Science & Technology
Facilities Council

Technical Report

Least-squares problems, normal equations, and stopping criteria for the conjugate gradient method

Mario Arioli and Serge Gratton

December 9, 2008

RAL-TR-2008-008

Library and Information Services
SFTC Rutherford Appleton Laboratory
Harwell Science and Innovation Campus
Didcot
OX11 0QX
UK
Tel: +44 (0)1235 445384
Fax: +44(0)1235 446403
Email: library@rl.ac.uk

The STFC ePublication archive (epubs), recording the scientific output of the Chilbolton, Daresbury, and Rutherford Appleton Laboratories is available online at: http://epubs.cclrc.ac.uk/

ISSN 1358-6254

Neither the Council nor the Laboratory accept any responsibility for loss or damage arising from the use of information contained in any of their reports or in any communication about their tests or investigation

RAL-TR-2008-008 Revision

Least-squares problems, normal equations, and stopping criteria for the conjugate gradient method

Mario Arioli[1] and Serge Gratton[2]

ABSTRACT

The Conjugate Gradient method can be successfully used in solving the symmetric and positive definite normal equations obtained from least-squares problems. Taking into account the results of Hestenes and Stiefel (1952), Golub and Meurant (1997), and Strakoš and Tichy (2002), which make it possible to approximate the energy norm of the error during the conjugate gradient iterative process, we adapt the stopping criterion introduced by Arioli (2005). Moreover, we show how the energy norm of the error is linked to the statistical properties of the least-squares problem and to the χ^2-distribution and to the Fisher-Snedecor distribution. Finally, we present the results of several numerical tests that experimentally validate the effectiveness of our stopping criteria.

Keywords: Conjugate gradient method, least squares problem, normal equations .

AMS(MOS) subject classifications: 65F10, 65N30.

[1] Current reports available by anonymous ftp to `ftp.numerical.rl.ac.uk` in directory pub/reports. The report also available through the URL `www.numerical.rl.ac.uk/reports/reports.html`. This work was supported by the EPSRC Grants EP/E053351/1 and EP/F006535/1.
Current reports available by anonymous ftp to ftp.numerical.rl.ac.uk in directory pub/reports.

[1] M.Arioli@rl.ac.uk, Rutherford Appleton Laboratory,

[2] Serge.Gratton@cnes.fr, Serge.Gratton@cerfacs.fr, CNES, 18 avenue Edouard Belin, Toulouse cedex 04, 31401, France, CERFACS, 42 ave G. Coriolis, Toulouse, 31057, France

The work of the first author was supported by the EPSRC Grants EP/E053351/1 and EP/F006535/1.

Computational Science and Engineering Department
Atlas Centre
Rutherford Appleton Laboratory
Oxon OX11 0QX

December 9, 2008

Contents

1 Introduction

The least-squares method can be used to compute reliable realizations of the minimum-variance unbiased estimates for linear regression models, and the conjugate gradient method is a very effective iterative algorithm for solving the related normal equations system.

In this paper, we want to obtain reliable stopping criteria for the conjugate gradient algorithm when it is applied to the normal equations of a least-squares problem in order to compute the realizations of the estimators related to a linear regression model.

We will take advantage of the stochastic properties of the linear regression model in order to introduce stopping criteria that, given an a-priori probability η, will stop the conjugate gradient method when the current iteration and the norm of the corresponding residual are reliable realizations with probability η. We will focus only on stopping criteria based on the energy norm: $||x||^2_{A^TA} = x^T A^T Ax$ where $A \in \mathbb{R}^{m\times n}$ is a full rank n matrix. Moreover, we will explain the link between the energy norm of the error and the χ^2 and F distribution and use it to define reliable probabilistic thresholds in the stopping criteria.

Recently, several authors have proposed rules that compute error bounds for the conjugate gradient method (Ashby, Holst, Manteuffel and Saylor 2001, Axelsson and Kaporin 2001, Calvetti, Morigi, Reichel and Sgallari 2000, Calvetti, Morigi, Reichel and Sgallari 2001, Golub and Meurant 1997, Golub and Strakos 1994, Meurant 1997, Meurant 1999*a*, Meurant 1999*b*, Strakoš and Tichy 2002). Some of these rules compute estimates of the error in Euclidean norm, and others compute estimates related to the energy norm. In their historical paper, Hestenes and Stiefel 1952 proposed a method to estimate the energy norm of the error that uses the values computed during the conjugate gradient method. Strakoš and Tichý 2002, 2005, studied the relations between the estimates proposed by Hestenes and Stiefel (1952), Golub and Meurant (1997), Golub and Strakos (1994), Meurant (1997), Meurant (1999*a*), and Meurant (1999*b*) and proved that the Hestenes-Stiefel estimate (1952) is numerically stable. The results of the previous papers have been used to introduce reliable stopping criteria for the conjugate gradient method when the linear systems arise from the approximation of elliptic variational problems. We will show that the results presented in the literature can be used in order to numerically evaluate our probabilistic stopping criteria.

We shall first summarise the principal properties of the linear regression model in Section 2. In Section 3, we discuss some relations between statistical tests and perturbed solutions of a least-squares problem. Then, in Section 4, we will use the recent results of Arioli (2005), Strakoš and Tichy (2002), Strakoš and Tichy (2002), Strakos and Tichý (2005) to build reliable stopping criteria and to analyse their properties. Finally, in Section 5 and Section 6, we will present the numerical experiments we performed on selected ill-conditioned test problems, and, in Section 7, we will present our conclusions.

In the following, we will denote stochastic variables by bold symbols. We also warn the reader that we will normally make a distinction between random variables, their estimations, and their realizations. The first two are stochastic but the latter ones are numbers.

2 Problem description

2.1 Linear regression

In this section, for any random vector $\mathbf{z}$, we denote by $E[\mathbf{z}]$ its mean and by $V[\mathbf{z}] = E[(\mathbf{z} - E[\mathbf{z}])(\mathbf{z} - E[\mathbf{z}])^T]$ its covariance matrix. The notation $\mathbf{z} \sim \mathcal{N}(z, C)$ means that $\mathbf{z}$ follows a Gaussian distribution with mean z and covariance matrix C. Let $A \in \mathbb{R}^{m\times n}, \;\; m \geq n$, with $\text{Rank}(A) = n$. We consider the linear regression model

$$\mathbf{y} = Ax + \mathbf{e}, \tag{2.1}$$

where $E[\mathbf{e}] = 0$ and $V[\mathbf{e}] = \sigma^2 I_m$. We point out that A defines a given model and x is an unknown deterministic value. The best minimum-variance unbiased (MVU) estimator of x is related to $\mathbf{y}$ by the Gauss-Markov theorem

Theorem 2.1. *For the linear model (2.1) the minimum-variance unbiased estimator of x is given by*

$$\mathbf{x}^* = (A^T A)^{-1} A^T \mathbf{y}.$$

The variance of the estimation error $V[\mathbf{x}^*]$ satisfies $V[\mathbf{x}^*] = \sigma^2 (A^T A)^{-1}$. If in addition, $\mathbf{e} \sim \mathcal{N}(0, \sigma^2 I_m)$, and if we set

$$\mathbf{s^2} = \frac{1}{m-n} ||\mathbf{r}||_2^2, \tag{2.2}$$

where $||\mathbf{r}||_2 = ||\mathbf{y} - A\mathbf{x}^*||_2$, we have

$$\mathbf{x}^* \sim \mathcal{N}(x, \sigma^2 (A^T A)^{-1}),$$

and

$$\mathbf{s^2} \sim \frac{\sigma^2}{m-n} \chi^2(m-n).$$

Moreover, the predicted value $\hat{\mathbf{y}} = A\mathbf{x}^*$ and the residual $\mathbf{r}$ are independently distributed as

$$\hat{\mathbf{y}} \sim \mathcal{N}(Ax^*, \sigma^2 A (A^T A)^{-1} A^T) \qquad \textit{and} \qquad \mathbf{r} \sim \mathcal{N}(0, \sigma^2 (I - A(A^T A)^{-1} A^T))$$

Proof. See Theorem 3.1 and Corollary 3.1 in (Hocking 1996, Ch 3, pp 69–70) or (Magnus and Neudecker 1999). □

2.2 Least squares problem

The best minimum-variance unbiased (MVU) estimators of x and σ^2 are closely related to the solution of the least-squares problem (LSP),

$$\min_x ||y - Ax||_2^2 \tag{2.3}$$

where y is a realization $\mathbf{y}$. The least-squares problem (LSP) has a unique solution

$$x^* = (A^T A)^{-1} A^T y,$$

and the corresponding minimum value is achieved by the square of the euclidean norm of

$$r = y - Ax^* = (I - P)y$$

where the matrix $I - P = I - A(A^TA)^{-1}A^T$ is the orthogonal projector onto $\mathrm{Ker}(A^T)$ and P is the orthogonal projector onto $\mathrm{Range}(A)$.

We remark here that the solution of LSP is deterministic and, therefore, supplies only a realisation of the MVU $\mathbf{x}^*$ and of $\mathbf{s^2}$ the corresponding estimator of σ^2.

The vector x^* is also the solution of the normal equations, i.e. it is the unique stationary point of $||y - Ax||_2^2$:

$$A^TAx^* = A^Ty. \tag{2.4}$$

We will denote in the following by

$$R(x) = A^T(y - Ax)$$

the residual of (2.4). Given a vector $\tilde{x} \in \mathcal{W}$, the following relations are satisfied:

$$\begin{aligned}\big(I - P\big)\,\big(y - A\tilde{x}\big) &= \big(I - P\big)y \\ \big(y - A\tilde{x}\big) &= \big(y - Ax^*\big) + A(A^TA)^{-1}A^T\big(y - A\tilde{x}\big) \\ &= \big(y - Ax^*\big) + A(A^TA)^{-1}R(\tilde{x}),\end{aligned}$$

and, then, we have

$$||y - A\tilde{x}||_2^2 = ||y - Ax^*||_2^2 + ||R(\tilde{x})||_{(A^TA)^{-1}}^2, \tag{2.5}$$

owing to the orthogonality between $y - Ax^*$ and $A(A^TA)^{-1}R(\tilde{x})$.

From the orthogonality of the projector P, the following are satisfied

$$\begin{aligned}y &= Py + \big(I - P\big)y \\ ||y||_2^2 &= ||Py||_2^2 + ||\big(I - P\big)y||_2^2 \\ ||y||_2^2 - ||Py||_2^2 &= ||\big(I - P\big)y||_2^2 = ||y - Ax^*||_2^2.\end{aligned} \tag{2.6}$$

Moreover, we have

$$||Py||_2^2 = y^TA\big(A^TA\big)^{-1}A^Ty = x^{*T}A^TAx^*,$$

and, then, from (2.5) and (2.6) we conclude that

$$||y||_2^2 - ||x^*||_{A^TA}^2 = ||\big(I - P\big)y||_2^2 = ||y - Ax^*||_2^2. \tag{2.7}$$

Finally, it is easy to verify that, given $\tilde{x}$ as an approximation of x^*,

$$\delta y = -A(A^TA)^{-1}R(\tilde{x}) \tag{2.8}$$

is the minimum norm solution of

$$\min_{w} ||w||_2^2 \qquad \text{such that} \qquad A^T A\tilde{x} = A^T(y+w). \tag{2.9}$$

Moreover, using $R(\tilde{x}) = A^T(y - A\tilde{x}) = A^T A(x^* - \tilde{x})$, we have

$$||\delta y||_2^2 = ||R(\tilde{x})||_{(A^TA)^{-1}}^2 = ||x^* - \tilde{x}||_{A^TA}^2. \tag{2.10}$$

Remark 1. *Owing to our focus on linear regression, we must assume that the perturbations are only relative to the vector y. Any perturbation of the matrix defining the linear regression model would be related to the study of a Total Least Squares Problem (TLSP).*

3 Statistical tests and perturbation theory

In this section, we summarize some results in the statistical significance test theory. Our aim is to identify when a solution of a perturbed LSP can be interpreted as a faithful realization of the stochastic variable $\mathfrak{x}$.

Furthermore, we would like to identify a subspace of Range(A) where the solution of a perturbed LSP is again a realization of $\mathfrak{x}$ with an assigned probability.

In order to achieve these two goals, we will confine our analysis to the χ^2 distribution test and to the overall F-test.

3.1 χ^2 distribution test

Introducing perturbations in the right-hand side, we can interpret any approximate solution $\tilde{x}$ of the LSP as being an exact solution of a perturbed LSP

$$\min_{x} ||y - Ax + \delta y||_2. \tag{3.11}$$

Among other possible choices, the vector δy defined by $\delta y = -(y - A\tilde{x})$ is such that $\tilde{x}$ exactly solves (2.9). Notice that for any u orthogonal to Range(A), the perturbation $\delta y = -(y - A\tilde{x}) + u$ is also such that $\tilde{x}$ exactly solves (2.9). To assess the quality of $\tilde{x}$ as an approximate solution of (2.1) it is reasonable to consider $\tilde{x}$ as a satisfactory solution if there exists a δy such that δy does not dominate, in some sense to be determined, the Gaussian noise $\mathbf{e} \sim \mathcal{N}(0, \sigma^2 I_n)$. Therefore, we can assume that δy be a sample of the stochastic variable $\mathbf{e}$, and that δy is dominated by $\mathbf{e}$ if for some small enough η,

$$Prob(\|\mathbf{e}\|_2^2 \geq \|\delta y\|_2^2) \geq 1 - \eta.$$

where the random variable $\frac{\|\mathbf{e}\|_2^2}{\sigma^2}$ follows a centered χ^2 distribution with m degrees of freedom. Thus, we can formulate our criterion as

$$p_\chi\left(\frac{\|\delta y\|_2^2}{\sigma^2}, m\right) \equiv Prob\left(\frac{\|\mathbf{e}\|_2^2}{\sigma^2} \leq \frac{\|\delta y\|_2^2}{\sigma^2}\right) \leq \eta,$$

where, since $\mathbf{e}$ is a Gaussian distribution with covariance matrix $\sigma^2 I$, the value of $p_\chi\,(.,m)$ is the so-called cumulative distribution function of the χ^2 distribution.

3.2 The overall F-test

In this section, we summarize and prove a few results on **F**-distribution and F-test that appear in the literature but in a format inconsistent with our notation. We hope that this will help some of the readers unfamiliar with Statistics in understanding our approach.

Given the model (2.1) and assuming that $\mathbf{e} \in \mathcal{N}\big(0, \sigma^2 I_n\big)$ we denote by $\Omega = \text{Range}(A)$ the n-dimensional subspace generated by the columns of A. For any realization y of $\mathbf{y}$, the corresponding realisation of the MVU estimator, that is $x^* = (A^T A)^{-1} A^T y$, belongs to $\mathbb{R}^n$. We would like to know whether it is statistically reasonable to look for x in the smaller subspace $\{z \in \mathbb{R}^n, Q^T z = 0\}$, where Q is a given $n \times (n-k)$ orthogonal matrix; we shall see that this problem is equivalent to the problem of truncating the CG iterations. Back to statistics, we want to test whether the following assumption is reasonable in the model (2.1) :

$$\text{(H)} : Q^T \mathfrak{x} = 0.$$

We will use the theory of statistical inference to accept or reject assumption (H). In this framework, we denote by **RSS** the *residual sum of squares* for the (full) model (2.1), as it is obtained when the MVU solution is considered. We have $\mathbf{RSS} = \|A\mathbf{x}^* - \mathbf{y}\|_2^2 = \|(I_m - P)\,\mathbf{y}\|_2^2$.

Let now $\mathbf{RSS_H}$ be the *residual sum of squares* for the model under the (H) assumption. Let $\underline{Q}$ be an orthogonal matrix such that the columns of $[\underline{Q}, Q]$ form an orthonormal basis of $\mathbb{R}^n$. We set

$$\mathfrak{x} = \begin{bmatrix} \underline{Q} & Q \end{bmatrix} \begin{bmatrix} \mathfrak{x}_1 \\ \\ \mathfrak{x}_2 \end{bmatrix}. \tag{3.12}$$

The decomposition (3.12) gives the linear model

$$\mathbf{y} = A[\underline{Q}, Q] \begin{bmatrix} \mathfrak{x}_1 \\ \\ \mathfrak{x}_2 \end{bmatrix} + \mathbf{e}, \quad \text{and} \quad \mathfrak{x}_2 = 0,$$

which can be rewritten

$$\mathbf{y} = A\underline{Q}\mathfrak{x}_1 + \mathbf{e}, \quad \text{and} \quad \mathfrak{x} \;=\; \underline{Q}\mathfrak{x}_1. \tag{3.13}$$

Under assumption (H), we call (3.13) the *reduced linear model.* The corresponding MVU estimator is $\mathbf{x}^*_{\mathbf{H}} = \underline{Q}(\underline{Q}^T A^T A \underline{Q})^{-1} \underline{Q}^T A \mathbf{y}$ and the reduced residual sum of squares is $\mathbf{RSS_H} = \|(I - P_H)\mathbf{y}\|^2$ where $P_H = A\underline{Q}(\underline{Q}^T A^T A \underline{Q})^{-1} \underline{Q}^T A$ is the orthogonal projector onto span$(A\underline{Q})$.

Our procedure for testing whether (H) holds relies on a comparison between $\mathbf{RSS_H}$ and **RSS**.

Lemma 3.1. *In the framework of the minimum-variance unbiased estimation, the residual sum of squares* **RSS** *and* $\mathbf{RSS_H}$ *of models (2.1) and (3.13) satisfy*

$$\mathbf{RSS_H} - \mathbf{RSS} = \mathbf{y}^T A^{\dagger T} Q \left(Q^T (A^T A)^{-1} Q\right)^{-1} Q^T A^{\dagger} \mathbf{y}.$$

Proof. From the definition of $\mathbf{RSS}$ and $\mathbf{RSS_H}$, and the relations $(I-P)^2 = (I-P)$ and $(I-P_H)^2 = (I-P_H)$, we get

$$\begin{aligned}\mathbf{RSS_H} - \mathbf{RSS} &= \mathbf{y}^T(I-P_H)\mathbf{y} - \mathbf{y}^T(I-P)\mathbf{y} \\ &= \mathbf{y}^T(P-P_H)\mathbf{y} \\ &= \mathbf{y}^T A \left((A^TA)^{-1} - \underline{Q}\left(\underline{Q}^T A^T A \underline{Q}\right)^{-1} \underline{Q}^T\right) A^T\mathbf{y}\end{aligned}$$

Setting $W_1 = (A^TA)^{\frac{1}{2}}\underline{Q}$ and $W_2 = (A^TA)^{-\frac{1}{2}}Q$, we obtain

$$\begin{aligned}&\mathbf{RSS_H} - \mathbf{RSS} = \\ &\mathbf{y}^T A (A^TA)^{-\frac{1}{2}} \left(I_n - W_1\left(W_1^T W_1\right)^{-1} W_1\right)(A^TA)^{-\frac{1}{2}} A^T\mathbf{y}.\end{aligned} \tag{3.14}$$

From $Q^T\underline{Q} = 0$ it follows that the two matrices W_1, and W_2 are orthogonal, which shows that $I_n - W_1\left(W_1^TW_1\right)^{-1}W_1 = W_2\left(W_2^TW_2\right)^{-1}W_2$, and substituting this expression into (3.14), we get

$$\mathbf{RSS_H} - \mathbf{RSS} \;=\; \mathbf{y}^T A (A^TA)^{-\frac{1}{2}} W_2\left(W_2^TW_2\right)^{-1} W_2 (A^TA)^{-\frac{1}{2}} A^T\mathbf{y},$$

and using $W_2 = (A^TA)^{-\frac{1}{2}}Q$ gives the final result. □

Using Lemma 3.1, we prove some properties of $\mathbf{RSS_H} - \mathbf{RSS}$ in the following proposition.

Proposition 3.1. *Consider the residual sum of squares* $\mathbf{RSS}$ *and* $\mathbf{RSS_H}$ *of models (2.1) and (3.13) respectively.*

1. *We have that* $E[\mathbf{RSS}] = (m-n)\sigma^2$ *and that*

$$\begin{aligned}E[\mathbf{RSS_H} - \mathbf{RSS}] &= (n-k)\sigma^2 + \mathfrak{x}^T Q\left(Q^T(A^TA)^{-1}Q\right)^{-1} Q^T\mathfrak{x} \\ &\geq (n-k)\sigma^2.\end{aligned} \tag{3.15}$$

2. *Suppose the (H) assumption holds. We have that*

$$E[\mathbf{RSS_H} - \mathbf{RSS}] = (n-k)\sigma^2,$$

and if, in addition, $\mathbf{e} \sim \mathcal{N}\big(0, \sigma^2 I_m\big)$*, then the following probability distributions hold*

$$\mathbf{RSS_H} - \mathbf{RSS} \sim \sigma^2\chi^2(n-k) \quad \text{and} \quad \mathbf{RSS} \sim \sigma^2\chi^2(m-n).$$

Proof. For the first part of the proposition, starting from $\mathbf{x}^* = A^\dagger\mathbf{y}$, we get from Lemma 3.1 that

$$\mathbf{RSS_H} - \mathbf{RSS} = \mathbf{x}^{*T} M \mathbf{x}^*,$$

and

$$\mathbf{RSS_H} - \mathbf{RSS} = (\mathbf{x}^* - \mathfrak{x})^T M(\mathbf{x}^* - \mathfrak{x}) + 2\mathfrak{x}^T M(\mathbf{x}^* - \mathfrak{x}) + \mathfrak{x}^T M\mathfrak{x},$$

where $M = Q\left(Q^T(A^TA)^{-1}Q\right)^{-1}Q^T$. By using the fact that $\mathbf{x}^*$ is an unbiased estimation of $\mathfrak{x}$, (i.e. $E[\mathbf{x}^* - \mathfrak{x}] = 0$), we get

$$E[\mathbf{RSS_H} - \mathbf{RSS}] = E[(\mathbf{x}^* - \mathfrak{x})^T M(\mathbf{x}^* - \mathfrak{x})] + \mathfrak{x}^T M\mathfrak{x}. \tag{3.16}$$

Using the linearity of the trace and of the mathematical expectation, and the definition of the variance $Q[\mathbf{x}^*]$ of the estimation, we have that

$$\begin{aligned} E[(\mathbf{x}^* - \mathfrak{x})^T M(\mathbf{x}^* - \mathfrak{x})] &= E[\text{trace}(M(\mathbf{x}^* - \mathfrak{x})(\mathbf{x}^* - \mathfrak{x})^T)] \\ &= \text{trace}(ME[(\mathbf{x}^* - \mathfrak{x})(\mathbf{x}^* - \mathfrak{x})^T]) = \text{trace}(MV[\mathbf{x}^*]) \end{aligned}$$

Using Theorem 2.1, we know that $V[\mathbf{x}^*] = \sigma^2(A^TA)^{-1}$, which yields

$$\begin{aligned} E[(\mathbf{x}^* - \mathfrak{x})^T M(\mathbf{x}^* - \mathfrak{x})] &= \sigma^2\text{trace}(Q\left(Q^T(A^TA)^{-1}Q\right)^{-1}Q^T(A^TA)^{-1}]) \\ &= \sigma^2\text{trace}(I_{n-k}) = \sigma^2(n-k). \end{aligned}$$

Substituting this last expression into (3.16) completes the proof of (3.15).

For the second part, under the (H) assumption, we have $Q^T\mathfrak{x} = 0$, and also $M\mathfrak{x} = 0$, and, then, $E[\mathbf{RSS_H} - \mathbf{RSS}] = \sigma^2(n-k)$. Using $M\mathfrak{x} = 0$, and $A^\dagger A\mathfrak{x} = \mathfrak{x}$, we get $MA^\dagger A\mathfrak{x} = 0$ and letting $P_H = A^{\dagger T}MA^\dagger$, we have that $\mathbf{RSS_H} - \mathbf{RSS} = (\mathbf{y} - A\mathfrak{x})^T P_H(\mathbf{y} - A\mathfrak{x}) = \mathbf{e}^T P_H\mathbf{e}$. Denoting by $W_1 = A^{\dagger T}Q$, we obtain that

$$P_H = W_1\left(W_1^TW_1\right)^{-1}W_1^T.$$

This shows that P_H is the orthogonal projection onto Range(W_1). Since Ker(W_1) = Ker$(W_1^TW_1)$ = Ker$(Q^T(A^TA)^{-1}Q) = \{0\}$, W_1 is a full rank matrix and rank$(P_H) = n-k$. Therefore from $\mathbf{e} \sim \mathcal{N}\big(0, \sigma^2 I_m\big)$, we get that

$$\frac{\mathbf{RSS_H} - \mathbf{RSS}}{\sigma^2} = \frac{\mathbf{e}^T P_H\mathbf{e}}{\sigma^2} \sim \chi^2(n-k).$$

Similarly from $\mathbf{RSS} = \|\mathbf{y} - A\mathbf{x}^*\|^2 = \|(I_m - P)\mathbf{y}\|^2$, we get, since $(I_m - P)A\mathfrak{x} = 0$, that $\mathbf{RSS} = \mathbf{e}^T(I-P)\mathbf{e}$. As before,because $I-P$ is an orthogonal projector of rank $m-n$ and $\mathbf{e} \sim \mathcal{N}\big(0, \sigma^2 I_m\big)$, we get that

$$\frac{\mathbf{RSS}}{\sigma^2} = \frac{\mathbf{e}^T(I_m - P)\mathbf{e}}{\sigma^2} \sim \chi^2(m-n).$$

□

We are now in position to propose our test for assumption (H). Using Proposition 3.1 is is clear that when (H) is true, $E[\mathbf{RSS_H} - \mathbf{RSS}] = \sigma^2(n-k)$, therefore the ratio $\frac{\mathbf{RSS_H} - \mathbf{RSS}}{\sigma^2(n-k)}$ will be close to 1. When (H) is not true, we know from (3.15) that this ratio will be larger than 1. In practical situations, σ is not always known, and we replace this quantity by its MVU estimation. From Theorem 2.1, we get the following ratio:

$$\mathbf{F} = \frac{\frac{\mathbf{RSS_H} - \mathbf{RSS}}{n-k}}{\frac{\mathbf{RSS}}{m-n}}.$$

Under assumption (H) $\mathbf{F}$ is the ratio of the two χ^2 distributions $\chi^2(n-k)$ and $\chi^2(m-n)$ divided by their degrees of freedom $n-k$ and $m-n$. Therefore (see Hocking, 1996, Chapter 3.3), under the (H) assumption, $\mathbf{F}$ follows a Fisher-Snedecor distribution with $n-k$ and $m-n$ degrees of freedom. We explain now the principle of the F-test. Using the realization y of $\mathbf{Y}$, we compute two realizations RSS and RSS_H of the residual sum of squares, and forming their ratio

$$f = \frac{\frac{RSS_H - RSS}{n-k}}{\frac{RSS}{m-n}},$$

we compare the value of f with what can be obtained from a Fisher-Snedecor distribution with $n-k$ and $m-n$ degrees of freedom (i.e. with what we should have if assumption (H) holds). If the realization f is very different from what should be expected (i.e. is too large), assumption (H) is rejected.

More precisely, let us denote by $p_{FS}(., n-k, m-n)$ the Fisher-Snedecor cumulative distribution function, that is associated with $\mathbf{F}$ if the assumption (H) is true. In the analysis of statistical significance tests, a (usually small) value $\eta > 0$ is introduced, that is called the level of significance of the test. A value f_0 is computed such that $p_{FS}(f_0, n-k, m-n) = \eta$, and the region $\{\mathbf{F} > f_0\}$ is then called the critical region of the test. The assumption (H) is rejected whenever f is such that $f > f_0$; η therefore represents the probability of rejecting (H) when (H) is true (because f is considered to be too large). In the alternative case, $f \leq f_0$, (H) is accepted (there is no statistical reason for rejecting it).

Remark 2. *We point out that the F-test can be seen as a test on a possible model reduction of the original problem (2.1). The test aims to validate if the original number of parameters n can be drastically reduced.*

4 Stopping criteria for CGLS

If we use the conjugate gradient method in order to compute the solution of (2.4), it is quite natural to have a stopping criterion which takes advantage of the minimization property of this method and of the stochastic properties of the underpinning problem (2.1). Our analysis is based on the results of (Hestenes and Stiefel 1952, Strakoš and Tichy 2002, Strakos and Tichý 2005, Meurant 1999*b*, Arioli 2005) that are relative to the solution of linear systems arising in the approximation of PDEs.

At each step k the conjugate gradient method minimizes the energy norm of the error $\delta x^{(k)} = x^* - x^{(k)}$ on a Krylov space $x^{(0)} + \mathcal{K}_k$ (Greenbaum 1997):

$$\min_{x^{(k)} \in\, x^{(0)} + \mathcal{K}_k} \delta x^{(k)T} A^T A \delta x^{(k)}, \tag{4.17}$$

where $\mathcal{K}_k = span(A^T y, (A^T A)A^T y, \ldots, (A^T A)^{k-1}A^T y)$. Let $R^{(k)} = A^T\big(y - Ax^{(k)}\big)$ denote the normal equations residual at step k. Moreover, using $A^T A\delta x^{(k)} = A^T(y - Ax^{(k)}) = R^{(k)}$, the value $\|\delta x^{(k)}\|_{A^T A}$ is equal to the dual norm of the residual $\|R^{(k)}\|_{(A^T A)^{-1}}$ (2.10). The conjugate gradient iterates satisfy the following relations (Meurant 1999*a*):

$$\begin{aligned} x^{(k)} &= x^{(k-1)} + \alpha_{k-1}q^{(k-1)}, \quad \alpha_{k-1} = \frac{R^{(k-1)T}R^{(k-1)}}{q^{(k-1)T}A^T Aq^{(k-1)}}\,, \\ R^{(k)} &= R^{(k-1)} - \alpha_{k-1}A^T Aq^{(k-1)}\,, \\ q^{(k)} &= R^{(k)} + \beta_{k-1}q^{(k-1)}, \quad \beta_{k-1} = \frac{R^{(k)T}R^{(k)}}{R^{(k-1)T}R^{(k-1)}}, \end{aligned}$$

where $x^{(0)} = 0$ and $R^{(0)} = q^{(0)} = y$. The quantity α_{k-1} gives the step-size on the direction $q^{(k-1)}$ during the conjugate gradient algorithm. Therefore, in exact arithmetic, we have the final value

$$x^* = \sum_{j=0}^{n-1} \alpha_j q^{(j)}\ ,$$

and taking into account that

$$q^{(j)T}A^T Aq^{(i)} = 0, \qquad i \neq j,$$

the energy norm of the error $\delta x^{(k)} = x^* - x^{(k)}$ is

$$\|\delta x^{(k)}\|^2_{A^T A} = e_A^{(k)} = \sum_{j=k}^{n-1} \alpha_j R^{(j)T}R^{(j)}, \tag{4.18}$$

and the energy norm of x^* is

$$\|x^*\|^2_{A^T A} = \sum_{j=0}^{n-1} \alpha_j R^{(j)T}R^{(j)}, \tag{4.19}$$

Under the assumption that $e_{A^T A}^{(k+d)} << e_{A^T A}^{(k)}$, where the integer d denotes a suitable delay, the Hestenes and Stiefel estimate ξ_k of the energy-norm of the error (4.18) will be then computed by the formula

$$\xi_k = \sum_{j=k}^{k+d-1} \alpha_j R^{(j)T}R^{(j)}. \tag{4.20}$$

When the conjugate gradient method is applied to systems related to the approximation of PDEs, $d = 10$ is indicated as a successful compromise in order to compute a faithful estimate of (4.18) and (4.19) (Golub and Meurant 1997), and numerical experiments support this conclusion (Golub and Meurant 1997, Arioli 2005). In Section 6, we will indicate that the cheaper choice $d = 5$ can be reliable if a good preconditioner is available, and we will experimentally compare several choices for the value of d when the matrix A is ill conditioned. In this latter case, we may choose a larger value for d.

Finally, we must estimate $\|y - Ax^*\|_2$. It follows from (4.19) that

$$\|x^*\|^2_{A^TA} \geq \sum_{j=0}^{k-1} \alpha_j R^{(j)T} R^{(j)} = \nu_k.$$

Therefore, from (2.7) we have the following upper bound

$$\|y - Ax^*\|_2^2 \leq \|y\|_2^2 - \nu_k. \tag{4.21}$$

Introducing a preconditioner, we want to speed up the convergence rate of the conjugate gradient method but this will change the matrix and, therefore, the energy norm. However, we still want to estimate $e_A^{(k)}$. In (Meurant 1999*b*, Arioli 2005) it is proved that the energy norm of the preconditioned problem is equal to $e_A^{(k)}$. Moreover, when finite precision arithmetic is used, the Hestenes and Stiefel approach is stable (Strakoš and Tichy 2002, Strakos and Tichý 2005).

Finally, if we denote by M the preconditioner, we obtain the variant of the preconditioned conjugate gradient algorithm for normal equations CGLS2 (Björck, Elfving and Strakoš 1998) described in Figure 4.1, which incorporates a general stopping criteria, that will be specified in the following sections, for a suitable choice of d. However, our analysis can be extended to other algorithms based on Krylov spaces approximation where the normal equations are not explicitly computed such as LSQR (Paige and Saunders 1982).

4.1 Energy norm stopping criterion

In (Meurant 1999*b*, Arioli 2005), a stopping criterion such as the following was introduced:

$$\text{IF} \quad \|R^{(k)}\|_{(A^TA)^{-1}} \leq \eta \|y - Ax^*\|_2 \quad \text{THEN STOP} \ , \tag{4.22}$$

with $\eta < 1$ an a-priori threshold fixed by the user.

Taking into account (4.20) and (4.21), we could replace $\|y - Ax^*\|_2$ with its upper bound at the step k of the conjugate gradient method. Therefore, we can replace (4.22) with the implementable:

$$\text{IF} \qquad \xi_k \leq \eta(\|y\|_2^2 - \nu_k) \qquad \text{THEN STOP.} \tag{4.23}$$

4.2 χ^2 stopping criteria for CG

To detect the convergence as early as possible and avoid over-solving in the LSP, we consider a δy_0 with minimum Euclidean norm such that $\tilde{x}$ exactly solves (3.11). Using the (2.8), (2.9) and (2.10), we see that $||\delta y_0||_2^2 = ||R(\tilde{x})||^2_{(A^TA)^{-1}} = \|\delta x^{(k)}\|_{A^TA}$. Finally, using the estimation (4.20), we propose the following stopping criterion,

$$\text{IF} \ \ p_\chi\left(\frac{\xi_k}{\sigma^2}, m\right) \leq \eta \ \ \text{THEN STOP} \ . \tag{4.24}$$

```
Preconditioned CGLS2 (PCGLS)
Given an initial guess $x^{(0)}$, compute $R^{(0)} = A^T\big(y - Ax^{(0)}\big)$,
and solve $Mz^{(0)} = R^{(0)}$. Set $q^{(0)} = z^{(0)}$, $\beta_0 = 0$, $\alpha_{-1} = 1$,
$\nu_0 = 0$, $\chi_1 = R^{(0)T} z^{(0)}$, and $\xi_0 = \infty$.

$k = 0$
while $\mathbf{z}(\xi_{k-d}, \|y\|_2^2 - \nu_k, \sigma^2) > \eta$ do
        $k = k + 1$;
        $q^{(k-1)} = Aq^{(k-1)}$;
        $\alpha_{k-1} = \chi_k / ||q^{(k-1)}||_2^2$;
        $\psi_k = \alpha_{k-1}\chi_k$; $\nu_k = \nu_{k-1} + \psi_k$;
        $x^{(k)} = x^{(k-1)} + \alpha_{k-1} q^{(k-1)}$;
        $R^{(k)} = R^{(k-1)} - \alpha_{k-1} A^T q^{(k-1)}$;
        Solve $Mz^{(k)} = R^{(k)}$;
        $\chi_{k+1} = R^{(k)T} z^{(k)}$ ;
        $\beta_k = \chi_{k+1} / \chi_k$;
        $q^{(k)} = z^k + \beta_k q^{(k-1)}$;
        if $k > d$ then
            $\xi_{k-d} = \sum_{j=k-d+1}^{k} \psi_j$;
        else
            $\xi_{k-d} = \xi_{k-1}$;
        endif
end while.
```

Figure 4.1: Preconditioned Conjugate Gradient Algorithm Normal Equations (PCGLS2)

Alternatively, we can substitute σ^2 with its estimator s^2 and, then, approximate this by $||y||_2^2 - \nu_k$ using (4.21). Thus, we have the alternative stopping criterion

$$\text{IF} \;\; p_\chi\left(\frac{\xi_k}{||y||_2^2 - \nu_k}, m\right) \leq \eta \;\; \text{THEN STOP} \; . \tag{4.25}$$

Formula (4.25) can be seen as a non-linear version of the stopping criterion (4.23).

4.3 An F-test stopping rule for CG

In this framework, the k-th iterates of the conjugate gradient method belongs to the Krylov subspace

$$\mathcal{K}_k = span(A^T y, (A^T A) A^T y, \ldots, (A^T A)^{k-1} A^T y).$$

The finite termination property of the conjugate gradient method, implies that the best linear unbiased estimate of $\mathbf{x}^*$ belongs to $\mathcal{K}_n$. It is also well known that

- the Krylov spaces $\mathcal{K}_k$ form a sequence such that

$$\mathcal{K}_0 \subset \mathcal{K}_1 \subset \cdots \subset \mathcal{K}_k \subset \mathcal{K}_n \subset \mathrm{Range}(A^T A),$$

and (see (Greenbaum 1997)) there exists $Q_j \in \mathbb{R}^{n\times k} \quad k = 1,\ldots,n$ such that

$$\begin{array}{rcl} \mathcal{K}_k & = & \mathrm{Range}(Q_k) \\ A^T A Q_k & = & Q_{k+1} T_k \\ Q_k^T Q_k & = & I \qquad \forall k \\ Q_{k+1} & = & \big[Q_k; q^{(k+1)}\big], \qquad \Big(q^{(k+1)T} Q_k = 0\Big); \end{array}$$

where $T_k \in \mathbb{R}^{(k+1)\times k}$ has the following structure

$$T_k = \begin{bmatrix} T_{kk} \\ \\ \gamma e_k^T \end{bmatrix}$$

with T_{kk} symmetric positive definite tridiagonal matrix;

- the orthogonal projector P_k on $\omega_k = A\mathcal{K}_k$ can be expressed as

$$P_k = AQ_k \big(T_{kk}\big)^{-1} Q_k^T A^T.$$

Let us denote by $\mathfrak{Q}_k = [q^{(n-k+1)}, \ldots q^{(n)}]$. Stopping the CG iterations reduces to deciding whether the hypothesis

$$\text{(H):} \qquad \mathfrak{Q}_k^T \mathfrak{x} = 0$$

is statistically reasonable under the linear model assumptions $\mathbf{y} = A\mathfrak{x} + \mathbf{e}$. In practice, using the realization y of the random variable $\mathbf{y}$, the F-test sets the following *residual sum of squares definitions*, $RSS = y^T(I-P)y = \|(I-P)y\|^2$, and similarly $RSS_k = \|(I-P_k)y\|^2$ and considers the quantity

$$f_k = \frac{(RSS_k - RSS)/(n-k)}{RSS/(m-n)}.$$

More precisely, we decide to accept the hypothesis, if f_k is small enough, so that the probability of rejecting (H) when (H) is true is small enough, i.e if

$$Prob(\mathbf{F} \leq f_k) \leq \eta.$$

By definition of the cumulative distribution function $p_{FS}(., n-k, m-n)$, the criterion would read

$$\text{IF} \;\; p_{FS}\,(f_k, n-k, m-n) \leq \eta \;\; \text{THEN STOP} \; . \qquad (4.26)$$

To implement this criterion in practice, we must find reasonable approximations for the realizations of RSS and RSS_k. Clearly,

$$RSS = \|(I-P)y\|_2^2 = \|y - A(A^T A)^{-1} A^T y\|_2^2 = \|y - Ax^*\|_2^2.$$

Similarly,

$$RSS_k = \|(I - AQ_k\big(T_{kk}\big)^{-1}Q_k^T A^T)y\|_2^2 = \|y - Ax_k\|_2^2.$$

Therefore, we have

$$f_k = \frac{(\|y - Ax_k\|_2^2 - \|y - Ax^*\|_2^2)/(n-k)}{\|y - Ax^*\|_2^2/(m-n)},$$

and using (2.6), we get

$$\text{IF } \; p_{FS}\left(\frac{\|R^{(k)}\|_{(A^TA)^{-1}}^2/(n-k)}{||y - Ax^*||_2^2/(m-n)}, n-k, m-n\right) \leq \eta \;\; \text{THEN STOP} \; . \tag{4.27}$$

As for the χ^2 test, using ξ_k and $||y||_2^2 - \nu_k$ to respectively approximate $\|R^{(k)}\|_{(A^TA)^{-1}}^2$ and $||y - Ax^*||_2^2$, we get our F-test based stopping criterion

$$\text{IF } \; p_{FS}\left(\Big(\frac{m-n}{n-k}\Big)\frac{\xi_k}{||y||_2^2 - \nu_k}, n-k, m-n\right) \leq \eta \;\; \text{THEN STOP} \; . \tag{4.28}$$

Remark 3. *The values of $p_\chi\,(.,m)$ and $p_{FS}(.,n-k,m-n)$ can be computed using standard algorithms (Abramowitz and Stegun 1964, sec. 6.5 and sec.26.5). The cost of the numerical computation of the integrals involved is normally negligible compared to the cost of matrix by vector products performed during the conjugate gradient method. Moreover, reliable software exists in Matlab and in IMSL and NAG FORTRAN libraries.*

4.4 Choice of η

The choice of η will depend on the properties of the problem that we want to solve, and, in the practical cases, η can be frequently much larger than ε , the roundoff unit of the computer finite precision arithmetic.

The stopping criterion (4.23) depends on an η that it is quite difficult to choose. A priori there is not an evident link between η and the statistical nature of the original problem. The only reasonable choices could be related to roundoff perturbations and this implies that $\eta = \epsilon$ or $\eta = \sqrt{\epsilon}$ where $\epsilon \approx 10^{-16}$ is the machine precision. The stopping criterion (4.25) can be seen as a nonlinear statistical version of (4.23). In this case the choice of η is related to the probability the user would like to fix. The same probabilistic interpretation holds for stopping criteria (4.24) and (4.28). In both these cases the user can choose the value of η as a probability. In practice, $\eta \leq 10^{-3}$ is a suitable choice.

5 Test problems

5.1 Dense tests

In order to illustrate the behaviour of the stopping criteria proposed for CG on the model (2.1), we propose academic test cases. We consider the $m \times n$ matrix A that is given from its singular value decomposition $A = U\Sigma V^T$, where, using the Matlab notation:

- The $m \times m$ and $n \times n$ orthogonal matrices U and V are obtained in Matlab as Q factors in the QR decomposition of a random matrix (i.e. for the computation of V, `X=rand(n,n);` and `[V,R]=qr(X);`)
- The entries of the diagonal matrix Σ are nearly equally spaced between 1 and an assigned $\kappa(A)$ (i.e. `Sigma=diag(linspace(1,`κ`(A),n))`).

We set $m = 200$ and $n = 40$, and vary the condition number of the problem by considering $\kappa(A)$=10^2, 10^3 and 10^4 and precondition CG using one step of symmetric Gauss-Seidel iterative method. The particular realization of the right-hand side of the linear model is defined by $y = Ax + e$, where $x = (\cos(1), \ldots, \cos(n))^T$, and e is obtained from the Matlab command `randn`, which means that e is a realization of a pseudo-random vector drawn from a Gaussian distribution with mean zero and a standard deviation equal to one as obtained from the Marsaglia's Ziggurat algorithm, see `help randn` in Matlab.

5.2 Data assimilation test

Data assimilation problems constitute an important class of regression problems (Tshimanga, Gratton, Weaver and Sartenaer 2008). Their purpose is to reconstruct the initial conditions at $t = 0$ of a dynamical system based on knowledge of the system's evolution laws and on observations of the state at times t_i. More precisely, consider a linear dynamical system described by the equation $\dot{u} = f(t, u)$ whose solution operator is given by $u(t) = M(t)u_0$. Assume that the system state is observed (possibly only in parts) at times $\{t_i\}_{i=0}^N$, yielding observation vectors $\{y_i\}_{i=0}^N$, whose model is given by $y_i = Hu(t_i) + \epsilon$, where ϵ is a noise with covariance matrix $R_i = \sigma^2 I$. We are then interested in finding u_0 which minimizes

$$\frac{1}{2}\sum_{i=0}^{N} \|HM(t_i)u_0 - y_i\|^2_{R_i^{-1}}.$$

We consider here the case where the dynamical system is the linear heat equation in a two-dimensional domain, defined on $S_2 = [0, 1] \times [0, 1]$ by

$$\frac{\partial u}{\partial t} = -\Delta u \quad in \quad S_2, \quad u = 0 \quad on \quad \partial S_2, \quad u(.,0) = u_0 \quad in \quad S_2 \tag{5.29}$$

The system is integrated with timestep dt, using an implicit Euler scheme. In the physical domain, a regular finite difference scheme is taken for the Laplace operator, with same spacing h in the two spatial dimensions. The data of our problem is computed by imposing a solution $u_0(x, y, 0)$ computing the exact system trajectory and observing Hu at every point in the spatial domain and at every time step. In our application, $m = 8100$, $n = 900 = 30^2$, $dt = 1$, $h = 1/31$, $N = 8$ and $H = \text{diag}((1^{1.5}, 2^{1.5}, \ldots, n^{1.5})$. The observation vector y is obtained by imposing $u_0(x, y, 0) = \frac{1}{4}\sin(\frac{1}{4}x)(x-1)\sin(5y)(y-1)$, and by adding a random measurement error with Gaussian distribution with zero mean and covariance matrix $R_i = \sigma^2 I_n$, where $\sigma = 10^{-3}$. In our numerical experiments, we use CGLS2 without a preconditioner.

6 Numerical experiments

In the next series of experiments, rather than implementing a particular stopping criterion we run the PCGLS2 until the n-th step is reached, and report on the values taken by the various stopping criteria considered in this paper at x_k:

1. $\mu_1 = p_{FS}\left(\left(\frac{m-n}{n-k}\right)\frac{\xi_k}{\|y\|_2^2-\nu_k}, n-k, m-n\right)$ (see Equation (4.28)),

2. $\mu_2 = p_\chi\left(\frac{\xi_k}{\sigma^2}, m\right)$ (see Equation (4.24)),

3. $\mu_3 = p_\chi\left(\frac{(m-n)\xi_k}{\|y\|_2^2-\nu_k}, m\right)$, (see Equation (4.25)).

The value of p_{FS} and p_χ are respectively computed by the Matlab functions `fcdf`, and chis_cdf and gammainc that implement the algorithms presented in (Abramowitz and Stegun 1964). In all our numerical experiments, we choose $\eta = 10^{-6}$ in all our stopping criteria.

In Table 6.1, we report the iteration index for which each stopping criterion achieves convergence, both for $d = 5$ and $d = 10$ and for all out test problems.

	$\mathbf{d} = 5$			$\mathbf{d} = 10$		
Test problem	μ_1	μ_2	μ_3	μ_1	μ_2	μ_3
Dense2 ($\kappa(A) = 10^2$)	12	6	6	14	6	6
Dense3 ($\kappa(A) = 10^3$)	25	14	10	26	16	12
Dense4 ($\kappa(A) = 10^4$)	28	26	14	28	26	26
Data assimilation ($\kappa(A) = 10^3$)	24	15	10	25	16	10

Table 6.1: Number of iterations for each stopping criterion setting the parameter $\eta = 10^{-6}$.

The results of Table 6.1 indicate that the stopping criteria depend very mildly on the value of d. Because $\zeta_k = \frac{\|y\|_2^2-\nu_k}{m-n}$ overestimates σ^2 (in our case $\sigma^2 = 1$), the stopping criterion μ_3 accepts large residuals from CG, because the right hand side y is assumed to have a larger standard deviation. This explains why μ_3 detects convergence before μ_2. The criterion μ_1 based on the Fisher-Snedecor distribution is more conservative and seems to detect convergence after the two first stopping criteria. Moreover, the stopping criterion μ_3 stops the process too early and in the case of the dense test problem with $\kappa = 10^4$ the convergence oscillates quite dramatically as Figure 6.2 illustrates.

Our stopping criteria rely on the knowledge of an approximation ξ_k of the square of the energy norm of the error, that in turn relies on a suitable choice of the delay parameter d. In order to assess the quality of the approximation, we represent on the one hand $\xi_k^{1/2}$

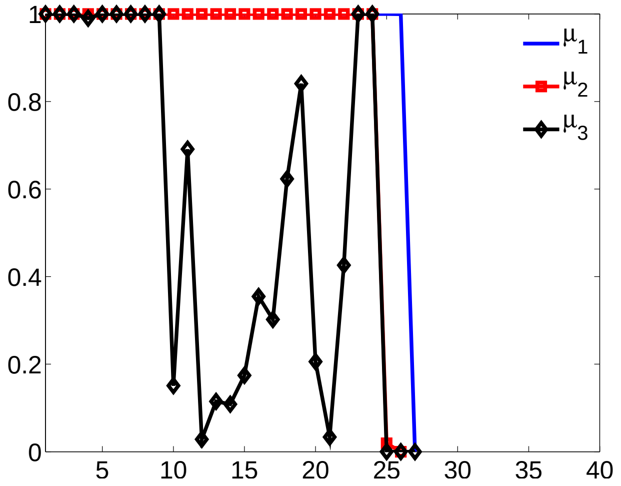

Figure 6.2: Stopping criteria versus CG iteration count for $\kappa = 10^4$

obtained with $d = 5$, and on the other hand, an approximation of the energy norm of the error, that we denote by $\delta_k^{1/2}$. The quantity δ_k is just $(x_k - x^*)^T(A^TA)(x_k - x^*)$, where x^* is approximated from the direct solution of the least squares problem from the Matlab backslash expression, i.e., from the QR method for least-squares. Note therefore that δ_k is subject to round-off errors in particular in the computation of the exact solution $x^* = A^\dagger y$. However, we observe on Figure 6.3 that $\xi_k^{1/2}$ is a reasonable approximation of the energy norm of the error $\delta_k^{1/2}$. If the standard deviation σ is unknown, we use Equation (2.2), and Equation (4.21), and approximate σ^2 by $\zeta_k = \frac{\|y\|_2^2 - \nu_k}{m-n}$. In Figure 6.4, we represent $\zeta_k^{1/2}$ along the CG iterations. The number of steps required in order to converge to $\sigma^2 = 1$ depends on the condition number of the problem and $\zeta_k^{1/2}$ decreases monotonically. In order to decide if our stopping criteria give relevant information on the convergence of our parameter estimation, we inspect the residual $r_k = y - Ax_k$ obtained at step k. Ideally, when the process has converged, i.e. when $k = n$ in exact arithmetic, the corresponding residual should be $r^* = y - Ax^* = A(\mathfrak{x} - x^*) + e$. Therefore, if x^* is close to the true $\mathfrak{x}$ of the linear model, and if x_k is sufficiently close to x^*, r_k should be a possible realization of a Gaussian vector. Figure 6.5 shows the residual histograms for 3 iteration numbers corresponding to the convergence detected by μ_3, μ_1 and to an iteration far away from the one where we stopped in case (d). In particular, we have represented the Gaussian with zero mean and standard deviation 1 in solid line. We see that at iteration 14 (middle graph in Figure 6.5(c)), the iterate has converged according to μ_3 (see Table 6.1), and the residual histogram is very far from being Gaussian.

It appears clearly in Figure 6.2 that of μ_2, which differs from μ_3 from the fact that the exact value of the variance is used, exhibit a smoother behaviour than μ_3. The criterion μ_1, that is relying on the F-test, and that does not assume that the variance is known, again gives a smoother behaviour than μ_3, and should be preferred. These observations

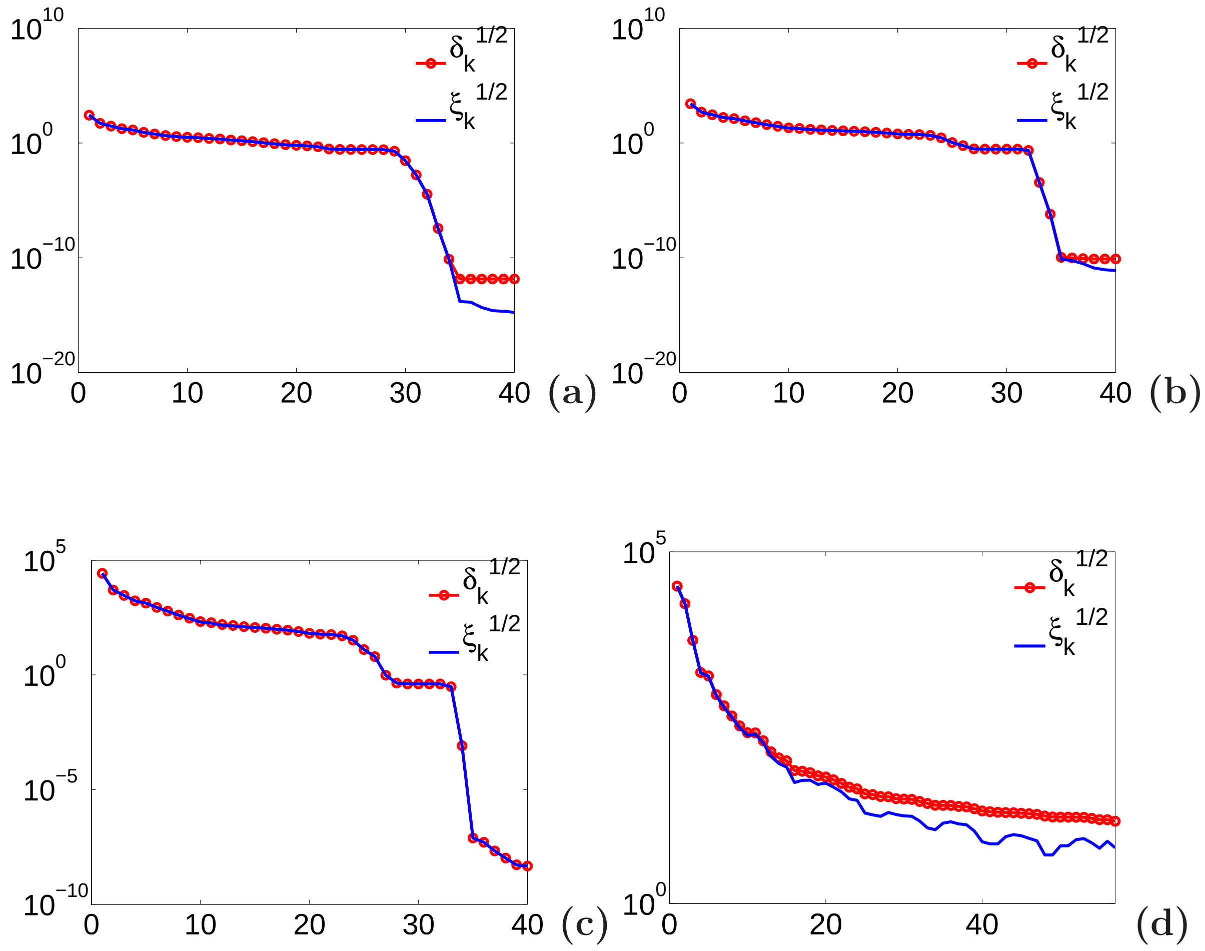

Figure 6.3: Energy norm of the error (d=5): (a) dense problem $\kappa(A) = 10^2$ (b) dense problem $\kappa(A) = 10^3$, (c) dense problem $\kappa(A) = 10^4$, (d) data assimilation problem.

are supported by all the convergence curves we obtained, even by those not reported in the present numerical section. Finally, we also observed that when the iterations are carried out until μ_1 is below 10^{-3}, the final residual r_k is close to a Gaussian vector.

Remark 4. *In the dense case, we experimented with several other distributions of the singular values. We used distributions such that the entries Σ_i on the diagonal of Σ are*

$$\log_{10} \Sigma_i = \pm c \left(\frac{i-1}{n-1} \right)^{\gamma},$$

and the results are similar to the ones described in the Section above.

7 Conclusions

Using available results on the energy norm for least-squares problems, we have proposed three possible choices for a stopping criterion based on statistical considerations.

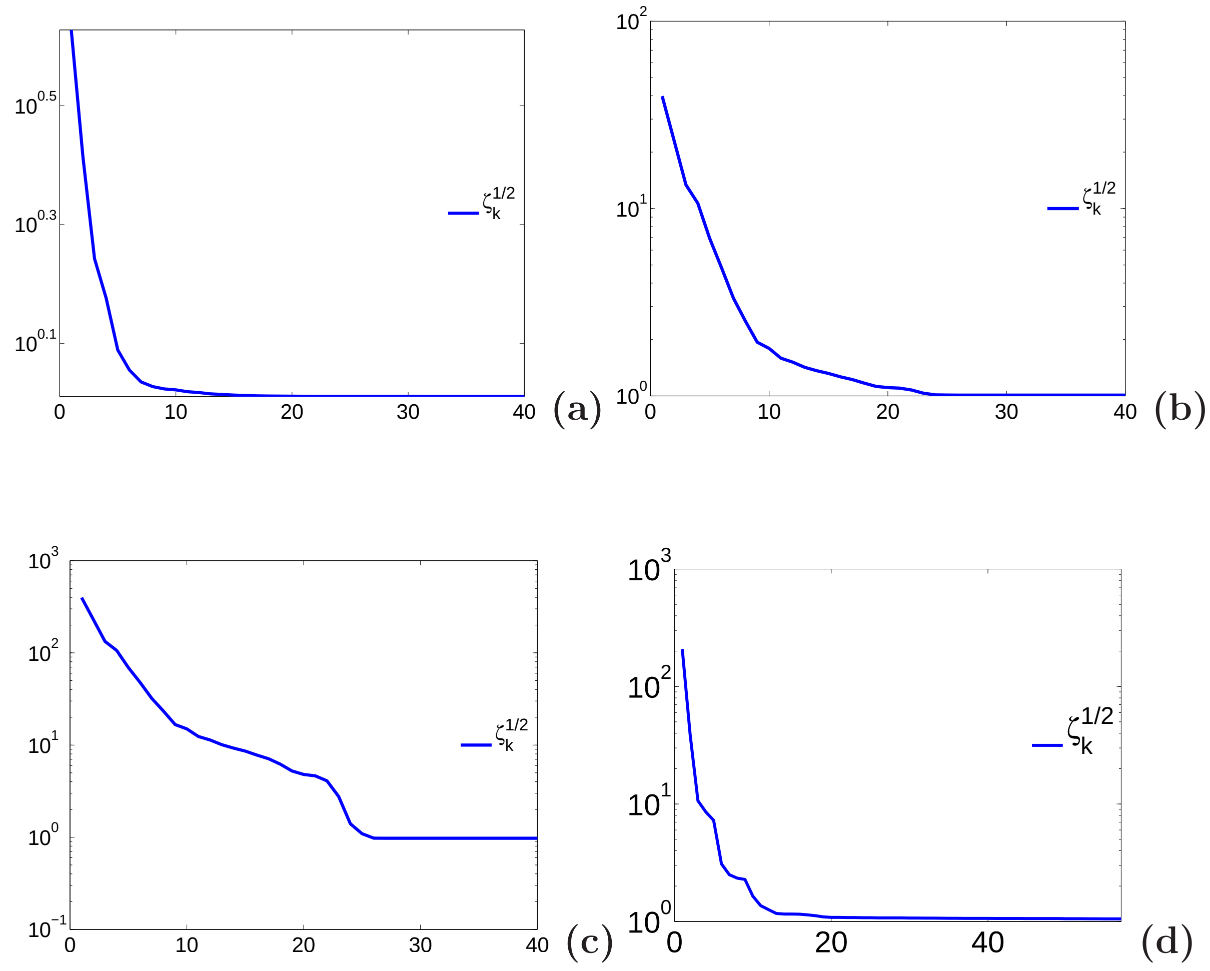

Figure 6.4: Representation of ζ_k (d=5): (a) dense problem $\kappa(A) = 10^2$ (b) dense problem $\kappa(A) = 10^3$, (c) dense problem $\kappa(A) = 10^4$, (d) data assimilation problem.

The slightly more conservative choice μ_1 (4.28) is the most reliable one. It will normally stop the conjugate gradient method after a number of iterations greater than the other choices μ_2 (4.24) and μ_3 (4.25). However, the numerical experiments show that the final residual has a distribution which fits the original Gaussian distribution more accurately. In particular, values of η close to 10^{-2} stop the PCGLS2 algorithm with a satisfactory residual r_k when μ_1 (4.28) is used and this is not always the case for the choices μ_2 and μ_3.

We want to stress the importance of using a robust preconditioner. The delay index d and the robustness of the stopping criteria depend on it. Our choice of the symmetric Gauss-Seidel preconditioner is not optimal, however, it is sufficient to allow for a value of $d = 5$ in our problems, and the residual histograms computed by using the stopping criterion μ_1 show a good fit with a Gaussian distribution.

Finally, our general conclusion is that for any iterative method, information on the accuracy of the data has to be taken into account to design best possible stopping criteria

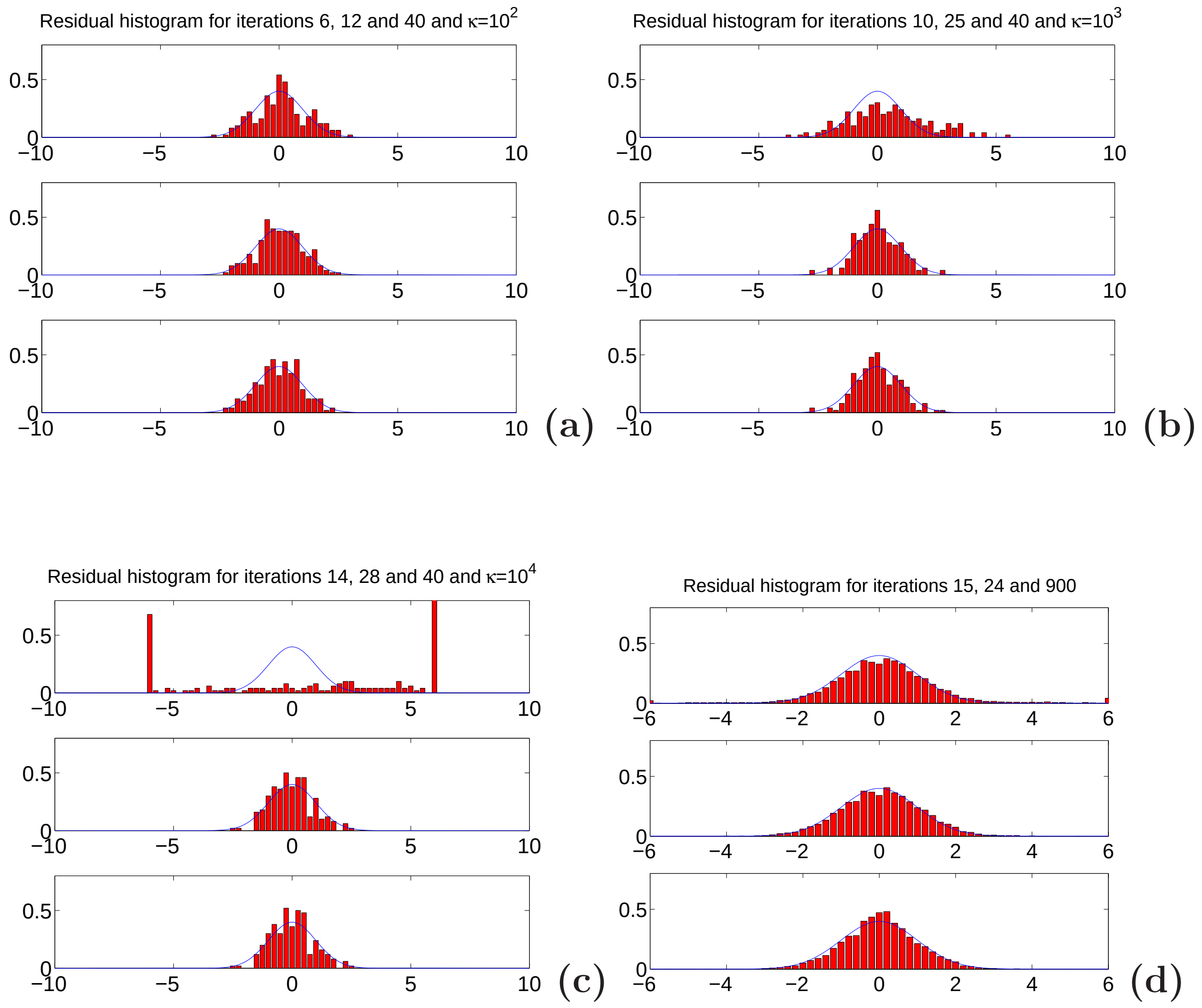

Figure 6.5: Residual histograms (d=5): (a) dense problem $\kappa(A) = 10^2$ (b) dense problem $\kappa(A) = 10^3$, (c) dense problem $\kappa(A) = 10^4$, (d) data assimilation problem.

that avoid both over- and under-solving, that respectively result in a too expensive or too inaccurate solution method. Owing to the variational properties of the conjugate gradients method for least-squares problems, and to the Gaussian nature of the residual at convergence, we propose stopping rules that rely on statistical tests involving the energy norm of the error. Further work could consist in comparing this approach with other tests of the Gaussianity of the residual, such as the periodogram technique (Rust 2000, Rust and O'Leary 2008).

References

Abramowitz, M. and Stegun, I. A. (1964), *Handbook of Mathematical Functions*, 55, National Bureau of Standards, Dover Publications.

Arioli, M. (2005), ‘A stopping criterion for the conjugate gradient algorithm in a finite element method framework’, *Numer. Math.* **97**, 1–24. Electronic version: DOI: 10.1007/s00211-003-0500-y.

Ashby, S. F., Holst, M. J., Manteuffel, T. A. and Saylor, P. E. (2001), ‘The Role of the Inner Product in Stopping Criteria for Conjugate Gradient Iterations’, *BIT* **41**(1), 26–52.

Axelsson, O. and Kaporin, I. (2001), ‘Error norm estimation and stopping criteria in preconditioned conjugate gradient iterations’, *Journal of Numerical Linear Algebra with Applications* **8**, 265–286.

Björck, A., Elfving, T. and Strakoš, Z. (1998), ‘Stability of conjugate gradient and Lanczos methods for linear least-squares problems’, *SIAM Journal on Matrix Analysis and Applications* **19**, 720–736.

Calvetti, D., Morigi, S., Reichel, L. and Sgallari, F. (2000), ‘Computable error bounds and estimates for the conjugate gradient method’, *Numerical Algorithms* **25**, 75–88.

Calvetti, D., Morigi, S., Reichel, L. and Sgallari, F. (2001), ‘An iterativa method with error estimators’, *J. Comp. Appl. Math.* **127**, 93–119.

Golub, G. and Meurant, G. (1997), ‘Matrices, moments and quadrature II; how to compute the norm of the error in iterative methods’, *BIT* **37**, 687–705.

Golub, G. and Strakos, Z. (1994), ‘Estimates in quadratic formulas’, *Numerical Algorithms* **8**, 241–268.

Greenbaum, A. (1997), *Iterative Methods for Solving Linear Systems*, Society for Industrial and Applied Mathematics, Philadelphia, PA, USA.

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

Hocking, R. R. (1996), *Methods and Applications of Linear models*, Wiley Series in Probability and Statistics, Wiley-Interscience, New York.

Magnus, J. R. and Neudecker, H. (1999), *Matrix Differential Calculus with Applications in Statistic and Econometrics*, Wiley Series in Probability and Statistics, revised edition edn, John Wiley & sons, Chichester, UK.

Meurant, G. (1997), ‘The computation of bounds for the norm of the error in the conjugate gradient algorithm’, *Numerical Algorithms* **16**, 77–87.

Meurant, G. (1999*a*), *Computer Solution of Large Linear Systems*, Vol. 28 of *Studies in Mathematics and its Application*, Elsevier/North-Holland, Amsterdam, The Netherlands.

Meurant, G. (1999*b*), 'Numerical experiments in computing bounds for the norm of the error in the preconditioned conjugate gradient algorithm', *Numerical Algorithms* **22**, 353–365.

Paige, C. C. and Saunders, M. (1982), 'LSQR: An algorithm for sparse linear equations and sparse least squares', *ACM Transactions on Mathematical Software* **8**, 43–71.

Rust, B. W. (2000), 'Parameter selection for constrained solutions to ill-posed problems', *SIAM Journal on Comput. Sci. Stat* **32**, 333–347.

Rust, B. W. and O'Leary, D. (2008), 'Residual periodograms for choosing regularization parameters for ill-posed problems', *Inverse Problems*.

Strakoš, Z. and Tichy, P. (2002), 'On error estimation in the conjugate gradient method and why it works in finite precision computations', *Electronic Transactions on Numerical Analysis (ETNA)* **13**, 56–80.

Strakos, Z. and Tichý, P. (2005), 'Error estimation in preconditioned conjugate gradients', *BIT Numerical Mathematics* **45**, 789–817.

Tshimanga, J., Gratton, S., Weaver, A. and Sartenaer, A. (2008), 'Limited-memory preconditioners with application to incremental four-dimensional variational data assimilation', *Q. J. Roy. Meterol. Soc.* **134**, 753–771.