Numerical linear algebra and some problems in computational statistics

Zdeněk Strakoš[1]

[1] Institute of Computer Science, Academy of Sciences, Pod Vod. věží 2, 182 07 Prague 8, and Faculty of Mathematics and Physics, Charles University, Sokolovská 83, 186 75 Prague 8, Czech Republic
E-mail: strakos@cs.cas.cz

Keywords: Moments, ordinary least squares, total least squares, errors-in-variables modeling, core problem theory, orthogonal bidiagonalization, orthogonality.

1 Introduction

In 1873 Chebyshev formulated the following problem. Let $f(u)$ be an unknown function which is nonnegative in the interval (a, b). Given the values of the integrals

$$\int_a^b f(u)du, \int_a^b uf(u)du, \ldots, \int_a^b u^{k-1}f(u)du, \tag{1}$$

it is required to determine the tight upper and lower bounds for the integral

$$\int_a^x f(u)du, \tag{2}$$

for any value of x in the interval (a, b). Chebyshev motivated his problem by investigation of limit theorems in probability theory, with some related work done even earlier by Heine. It was completely resolved by Markov; for more detailed comments and description of the related developments see Shohat and Tamarkin (1943), Akhiezer (1965) and Gautschi (1981). Here we will not describe the well known role of moments in approximation and determination of the characteristic and distribution functions of the random variable in probability theory. We point out, however, that an analogous approach is used in solving systems of linear algebraic equations using the matching moments model reduction described below.

Chebyshev noticed a possible mechanical interpretation of his problem, but he did not investigate it. The mechanical interpretation was investigated by Stieltjes, who gave a complete solution to the following problem. Given a sequence of numbers ξ_k, $k = 0, 1, \ldots$, a non-decreasing distribution function $\omega(\lambda)$ is sought such that the Riemann-Stieltjes integrals satisfy

$$\int_0^\infty \lambda^k d\omega(\lambda) = \xi_k, \quad k = 0, 1, \ldots , \tag{3}$$

where

$$\int_0^\infty \lambda^k d\omega(\lambda) \tag{4}$$

represents the k-th (generalized) moment of the distribution of positive mass on the half line $\lambda \geq 0$. Stieltjes based his investigation on continued fractions; for the instructive summary we refer to Appendix 2 in Gantmacher and Krein (2002). It is worth to point out, however, that continued fractions played an important role also in the work of Chebyshev and Markov.

The Stieltjes moment problem is also deeply related to the earlier work of Gauss and Jacobi, and to the later work of Christoffel, on the Gauss(-Christoffel) quadrature. Let the distribution function $\omega(\lambda)$, $\lambda \geq 0$ be given. It is required to determine a nondecreasing function $\omega^{(n)}(\lambda)$ with n points of increase $0 < \theta_1^{(n)} < \ldots < \theta_n^{(n)}$ and the positive weights $\omega_1^{(n)}, \ldots, \omega_n^{(n)}$, $\sum_{\ell=1}^n \omega_\ell^{(n)} = 1$ such that the first $2n$ moments are matched,

$$\int_0^\infty \lambda^k d\omega(\lambda) = \int_0^\infty \lambda^k d\omega^{(n)}(\lambda) = \sum_{\ell=1}^n \omega_\ell^{(n)} \{\theta_\ell^{(n)}\}^k, \quad k = 0, 1, \ldots, 2n-1\,. \tag{5}$$

Since (7) holds for every monomial λ^k, $k = 0, 1, \ldots, 2n-1$, it must also hold for any linear combination of the given monomials, i.e. for any polynomial $p(\lambda)$ of degree at most $2n-1$. The value of the integral

$$\int_0^\infty p(\lambda)\, d\omega(\lambda) \tag{6}$$

is given by the weighted sum of the polynomial values at n nodes $\theta_\ell^{(n)}, \ell = 1, \ldots, n$,

$$\int_0^\infty p(\lambda)\, d\omega(\lambda) = \sum_{\ell=1}^{n} \omega_\ell^{(n)} p(\theta_\ell^{(n)})\,. \tag{7}$$

Consequently, (7) represents the n-point Gauss-Christoffel quadrature, see, e.g., Gautschi (1981), which is intimately related to *orthogonal polynomials* and Jacobi matrices. For the recent surveys of the related algebraic relationships described in the polynomial as well as in the matrix language we refer to Strakoš (2008), Meurant and Strakoš (2006) and Strakoš and Tichý (2002). The underlying theory of orthogonal polynomials has extremely wide and deep connections; for the computational aspects we refer to Gautschi (2004). It should be pointed out that the fundamental concept of moments and its use in various applications was pursued by Golub throughout his professional life, which is nicely documented in Golub (2007).

Our historical introduction leads to two main observations which will further be illustrated in our exposition:

1. The idea of *matching moments*, present in many developments in computational statistics, plays an important role also in numerical linear algebra.

2. The idea of matching moments is in a fundamental way linked with the key concept of *orthogonality*.

Mathematically, orthogonal projections often helps to extract efficiently the required information from the data. Computationally, orthogonality is crucial in maintaining the influence of rounding errors at the minimal possible level.

Computational behaviour of methods and algorithms is of great importance in numerical linear algebra, and it might be beneficial to consider this to a to larger extent also in computational statistics.

2 Linear approximation problems

In numerical linear algebra, a linear approximation problem is in its general form represented by

$$Ax \approx b, \quad A \text{ is a nonzero } N \text{ by } M \text{ matrix}, \quad b \text{ is a nonzero vector of length } N\,, \tag{8}$$

where N can be greater than, equal to, or smaller than M. We consider, for simplicity of our exposition, A, x and b real, and b nonorthogonal to the range of A, i.e. $A^T b \neq 0$. The last assumption has a natural interpretation. If $A^T b = 0$, then it is meaningless to look for an approximation of b in the form of the linear combination of the columns of A.

Construction of methods and algorithms as well as terminology and notation reflects in computational statistics the focus on *applications with statistical interpretation of data.* In linear models, the book Rao and Toutenburg (1999) can be used as a starting point when proceeding towards more general numerical linear algebra settings. Starting from the computational statistics perspective, one should consider that numerical linear algebra deals with approximation problems which arise from many areas, including numerical solution of partial differential equations, numerical optimization, control theory and signal processing, image processing, information retrieval etc. Methods and algorithms in numerical linear algebra are *generally applicable to linear approximation problems arising from such different areas.* The methodological approach, the level of abstraction as well as notation must reflect this generality. As a positive consequence, this enables an easy transfer of knowledge developed while solving problems in one area to applications in a different area.

Clearly, the viewpoint of computational statistics with a specific application context in mind is inevitably different from the more abstract and more general viewpoint of numerical linear algebra. This difference makes the interactions between the fields difficult but also exciting. Despite the methodological differences, the interactions can be very useful *for both fields.* As an example we refer to the

paper Stewart (1987) on collinearity and linear regression published in Statistical Science, and to the comments of Marquardt, Belsley, Thisted, Hadi and Velleman, as well as the rejoinder of Stewart, published in the same issue of that journal following the original paper.

The author of this contribution works in numerical linear algebra, and is hardly able to formulate the points from the perspective of computational statistics. Therefore this contribution is written from the point of view of numerical linear algebra, and it follows the notation typical in that field. Nevertheless, it aims to indicate through several examples the close relationship between corresponding problems and methods in computational statistics and numerical linear algebra, and it attempts to contribute in this way to development of mutual understanding.

In our contribution we will distinguish three main classes of problems.

2.1 Systems of linear algebraic equations

In the first class the matrix A is square and numerically nonsingular, and the unknown vector x can be found for the given b by solving the system of linear algebraic equations

$$Ax = b, \quad A \text{ is a nonsingular } N \text{ by } N \text{ matrix}\,. \tag{9}$$

Such systems arise, e.g., from discretization of continuous problems, as in numerical solution of partial differential equations. The data A and b do not represent the original problem accurately due to errors in modeling and discretization. The goal is to approximate efficiently the exact solution of (9) to the sufficient accuracy, so that the computational error does not contribute significantly to the total error in solving the original continuous problem, see, e.g., Strakoš and Liesen (2005).

2.2 Ordinary least squares problems

In the second class we consider the generally *error free* rectangular N by M matrix A and the right hand side (the observation vector) significantly affected by errors. Then the unknown vector x is sought as the solution of the ordinary least square problem (OLS)

$$Ax = b + r, \quad \min \|r\|\,. \tag{10}$$

Here b is orthogonally decomposed

$$b = b|_{\mathcal{R}(A)} + b|_{\mathcal{N}(A^T)}$$

into parts $b|_{\mathcal{R}(A)}$ in the range of A and $b|_{\mathcal{N}(A^T)}$ in the nullspace of A^T, and the unknown vector x is obtained as the unique minimum norm solution of the following problem

$$Ax = b|_{\mathcal{R}(A)}, \quad x \in \mathcal{R}(A^T), \quad r = -b|_{\mathcal{N}(A^T)}\,. \tag{11}$$

In other words, b is orthogonally projected onto the subspace generated by the columns of A, and the errors in b are *assumed to be* orthogonal to that subspace.

This corresponds in computational statistics to standard linear regression, see, e.g., the description given in Rao and Toutenburg (1999), Chapter 3, in particular Sections 3.1 - 3.3.

2.3 Total least squares problems

In the third class the significant errors can be contained both in the general rectangular matrix A representing the model as well as in the observation vector b. Then the unknown solution vector x is sought such that

$$(A + E)x = b + r, \quad \min \|[r, E]\|_F\,, \tag{12}$$

where $\|\cdot\|_F$ means the Frobenius norm of the given matrix, see Rao and Toutenburg (1999), Section 3.12. This problem goes by many names, including errors-in-variables modeling in systems theory, orthogonal regression in computational statistics, and total least squares (TLS) in numerical linear algebra. Basic analysis of the problem was given in the seminal paper Golub and Van Loan (1980). Theory and computational algorithms were then extended to the so-called nongeneric case in the fundamental book Van Huffel and Vandewalle (1991), which covered also statistical properties of TLS. It is also worth to mention proceedings of the workshops on errors-in-variables modeling organized by Van Huffel in Leuven with the goal of bringing together statisticians and numerical analysts, see Van Huffel (1997) and

Van Huffel and Lemmerling (2002). The TLS theory has recently been further revised in a series of papers Paige and Strakoš (2006), Paige and Strakoš (2002a), Paige and Strakoš (2002b), with references to many relevant previous publications.

Though the TLS problem (12) seemingly appears to be a natural and benign generalization of the OLS problem (10), this generalization represents in fact a tremendous complication. Unlike (10), the solution of the TLS problem (12) may not exist, and the mathematical theory and also numerical computation becomes very intriguing, see Paige and Strakoš (2006). Some ideas are recalled below.

2.4 Regularization

In some applications the linear approximation problem (8) arises from discretization of the so-called *ill-posed problems*, where the solution of the original problem does not depend continuously on data. Models represented by Fredholm integral equations of the first kind with a non-degenerate kernel function give typical examples, with applications in image processing, signal processing etc. The related linear algebraic problem is practically rank deficient. Moreover, it cannot be efficiently replaced by a rank-reduced model without a substantial loss of information encoded in the data. Even worse, the data suffer from some inherent noise. Therefore, if the problem is ill-posed, then in any of the classes mentioned above it is meaningless to compute the exact solution. Instead, some form of stabilization technique, called *regularization*, must be applied.

Here we do not deal explicitly with ill-posed problems. Description of regularization techniques is out of the scope of our contribution. The point is, however, important, and several algorithms mentioned below, such as the truncated singular value decomposition, the conjugate gradient method, and the iterative Golub-Kahan bidiagonalization, exhibit regularization properties. For a basic reading on regularization and on solving discrete ill-posed problems we refer to Hansen (1998). Statistical concepts used in regularization are described in Sima (2006).

3 Matching moments model reduction

We first describe an algebraic formulation of the matching moments model reduction. The subsequent sections will then show how the concepts introduced in this section are present in several methods for solving linear approximation problems.

Consider the system of linear algebraic equations (9) and assume, in addition, A symmetric positive definite (SPD). Let x_0 be the initial approximation to the solution x (without additional information it is advisable to take $x_0 = 0$), $r_0 = b - Ax_0$ be the initial residual, $v_1 = r_0/\|r_0\|$. Let $\{\lambda_i, u_i\}$, $i = 1, \ldots, N$ be the eigenvalues and the corresponding orthonormal eigenvectors of A, where we assume, for the simplicity of notation, $\lambda_1 < \lambda_2 < \ldots < \lambda_N$ (from A SPD we have $0 < \lambda_1$). Let $U \equiv (u_1, \ldots, u_N)$ be the orthonormal matrix with the eigenvectors u_i as its columns. Then, using the spectral decomposition of A,

$$v_1^T A\, v_1 = v_1^T U \ \mathrm{diag}\,(\lambda_j)\ U^T v_1 = \sum_{j=1}^{N} |(v_1, u_j)|^2\, \lambda_j\,. \tag{13}$$

This inspires the following construction. Let $\omega(\lambda)$ be the distribution function with the (finite) points of increase λ_j equal to the eigenvalues of A, and the weights ω_j given by the size of the squared components of v_1 in the corresponding invariant subspaces, $\omega_j = |(v_1, u_j)|^2$, see Hestenes and Stiefel (1952), Strakoš and Tichý (2002). Then, using (13), the moments of the distribution function $\omega(\lambda)$ can be expressed in the matrix form as

$$\int_0^\infty \lambda^k d\omega(\lambda) = \sum_{j=1}^{N} \omega_j \{\lambda_j\}^k = v_1^T A^k v_1, \quad k = 0, 1, \ldots \tag{14}$$

Now consider the conjugate gradient method for solving (10), x_0 and r_0, see Hestenes and Stiefel (1952). The method constructs the sequence of approximations x_n to the solution x such that

$$x_n \in x_0 + \mathcal{K}_n(A, r_0), \quad \mathcal{K}_n(A, r_0) \equiv \mathrm{span}\,\{r_0, Ar_0, \ldots, A^{n-1} r_0\}\,, \tag{15}$$

i.e. $x_n - x_0$ belongs to the n-th Krylov subspace of A with respect to r_0 generated by the vectors $r_0, \ldots, A^{n-1} r_0$. The approximation x_n is determined from the (Galerkin) orthogonality condition

$$A(x - x_n) \perp \mathcal{K}_n(A, r_0)\,. \tag{16}$$

Using the *orthonormal basis* $V_n = (v_1, \ldots, v_n)$ of the Krylov subspace $\mathcal{K}_n(A, r_0)$ generated by the Lanczos algorithm applied to A with v_1, see Lanczos (1952), the conjugate gradient method is given by

$$x_n = x_0 + V_n y_n, \quad T_n y_n = \|r_0\| e_1, \quad n = 1, 2, \ldots, \tag{17}$$

where T_n represents the Jacobi matrix of the recurrence coefficients for the *orthonormal* Lanczos vectors $v_1, v_2, \ldots, v_n$, or, equivalently, the Jacobi matrix of the recurrence coefficients for the first n orthonormal polynomials corresponding to the inner product defined by the Riemann-Stieltjes integral with the distribution function $\omega(\lambda)$. Here the orthogonality of polynomials translates to the orthogonality of the Lanczos vectors and vice versa.

It is well known that the nodes and the weights of the n-th Gauss-Christoffel quadrature are given by the eigenvalues $\theta_\ell^{(n)}$ and the squared first components of the corresponding orthonormal eigenvectors of T_n respectively. Then, using the spectral decomposition of T_n analogously to the spectral decomposition of A in (14), we get for the term on the right hand side of (7)

$$\sum_{\ell=1}^{n} \omega_\ell^{(n)} \{\theta_\ell^{(n)}\}^k = e_1^T T_n^k e_1, \quad k = 0, 1, \ldots . \tag{18}$$

Summarizing, the n-point Gauss-Christoffel quadrature (7) for the given $\omega(\lambda)$ can be written in the matrix form as

$$v_1^T A^k v_1 = e_1^T T_n^k e_1, \quad k = 0, 1, \ldots, 2n-1\,, \tag{19}$$

and the first n steps of the conjugate gradient method can be seen as the *model reduction* from $Ax = b$ to $T_n y_n = \|r_0\| e_1$ such that the first $2n$ moments (19) are matched. For details we refer to the survey paper Meurant and Strakoš (2006) and in particular to Strakoš (2008), where the relationship between the matching moments model reduction and Krylov subspace methods is extended to more general matrices and methods.

Krylov subspace methods can be used for solving all linear approximation problems from Section 2, including discrete ill-posed problems. They represent one of the main iterative tools of modern numerical linear algebra. Here we have demonstrated that they are intimately linked with the problem of moments. Whenever the approximation problem contains a significant dominance which can be captured *mathematically* by matching a number of moments which is small with respect to the size of the problem, Krylov subspace methods may represent *computationally* efficient tools for approximating the solution.

In the remaining two sections we will throughout examples demonstrate relationship of Krylov subspace methods to some methods in computational statistics, and, in particular, their fundamental role in theoretical analysis and numerical solution of the OLS problem (10) and the TLS problem (12).

4 Linear regression and ordinary least squares

From the computational statistics perspective, linear regression is described, e.g., in Chapter 3 of the book Rao and Toutenburg (1999). From the numerical linear algebra perspective, the standard monograph Björck (1996) describes the basic theory of the least squares problem as well as the methods for its solution, including their algorithmic realizations and analysis of their numerical behavior.

Here we will not review a variety of methods. Our goal is to point out that in computational statistics and numerical linear algebra there exist *mathematically equivalent approaches for solving essentially the same problems.* The statistical approach is typically motivated by exploitation and interpretation of statistical relationships between different variables. In finite precision computations, algorithmic implementations of mathematical concepts should also take into account possible numerical instabilities. *One of the main goals in numerical linear algebra is to identify methods and algorithms which are numerically stable. Briefly, this means that the computational error due to roundoff in finite precision arithmetic computations is (in some rigourously defined meaning) under control,* see, e.g., Higham (2002), Strakoš and Liesen (2005). In this way, numerical linear algebra can perhaps in some cases provide computationally interesting methods or implementations for solving problems in computational statistics.

Properties of the solution (11) of the OLS problem (10) can conveniently be analyzed using the singular value decomposition (SVD) of the system matrix,

$$A = S\Sigma W^T = \sum_{\ell=1}^{r} s_\ell \sigma_\ell w_\ell^T, \quad S \equiv (s_1, \ldots, s_N), \quad W \equiv (w_1, \ldots, w_M)\,. \tag{20}$$

Here r represents the rank of A, $\sigma_1 \geq \sigma_2 \geq \ldots \geq \sigma_r > 0$ are the singular values, s_ℓ and w_ℓ are the corresponding left and right singular vectors, Σ is the N by M rectangular diagonal matrix with the left principal block equal to $D \equiv \mathrm{diag}\,(\sigma_1, \ldots, \sigma_r)$, and S and W are orthogonal matrices. The SVD decomposition means that any matrix can be *orthogonally diagonalized*, where the resulting diagonal consists of the positive singular values in descending order, complemented by zeros. The minimum norm solution (11) of (10) is then given by

$$x = \sum_{\ell=1}^{r} \frac{s_\ell^T b}{\sigma_\ell}\, w_\ell\,. \tag{21}$$

It might be convenient to approximate the solution (21) by considering the principal components corresponding to the several largest singular values,

$$x_k^{\mathrm{PCR}} = \sum_{\ell=1}^{k} \frac{s_\ell^T b}{\sigma_\ell}\, w_\ell, \quad k \ll r\,. \tag{22}$$

This gives in the statistical language the principal component approximation (PCA or PCR), in the numerical linear algebra language the truncated SVD approximation (TSVD). Here it should be emphasized that it *does matter* how the solution (22) is computed in finite precision arithmetic. Many modern software tools, including SAS and SPSS, contain PCR implementations based on efficient and numerically stable algorithms. On the other hand, some methods which can be found in the literature can be inefficient or numerically unstable. For the discussion of the NIPALS method for computing principal components, e.g., we refer to Eldén (2004).

In numerical linear algebra, the SVD is computed in two steps.

- First, A is orthogonally bidiagonalized,

$$A = PBQ^T\,, \tag{23}$$

  where P and Q are orthogonal matrices and B is a lower bidiagonal matrix (i.e., B has only the main diagonal and the first lower diagonal nonzero). Here the lower bidiagonal form is chosen for later convenience. The SVD decomposition works with the upper bidiagonal form analogously.

- Second, the SVD of B is computed using the implicit QR algorithm.

The procedure was proposed in a remarkable paper Golub and Kahan (1965). Important refinements of the SVD decomposition of the bidiagonal matrix B were proposed later by various authors, see, e.g., the basic textbook Watkins, (2002), in particular Chapter 4 and Section 5.9.

Golub and Kahan (1965) suggested two approaches for computing the bidiagonalization (23). The direct approach is based on Householder reflections. The iterative approach is well-suited for large problems. Starting from a normalized vector p_1, $\|p_1\| = 1$, it computes two sequences of orthonormal vectors $p_1, p_2, \ldots, p_{k+1}$ and $q_1, \ldots, q_k$ such that, in the matrix form,

$$A^T P_k = Q_k B_k^T, \quad AQ_k = P_{k+1} B_{k+}\,, \tag{24}$$

where the matrices $P_{k+1} \equiv (p_1, \ldots, p_{k+1})$ and $Q_k \equiv (q_1, \ldots, q_k)$ have orthonormal columns,

$$B_k = \begin{pmatrix} \alpha_1 & & & \\ \beta_2 & \ddots & & \\ & \ddots & \ddots & \\ & & \beta_k & \alpha_k \end{pmatrix}, \quad B_{k+} = \begin{pmatrix} \alpha_1 & & & \\ \beta_2 & \ddots & & \\ & \ddots & \ddots & \\ & & \ddots & \alpha_k \\ & & & \beta_{k+1} \end{pmatrix},$$

and α_ℓ, β_ℓ represent the normalization coefficients, $\alpha_\ell \geq 0$, $\beta_\ell \geq 0$, $\ell = 1, \ldots$, with all other entries of B_k and B_{k+} equal to zero. The iterative form (24) is called the Golub-Kahan iterative bidiagonalization (in many references it is also called the Lanczos bidiagonalization).

In Paige and Saunders (1982a) the Golub and Kahan iterative bidiagonalization was used for an approximation of the solution (21) in the following way. Let $p_1 \equiv b/\|b\|$. Then $q_1, \ldots, q_k$ represent an orthonormal basis of the Krylov subspace

$$\mathcal{K}_k(A^T A, A^T b) \equiv \mathrm{span}\,\{A^T b, (A^T A)A^T b, \ldots, (A^T A)^{k-1} A^T b\}\,.$$

Considering an approximation x_k to the solution of (11) in this subspace, $x_k \in \mathcal{K}_n(A^TA, A^Tb)$, we get $x_k \in \mathcal{R}(A^T)$. Using the orthonormal basis $q_1, \ldots, q_k$ of $\mathcal{K}_k(A^TA, A^Tb)$ we can write $x_k = Q_k y_k$, and the OLS problem (10) can be approximated by

$$AQ_k y_k = b + \hat{r}_k, \quad \min \|\hat{r}_k\| \,. \tag{25}$$

Using (24), this gives $P_{k+1}(B_{k+}y_k - \|b\|e_1) \equiv P_{k+1}r_k = \hat{r}_k$, $\|r_k\| = \|\hat{r}_k\|$. Consequently, (25) can be equivalently written in the form

$$B_{k+}y_k = \|b\|e_1 + r_k, \quad \min \|r_k\|, \quad x_k = Q_k y_k \,. \tag{26}$$

This represent the mathematical formulation of the LSQR method, with the detailed theory and implementation given in Paige and Saunders (1982a) and Paige and Saunders (1982b). Please note that the presence of Krylov subspaces remind us again about moments, which are indeed implicitly present in the algebraic formulation. The implementation given in Paige and Saunders (1982b) focuses on efficiency in storage and number of arithmetic operations, and it considers numerical stability. As above, the role of *orthogonality* is fundamental.

The description given here may at first sight not resemble any method in computational statistics. However, as stated already in Paige and Saunders (1982b), and as thoroughly explained with relationship to wider context in Eldén (2004), LSQR is *mathematically equivalent to the partial least squares method* of Wold (1975), Wold et al. (1984), see also Rao and Toutenburg (1999), Section 3.10.4. It should be emphasized that LSQR and PLS are not *computationally equivalent*. It might be useful to compare various implementations described in literature with the numerically well understood and efficient implementations of LSQR.

We finish this section with giving reference to the recent work of Arioli and Gratton (2008) which gives statistical interpretation of the energy norm in solving the OLS problem via the Krylov subspace methods related to LSQR.

5 Orthogonal regression and total least squares

As we have seen above, the Golub-Kahan iterative bidiagonalization is very useful in solving large scale OLS problems. We will conclude our short essay by illustrating the fact that the Golub-Kahan iterative bidiagonalization represents one of the trully fundamental building blocks in analysis and solution of the general linear approximation problem (8). Let (8) be orthogonally invariant, i.e., for any orthogonal P and Q let

$$P^TAQ\,(Q^Tx) \approx P^Tb \tag{27}$$

be equivalent to (8). Since the transformation by orthogonal matrices does not change the unitarily invariant norms, (10)-(12) are orthogonally invariant.

Consider now P and Q given by the (full) Golub-Kahan iterative bidiagonalization of A with $p_1 = b/\|b\|$. If at any step the normalization coefficient β_{k+1} or α_{k+1} become zero, the Golub-Kahan bidiagonalization is simply restarted with a new starting vector orthogonal to the corresponding vectors computed previously. Such continuation was proposed in Golub and Kahan (1965). The significance of the occurrence $\beta_{k+1} = 0$ or $\alpha_{k+1} = 0$ was not examined there. The approach of the authors was natural, because their goal was to compute the (full) singular value decomposition of A and therefore they focused on full bidiagonalization. In solving the approximation problems (27), the situation $\beta_{k+1} = 0$ for some $k+1 < N$ or $\alpha_{k+1} = 0$ for some $k+1 < M$ has, however, fundamental consequences.

Suppose $\beta_{k+1} = 0$ or $\alpha_{k+1} = 0$ occur for the first time, i.e., let all previously computed normalization coefficients be greater than zero. Then (27) can be written as

$$\left[\begin{array}{c|c} B_{11} & 0 \\ \hline 0 & B_{22} \end{array}\right] \left[\begin{array}{c} y_1 \\ y_2 \end{array}\right] \approx \left[\begin{array}{c} b_1 \\ 0 \end{array}\right], \tag{28}$$

where B_{11} has all elements on its main diagonal and on its first lower diagonal positive, and $b_1 = \|b\|e_1$. This means that only the problem

$$B_{11}y_1 \approx b_1 \tag{29}$$

needs to be solved, with B_{11} being equal to B_k or B_{k+} respectively. With $y_2 = 0$ a solution of $B_{22}y_2 \approx 0$, the solution of the original linear approximation problem (8) is given by

$$x = Q \left[\begin{array}{c} y_1 \\ 0 \end{array}\right]. \tag{30}$$

As shown in Paige and Strakoš (2006), see also Hnětynková and Strakoš (2007), the matrix B_{11} in (28) obtained from the partial Golub-Kahan iterative bidigonalization has minimal dimensions, and the matrix B_{22} has maximal dimensions, over all orthogonal transformations (27) giving the block structure (28), without *any* additional assumptions on the structure of B_{11} and b_1. The partial Golub-Kahan iterative bidiagonalization described above therefore represents a fundamental decomposition of data for any orthogonally invariant linear approximation problem (8). It determines the *core problem* (29) within (8) which contains all necessary and sufficient information for solving the original problem.

Without giving details, any left principal part of the matrix B_{11} in (28) can be seen as a result of the matching moments model reduction. Please note that the idea of matching moments links essentially all parts of our contribution.

The core problem theory developed in Paige and Strakoš (2006) revised, based on the earlier publications Paige and Strakoš (2002a) and Paige and Strakoš (2002b), both the theory and computations of the total least squares problem (12). It is relevant to all forms of the linear approximation problems mentioned above, and also to regularization techniques for solving ill-posed problems using the Golub-Kahan iterative bidiagonalization. Theory and some algorithmic ideas can be found in Paige and Strakoš (2006) and Hnětynková and Strakoš (2007). Important implementation issues and extensions to approximation problems with multiple right hand sides still need to be worked out, see also Sima (2006), Plešinger (2008).

Since the *TLS problem and the linear orthogonal regression are equivalent,* see the instructive description in Van Huffel and Vandewalle (1991), the core problem approach can also be found to be relevant in statistical computations.

6 Concluding remarks

Numerical linear algebra and computational statistics have much in common. Numerical linear algebra offers some generally applicable tools such as the singular value decomposition, numerically stable algorithms for computing orthogonal projections, and various direct and iterative methods, which can be applied also to problems in computational statistics. We have demonstrated through the example of linear approximation problems that the relationship between numerical linear algebra and computational statistics goes much deeper. They share some common roots, which has influenced developments in both fields. In order to demonstrate that, we have briefly recalled the role of moments in modern iterative methods of numerical linear algebra. Though computational statistics and numerical linear algebra have different goals and use different mathematical languages, their common roots lead to analogies among the results in both fields and also in recent practice. We have recalled several examples and pointed out some recent references.

Orthogonality is mentioned many times throughout this contribution. It plays a fundamental role in the description of many problems, in their analysis and understanding, as well as in interpretation of the relationship between individual variables. But it also plays a fundamental role in *computations*. Enforcing orthogonality prevents in many cases loss of information due to roundoff (as in computing projections onto Krylov subspaces, which is relevant, e.g., to both LSQR and PLS). Mathematical properties of results of finite precision computations depend to a large extent on preserving orthogonality, or on the way the orthogonality is lost, and the related issues should be thoroughly investigated. Concluding, orthogonality is not merely a theoretical mathematical concept, it is also a fundamental computational concept.

Though historical developments and the current general trends have resulted in narrow specializations and sometimes even fragmentation of sciences, we believe that the future progress of disciplines which are related depends largely on *a mutual transfer of knowledge*. Interactions between numerical linear algebra and computational statistics, which despite their differences try to exploit knowledge from their related results, can bring an enormous benefit to both fields of mathematics.

Acknowledgments. The author thanks to Rosemarie Renaut, Jaromír Antoch, Petr Tichý and Iveta Hnětynková for useful comments. This work was supported by the GA AS under project IAA100300802, and by the Institutional Research Plan AV0Z10300504.

References

Akhiezer, N. I. (1965). *The classical moment problem and some related questions in analysis*, Translated by N. Kemmer, Hafner Publishing Co.

Arioli, M. and Gratton, S. (2008). *Least squares problems, normal equations and stopping criteria for the conjugate gradient method*, submitted for publication.

Björck, Å(1996). *Numerical methods for least squares problems*, SIAM.

Eldén, L. (2004). *Partial least-squares vs. Lanczos bidiagionalization–I: analysis of a projection method for multiple regression*, Computational Statistics and Data Analysis 46, pp. 11–31.

Gantmacher, F. R. and Krein, M. G. (1941). *Oscillation matrices and kernels and small vibrations of mechanical systems*, English translation based on the 1941 Russian original edited and with a preface by Alex Eremenko, AMS Chelsea Publishing.

Gautschi, W. (1981). *A survey of Gauss-Christoffel quadrature formulae*, E.B. Christoffel, pp. 72–147, Birkhäuser.

Gautschi, W. (2004). *Orthogonal polynomials, computation and approximation*, Numerical Mathematics and Scientific Computation, Oxford University Press.

Golub, G. H. (2007). *Milestones in matrix computation: selected works of Gene H. Golub, with commentaries*. Raymond H. Chan, Chen Greif and Dianne P. O'Leary (eds), Oxford University Press.

Golub, G. H. and Van Loan, C. F. (1980). *An analysis of the total least squares problem*, SIAM J. Numer. Anal. 17, pp. 883–893.

Golub, G. H. and Kahan, W. (1965). *Calculating the singular values and pseudo-inverse of a matrix*, SIAM J. Numer. Anal. Ser B 2, pp. 205–224.

Hansen P. C. (1998). *Rank deficient and discrete ill-posed problems*, SIAM.

Higham N. J. (2002). *Accuracy and stability of numerical algorithms, Second edition*, SIAM.

Hestenes, M. R. and Stiefel, E. (1952). *Methods of conjugate gradients for solving linear systems*, J. Research Nat. Bur. Standards 49, pp. 409–436.

Hnětynková, I. and Strakoš, Z. (2007). *Lanczos tridiagonalization and core problems*, Linear Algebra and Its Applications 421, pp. 243–251.

Lanczos, C. (1952). *Solution of systems of linear equations by minimized iterations*, J. Research Nat. Bur. Standards 49, pp. 33–53.

Meurant, G. and Strakoš, Z. (2006). *The Lanczos and conjugate gradient algorithms in finite precision arithmetic*, Acta Numerica 15, pp. 471–542.

Paige, C. C. and Saunders, M. A. (1982a). *LSQR: An algorithm for sparse linear equations and sparse least squares*, ACM Trans. Math. Software 8, pp. 43–71.

Paige, C. C. and Saunders, M. A. (1982b). *Algorithm 583 LSQR: sparse linear equations and least squares problem*, ACM Trans. Math. Software 8, pp. 195–209.

Paige, C. C. and Strakoš, Z (2006). *Core problems in linear algebraic systems*, SIAM J. Matrix Anal. Appl. 27, pp. 861–875.

Paige, C. C. and and Strakoš, Z. (2002a). *Scaled total least squares fundamentals*, Numer. Math. 91, pp. 117–146.

Paige, C. C. and and Strakoš, Z. (2002b). *Unifying least squares, total least squares and data least squares,* Van Huffel, S. and Lemmerling, P. (eds), Total Least Squares and Errors-in-Variables Modeling, pp. 25–34, Kluwer.

Plešinger, M. (2008). *The total least squares problem and reduction of data in* $AX \approx B$, Ph.D thesis, Technical University of Liberec and Institute of Computer Science, Academy of Sciences of the Czech Republic.

Rao, R. C. and Toutenburg, H. (1999). *Linear models: least squares and alternatives, second edition.* Springer.

Sima, D. (2006). *Regularization techniques in model fitting and parameter estimation*. Ph.D thesis, Catholic University Leuven.

Shohat, J. A. and Tamarkin, J. D. (1943). *The problem of moments*. American Mathematical Society Mathematical surveys, vol. II, American Mathematical Society.

Stewart, G. W. (1987). *Collinearity and least squares regression*, Statistical Science 2, pp. 68–84.

Strakoš, Z. (2008). *Model reduction using the Vorobyev moment problem*, Accepted for publication in Numerical Algorithms.

Strakoš, Z. and Liesen, J. (2005). *On numerical stability in large scale linear algebraic computations*, Z. Angew. Math. Mech. 85, pp. 307–325.

Strakoš, Z. and Tichý, P. (2002). *On error estimation in the conjugate gradient method and why it works in finite precision computations*, Electron. Trans. Numer. Anal. 13, pp. 56–80 (electronic).

Van Huffel, S. and Vandewalle, J. (1991). *The total least squares problem: computational aspects and analysis*, SIAM.

Van Huffel, S. (ed) (1997). *Recent advances in total least squares techniques and errors-in-variables modelling*, SIAM.

Van Huffel, S. and Lemmerling, P. (eds) (2002). *Total least squares techniques and errors-in-variables modelling*, Kluwer.

Watkins, D. (2002). *Fundamentals of matrix computations*, J. Wiley.

Wold, H. (1975). *Soft modeling by latent variables; the nonlinear iterative partial least squares approach*, Gani, J. (ed), Perspectives in Probability and Statistics, Papers on Honour of M.S. Barlett, Academic Press.

Wold, S., Ruhe, A., Wold, H. and Dunn, W. J. (1984). *The collinearity problem in linear regression. The partial least squares (PLS) approach to generalized inverses*, SIAM J. Sci. Stat. Comput. 5, pp. 735–743.