ACTA ET COMMENTATIONES UNIVERSITATIS TARTUENSIS DE MATHEMATICA
Volume 15, Number 1, 2011
Available online at www.math.ut.ee/acta/

On the relationship between the method of least squares and Gram–Schmidt orthogonalization

Hilmar Drygas

Abstract. A method for solving Least Squares Problems is developed which automatically results in the appearance of the Gram–Schmidt orthogonalizers. Given these orthogonalizers an induction-proof is available for solving Least Squares Problems.

1. Introduction

The method of Least Squares consists in the following minimization problem. For given vectors $y, x_1, \dots x_k \in \mathbb{R}^n$ find numbers $\beta_1, \dots, \beta_k$ such that

$$\| y - \sum_{i=1}^{k} \beta_i x_i \|$$

is minimized. The underlying linear model is

$$y = x_1\beta_1 + \ldots + x_k\beta_k + \epsilon,$$

where ϵ is a disturbance term. Mostly, it is assumed that ϵ is a random vector with expectation 0 and covariance matrix $\sigma^2 I_n$, where $\sigma > 0$ is an unknown parameter. The method of Least Squares is therefore also described by

$$\| \epsilon \| = \min .$$

The simplest linear model is $y = a\,1_n + \epsilon$, where 1_n is the vector of ones. The Least Squares problem for estimating a can be solved by Steiner's theorem.

Steiner's theorem is essentially a theorem in Mechanics ("Trägheits-Momente"). The following theorem is a special case of the equation

$$E(X-a)^2 = (a - EX)^2 + \mathrm{Var}\,(X)$$

Received January 15, 2010.
2010 *Mathematics Subject Classification.* 15A63, 62J05, 65F25.
Key words and phrases. Gram–Schmidt orthogonalization, Least Squares method, linear models, orthogonal projection, simple regression, Steiner's theorem.

for a random variable X .

Theorem 1.1 (Steiner's theorem). *Let* $w_i \geq 0$, $\sum\limits_{i=1}^{n} w_i > 0$ *and*

$$\sum_{i=1}^{n} w_i(y_i - a)^2 = \sum_{i=1}^{n} w_i(y_i - \overline{y}_{wgh})^2 + (\sum_{i=1}^{n} w_i)(a - \overline{y}_{wgh})^2.$$

Then $\overline{y}_{wgh} = (\sum\limits_{i=1}^{n} w_i)^{-1}(\sum\limits_{i=1}^{n} w_i y_i)$.

Proof. The proof follows from the Pythagoras theorem since

$$\sum_{i=1}^{n} w_i(y_i - \overline{y}_{wgh})(a - \overline{y}_{wgh}) = 0 .$$

□

Thus $a = \overline{y}_{wgh}$ solves the Least Squares problem

$$\sum_{i=1}^{n} w_i(y_i - a)^2 = \min.$$

This theorem can also be used to find estimates for the regression model

$$y_i = \alpha + \beta x_i + \epsilon_i \, , \, i = 1, \ldots, n \, , \, y = \alpha 1_n + \beta x + \epsilon.$$

The task consists in minimizing

$$Q = \sum_{i=1}^{n} (y_i - \alpha - \beta x_i)^2 .$$

By Steiner's theorem we get the solution

$$\hat{\alpha} = \overline{y} - \beta \overline{x} \, , \, \overline{x} = \frac{1}{n} \sum_{i=1}^{n} x_i \, , \, \overline{y} = \frac{1}{n} \sum_{i=1}^{n} y_i \, .$$

By plugging in the obtained estimates we get

$$\begin{aligned} Q &= \sum_{i=1}^{n} \Big(y_i - \overline{y} - \beta(x_i - \overline{x}) \Big)^2 \\ &= \sum_{i=x_i \neq \overline{x}} (x_i - \overline{x})^2 \left(\frac{y_i - \overline{y}}{x_i - \overline{x}} - \beta \right)^2 + \sum_{i:x_i = \overline{x}} (y_i - \overline{y})^2 \, . \end{aligned}$$

According to Steiner's theorem the estimate of β from minimization is a weighted mean of the slopes $\frac{y_i-\overline{y}}{x_i-\overline{x}}$, namely

$$\hat{\beta} = \frac{\sum\limits_{i=x_i\neq\overline{x}} (x_i-\overline{x})^2 \frac{(y_i-\overline{y})}{(x_i-\overline{x})}}{\sum\limits_{i=x_i\neq\overline{x}} (x_i-\overline{x})^2} = \frac{\sum\limits_{i=1}^{n} (x_i-\overline{x}) y_i}{\sum\limits_{i=1}^{n} (x_i-\overline{x})^2} .$$

If all x_i are equal to $\overline{x}$, then β is arbitrary since Q does not depend on β .

This above used procedure can be extended to the general case as will be shown in the next section.

2. Generalization of successive estimation

Theorem 2.1 (Generalized Steiner's theorem). *For vectors x and y the following equality holds:*

$$\| y - ax \|^2 = \| y - \frac{(y,x)}{(x,x)} x \|^2 + \| x \|^2 \left(a - \frac{(x,y)}{(x,x)} \right)^2$$

if $x \neq 0$.

Proof. The vectors $\left(y - \frac{(y,x)}{(x,x)} x \right)$ and x are orthogonal. The Pythagoras theorem therefore yields the result. □

Corollary 2.2. *The Least Squares solution of*

$$\| y - ax \| = \min.$$

is obtained when $a = \frac{(y,x)}{(x,x)}$.

Now we want to minimize

$$\| y - \sum_{i=1}^{k} \beta_i x_i \|^2 \ . \tag{2.1}$$

If $x_1 = 0$, then β_1 does not appear in (2.1) and it is therefore arbitrary. If $x_1 \neq 0$, then according to Theorem 2.1

$$\hat{\beta}_1 = \frac{(y - \sum\limits_{i=2}^{k} \beta_i x_i, x_1)}{(x_1, x_1)} .$$

By plugging in this estimate into (2.1) we get the new minimization problem

$$\| y^{(2)} - \sum_{i=2}^{k} \beta_i x_i^{(2)} \| = \min ,$$

where

$$y^{(2)} = y - \frac{(y, x_1)}{(x_1, x_1)} x_1 = P_{\{x_1\}^\perp} y, x_i^{(2)} = x_i - \frac{(x_i, x_1)}{(x_1, x_1)} x_1 = P_{\{x_1\}^\perp} x_i \,.$$

If $x_2^{(2)} \neq 0$ – otherwise β_2 is arbitrary – we obtain

$$\hat{\beta}_2 = \frac{(y^{(2)} - \sum_{i=3}^{k} \beta_i x_i^{(2)}, x_2^{(2)})}{(x_2^{(2)}, x_2^{(2)})} = \frac{(y - \sum_{i=3}^{k} \beta_i x_i, x_2^{(2)})}{(x_2^{(2)}, x_2^{(2)})} \,.$$

and by plugging in the obtained estimate we get a new minimization problem with $y^{(3)}, x_i^{(3)}$, $i = 3, \ldots, k$. Continuing with this procedure we get successively the solutions $(j = 3, \ldots, k)$

$$\hat{\beta}_j = \frac{(y - \sum_{i=j+1}^{k} \beta_i x_i^{(j)}, x_j^{(j)})}{(x_j^{(j)}, x_j^{(j)})} \,, \text{ if } x_j^{(j)} \neq 0$$

and finally

$$\hat{\beta}_k = \frac{(y, x_k^{(k)})}{(x_k^{(k)}, x_k^{(k)})} \,, \text{ if } x_k^{(k)} \neq 0 \,.$$

In order to simplify the notation we define

$$q_1 = x_1 \,,\, q_j = x_j^{(j)} \,,\, j = 2, \ldots, k \,.$$

Then

$$\begin{aligned} x_i^{(l)} &= x_i^{(l-1)} - \frac{(x_i^{(l-1)}, q_{l-1})}{(q_{l-1}, q_{l-1})} q_{i-1} \\ &= P_{\{q_{l-1}\}^\perp} x_i^{(l-1)} \,,\, i = l, \ldots, k \,,\;\; l = 1, \ldots, k_i \end{aligned}$$

where, of course, $x_i^{(1)} = x_i \,,\, i = 1, \ldots, k$. Therefore

$$q_l = P_{\{q_{i-1}\}^\perp} x_i^{(l-1)}$$

and

$$\begin{aligned} x_i^{(l)} &= P_{\{q_{l-1}\}^\perp} P_{\{q_{l-2}\}^\perp} \ldots P_{\{q_1\}^\perp} x_i \\ q_l &= P_{\{q_{l-1}\}^\perp} \ldots P_{\{q_l\}^\perp} x_l \,,\, l = 2, \ldots, k \,. \end{aligned}$$

The next step consists in proving that

$$\prod_{j=1}^{i-1} P_{\{q_{i-j}\}^\perp} = P_{\{q_1, \ldots, q_{i-1}\}^\perp} \,.$$

By Achieser, Glasmann (1981), p. 97, the product of projections is a projector if and only if the projectors commute. By Rao, Mitra (1971), p. 189, the projection onto the intersection of the subspaces M and N is given by

$$2P(P+Q)^{-}Q$$

where P is the projection onto M and Q the projection onto N. If P and Q commute, there must be a simple formula for the Moore–Penrose generalized inverse of $(P+Q)$, namely $(P+Q)^{+}$. This formula will be given by the following theorem.

Theorem 2.3. *If* $PQ = QP$, *then* $(P+Q)^{+} = P + Q - \frac{3}{2}PQ$.

Proof. The proof follows from verification of the defining equalities of the Moore–Penrose inverse. An alternative is that P and Q are jointly diagonalizable if $PQ = QP$:

$$P = C\,\mathrm{diag}(\lambda_1, \dots, \lambda_n)C'\,,\; Q = C\;\mathrm{diag}(\mu_1, \dots, \mu_n)C'$$

and the eigenvalues λ_i and μ_i are either 0 or 1. Then

$$P + Q = C\left(\mathrm{diag}(\lambda_1 + \mu_i), \dots, \big(\lambda_\mu + \mu_n)\right)C'\,,$$
$$(P+Q)^{+} = C\,\mathrm{diag}\big((\lambda_1+\mu_1)^{+}, \dots, (\lambda_n+\mu_n)^{+}\big)C'.$$

But

$$(\lambda_i + \mu_i)^{+} = \lambda_i + \mu_i - \frac{3}{2}\,\lambda_i\mu_i$$

in all possible cases. □

Theorem 2.4. *The product of projections* PQ *is the projection onto* $\mathrm{im}(P) \cap \mathrm{im}(Q)$ *if and only if* $QM^{\perp} \subseteq M^{\perp}$. *A sufficient condition for this is* $M^{\perp} \subseteq N$.

Proof. The product of projections PQ is the projection onto $M \cap N$ if and only if it is the identity on $N \cap M$ and vanishes on $(M \cap N)^{\perp} = M^{\perp} + N^{\perp}$, the other properties are satisfied straightforwardly, only $PQM^{\perp} = 0$ must be examined. This is equivalent to the inclusion $QM^{\perp} \subset M^{\perp}$. This condition is met if $M^{\perp} \subseteq N$. □

Theorem 2.5. *In the previous notation the following equality holds:*

$$\prod_{j=1}^{i-1} P_{\{q_{i-j}\}^{\perp}} = P_{\{q_1, \dots, q_{i-1}\}^{\perp}} \quad \textit{and} \quad q_i \in \{q_1, \dots, q_{i-1}\}^{\perp}.$$

Proof. The proof follows the mathematical induction method. The first assertion of the theorem is correct for $i = 2$ and $q_2 = P_{\{q_1\}^{\perp}} x_2 \in \{q_1\}^{\perp}$. Let us assume by induction

$$\prod_{j=1}^{i-1} P_{\{q_j\}^{\perp}} = P_{\{q_1, \dots, q_{i-1}\}^{\perp}} \quad \text{and} \quad q_i \in \{q_1, \dots, q_{i-1}\}^{\perp}.$$

Then

$$\prod_{j=1}^{i} P_{\{q_{i-j}\}^\perp} = P_{\{q_i\}^\perp} P_{\{q_1,\ldots,q_{i-1}\}^\perp} .$$

Since $q_i \in \{q_1, \ldots, q_{i-1}\}^\perp$, it follows from Theorem 2.4 $(M = \{q_i\}^\perp$, $M^\perp = \{\lambda q_i \,;\, \lambda \in \mathbb{R}\})$ that

$$\prod_{j=1}^{i} P_{\{q_j\}^\perp} = P_{\{q_i\}^\perp} P_{\{q_1,\ldots,q_{i-1}\}^\perp} = P_{\{q_i\}^\perp \cap \{q_1,\ldots,q_{i-1}\}^\perp} = P_{\{q_1,\ldots,q_i\}^\perp} .$$

Since $q_{i+1} = P_{\{q_1,\ldots,q_i\}^\perp} x_{i+1} \in \{q_1, \ldots, q_i\}^\perp$, also the second assertion is proved. □

Corollary 2.6. *If* $q_0 = 0$, *then* $q_i = P_{\{q_0,\ldots,q_{i-1}\}^\perp} x_{i,i=1,\ldots,k}$ *and* $x_i^{(l)} = P_{\{q_0,q_1,\ldots,q_{l-1}\}^\perp} x_i$.

Since

$$q_i = P_{\{q_0,\ldots,q_{i-1}\}^\perp} x_i = x_i - P_{\text{span}\{q_1,\ldots,q_{i-1}\}} x_i = x_i - \sum_{j:q_j \neq 0}^{i-1} \frac{(q_j, x_i)}{(q_j, q_j)} q_j ,$$

the q_i describe the Gram–Schmidt orthogonalization procedure. It follows that from the principle of Least Squares the Gram–Schmidt orthogonalization procedure could be invented.

3. An induction argument for the Least Squares method

Since the Gram–Schmidt orthogonalization procedure is well-known, the estimates by the Least Squares method can also be obtained by mathematical induction. We determine by the induction method m linearly independent vectors among $x_1, \ldots, x_k$. We assume that $x_1, \ldots, x_m$ are linearly independent and $\text{Rank}(x_1, \ldots, x_k) = m$. Therefore $x_{m+1}, \ldots, x_k$ are linear combinations of $x_1, \ldots, x_m$. As we have seen in the last section

$$\hat{\beta}_1 = \frac{(y - \sum_{i=2}^{k} \beta_i x_i, x_1)}{(x_1, x_1)} .$$

After plugging into we get the new minimization problem: minimize

$$\| y^{(2)} - \sum_{i=2}^{k} \beta_i x_i^{(2)} \|^2,$$

where

$$y^{(2)} = P_{\{x_1\}^\perp y} ,\; x_i^{(2)} = P_{\{x_1\}}^\perp x_i ,\; i = 2, \ldots, k .$$

In the following $\operatorname{span}\{x_1, \dots, x_k\}$ denotes the subspace generated by the vectors $x_1, \dots, x_k$.

Lemma 3.1. *Let* $\operatorname{Rank}(x_1, \dots, x_k) = m$ *and let* $x_1, \dots, x_m$ *be linearly independent. Then* $x_i^{(2)}$, $i = 2, \dots, m$, *are linearly independent and* $x_i^{(2)}$, $i > m$, *are linear combinations of the* $x_i^{(2)}$, $i = 2, \dots, m$.

Proof. We write P for short instead of $P_{\{x_1\}^\perp}$. From

$$\sum_{i=2}^{m} \lambda_i x_i^{(2)} = P(\sum_{i=2}^{m} \lambda_i x_i) = 0$$

it follows that $\sum\limits_{i=2}^{m} \lambda_i x_i \in \operatorname{span}\{x_1\}$ and hence $\lambda_2 = \ldots = \lambda_m = 0$ from the linear independence of $x_1, \dots, x_m$.

For $i > m$ we get $x_i^{(2)} = \sum\limits_{j=2}^{m} \lambda_{ij} x_j^{(2)}$ if $x_i = \sum\limits_{j=1}^{m} \lambda_{ij} x_j$. □

Theorem 3.2. *Let* $\operatorname{Rank}(x_1, \dots, x_k) = m$ *and let, moreover,* $x_1, \dots, x_m$ *be linearly independent. Futhermore, let* $q_1, \dots, q_m$ *be the pairwise orthogonal vectors obtained from* $(x_1, \dots, x_m)$ *by applying the Gram–Schmidt orthogonalization procedure. Then the Least Squares solutions* $\hat{\beta}_1, \dots, \hat{\beta}_m$ *are recursively given by*

$$\hat{\beta}_m = \frac{(q_m, y - \sum\limits_{i=m+1}^{k} \beta_i x_i)}{(q_m, q_m)},$$

$$\hat{\beta}_i = \frac{(q_i, y - \sum\limits_{j=i+1}^{m} \hat{\beta}_j x_j - \sum\limits_{j=m+1}^{k} \beta_j x_j)}{(q_i, q_i)},$$

$i = m-1, m-2, \dots, 1$. *Here* $\beta_{m+1}, \dots, \beta_k$ *are arbitrary and*

$$y - \sum_{i=1}^{m} \hat{\beta}_i x_i - \sum_{j=m+1}^{k} \beta_i x_i$$

does not depend on $\beta_{m+1}, \dots, \beta_k$.

Proof. We shall follow mathematical induction on m. If $m = 1$, then

$$\hat{\beta}_1 = \frac{(x_1, y - \sum\limits_{i=2}^{k} \beta_i x_i)}{(x_1, x_1)}$$

and

$$y - \hat{\beta}_1 x_1 - \sum_{i=2}^{k} \beta_i x_i = y^{(2)} - \sum_{i=2}^{k} \beta_i x_i^{(2)} .$$

Since $x_i \in \operatorname{span}\{x_1\}$, it follows that $x_i^{(2)} = 0$, $i = 2, \ldots, k$, and therefore

$$y - \hat{\beta}_1 x_1 - \sum_{i=2}^{k} \beta_i x_i = y^{(2)}$$

which does not depend on $\beta_2, \ldots, \beta_k$.

We now arrive at the problem of minimizing

$$\| \, y^{(2)} - \sum_{i=2}^{k} \beta_i x_i^{(2)} \, \| \ .$$

By the induction assumption we use that $x_2^{(2)}, \ldots, x_2^{(m)}$ are linearly independent and $\operatorname{Rank}(x_2^{(2)}, \ldots, x_k^{(2)}) = m - 1$. Then the solutions are

$$(q_m^{(2)}, q_m^{(2)}) \, \hat{\beta}_m = (q_m^{(2)}, y^{(2)} - \sum_{i=m+1}^{k} \beta_i x_i^{(2)})$$

and

$$(q_i^{(2)}, q_i^{(2)}) \, \hat{\beta}_i = (q_i^{(2)}, y^{(2)} - \sum_{j=i+1}^{m} \hat{\beta}_j x_j - \sum_{j=m+1}^{k} \beta_j x_j) \, ,$$

$i = m{-}1, \ldots, 2$. Here $\beta_{m+1}, \ldots, \beta_k$ are arbitrary numbers and $q_2^{(2)}, \ldots, q_m^{(2)}$ are obtained by applying the Gram–Schmidt orthogonalization procedure to $x_2^{(2)}, \ldots, x_m^{(2)}$. Moreover, $y^{(2)} - \sum\limits_{i=2}^{m} \hat{\beta}_i \hat{x}_i - \sum\limits_{i=m+1}^{k} \beta_i x_i$ does not depend on $\beta_{m+1}, \ldots, \beta_k$. From this it follows that

$$y - \sum_{i=1}^{m} \hat{\beta}_i x_i - \sum_{j=m+1}^{k} \beta_i x_i = y^{(2)} - \sum_{i=2}^{k} \hat{\beta}_i x_i^{(2)} - \sum_{i=m+1}^{k} \beta_i x_i^{(2)}$$

does not depend on $\beta_{m+1}, \ldots, \beta_k$ either.

We now prove by mathematical induction that $q_i^{(2)} = q_i$, $i = 2, \ldots, m$. This is correct for $i = 2$ since $x_2^{(2)} = q_2^{(2)} = x_2 - \frac{(x_1, x_2)}{(x_1, x_1)} x_1 = q_2$ and by using the induction assumption we get

$$q_i^{(2)} = x_2^{(2)} - \sum_{j=2}^{i-1} \frac{(x_i^{(2)}, q_j^{(2)})}{(q_j^{(2)}, q_j^{(2)})} q_j^{(2)} = x_i - \frac{(x_i, x_1)}{(x_1, x_1)} x_1 - \sum_{j=2}^{i-1} \frac{(x_i^{(2)}, q_j)}{(q_j, q_j)} q_j \, .$$

Since $(x_i^{(2)}, q_j) = (x_i, q_j)$ for $j \geq 2$, it follows that $q_i^{(2)} = q_i$, $i = 2, \ldots, m$.

From $(q_m, q_1) = 0$ for $i \geq 2$ we finally get

$$\hat{\beta}_m = \frac{(q_m, y^{(2)} - \sum_{i=m+1}^{k} \beta_i x_i^{(2)})}{(q_m, q_m)} = \frac{(q_m, y - \sum_{i=m+1}^{k} \beta_i x_i)}{(q_m, q_m)}$$

and for $i = 2, \ldots, m-1$,

$$\begin{aligned} \hat{\beta}_i &= \frac{(q_i, y^{(2)} - \sum_{j=i+1}^{m} \hat{\beta}_j x_j^{(2)} - \sum_{j=m+1}^{k} \beta_j x_j^{(2)})}{(q_i, q_i)} \\ &= \frac{(q_i, y - \sum_{j=i+1}^{m} \hat{\beta}_j x_j - \sum_{j=m+1}^{k} \beta_j x_j)}{(q_i, q_i)}. \end{aligned}$$

This is completed by

$$\begin{aligned} \hat{\beta}_1 &= \frac{(x_1, y - \sum_{j=2}^{m} \hat{\beta}_j x_j - \sum_{j=m+1}^{k} \beta_j x_j)}{(x_1, x_1)} \\ &= \frac{(q_1, y - \sum_{j=2}^{m} \hat{\beta}_j x_j - \sum_{j=m+1}^{k} \beta_j x_j)}{(q_1, q_1)}. \end{aligned}$$

□

4. Historical remarks

The Gram–Schmidt procedure goes back to the papers by Gram (1883) and Schmidt (1907). Since the paper of Wong (1935), the name "Gram–Schmidt orthogonalization" has become common in mathematical literature. However, already Laplace (1812, 1814 and 1820) has invented this algorithm though he did not recognize it as an orthogonalization procedure. What Laplace did, was the following. He considered the linear model

$$y = \beta_1 x_1 + \ldots + \beta_k x_k + \epsilon$$

under the usual standard assumptions. He was only interested in the estimation of one of the β_i. Without loss of generality it can be assumed that β_k is the parameter to be estimated. In Drygas (1976) and Drygas (2008) it was shown that this estimator is given by

$$\frac{(y, (I-P)x_k)}{\| (I-P)x_k \|^2}$$

where P is the orthogonal projection onto span $\{x_1, \ldots, x_{k-1}\}$ and the variance of the estimator is equal to

$$\frac{\sigma^2}{\| (I-P)x_k \|^2} .$$

The parameter β_k is estimable if and only if $x_k \notin \text{span}\{x_1, \ldots, x_{k-1}\}$. Laplace proceeded as follows: he formed the orthogonal projection P_1 onto the orthogonal complement of span $\{x_1\}$ and arrived at the model

$$P_1 y = \beta_2 P_1 x_2 + \ldots + \beta_k P_1 x_k + P_1 \epsilon .$$

In order to eliminate x_2 he formed P_2, the projection onto the orthogonal complement of span $\{P_1 x_2\} =:$ span $\{q_2\}$. Continuing in this way he finally arrived at the model

$$P_{k-1} \ldots P_2 P_1 y = \beta_k P_{k-1} \ldots P_2 P_1 x_k + P_{k-1} \ldots P_1 \epsilon .$$

The estimator of β_k was then

$$\frac{(y, P_{k-1} \ldots, P_2 P_1 x_k)}{\| P_{k-1} \ldots P_2 P_1 x_k \|^2}$$

and the variance was

$$\frac{\sigma^2}{\| P_{k-1} \ldots P_2 P_1 x_k \|^2} .$$

It is evident that $P_{k-1} \ldots, P_2 P_1 = (I-P)$. This is proved by showing that $P_{k-1} \ldots P_1 y = y$ if $y \in \{x_1, \ldots, x_{k-1}\}^\perp = \{q_1, \ldots, q_{k-1}\}^\perp$ and $P_{k-1} \ldots P_1 y = 0$ if $y \in \text{span}\{x_1, \ldots, x_{k-1}\} = \text{span}\{q_1, \ldots, q_{k-1}\}$. In the same way it can be shown that

$$P_i \ldots P_1 = P_{\{q_1, \ldots q_i\}^\perp} .$$

The vectors $x_1 = q_1, q_2, \ldots, q_{k-1}$ and $q_k = (I-P)x_k$ constitute the orthogonalizers of $\{x_1, \ldots, x_k\}$.

A translation of the work by Laplace together with a detailed discussion can be found in Langou (2009).

A comprehensive discussion of the history of Gram–Schmidt procedure can be found in Leon, Björck and Gander (2009) and also in Björck (2010).

It is generally said that the Method of Least Squares is due to Carl Friedrich Gauss (see Gauss (1973)) but also Legendre (1805) published a paper about this method.

Acknowledgements. I thank the anonymous referee for the constructive comments and the additional literature, in particular the hint to the work of Laplace. I also thank the organizers of the XVIII International Workshop on Matrices and Statistics, Julia Volaufova and Viktor Witkovsky for their support.

References

Achieser, N. I. and Glasmann, I. M. (1981), *Theorie der linearen Operatoren im Hilbert-Raum*, Verlag Harri Deutsch, Thun.

Björck, Å. (2010), *Gram–Schmidt orthogonalization, 100 years and more*, Paper presented at IWMS, Shanghai (China), 5. - 8. Janue, 2010.

Drygas, H. (1976), *Weak and strong consistency of the least squares estimators in regression models*, Z. Wahrscheinlichkeitstheorie Verw. Gebiete **3**4, 119–127.

Drygas, H. (2008), *QR-decomposition from the statistical point of view*; In: Recent Advances in Linear Models and Related Areas, Essays in Honour of Helge Toutenburg, Shalab and Heumann (Eds), Physica-Verlag, Springer, Heidelberg, pp. 293–311.

Gauss, C. F. (1973), *Werke. Band IV*, Reprinted of the 1873 original, Georg Olms Verlag, Hildesheim.

Gram, J. P. (1883), *Über die Entwickelung reeller Funktionen in Reihen mittelst der Methode der kleinsten Quadrate*, J. Reine Angew. Math. **9**4, 41–73.

Langou, J. (2009), *Translation and modern interpretation of Laplace's théorie analytique des probabilités*, Technical Report CCM 280, UC Denver, Denver, Colerado, pp. 505–512, 516–520.

Laplace, P.-S. (1812, 1814, 1820), *Théorie Analytique des Probabilités, Troisième Edition, Premier Supplément Sur l'application du calcul des probabilités à la philosophie naturelle*, Paris, Courcier.

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

Leon, S. J. Björck, Å., and Gander, W. (2009), *Gram–Schmidt orthogonalization: 100 years and more*, SIAM Review, submitted.

Rao, C. R. and Mitra, S. K. (1971), *Generalized Inverse of Matrices and its Applications*, Wiley, New York – London – Sydney – Toronto.

Schmidt, E. (1907), *Zur Theorie der linearen und nichtlinearen Integralgleichungen*, Math. Ann. **4** , 161–174.

Wong, Y. K. (1935), *An application of orthogonalization to the theory of least squares*, Ann. Math. Stat. **6**, 53–75.

University of Kassel, Fachbereich 10, Institut für Mathematik, 34109 Kassel, Germany

E-mail address: hilmar.drygas@onlinehome.de