GAUSS'S LEAST SQUARES CONJECTURE

JOAKIM EKSTRÖM

Abstract. This article investigates the claim of Gauss, made in Theoria Motus, that deviations from the normal distribution is of no importance in practice relative to whether the method of generalized least squares yields the most probable value under the density criterion. The main result of the article is that given an independence assumption the method of generalized least squares yields the most probable value under the density criterion if and only if the observational errors are median zero normally distributed. As a consequence, the method of generalized least squares carries within itself a normal distribution assumption, in this sense. The corresponding claim under Pearson's distance criterion is also studied, yielding a similar conclusion.

1. Introduction

In *Theoria Motus* (1809), Carl Friedrich Gauss sought to determine Kepler orbits of observed heavenly bodies. To this end, he employed the fifth axiom of Bernoulli (1713), *"Between two, the one that seems more probable should always be chosen"*, the density criterion of Lambert (1760, 1765), *"Given a probability distribution, the value with the higher probability density is deemed to be more probable than the value with the lower probability density"*, and his own theory of observational errors, by which hypotheses can be empirically evaluated through their corresponding residuals.

Gauss's work yielded the method of generalized least squares. If the joint probability distribution of the observational errors has covariance matrix Σ, then the most probable Kepler orbit is that whose residual vector x minimizes the generalized least squares expression $x^{\mathrm{t}}\Sigma^{-1}x$, where x^{t} denotes matrix transpose of x. The expression can be rewritten through $x^{\mathrm{t}}\Sigma^{-1}x = \|\Sigma^{-1/2}x\|^2 = \|x\|_\star^2$, where $\|x\|^2 = x^{\mathrm{t}}x$ denotes the squared Euclidean norm, and $\|x\|_\star$ is a separate covariance-adjusted norm. As a consequence the method of generalized least squares is intuitively interpretable in geometric terms. Furthermore, Legendre (1805) had shown that the minimum of the generalized least squares expression is easily found by solving his system of normal equations.

The method is remarkably elegant; it has a probabilistic foundation, it is geometrically interpretable, conceptually intuitive, and it was easy to apply in practice during the time when there were no electronic computers. However, there is one circumstance that detracts from the elegance of the method: the derivations of Gauss rest on an arbitrary normal distribution assumption.

Not only is the assumption of normally distributed observational errors arbitrary, it is quite unreasonable. It is difficult to imagine that any astronomer who is measuring angles on the celestial sphere would obtain observational errors as great as 180 degrees, corresponding to the antipodal point, not to mention 500 degrees or greater. Gauss (1809, §178) acknowledges that the observational errors ought to be bounded by certain limits, but claims that the misspecified probability distribution of the observational errors is of no importance in practice. Additionally, he claims (§179) that as a consequence the method of generalized least squares should be considered an axiom with the same propriety as the method of the arithmetic mean.

This short article aims to thoroughly investigate Gauss's claim that the distribution assumption is unimportant relative to whether the method of generalized least squares yields the most probable value under the density criterion. Since Gauss did not provide any formal proof, the claim is referred to as *Gauss's least squares conjecture* and is, for the purpose of the present investigation, formalized as follows.

Gauss's least squares conjecture. *The value obtained through the method of generalized least squares is the most probable under the density criterion.*

An immediate historical remark is that Gauss's least squares conjecture is falsified by Daniel Bernoulli (1778). The nephew of Jakob Bernoulli conducted an investigation basically identical to that of Gauss (1809), applying Bernoulli's fifth axiom and the density criterion to determine the most probable value, but concluded that the arithmetic mean, a special case of the method of generalized least squares, does not yield the most probable value under the assumption of independent and

identically semi-circle distributed observational errors. Coincidentally, the analytically inconvenient non-normal distribution assumption made Daniel Bernoulli's method practically unusable, as noted in a comment by Leonhard Euler.

2. Investigation and results

The present section aims to find necessary and sufficient conditions for Gauss's least squares conjecture under the assumption of independent observations (Gauss, 1809). To this end, the result that the method of generalized least squares yields the most probable value if and only if the probability distribution is unimodal and median zero elliptical with a density function (see Ekström, 2012a, Theorem 10) is used.

The following lemma is shown through a standard argument; its inclusion in the present text does not constitute a claim of originality. For a random variable U, the notation $\mathcal{L}(U)$ denotes the probability distribution (or *law*) of U.

Lemma 1. *Suppose the random variable U has two or more components, at least two of which are pair-wise independent, and $\mathcal{L}(U)$ is spherical, then $\mathcal{L}(U)$ is normal.*

Proof. Since $\mathcal{L}(U)$ is spherical, $\mathcal{L}(AU) = \mathcal{L}(U)$ for all orthogonal transformations A. By taking permutations, which are orthogonal, it follows that all components of U are identically distributed and pair-wise independent.

Let the characteristic function of the component distribution be denoted ϕ, then $\phi_{\mathcal{L}(U)}(\vec{t}) = \prod_{k=1}^{p} \phi(t_k)$ by independence. Additionally, $\phi_{\mathcal{L}(U)}(A\vec{t}) = \phi_{\mathcal{L}(U)}(\vec{t})$ for all orthogonal transformations A, thus ϕ is even and hence real-valued. By taking A such that $A\vec{t} = ce_k$, some $c \in \mathbb{R}$ and normalized basis vector e_k, it follows that $\phi_{\mathcal{L}(U)}(\vec{t}) = \phi(\|A\vec{t}\|) \prod_{l \neq k} \phi(0) = \phi(\|\vec{t}\|)$ since $\phi(0) = 1$ for all characteristic functions.

By continuity, $\phi(x) > 0$ for $x \in B_r(0)$, some $r > 0$. Let $h(y) = \log \phi(y^{1/2})$, $y \in [0, r^2)$, and take $\vec{t} \in B_{r^2}(0)$ and note that

$$h(\sum_{k=1}^{p} t_k^2) = h(\|\vec{t}\|^2) = \log \phi(\|\vec{t}\|) = \log \prod_{k=1}^{p} \phi(t_k) = \sum_{k=1}^{p} \log \phi(t_k) = \sum_{k=1}^{p} h(t_k^2).$$

By induction, $h(y)$ is defined for all $y \in [0, \infty)$ and $h(ny) = nh(y)$ for all $n \in \mathbb{N}$. Since $nh(y/n) = h(y)$, $h(qy) = qh(y)$ for all non-negative rational numbers q, and thus for all non-negative real numbers since $\mathbb{Q}$ is dense in $\mathbb{R}$ and h continuous. Consequently, $h(y) = yh(1) = ay$ some $a \in \mathbb{R}$, and thus $\phi(t) = e^{at^2}$ for all $t \in \mathbb{R}$ since ϕ is even. The conclusion follows from uniqueness of characteristic functions. □

Theorem 2. *Suppose the random variable U has two or more components, at least two of which are pair-wise independent and non-degenerate, and $\mathcal{L}(U)$ is elliptical, then $\mathcal{L}(U)$ is normal.*

Proof. By definition of elliptical random variables, $\mathcal{L}(U) = \mathcal{L}(AV)$ where $\mathcal{L}(V)$ is spherical and A some affine transformation. Transform each of the two independent non-degenerate components affinely by subtracting the median and normalizing the difference by some non-zero inter-percentile range; the two non-zero inter-percentile ranges exist because the random variables are non-degenerate.

Then the two are jointly spherical and independent, and thus the spherical distribution $\mathcal{L}(V)$ is normal by Lemma 1. Hence $\mathcal{L}(U) = \mathcal{L}(AV)$ is normal. □

In addition to Lambert's density criterion, a statistical criterion proposed by Karl Pearson (1900) is widely used. Pearson's *distance criterion* utilizes his chi distance, which was later extended into the Mahalanobis distance. The distance criterion is formalized: *"Given a probability distribution, the value with smaller Mahalanobis distance to the distribution point of reference is deemed to be more probable than the value with greater Mahalanobis distance to the distribution point of reference."* The distribution point of reference is typically the median, particularly for elliptical distributions, but can also be taken as an extreme value, for instance the value zero relative to the chi distribution.

For completeness, Gauss's least squares conjecture is investigated also relative to the distance criterion. The value that is most probable under the distance criterion is sometimes referred to as the *minimum chi-square estimate*. The following theorem is of interest.

Theorem 3. *Suppose the random variable U has two or more components, all of them independent, and $\mathcal{L}(U)$ is such that the generalized method of least squares yields the most probable value under the distance criterion, then $\mathcal{L}(U)$ is normal.*

The proof of Theorem 3 uses the following lemma relating to Mahalanobis balls, i.e. the set of points that have a smaller than the radius Mahalanobis distance to the center-point of the ball.

Lemma 4. *Suppose the random variable U has two or more components, all of them independent, and the Mahalanobis balls under $\mathcal{L}(U)$ at some center-point are Euclidean, then $\mathcal{L}(U)$ is normal.*

Proof. Suppose initially that the Euclidean Mahalanobis balls have center-point zero. For any orthogonal transformation A and any $r \geq 0$ it holds that $B_r(0) = E_s(0) = AE_s(0) = AB_r(0)$ for some $s \geq 0$, where $B_r(0)$ and $E_s(0)$ denote the Mahalanobis and Euclidean balls, respectively, with center-point 0 and radii r and s. It follows that $\partial B_r(0) = A\partial B_r(0)$, and since orthogonal transformations have orthogonal inverses the Mahalanobis distance under $\mathcal{L}(U)$ satisfies $d(Ax, 0) = d(x, 0)$ for all orthogonal transformations A.

Since the components of U are independent by assumption, by the Pythagorean property the Mahalanobis distance satiesfies for all $x \in \mathbb{R}^p$

$$d(x, 0)^2 = \sum_{k=1}^{p} d_k(x_k, 0)^2,$$

where p is the number of components of U and $d_1, \ldots, d_p$ are the Mahalanobis distances under the distributions of each component. Under necessary uniqueness conditions (Ekström, 2011a) those univariate Mahalanobis distances satisfy

$$d_k(x_k, 0) = |\Phi^{-1} \circ F_k(x_k) - \Phi^{-1} \circ F_k(0)|,$$

where F_k is the distribution function of the distribution of the k:th component and Φ^{-1} is the inverse standard normal distribution function. Let $h_k(z) = d_k(z, 0)$, then since $d(Ax, 0) = d(x, 0)$ for all orthogonal transformations A, including permutations, it follows that $h_1, \ldots, h_p$ are all identical and even.

By construction of the Mahalanobis distance, the Mahalanobis transformation $\Phi^{-1} \circ F_k$ maps the k:th component of U to a standard normally distributed random variable. Because both Φ^{-1} and F_k are non-decreasing, the Mahalanobis transformation is non-decreasing, and since it also is real-valued it follows that every discontinuity is a positive jump. However, if it had a jump discontinuity then the transformation could not map a random variable to a normally distributed ditto; hence the Mahalanobis transformation and h_k are continuous. Since F_k is a distribution function, its limits at $\pm\infty$ are zero and one, respectively, and since h_k is continuous and even with a fixed point at zero, the positive and negative limits of h_k are both ∞. Hence, there is a $z \in (0, \infty)$ such that $h(z) \in (0, \infty)$, where $h = h_k$ all $k = 1, \ldots, p$.

Take $x = (z, \ldots, z)^{\mathrm{t}}$, then $d(x, 0)^2 = d(\|x\| e_k, 0)^2 = h(\sqrt{p}z)^2$, where e_k denotes a normalized basis vector, and by the Pythagorean property $d(x, 0)^2 = ph(z)^2$. By iteration, $h(nz) = nh(z)$ for all positive integer times $\sqrt{p}$ multiples n. Since $nh(z/n) = h(z)$, the equality holds for all positive rational times $\sqrt{p}$ multiples n, and because h is non-decreasing on the positive half line, $h(nz) = nh(z)$ for all positive reals n. Because the functions $\Phi^{-1} \circ F_k(z) - \Phi^{-1} \circ F_k(0)$, $k = 1, \ldots, p$, are non-decreasing, they are odd, and thus $\Phi^{-1} \circ F_k(z) = az + b$ for some real $a > 0$ and $b = \Phi^{-1} \circ F_k(0)$.

Consequently, all component distributions are normal and since they are independent by assumption, $\mathcal{L}(U)$ is normal. Lastly, if the center-point is some non-zero m, then it holds by the homogeneity property of Mahalanobis balls that $B_r^{\mathcal{L}(U-m)}(0) = B_r^{\mathcal{L}(U)}(0 + m) - m = E_s(m) - m = E_s(0)$, and thus $\mathcal{L}(U - m)$ is normal by the preceding argument. Therefore $\mathcal{L}(U)$ is also normal, which is the conclusion of the lemma. □

As a remark prior to the proof of Theorem 3, the method of generalized least squares can in some instances be conceived even though no covariance matrix exists. For example, suppose a random variable U has Cauchy distributed components and there exists a matrix A such that $A^{-1}U$ has pair-wise uncorrelated components that are identically (Cauchy) distributed, then the method of generalized least squares may well be defined by minimization of $x^{\mathrm{t}} A^{-2} x$, i.e. in this example there is a matrix that can substitute for the non-existent covariance matrix. To obtain the greatest degree of generality, the proof of Theorem 3 does not presume existence of a covariance matrix, but only a linear injective transformation L such that the inversely transformed random variable has pair-wise uncorrelated components.

Proof of Theorem 3. Suppose that the method of generalized least squares yields the most probable value under the distance criterion; i.e. suppose $\|a\|_\star = \inf_{x \in A} \|x\|_\star \implies d(a, m) = \inf_{x \in A} d(x, m)$ for all sets A, where d is the Mahalanobis distance under $\mathcal{L}(U)$, m a point of reference, and $\|x\|_\star = \|L^{-1}(x)\|$ given some linear injective L. This implication holds if and only if $a \in \partial E_r^\star(0)$ and $E_r^\star(0) \cap A = \varnothing$ implies for $s = d(a, m)$ that $B_s(m) \cap A = \varnothing$, for all sets A, where $E^\star$ and B denotes the balls under $\|\cdot\|_\star$ and d, respectively. For any $s > 0$ and $a \in \partial B_s(m)$, take $A = E_r^\star(0)^c$ where $r = \|a\|_\star$, then the condition $B_s(m) \cap A = \varnothing$ yields for all $s \geq 0$, $B_s(m) = E_r^\star(0)$ some $r \geq 0$. Additionally, $m = 0$ follows.

By construction of the method of generalized least squares, the linear, injective L is such that distinct components of $\mathcal{L}(L^{-1}(U))$ are uncorrelated. Since the components of $\mathcal{L}(U)$ are independent by assumption, L may without loss of generality be taken such that, for $k = 1, \ldots, p$, its k:th component

is a function only of its argument's k:th coordinate, and as a consequence it follows that components of $\mathcal{L}(L^{-1}(U))$ are independent.

By the homogeneity property of Mahalanobis balls, $B_s^{\mathcal{L}(L^{-1}(U))}(m) = L^{-1}(B_s^{\mathcal{L}(U)}(L(m))$. Since $B_s^{\mathcal{L}(U)}(m) = E_r^\star(0)$ and $L^{-1}(E_r^\star(0)) = E_r(0)$, it follows since $m = 0$ and L linear that $B_s^{\mathcal{L}(L^{-1}(U))}(m) = E_r(0)$. Lemma 4 then yields that $\mathcal{L}(L^{-1}(U))$ is normal. Hence, $\mathcal{L}(U)$ is normal. □

Remark 1. By the proof of Theorem 3, it holds that the method of generalized least squares yields the most probable value under the distance criterion only if the distribution point of reference used for distance minimization is zero. If the distribution point of reference is set to equal the median, as is conventional for all elliptical distributions, then, in the conclusion of Theorem 3, $\mathcal{L}(U)$ is median zero normal; thus making the conclusion identical to the conclusion of Theorem 2. Gauss (1821) argues that all observational error distributions should be assumed to have median zero, since the individuals taking measurements must be presumed competent and able to properly calibrate their measurement equipment; otherwise an analysis would be meaningless from the outset. By the same rationale, the observational errors are assumed independent and unimodal.

The following two theorems are the main results of the present article.

Theorem 5. *Suppose the random variable U has two or more components, at least two of which are pair-wise independent, then given $\mathcal{L}(U)$ the method of generalized least squares yields the most probable value under the density criterion if and only if $\mathcal{L}(U)$ is median zero normal.*

Proof. By Ekström (2012a, Theorem 10), the method of generalized least squares yields the most probable value under the density criterion if and only if $\mathcal{L}(U)$ is unimodal, median zero elliptical and has a density function. Since two components of U are pair-wise independent by assumption, the existence of a density function and Theorem 2 yield that $\mathcal{L}(U)$ is normal. The conclusion then follows by Ekström (2012a, Theorem 10). □

Theorem 6. *Suppose the random variable U has two or more components, all of them independent, then given $\mathcal{L}(U)$ the method of generalized least squares yields the most probable value under the distance criterion if and only if $\mathcal{L}(U)$ is normal with distribution point of reference zero.*

Proof. By Theorem 3, under the independence assumption the method of generalized least squares yields the most probable value under the distance criterion only if $\mathcal{L}(U)$ is normal. Suppose that $\mathcal{L}(U)$ is normal, with covariance matrix Σ, then the Mahalanobis distance under $\mathcal{L}(U)$ satisfies $d(x, m) = \|\Sigma^{-1/2}x - \Sigma^{-1/2}m\|$. Consequently, the expressions $d(x, m)$ and $\|x\|_\star$ have identical minima only if m, the distribution point of reference, is zero. In this case $d(x, m) = \|x\|_\star$, so the converse is immediate. □

As per Theorem 5, under the independence assumption of Gauss (1809), Gauss's least squares conjecture holds only if the probability distribution of the observational errors is normal. Similarly, by Theorem 6 the method of generalized least squares yields the most probable value under the distance criterion only if the observational errors are normal. As a result, the probability distribution of the observational errors is of critical importance relative to the probabilistic properties of the method

of generalized least squares. Furthermore, under the assumptions of Gauss (1809) the method of generalized least squares is a special case of statistical hypothesis generation under either of the density or the distance criteria (cf. Ekström, 2012b).

3. Concluding remarks

Contrary to the discussed claim of Gauss (1809, §178), that the probability distribution of the observational error is of no importance in practice, the probability distribution is crucial as to whether the method of generalized least squares yields the most probable value under the density criterion or not. Given an independence assumption, the normal distribution is not one of many, but the only probability distribution under which the method of generalized least squares yields the most probable value.

An implication of Theorems 5 and 6 is that given the independence assumption, the method of generalized least squares and the normal distribution assumption are inherently interconnected. Given the independence assumption, applying the method of generalized least squares, under the implicit assumption that it yields the most probable value, is equivalent to standard statistical hypothesis generation under a median zero normal distribution assumption.

While the method of generalized least squares in practice is applied without specification of a probability distribution, in the sense of Theorems 5 and 6 the method still carries within itself a normal distribution assumption. In particular, application of the method of generalized least squares does not in and of itself relieve the burden of a probability distribution assumption on the observational errors; rather the circumvention of sorts effectively consists of a median zero normal distribution assumption. A general understanding of this aspect among the members of the scientific community is desirable, not least since the method of generalized least squares probably is one of the most commonly used methods in modern science.

In addition to *Theoria Motus*, Gauss wrote two other pieces on statistical hypothesis generation; *Bestimmung der Genauigkeit* (1816) and *Theoria Combinationis* (1821). In *Theoria Combinationis*, Gauss revised his rationale for favoring the method of generalized least squares, instead making the substantially less ambitious argument that when restricting to linear combinations of the observations, $a^{\mathrm{t}}x$, taking $a = (1, \ldots, 1)^{\mathrm{t}}/n$ minimizes $a^{\mathrm{t}}a$ while satisfying $a^{\mathrm{t}}(1, \ldots, 1)^{\mathrm{t}} = 1$, i.e. minimizing variance while being unbiased given uncorrelated observations of equal mean and variance. The fact that Gauss revised his rationale may be taken as circumstantial evidence that he recognized weaknesses in the argument of *Theoria Motus*, presumably his least squares conjecture.

During the time when there were no electronic computers, few practically feasible alternatives to the method of generalized least squares existed. Therefore arbitrarily restricting estimators to linear combinations or assuming normally distributed observational errors, while undesirable theoretically, could be well motivated on pragmatical grounds. In the twenty-first century, by contrast, computer assisted numerical optimization facilitates quick and all but effortless determination of most probable values given practically any continuous probability density function, thus removing the rationale for imposing the arbitrary restrictions or assumptions. For instance, the optimization problem of

Bernoulli (1778), maximizing a polynomial of degree $2n - 1$ where n is the number of observations, is solved easily through computer assisted numerical optimization.

A final historical remark is that in the efforts to motivate a normal distribution assumption, Gauss (1809) discussed the circumstance that a median zero normal distribution assumption yields the arithmetic mean as the most probable value under the density criterion, and since the arithmetic mean is accepted with little objection, there is reason to accept the normal distribution assumption similarly. During the nineteenth century, this argument was in part misinterpreted as a proof that all observational errors are median zero normal (see, e.g., Airy, 1875; Merriman, 1884), and the result was subsequently termed the *law of error*. Towards the end of the century, falsifying the law of error in a convincing fashion was a primary motivation for Karl Pearson's contributions to statistical methodology. For example, Pearson's seminal article *On the criterion* (1900) develops statistical hypothesis testing, the p-value, and the chi-square test as means towards demonstrating that the law of error is a falsity. In this fortuitous way, *Theoria Motus* and Gauss's least squares conjecture continued to contribute to the development statistical methodology well into the twentieth century.

Acknowledgements

This work was supported by the Swedish Council for Working Life and Social Research, project 2010-1406. The author is grateful to Joshua EmBree for valuable comments and suggestions for improvements.

References

Airy, G. B. (1875). *Algebraical and numerical theory of errors of observations and the combination of observations*. London: MacMillan and Co.

Bernoulli, D. (1778). Diiudicatio maxime probabilis plurium obseruationum discrepantium atque verisimillima inductio inde formanda. *Acta Acad. Scient. Imper. Petrop.*, *pro Anno 1777, Pars prior*, 3–23. English translation by C. G. Allen, 1961.

Bernoulli, J. (1713). *Ars Conjectandi*. Basel: Thurnisiorum Fratrum. English translation by E. D. Sylla, 2006.

Ekström, J. (2011a). Mahalanobis' distance beyond normal distributions. *UCLA Statistics Preprints*, *624*.

Ekström, J. (2011b). The Gauss-Pearson decomposition, and the link between classical philosophy and present day statistics. *UCLA Statistics Preprints*, *637*.

Ekström, J. (2012a). On statistical criteria: Theory, history, and applications. *UCLA Statistics Preprints*, *640*.

Ekström, J. (2012b). Statistical hypothesis generation: Determining the most probable subset. *UCLA Statistics Preprints*, *641*.

Gauss, C. F. (1809). *Theoria Motus Corporum Coelestium in sectionibus conicis solem ambientium*. Hamburg: F. Perthes und I. H. Besser. English translation by C. H. Davis, 1858.

Gauss, C. F. (1816). Bestimmung der Genauigkeit der Beobachtungen. *Zeitschrift für Astronomie und verwandte Wissenschaften*, *1*, 185–197.

Gauss, C. F. (1821). Theoria combinationis observationum erroribus minimis obnoxiae: Pars prior. *Commentationes societatis regiae scientarium Gottingensis recentiores*, *5*, 33–62. English translation by G. W. Stewart, 1995.

Lambert, J. H. (1760). *Photometria, sive de mensura et gradibus lumnis, colorum et umbrae*. Augsburg: Typis Christophori Petri Detleffsen. English translation by D. L. DiLaura, 2001.

Lambert, J. H. (1765). Das Mittel zwischen den Fehlern. In *Beiträge zum Gebrauche der Mathematik und deren Anwendung*, (pp. 296–313). Berlin: Verlage des Buchladens der Realschule.

Legendre, A. M. (1805). *Nouvelles méthodes pour la détermination des orbites des comètes*. Paris: Firmin Didot.

Merriman, M. (1884). *A Text Book on the Method of Least Squares*. New York: John Wiley & Sons.

Pearson, K. (1900). On the criterion that a system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. *Phil. Mag. Ser. 5*, *50*, 157–175.

UCLA Department of Statistics, 8125 Mathematical Sciences Building, Box 951554, Los Angeles CA, 90095-1554

E-mail address: joakim.ekstrom@stat.ucla.edu