military cartographer. Cholesky (1875 - 1918) was born in France and died from his wounds in the last months of the First World War, just a hundred years ago. In this commemorative essay, I will very briefly outline his career and that of Ernest Benoit (1873 - 1956) who coined the name Cholesky's method (*la méthode du commandant Cholesky*). I will

[*] I wish to express my deepest gratitude to the Basque Center for Applied Mathematics in Bilbao, and in particular to Maria Xosé Rodriguez Álvarez, for having given me the wonderful opportunity to commemorate the hundredth anniversary of Cholesky's death in such positive environment. My warm thanks to Professor Claude Brezinski who kindly encouraged me to write my statistically biased vision of a particular work of Cholesky. I understood from his book on Cholesky that the invention of the decomposition was just a topic in his wide spectrum of interests. My deep gratitude goes to Camilla Collis who greatly helped me to structure this article. I wish also to thank the anonymous referees for their many helpful suggestions or remarks. Errors or approximations remain absolutely mine.

[1] Honorary reserve artillery lieutenant. Retired senior lecturer in statistics, université de Toulouse (III).
E-mail: antoine@falguerolles.net

 ISSN: 2102-6238

also recall some landmarks in the history of regression in France at the dawn of the First World War. The position adopted in this paper is that Augustin-Louis Cauchy (1789-1857) computed approximations of what the Cholesky decomposition would have obtained years before the birth of Cholesky. Actually, his algorithm consists of estimating the unknown coefficients of a multiple regression by repeated use of simple regressions (and repeated recalculations of the variables). In this process a system of linear equations in upper-triangular form is produced whose solution gives the estimated regression coefficients. It thus mimics the triangular system obtained by applying a closely related variant of the Cholesky decomposition to the normal equations when performing least-squares estimation in regression and, under some implementation, derives it exactly. This recycling of Cauchy's computing strategy for multiple regression will not revolutionize the actual numerical computation of Cholesky's decomposition. At most, it will throw a statistical interpretation on computations usually hidden in the backstage of statistical packages. It must be emphasized that this paper does not consider the numerical aspects involved in modern regression analysis of large data sets.

I will also call onto the stage Moise-Emmanuel Carvallo (1856-1945), the author of a useful book on statistics and probability (1912), whose path may have crossed that of Cholesky at the *École polytechnique*. Carvallo proved that Cauchy's approach led to the least-squares solution under suitable specification. In this case, it turns out that the triangular system obtained by Cauchy's algorithm is exactly the triangular form of the normal equations obtained by applying a closely related variant of a variant of the Cholesky decomposition.

2. Cholesky's decomposition for Dummies

Let S be a positive definite matrix. The Cholesky decomposition states that there is a unique decomposition of S into the product AA' where A is a lower–triangular matrix and A' its transpose. This decomposition goes back to the 2nd December 1910 as it can be read from a hand-written manuscript of André-Louis Cholesky deposited by his family in the Archives of the École polytechnique. The manuscript is reproduced and discussed in Brezinski (2005). The manuscript had not been previously published before and the decomposition was known from secondary sources. A referee remarked that one of the numerical difficulty in the Cholesky decomposition is the computing of square roots. In Cholesky's time, the use of a logarithmic table was certainly a possibility. But this would have slowed the numerical burden which was then alleviated by the use of rudimentary calculators. Therefore Cholesky used a more manageable numerical approximation (ascribed to Heron of Alexandria) which is also detailed with its implementation in Brezinski (2005, p. 218–220).

A closely related variant of the classical Cholesky decomposition, which avoids the difficulty above, is the LDL' decomposition where L is a lower–triangular matrix with all elements on its main diagonal equal to 1, L' its transpose, D a diagonal matrix of strictly positive elements.

As an old applied statistician I often had to solve, either by hand or with early numerically unreliable computer programs, linear systems of the form $Sb = s$ where S was positive definite. A direct attack was to compute the inverse S^{-1} of S and then the product $S^{-1}s$. An alternative strategy was to compute a equivalent upper-triangular system $U\tilde{b} = u$, where the tilde on the b means that the coordinates of b were possibly reordered in the process. The values of $\tilde{b}$ (and consequently b) were then easily obtained by bottom-up computation. The introduction of the variant of the Cholesky decomposition of S simplifies the procedure above. The upper-triangular system looked for is then given by $L'b = \ell$ where $\ell = (LD)^{-1}s$. This may look like a typical recipe in numerical analysis which nowadays statisticians use without their knowledge most of the time. The situation will hopefully be reversed hereafter: elementary statistical considerations do lead to construct upper-triangular systems of the form above.

3. The characters on the stage

3.1. Cholesky (1875 - 1918) and Benoit (1873 - 1956)

In *Leonore*, the web database of the *Légion d'honneur*, Archives Nationales (2018), Artillery Squadron Commander André-Louis Cholesky [1] is reported killed in action on the 31 August 1918 near Bagneux (Aisne, France) during the First World War. The terminology of Cholesky decomposition comes from the name given in 1924 by a colleague, Ernest Benoit [2] a Colonial Artillery Major in a published article, to a transformation useful in cartography (see Benoit, 1924). The title of Benoit's publication translates as follows: *Note on a method for solving the normal equations arising in the application of least-squares to a system of linear equations the number of which is lower than that of unknowns - [...] (method of Major Cholesky)*. Both officers had graduated from the *École Polytechnique* and the *École d'Artillerie et du Génie* (Artillery and Engineers Academy). Both held alternately regimental positions and detachments to the *Service Cartographique des Armées* with missions in Algeria, Tunisia and Romania, for the former, and in Africa, Indochina, and Greece for the latter. Cholesky's life and work (published and unpublished) are thoroughly documented by Brezinski in several papers (see for instance Brezinski, 2005; Brezinski and Gross-Cholesky, 2005), and in the only existing book on Cholesky which he has co-authored with Dominique Tournès (see Brezinski and Tournès, 2014). Here I will just extract from Claude Brezinski's publications that Cholesky never published what is known nowadays as his decomposition (although he had introduced it the geographic services of the army) and that the British Professor John (Jack) Todd (1911 - 2007), a numerical mathematics pioneer, was central in disseminating the decomposition outside geodetic circles.

As a former graduate of the École Polytechnique, often nicknamed l'X, X 1895 is attached to Cholesky's name to indicate the year of his admission to this prestigious military school. Benoit is X 1892. Both were awarded the order of the *Légion d'Honneur*: Cholesky in 1915, Benoit in 1913.

In the nomadic life of an officer and a cartographer, the protestant community of La Roche-Chalais was a fixed point for the household of André-Louis Cholesky [3]. His civil marriage to

[1] André-Louis Cholesky (Montguyon, 15 October 1875 – *Mort pour la France au Nord de Bagneux (Aisne) à la Carrière*, August 1818 is the son of André Cholesky, innkeeper, and Marie Garnier.

[2] Ernest Benoit (Morez, 15 July 1873 – La Garde, 28 February 1956) is the son of Charles Benoit, house painter, and Louise Romand.

[3] I wish to express my warmest gratitude to Mrs. Dominique Mignon, president of the Society for the History of Protestantism in the Dordogne Valley. With her help, I was able to contact one of André-Louis Cholesky's grandson, Pastor Philippe Gross now retired and living near Bordeaux. Very kindly he sent me copies of documents establishing the close ties that the Cholesky family had with the small city of La Roche-Chalais and its Protestant community. The Protestant church (*temple protestant*) is nowadays a city cultural hall. When used as a church, a plaque commemorating André-Louis Cholesky's death in action in August 1918 was affixed on its walls. My thanks

Henriette Brunet, a first cousin, took place in the Town Hall of La Roche-Chalais (10 May 1907). Their respective mothers, Marie and Jeanne Garnier, were born in La Roche-Chalais in a protestant family. (The wedding was duly approved by the military authorities. The head of the Ministry of War was then General Georges Picquart, a central character in the *Affaire Dreyfus* and in the 2013 historical fiction thriller by Robert Harris entitled *An Officer and a Spy*.) The blessing of their marriage took place on the following day in the protestant church (*temple protestant*) of La Roche-Chalais (11 May 1907). As carefully inscribed in the Bible that was given to them on this occasion, three of their four children were baptized at La Roche-Chalais.

3.2. Cauchy (1789 - 1857)

Augustin Cauchy is here the central character who introduced in multiple linear regression an algorithm for constructing an upper-triangular system whose solution gave the estimations looked for. A devout catholic and a royalist legitimist (faithful to the divine right of the current King and to the rule of dynastic succession in the eldest branch to the French crown), he could not swear the administrative oath of fidelity to the new King, from a lower- branch, elevated in France after the 1830 July Revolution. Like the dethroned king and family, Cauchy went into exile. While in Prague, he eventually took the position of preceptor to the *true* heir to the French throne. There, Cauchy investigated a multiple linear regression problem (Cauchy, 1836, page 195) for which he proposed a general algorithm based on the repetitive fit of simple linear regressions (without intercept) and a simplified computation of the slope estimate. Cauchy's exile did not last long and he returned to France where he resumed official teaching and researching activities (Falguerolles, 2012).

Rarely considered nowadays, Cauchy's approach was nevertheless used in parallel with least-squares for some years. An example is given by Vilfredo Pareto (Pareto, 1897, page 371) . (This is also the paper where he describes Iteratively (re-)Weighted Least-Squares in the presence of a logarithmic link function.)

3.3. Carvallo (1856 - 1945)

Moïse-Emmanuel Carvallo[4] is the son of Jacob-Jules Carvallo and Élodie-Sara Rodrigues. Both parents came from Marrano families who had obtained regular French citizenship under Napoléon (*décret de Bayonne*, 28 juillet 1808). These families had emigrated earlier from the Iberian peninsula and had settled in the Bordeaux region. They belonged to a talented network with strong connections to the Saint-Simon utopists: Olinde Rodrigues (1795 - 1851), the Pereire

also go to the town hall of La Roche-Chalais for sending me a full copy of the marriage certificate of André-Louis Cholesky and Henriette Brunet.

Montguyon was André-Louis Cholesky's birthplace and the place where his parents currently lived. In Montguyon, Mr. Raymond Nuvet played a central role for giving the name of André-Louis Cholesky to the cultural centre and for having a plaque affixed on the First World War memorial. I thank him for sending me details and a picture of the plaque. My thanks also to Mr. Nicolas Champ who informed me that Pastor Pierre Guiraud had presided over the funeral services of the parents of André-Louis Cholesky. They both died in 1929 in Montguyon and were buried in La Roche-Chalais.

[4] Narbonne, 17 october 1856 – Unknown, 30 janvier 1945.

brothers Émile (1800 -1875) and Isaac (1806 - 1880), ... who did much for the development of banking, insurance and railways in France (see Altmann and Ortiz, 2005). Emmanuel's father, X 1840 and *École des ponts et chaussées*, was a renowned civil engineer (see Cohen, 1988, note 30 on page 209) with an international career in France, Spain, and Italy.

Emmanuel Carvallo is X 1877. He defended a doctorate thesis in mathematics (theoretical optics) in 1890. In this work he came across regression problems (Carvallo, 1890). He carefully investigated the least-squares and the Cauchy approaches to regression (see below subsection 6.1). Although not completely convinced by the conclusions of Carvallo, Henri Poincaré wrote in his report: "his thesis is likely to help making a serious progress for two of the most interesting branches of mathematics, probability calculus on the one hand, and mathematical physics on the other hand" (Publication des archives Henri Poincaré, 2007, Chapter 62, p.384). (*Sa thèse est de nature à faire faire un progrès sérieux à deux parties les plus intéressantes des mathématiques, au calcul des pobabilités d'une part, et d'autre part à la physique mathématique.*)

Emmanuel Carvallo spent most of his career teaching and in particular at l'X of which he became Director of Studies. He was appreciated by the students who nicknamed their school *carva* rather than l'X. This is recalled in particular by Benoit Mandelbrot, X 1944, in his posthumous memoirs (see Mandelbrot, 2013, page 86).

Carvallo also entertained scientific relations with Spain and became a corresponding member of the Royal Academy of Madrid (1893). He published a useful and successful statistical textbook [5] which, as he wrote in the Preface, could well be the answer to the competition launched by the Royal Academy in 1910 for a book on probability and statistics addressed to anyone with a general education (see Carvallo, 1912, Préface, p. VI).

4. Why multiple linear regression?

The title of the article announces that the views adopted here are those of an applied statistician. One frequent task in applied statistics is that of regression. Regression expresses a relationship between some explanatory variable(s) and the expected value of a response variable. Still regression may arise in different contexts worth examining.

4.1. Different paradigms?

For an applied statistician it is common say that $observed = model + error$ (or $observed = expected + error$) which can be reformulated as $observed = theoretical + error$. For a cartographer $true\ value = observed + error$ which can be reformulated as $theoretical = observed + error$. Note that in the formulas above the *error* is assumed to be symmetrically distributed with respect to 0.

Both the statistician and the cartographer want to estimate the *theoretical*. The former knows that, even in the case of an exact estimation, the *theoretical* will never be observed in non triv-

[5] This book is difficult to read today because of the notations and the computational approach. Carvallo's reference distribution law is the normal with mean 0 and variance $\frac{1}{2}$. This simplifies the density of the associated folded distribution ($\frac{2}{\sqrt{\pi}} exp(-t^2)$ for $t > 0$ and 0 elsewhere). The price to pay is the introduction of an index of deviation (*écart étalon*) which is equal to $\sqrt{2}$ times the standard deviation commonly defined nowadays.

ial cases while the latter assume that it could be observed under perfect conditions. Still the estimating tools are the same, the error term just changing side.

4.2. Linearization

How does linearity arise in the general problem of regression? In many situations, the *theoretical* is a non-linear function, say f, of several unknowns parameters, say b, and of several known parameters. Under some regularity conditions, the first order Taylor expansion of f about a starting approximation b_0 of b gives $f(b) \approx f(b_0) + \nabla_f(b_0)'(b - b_0)$, where the value of $\nabla_f(b_0)$ also depends on the known parameters. This is of the general form $x'b$ (possibly at the price of introducing an unknown intercept parameter). Usually there are several (n) observations with different values of the known parameters such that $x_i'b \approx y_i$ for $i \in \{1, \dots, n\}$. These define the linear system considered in linear regression: $Xb \approx y$.

Attached to each observation i, $i \in \{1, \dots, n\}$, is possibly a positive weight which often depends on its precision. If considered, the weights will be denoted by w_i.

When solved, the linearization process can be reiterated to obtain more precise values of the unknown parameters. A typical example is the case of generalized linear models where the expected value for the observed response y_i is of the form $f(x_i'b)$, e.g. $\exp(x_i'b)$, and the associated reciprocal function is called the link function, log in the given example. In this case the well-known Iteratively (re-)Weighted Least Squares (IWLS) algorithm does the trick (see McCullagh and Nelder, 1991). But the computational aspects of IWLS will not be discussed further for simplification purpose.

5. Multiple linear regression: combining or compensating

Two situations arise which impact the estimation of the unknown coefficients. In the first, the most common to statisticians, which will be called here **combining**, the design matrix X has more rows (n) than columns (q) and in most situations many more; the symbol $\approx$ means that the associated system of equations is (usually) inconsistent. In the second, X has strictly fewer rows than columns and is named **compensating**; the symbol $\approx$ can then be read as an equality but the associated system of equation is indeterminate. This is exactly the case treated by Cholesky which Ernest Benoit published in 1924.

In this section, several available regression methods will be considered. Most of them had arisen in discussions following its oral presentation [6].

5.1. Combining: X is $n \geq q$ and rank$(X) = q$

5.1.1. Least-squares

Minimizing $||Xb - y||_2^2$ leads to the normal equations $X'Xb = X'y$ where X' denotes the transpose of X. The normal equations have a unique solution with well-known properties: linearity, unbiasedness, ... There are several ways to solve the latter depending on the length of b and

[6] A preliminary version of this paper was also presented at the *Journées de Statistique* 2018, Paris-Saclay.

the technology at hand. A standard method is to construct an equivalent upper-triangular linear system $Ub = u$. Indeed there are several ways of doing it. The Cholesky decomposition or its variant offer one possibility at least theoretically.

Note that if the method of least-squares is the automobile of statistics, as humorously written by Stephen Stigler (see Stigler, 1999, page 320), Adrien-Marie Legendre (1755-1833) is certainly its Henry Ford. The French Adrien-Marie Legendre is indeed a pioneer in the use least-squares for estimating the unknown parameters of a linear model of regression: in 1805 he did publish an analytical description of the method of least-squares along with a sophisticated example with several explanatory variables (see Falguerolles and Pinchon, 2006). .

5.1.2. *Least absolute values and minimax value*

The idea of minimizing $||Xb-y||_1$ preceded that of least-squares. The Dalmatian Jesuit Roger-Joseph Boscovich (1711-1787) and Pierre-Simon Laplace (1749-1827) had considered simple cases of L_1 regression. A century and a half latter the foundation of linear programming and, in particular, the insight provided by the Simplex method have greatly facilitated the use of this choice of objective function and its implementation for large statistical problems. Still the solution may not be unique.

A referee recalled that minimizing $\sum |x_i'b-y_i|$ could be also obtained by a simple use of IWLS: $\sum \frac{1}{|x_i'b^{(k)}-y_i|}(x_i'b^{(k+1)}-y_i)^2$.

The minimax approach, namely minimizing the maximum value of the $|x_i'b-y_i|$ over $i \in \{1,\dots,n\}$, is also ascribed to Laplace (see Chapter 4 in Farebrother, 1999). It also absorbs into linear programming.

5.1.3. *Total least-squares*

The total least-squares approach consists of minimizing the Frobenius norm of a matrix obtained by concatenation of a matrix Ξ and a vector ξ of same dimensions as X and y, $Z = ||[\Xi|\xi]||_F^2$, subject to $(X+\Xi)b = y+\xi$. Total least squares was also discussed as early as the last quarter of the 19th century. For a historical account see Markovsky and Van Huffel (2007).

5.2. ***Compensating:*** X ***is*** $n < q$ ***and rank***$(X) = n$

5.2.1. *Least-squares*

Confronted with the non uniqueness of solutions of $Xb = y$ and therefore of the normal equations, cartographers minimize $||b||_2^2$ subject to $Xb-y = 0$. Thus they introduce Lagrange multipliers λ the values of which are obtained by solving $XX'\lambda = y$. Thus $\lambda = (XX')^{-1}y$ and $b = X'\lambda$. In passing, it can be seen that λ minimizes $||(XX')\lambda-y||_2^2$, a linear regression problem with unusual definite positive design matrix XX' and associated normal equations $(XX')(XX')\lambda = (XX')y$ which obviously simplifies into $XX'\lambda = y$. Again, the use of Cholesky decomposition or its variant for solving the linear systems involved is a good choice.

5.2.2. *Least absolute values*

A similar approach could be entertained but is not used in practice. One of its drawbacks would be that minimizing $||b||_1$ subject to $Xb - y = 0$ would lead in general to a solution with at most n non null regression coefficients.

5.3. *Remarks*

The now well-known Lasso regression and Ridge regression share some vague common features with the situation above. Lasso regression attempts to contract the number of non null regression coefficient. Ridge regression is more in line with the above. By substituting $X'X + \rho I_q$ for the rank deficient $X'X$ in the normal equations, Ridge regression leads to unique solutions b_ρ which depend on the tuning parameter ρ, $\rho > 0$. However compensating offers a closed-form solution. Their comparison is not pursued further since it is not central to the matter presented in this article.

Initialization
- $Z = [E|y]$ — $n \times (q+1)$ matrix concatenating matrix E and vector y; E (design matrix) and y (response)
- $M = [0]$ — null matrix of size $q \times (q+1)$

Algorithm

```
For I = 1,...,q
    M[I,I] = 1
        For J = I+1,...,q+1
            Regress without intercept column J of Z onto column I
            Extract slope b_(I),J and residuals Z[,J] - b_(I),J Z[,I]
                M[I,J] = slope above
                Z[,J] = residuals above
        End J
End I
```

System in upper-triangular form $Ub = u$
- $U = M[,1:q]$
- $u = M[,q+1]$

FIGURE 1. *Cauchy's algorithm*

---

6. Cauchy's algorithm for multiple linear regression

The algorithm is described in Figure 1. Cauchy introduced it for the case where $n \geq q$ (see subsection 5.1). Cauchy's aim was to obtain an upper-triangular system of linear equations (with 1 on the main diagonal) $Ub = u$ obtained by cleverly repeating simple linear regressions without intercept (one response y, one explanatory variable x, one unknown parameter also called slope b, no constant term also called intercept: $x_i b \approx y_i$). Interestingly any parametric method can be selected to perform the simple regressions.

Note that if there is a constant term in the starting multiple linear regression formula, the variables must be centered to their means, a property which is preserved for the residuals whatever selection of parametric method. Then the constant term is given *in fine* by the usual formula: $\overline{y} - \sum_{j=1}^{q} b_j \overline{X[,j]}$, where $\overline{y}$, $\overline{X[,1]}$, ..., $\overline{X[,q]}$ denote the arithmetic means of variables y, $X[,1]$, ..., $X[,q]$.

6.1. Estimator in simple linear regression

To speed up the algorithm and using centered explanatory variable x and response variable y, Cauchy suggested that the least-squares estimation of the unknown parameter (slope parameter)

$$\frac{\sum x_i y_i}{\sum x_i^2}$$

be approximated by

$$\frac{\sum sign(x_i) y}{\sum |x_i|}.$$

It turns out that this fast estimator (assuming no value of x equal to its mean 0) goes back to the German Tobias Mayer (1723-1762) and followers like Simon de Laplace (1749 - 1827). Their principle is to substitute the system $LXb = Ly$, where LX is $q \times n$ and rank q, for $Xb \approx y$ (see Stigler, 1986, pages 147-148). The construction of L is central to the method. An obvious choice for a modern statistician is $L = X'$ but this is least-squares with its heavy computation of sum of cross products. An alternative choice is to construct a matrix L with elements in $\{-1, 0, 1\}$. A referee recalled the use of random matrices L to assess the variability of the estimations. For a discussion of the simpler case of Mayer's procedure in the context of classification and regression trees see Falguerolles (2009).

Mayer's method for simple linear regression is still mentioned in the second half of last century in some elementary textbooks. In one (see Louquet and Vogt, 1971, page 45) the method is called *méthode des moyennes discontinues* (method of discontinuous means).

Emmanuel Carvallo noticed that the two formulas above can be seen as particular cases of a common weighted formula. Consider the weighted least-squares estimation formula, $\frac{\sum w_i\ x_i y_i}{\sum w_i\ x_i^2}$, the choice of either $w_i = 1$ or $w_i = \frac{1}{|x_i|}$ for observations i leads to the first formula or to the second. Consider now the weighted Cauchy estimation formula, $\frac{\sum w_i\ sign(x_i) y_i}{\sum w_i\ |x_i|}$, the choice of $w_i = 1$ or $w_i = |x_i|$ for observations i leads to the second formula or the first. Had Carvallo in mind the difficult question of sensitivity/robustness of the estimations with respect to the metric? Or simply the ease of computation by hand or by rudimentary calculator? He suggests a compromise: using as weights w_i rounded values of the $|x_i|$ in the Cauchy approximation, e.g. $w_i = 5000$ when $|x_i| = 5274$ (see Carvallo, 1890, page 14). This mixed strategy will not be further considered.

6.2. Algorithm

To formalize Cauchy's algorithm, the real response (centered) is denoted by y a vector of size n. E is a full rank matrix column (column centered) of size n (rows) and q (columns). E will be tailored to suit both situations (combining, $n \geq q$, and compensating, $n < q$). The algorithm, see Figure 1, is simple and readily implementable in any mathematical software package.

Note that the method can be adapted to several uni-variate response variables with common design matrix.

6.3. Initialization

Still what goes into E?
— $E = X$ when $n \geq q$.
— $E = XX'$ when $n < q$.

6.4. Interpretation

At the end of the iterative procedure, the first q columns of M contain the coefficients of an upper-diagonal system U ($U = M[, 1 : q]$) and the last column of M (the $q+1$th) a vector u ($u = M[, q+1]$). These define a linear system $Ub = u$ which mimics (or is in some cases) the set of normal equations in upper-triangular form (see Figure 2). Furthermore, the last column of Z (the $q+1$th) contains the residuals from the fitted values as obtained by solving $Ub = u$ and computing $y - XU^{-1}u$, obviously zeros in the case of compensation.

$$Ub = u$$

$$\begin{bmatrix}
1 & b_{(1),2} & b_{(1),3} & b_{(1),4} & \cdots & b_{(1),p} & \cdots & b_{(1),q} \\
 & 1 & b_{(2),3} & b_{(2),4} & \cdots & b_{(2),p} & \cdots & b_{(2),q} \\
 & & 1 & b_{(3),4} & \cdots & b_{(3),p} & \cdots & b_{(3),q} \\
 & & & 1 & \cdots & b_{(4),p} & \cdots & b_{(4),q} \\
 & & & & \ddots & \vdots & & \vdots \\
 & & & & & 1 & \cdots & b_{(p),q} \\
 & & & & & & \ddots & \vdots \\
 & & & & & & & 1
\end{bmatrix}
\begin{bmatrix} b_1 \\ b_2 \\ b_3 \\ b_4 \\ \vdots \\ b_p \\ \vdots \\ b_q \end{bmatrix}
=
\begin{bmatrix} b_{(1),q+1} \\ b_{(2),q+1} \\ b_{(3),q+1} \\ b_{(4),q+1} \\ \vdots \\ b_{(p),q+1} \\ \vdots \\ b_{(q),q+1} \end{bmatrix}$$

FIGURE 2. *The 'normal equations' in upper-triangular form for the nested regressions (here $p < q$). The estimations are not imbedded in most cases.*

In particular any upper-left corner block of M of size p ($p = 1, \ldots, q$) and the associated first p lines of the last column of M contain the upper-triangular system needed to compute the coefficients of the regression of y onto the first p explanatory variables (see Figure 2).

6.5. Variable selection

Cauchy realized that his algorithm could be easily modified to do some sort of forward variable selection. Indeed, for a fixed value I ($I = 1, \ldots, q$) and at the end of the nested J loop, columns $I+1, \ldots, q$ of Z contain the residuals from the regression of variables $I+1, \ldots, q$ and of the response onto the first I explanatory variables. Any variable not yet introduced as a regressor other than the pre-specified $(I+1)$th in the next step could be considered. It would then suffice to renumber these two variables and to permute the associated columns in matrix Z. What could clarify this swap? Expertise? There is no definite answer.

7. Pros and cons; an example

The good news was that any parametric simple regression can be used in the process, e.g. Cauchy's. But there are at least two associated drawbacks. The numerical values of the final estimations usually depend on the ordering of the columns of block E in matrix Z (see Figure 1). Furthermore, when an overall method of fitting exists, e. g. L_1 regression, the global estimates may not coincide with those given by the algorithm, even when using the same method in the $\frac{q(q+1)}{2}$ elementary linear regressions.

As an example, a data set investigated by Pareto is considered (Pareto, 1897, see page 378). It consists of yearly data (1855-1895): number of weddings (the response) to be related to the value of exports and coal production (the explanatory variables). The theme of this data set certainly illustrates the emerging use of statistics in the quest for establishing the 'laws' of social physics. Along this line Charles-Joseph Minard (1781-1870) in his short book *La Statistique* (see Minard, 1869) mentions the work of the British Henry-Thomas Buckle (1821-1862) and among other examples, his theory that marriages depends on the rate of salaries and income (see Buckle, 1861). (Buckle supported an uncompromising determinism in the two volumes of his *History of Civilization in England*!) Minard, negative about Buckle's system, criticizes this particular example on the ground that mores, thoughts, and common feelings, might be more relevant (see his discussion on pages 12-14) and possibly more elusive too.

Pareto aptly computed the order-one differences of the data, centered the transformed data to their means, and applied Cauchy algorithm. Table 1 presents the estimated regression coefficients obtained according to different methods of simple regression (least-squares, Cauchy-Mayer, least absolute value, total least-squares). Values obtained by direct multiple regression are also reported, whenever possible. The absence of a proper metric to measure distances between the different estimations blurs an overall comparison. Still, one can check that the order of introduction of the variables has often an impact on the estimated values and that Cauchy's algorithm does not always lead to the overall estimates. A noticeable exception is the use of least-squares.

8. Using least-squares in Cauchy's approach

In this section, only the repeated use of least-squares simple regressions in Cauchy's approach to multiple regression is considered. This may seem odd since nowadays multiple regression is well mastered theoretically and computationally. But this gives the opportunity to mention further properties of Cauchy's algorithm.

$$b_0 + b_1\Delta_u + b_2\Delta_t \approx \Delta_v$$

	b_0	b_1	b_2
Cauchy's algorithm with Cauchy-Mayer estimation			
Δ_u, Δ_t	-0.57028	0.25798	0.05092
Δ_t, Δ_u	-0.37727	0.29514	0.04113
Cauchy's algorithm with least-squares			
Δ_u, Δ_t	-0.06820	0.22339	0.03876
Δ_t, Δ_u	-0.06820	0.22339	0.03876
full least-squares			
Δ_u and Δ_t	-0.06820	0.22339	0.03876
Cauchy's algorithm with least absolute values			
Δ_u, Δ_t	-0.10749	0.23655	0.03865
Δ_t, Δ_u	0.60129	0.18232	0.02203
full least absolute values			
uncentered Δ_u and Δ_v	-0.21853	0.24451	0.03760
Δ_u and Δ_v	-0.14260	0.26249	0.03711
Cauchy's algorithm with total least-squares			
Δ_u, Δ_t	0.86074	0.22036	0.01008
Δ_t, Δ_u	-0.04692	0.24194	0.03621
full total least-squares			
Δ_u, Δ_t	-0.07461	0.24169	0.03710

TABLE 1. *Pareto's investigation of the relationship between marriages and prosperity in England (1855-1895). Proxy variables for prosperity are exports and coal production. Order one differenced annual data are considered. The weddings annual variation (Δ_v) is regressed onto the annual variation of exports (Δ_u) and the annual variation of coal extraction (Δ_t). In all cases the estimated values for b_1 and b_2 are positive*

8.1. Carvallo's contribution

Moïse-Emmanuel Carvallo has proven that the least-squares estimates obtained from the normal equations and those obtained from the upper-triangular system produced by Cauchy's algorithm are identical (Chapter III, pp.103–143 Carvallo, 1912). Additionally it can be shown that this upper-triangular system is exactly the system obtained by applying the Cholesky decomposition to the normal equations. The proofs, quite elementary, rest upon properties of block matrix and induction. They are not reported here.

8.2. Variable selection

If least-squares estimation is used, the variable selection aspect in Cauchy's algorithm can be reformulated. For a given I ($I \in \{1,\ldots,q\}$), and at the end of the nested loop J ($J \in \{I+1,\ldots,q+1\}$), Z stores the residuals of the regression of $J \in \{I+1,\ldots,q+1\}$ and y onto variables $1,\ldots,I$. Then the $q-I$ correlation coefficients between columns $I+1$, ..., q of Z and column $q+1$ of Z are the conditional correlation coefficients given the explanatory variables 1, ..., I between these variables and the response. Introducing the variable with the largest conditional correlation with the response might then be a sensible choice. This is indeed very easy to implement in modern computing environments.

9. Concluding remarks

In the light of the work of Claude Brezinski there are no reason to believe that the Cholesky decomposition of a positive definite matrix S into a product AA' where A is lower–triangular matrix, is inaccurately named. It seems inconceivable nowadays that Cholesky never published it although it was in use in the French military geographic circles as recalled by Ernest Benoit in 1924. Had André-Louis Cholesky not been killed in action during the First World War, his 1910 notes would have been eventually published. This is indeed a sad (mis)interpretation of academic precept "publish or perish".

Associated with Cholesky decomposition is the LDL' decomposition. In a multiple regression context ($X_i'b \approx y_i$, $i \in \{1,\ldots,n\}$), the algorithm introduced by Cauchy computes an upper-linear system whose solution gives an estimation of the unknown regression coefficients. This system mimics the $L'S = (LD)^{-1}s$ re-expression of the normal equations considered in Least-squares. Cauchy's algorithm is appealing since it involves the repeated use of simple regressions (one explanatory variable, no intercept but just a slope: $x_i b \approx y_i$, $i \in \{1,\ldots,n\}$) at the price of iterative re-computation of the variables. Any parametric method can be used for these elementary regressions. For computing ease, Augustin-Louis Cauchy proposed one. Emmanuel Carvallo realized that if Cauchy had used least-squares for the simple regressions, he would have obtained a linear system in upper-diagonal form ($Ub = u$) equivalent to the normal equations ($Sb = s$ where $S = X'X$ and $s = X'y$). It turns out that this upper system is nothing else than $L'b = (LD)^{-1}s$ where the decomposition $LDL' = S$ is associated to Cholesky.

References

Altmann, S. and Ortiz, E. L., editors (2005). *Mathematics and social utopias in France: Olinde Rodrigues and his times*, volume 28 of *History of Mathematics*. American Mathematical Society and London Mathematical Society.

Archives Nationales (2018). Base de donnée léonore, Patronyme des légionaires. http://www2.culture.gouv.fr/documentation/leonore/recherche.htm, Last accessed on 2018-09-19.

Benoit, E. (1924). Note sur une méthode de résolution des équations normales provenant de l'application de la méthode des moindres carrés à un système d'équations linéaires en nombre inférieur à celui des inconnues – application de la méthode à la résolution d'un système défini d'équations linéaires (procédé du commandant Cholesky). *Bulletin géodésique*, 2:67–77.

Brezinski, C. (2005). La méthode de Cholesky. *Revue d'histoire des mathématiques*, 11:205–238.

Brezinski, C. and Gross-Cholesky, M. (2005). André-Louis Cholesky. *Bulletin de la Société des amis de la bibliothèque de l'École polytechnique*, 39.

Brezinski, C. and Tournès, D. (2014). *André-Louis Cholesky, mathematician, topographer and army officer*. Birkhäuser, Springer International Publishing, Switzerland.

Buckle, H.-T. (1857-1861). *History of Civilization in England (2 volumes)*. J. W. Parker & Son, London.

Carvallo, E. (1890). Mémoire sur l'optique : Influence du terme de dispersion de Briot sur les lois de la double réfraction. *Annales Scientifiques de l'École Normale Supérieure*, 7.

Carvallo, M.-E. (1912). *Le calcul des probabilités et ses applications*. Gauthier-Villars, Paris.

Cauchy, A. L. (1836). *Mémoire sur la dispersion de la lumière, publié par la Société Royale des Sciences de Prague*. J. G. Calve, Prague.

Cohen, D. (1988). Juifs allemands et juifs portuguais à Paris sous Napoléon III. In Gili, J. A. and Schor, R., editors, *Hommes, idées, journaux : mélanges en l'honneur de Pierre Guiral*, Paris. Publication de la Sorbonne.

Falguerolles, A. d. (2009). Quelques remarques sur la méthode d'ajustement de Mayer : lien avec les méthodes de classification. *Math. & Sci. hum. (Mathematics and Social Sciences)*, 187:43–58.

Falguerolles, A. d. (2012). Cauchy, Prague and multiple regression. In Komárek, A. and Nagy, S., editors, *Proceedings of the 27th International Workshop on Statistical Modelling. Prague, Czech Republic, 16-20 July 2012*, pages 96–98. http://www.statmod.org/files/proceedings/iwsm2012_proceedings.pdf.

Falguerolles, A. d. and Pinchon, D. (2006). Une commémoration du bicentenaire de la publication (1805-1806) de la méthode des moindres quarrés[7] par Adrien-Marie Legendre. *Journal de la société française de statistique*, 147(2):81–105.

Farebrother, R. W. (1999). *Fitting Linear Relationship, A history of the calculus of observations 1750-1900*. Springer, New York.

Louquet, P. and Vogt, A. (1971). *Probabilités - Combinatoire - Statistiques*. Armand Colin, Paris, 2 edition.

Mandelbrot, B. B. (2013). *The fractalist: memoir of a scientific maverick*. Random House, New York.

Markovsky, I. and Van Huffel, S. (2007). Overview of total least-squares methods. *Signal Processing*, 87(10):2283 – 2302.

McCullagh, P. and Nelder, J. A. (1991). *Generalized Linear Models*. Chapman and Hall, London, 2 edition.

Minard, C.-J. (1869). *La Statistique*. Cusset, Paris.

Pareto, V. (1897). Quelques exemples d'application des méthodes d'interpolation à la statistique. *Journal de la Société de statistique de Paris*, 38:36–379.

Publication des archives Henri Poincaré (2007). *La correspondance entre Henri Poincaré et les physiciens, chimistes et ingénieurs*. Birkhäuser, Basel.

Stigler, S. M. (1986). *The History od Statistics, The Measurement of Uncertainty before 1900*. Harvard University Press, Cambridge, Massachusetts.

Stigler, S. M. (1999). *Statistics on the Table: the History of Statistical Concepts and Methods*. Harvard University Press, Cambridge, Massachusetts.

[7] The change in spelling of *quarré* in *carré* occured later in the 19th century.