Hybrid Methods for Nonlinear Least Squares Problems

Lukšan, Ladislav
2019

Institute of Computer Science
Academy of Sciences of the Czech Republic

Hybrid methods for nonlinear least squares problems

L.Lukšan, C.Matonoha, J.Vlček

Technical report No. 1246

May 2019

Pod Vodárenskou věží 2, 182 07 Prague 8 phone: +420 2 688 42 44, fax: +420 2 858 57 89,
e-mail:e-mail:ics@cs.cas.cz

Institute of Computer Science
Academy of Sciences of the Czech Republic

Hybrid methods for nonlinear least squares problems

L.Lukšan, C.Matonoha, J.Vlček

Technical report No. 1246

May 2019

Abstract:

This contribution contains a description and analysis of effective methods for minimization of the nonlinear least squares function $F(x) = (1/2)f^T(x)f(x)$, where $x \in R^n$ and $f \in R^m$, together with extensive computational tests and comparisons of the introduced methods. All hybrid methods are described in detail and their global convergence is proved in a unified way. Some proofs concerning trust region methods, which are difficult to find in the literature, are also added. In particular, the report contains an analysis of a new simple hybrid method with Jacobian corrections (Section 8) and an investigation of the simple hybrid method for sparse least squares problems proposed previously in [33] (Section 14).

Keywords:
Numerical optimization, nonlinear least squares, trust region methods, hybrid methods, sparse problems, partially separable problems, numerical experiments.

Content

1 Introduction

Consider an objective function of the form

$$F(x) = \frac{1}{2} f^T(x) f(x) = \frac{1}{2} \sum_{k=1}^{m} f_k^2(x), \tag{1}$$

where $f : \mathcal{D}_F \to R^m$ is a mapping defined on the set $\mathcal{D}_F \subset R^n$ (the mapping f is defined on the same region as the function F). We use the notation

$$f(x) = \begin{bmatrix} f_1(x) \\ \vdots \\ f_m(x) \end{bmatrix}, \qquad J(x) = \frac{\partial f(x)}{\partial x} = \begin{bmatrix} \frac{\partial f_1(x)}{\partial x_1}, & \ldots, & \frac{\partial f_1(x)}{\partial x_n} \\ \vdots & \vdots & \vdots \\ \frac{\partial f_m(x)}{\partial x_1}, & \ldots, & \frac{\partial f_m(x)}{\partial x_n} \end{bmatrix}$$

for the mapping f and its Jacobian matrix J. If the mapping f is continuously differentiable on an open set $\mathcal{D} \subset \mathcal{D}_F$, then

$$g(x) = J^T(x) f(x) = \sum_{k=1}^{m} f_k(x) h_k(x) \tag{2}$$

for $x \in \mathcal{D}$, where $h_k(x) = \nabla f_k(x)$ are gradients of the functions f_k, $1 \le k \le m$, and $g(x) = \nabla F(x)$ is the gradient of the function F (all computed at the point x). If the mapping f is twice continuously differentiable on $\mathcal{D}$, then

$$G(x) = J^T(x) J(x) + C(x) = \sum_{k=1}^{m} h_k(x) h_k^T(x) + \sum_{k=1}^{m} f_k(x) H_k(x), \tag{3}$$

for $x \in \mathcal{D}$, where $H_k(x) = \nabla^2 f_k(x)$ are Hessian matrices of the functions f_k, $1 \le k \le m$, and $G(x)$ is a Hessian matrix of the function F (all computed at the point x). Almost all investigated methods generate a sequence of points x_i, $i \in N$, such that $F_{i+1} \le F_i$, $i \in N$ (the sequence F_i, $i \in N$, is nonincreasing). Then $x_i \in \mathcal{D}(\overline{F}) = \{x \in R^n : F(x) \le \overline{F}\}$, where $\overline{F} \ge F(x_1)$. Investigating global convergence, we choose a set $\mathcal{D}$ in such a way that $\mathcal{D}(\overline{F}) \subset \mathcal{D}$ and assume that the set $\overline{\mathcal{D}}$ (closure) is compact, which is satisfied, for example, if the function F is coercive ($F(x) \to \infty$, if $\|x\| \to \infty$). Investigating the asymptotic rate of convergence, we assume that $\mathcal{D} = \mathcal{U}(x_*, \varepsilon) = \{x \in R^n : \|x - x_*\| < \varepsilon\}$, where x_* is the limit point of a sequence x_i, $i \in N$, and $\varepsilon > 0$.

Let $\mathcal{D}$ be the open set used in the above considerations. In the investigation of numerical methods for minimization of sums of squares, we use the following assumptions:

Assumption A1 The set $\overline{\mathcal{D}}$ is compact and the functions $f_k(x)$, $1 \le k \le m$, are twice continuously differentiable on $\overline{\mathcal{D}}$ (then there are constants $\overline{f}$, $\overline{h}$, $\overline{H}$ such that $f_k(x) \le \overline{f}$, $\|h_k(x)\| \le \overline{h}$, $\|H_k(x)\| \le \overline{H}$, $1 \le k \le m$, if $x \in \overline{\mathcal{D}}$).

Remark 1. Note that Assumption A1 implies inequalities $F(x) \le (m/2)\overline{f}^2$, $\|J(x)\| \le \overline{J} = \sqrt{m}\,\overline{h}$, $\|g(x)\| \le \overline{g} = m\overline{f}\,\overline{h}$ and $\|G(x)\| \le \overline{G} = m(\overline{h}^2 + \overline{f}\,\overline{H})$, if $x \in \overline{\mathcal{D}}$.

Assumption A2 The Jacobian matrix $J(x)$ has full column rank on $\overline{\mathcal{D}}$, i.e., there is a constant $\underline{J}$ such that $\|J(x)d\| \ge \underline{J}\|d\|$, if $x \in \overline{\mathcal{D}}$ and $d \in R^n$.

Assumption A3 The point $x_* \in R^n$ is a local minimizer of the function $F(x)$. This function is twice continuously differentiable at a neighborhood of x_* and the Hessian matrix $G(x_*)$ is positive definite.

Remark 2. Assumption A3 implies the existence of numbers $0 < \underline{G} \le \overline{G}$ and $\varepsilon > 0$ such that

$$\underline{G}\|d\|^2 \le d^T G(x) d \le \overline{G}\|d\|^2 \tag{4}$$

for an arbitrary nonzero vector $d \in R^n$ and $x \in \mathcal{U}(x_*, \varepsilon) = \{x \in R^n : \|x - x_*\| < \varepsilon\}$. If Assumption A3 is satisfied, then

$$\frac{1}{2}\underline{G}\|x - x_*\|^2 \leq F(x) - F(x_*) \leq \frac{1}{2}\overline{G}\|x - x_*\|^2, \tag{5}$$

$$\underline{G}\|x - x_*\| \leq \|g(x)\| \leq \overline{G}\|x - x_*\| \tag{6}$$

for all $x \in \mathcal{U}(x_*, \varepsilon)$ (it follows from the mean value theorem).

Assumption A4 The Hessian matrix $G(x)$ is Lipschitz continuous at a neighborhood of $x_* \in R^n$, i.e., there are numbers $\overline{L} > 0$ and $\varepsilon > 0$ such that

$$\|G(x_2) - G(x_1)\| \leq \overline{L}\|x_2 - x_1\|, \tag{7}$$

if $\|x_1 - x_*\| < \varepsilon$ and $\|x_2 - x_*\| < \varepsilon$.

2 Trust region methods

There are two basic classes of methods for unconstrained minimization: line-search methods and trust region methods [42], [36]. Numerical experiments show that trust region methods are more efficient for minimizing sum of squares than line-search methods. Therefore, we focus our attention on trust region methods. These methods are based on the idea introduced in [44]. In the description of trust region methods we utilize the knowledge of gradients $g_i = \nabla F(x_i)$, $i \in N$, and denote

$$Q_i(d) = \frac{1}{2}d^T B_i d + g_i^T d \tag{8}$$

(where $B_i \approx G(x_i)$) for the predicted decrease and

$$\rho_i(d) = \frac{F(x_i + d) - F(x_i)}{Q_i(d)} \tag{9}$$

for the ratio of both the actual and the predicted decreases of the objective function. Furthermore, we use the quantities

$$\vartheta_i = \frac{(B_i - G(x_i))d_i}{\|d_i\|}, \qquad \omega_i = \frac{B_i d_i + g_i}{\|g_i\|}, \qquad \tau_i = \frac{g_{i+1} - g_i - B_i d_i}{\|g_i\|}. \tag{10}$$

A detailed description of trust region methods is introduced in [12].

Definition 1. *We say that an iterative method $x_{i+1} = x_i + \alpha_i d_i$, $i \in N$, for unconstrained minimization of function $F : R^n \to R$, is a trust region method, if $0 < \Delta_1 \leq \overline{\Delta}$ and the following conditions hold.*

(T1) The direction vectors $d_i \in R^n$, $i \in N$, are determined in such a way that

$$\|d_i\| \leq \Delta_i, \tag{11}$$

$$\|d_i\| < \Delta_i \quad \Rightarrow \quad \omega_i \leq \overline{\omega}, \tag{12}$$

$$-Q_i(d_i) \geq \frac{\sigma}{2}\|g_i\| \min\left(\Delta_i, \frac{\|g_i\|}{\|B_i\|}\right) \tag{13}$$

where $0 \leq \overline{\omega} < 1$ and $0 < \underline{\sigma} \leq 1$.

(T2) The step-sizes $\alpha_i \geq 0$, $i \in N$, are selected in such a way that

$$\rho_i(d_i) < \underline{\rho} \quad \Rightarrow \quad \alpha_i = 0, \tag{14}$$

$$\rho_i(d_i) \geq \underline{\rho} \quad \Rightarrow \quad \alpha_i = 1, \tag{15}$$

$0 < \underline{\rho} < 1$.

(T3) The trust region radii $0 < \Delta_i \le \overline{\Delta}$, $i \in N$, *are chosen by the rule*

$$\rho_i(d_i) < \underline{\rho} \quad \Rightarrow \quad \underline{\beta}\|d_i\| \le \Delta_{i+1} \le \overline{\beta}\|d_i\|, \tag{16}$$

$$\underline{\rho} \le \rho_i(d_i) \le \overline{\rho} \quad \Rightarrow \quad \Delta_{i+1} = \min(\Delta_i, \overline{\gamma}\|d_i\|), \tag{17}$$

$$\rho_i(d_i) > \overline{\rho} \quad \Rightarrow \quad \Delta_{i+1} = \min(\underline{\gamma}\Delta_i, \overline{\gamma}\|d_i\|, \overline{\Delta}), \tag{18}$$

where $0 < \underline{\beta} \le \overline{\beta} < 1 < \underline{\gamma} < \overline{\gamma}$ *and* $0 < \underline{\rho} < \overline{\rho} < 1$.

Remark 3. Definition 1 is slightly complicated. The condition $\Delta_{i+1} \le \overline{\gamma}\|d_i\|$ is usually omitted, since Theorems 1 and 4 hold without this assumption. This condition is necessary for obtaining inequality (36) and, therefore, for correctness of Theorems 2 and 3. The rule (T2) is frequently replaced by a similar rule with $\underline{\rho} = 0$ [47], for which Theorems 1 and 4 hold again. Besides the assertion $\lim_{i\to\infty}\|g(x_i)\| = 0$ in case $\|B_i\| \le \overline{B}$, $i \in N$, (T2) is necessary for obtaining inequality (38) and, therefore, for correctness of Theorems 2 and 3.

A direction vector $d_i \in R^n$ satisfying conditions (11)–(13) can be computed in various ways. We have advantageously used the dog-leg method, introduced in [44] and improved in [17]. This method uses the formulas

$$d_i = -\frac{\Delta_i}{\|g_i\|}, \qquad \|d_i^C\| \ge \Delta_i, \tag{19}$$

$$d_i = d_i^C + \lambda_i(d_i^N - d_i^C), \qquad \|d_i^C\| < \Delta_i < \|d_i^N\|, \tag{20}$$

$$d_i = d_i^N, \qquad \|d_i^N\| \le \Delta_i, \tag{21}$$

where

$$d_i^C = -\frac{\|g_i\|^2}{g_i^T B_i g_i} g_i, \qquad d_i^N = -B_i^{-1} g_i \tag{22}$$

and λ_i is a number selected in such a way that $\|d_i\| = \Delta_i$. It is known (see [12]) that direction vector d_i computed by (19)–(22) satisfies conditions (11)–(13) with $\overline{\omega} = 0$ and $\underline{\sigma} = 1$. Moreover, this vector satisfies the additional condition

$$-d_i^T g_i \ge \underline{\sigma}\|g_i\| \min\left(\Delta_i, \frac{\|g_i\|}{\|B_i\|}\right) \tag{23}$$

with $\underline{\sigma} = 1$, so it is a descent direction vector. Note that the optimum step method introduced in [41] and the iterative method introduced in [52] and [56] are not suitable for solving least squares problems (which was confirmed by our numerical experiments).

The following four theorems are essential for investigation of convergence properties concerning trust region methods. The first part of Theorem 1 is proved in [47] and the second part in [53]. The proofs of the remaining theorems are based on ideas presented in [18] and [46]. We use the notation

$$N_1 = \{i \in N : \|d_i\| < \Delta_i\}, \qquad N_2 = \{i \in N : \rho_i \ge \underline{\rho}\}$$

in the subsequent considerations.

Theorem 1. *(Global convergence) Let the mapping* $f : R^n \to R^m$ *satisfy Assumption A1 and* $x_i \in R^n$, $i \in N$, *be a sequence generated by the trust region method (T1)–(T3) such that*

$$\sum_{i=1}^{\infty} \frac{1}{M_i} = \infty, \tag{24}$$

where

$$M_i = \max_{1\le j\le i} \|B_j\|. \tag{25}$$

Then $\liminf_{i\to\infty}\|g(x_i)\| = 0$. *If* $\|B_i\| \le \overline{B}$, $i \in N$, *then* $\lim_{i\to\infty}\|g(x_i)\| = 0$.

Remark 4. The proof of the first part of Theorem 1, given in [47], is based on the inequality

$$\|d_i\| \geq \underline{c}\,\frac{\|g_k\|}{M_k}, \qquad \underline{c} = \min\left(1-\overline{\omega},\ \frac{\underline{\sigma}(1-\underline{\rho})\|B_1\|}{2(\overline{G}+\|B_1\|)}, \frac{\|d_1\|\|B_1\|}{\|g_1\|}\right) < 1, \tag{26}$$

where $k = i$, if $i \in N_1$ or $i \notin N_2$ or $i = 1$, and $k < i$, if $i \notin N_1$ and $i \in N_2$ and $i \neq 1$.

Theorem 2. *(Linear convergence) Let x_i, $i \in N$, be a sequence generated by the trust region method (T1)–(T3) such that $\|B_i\| \leq \overline{B}$, $i \in N$. Let $x_i \to x_*$, where the point $x_* \in R^n$ satisfies Assumption A3. Then*

$$\sum_{i=1}^{\infty} \|e_i\| \triangleq \sum_{i=1}^{\infty} \|x_i - x_*\| < \infty. \tag{27}$$

Proof (a) We first prove that the sequence x_i, $i \in N_2$, converges linearly. Let $i \in N_2$. Remark 4 implies that there exists an index $k \leq i$ such that $\|d_i\| \geq \underline{c}\,\|g_k\|/M_k$, where the number $0 < \underline{c} < 1$ is given by (26). Since the sequence $F(x_i)$, $i \in N$, is non-increasing, we can write

$$1 \geq \frac{F(x_i)-F(x_*)}{F(x_k)-F(x_*)} \geq \frac{\underline{G}^2}{2}\frac{\|x_i-x_*\|^2}{\|g_k\|^2} \geq \frac{\underline{G}^2}{2\overline{G}^2}\frac{\|g_i\|^2}{\|g_k\|^2}$$

by (5)–(6), so

$$\|d_i\| \geq \underline{c}\,\frac{\|g_k\|}{M_k} \geq \frac{\underline{c}\,\underline{G}}{\sqrt{2}\,\overline{G}\,M_k}\|g_i\| \geq \frac{\underline{c}\,\underline{G}}{\sqrt{2}\,\overline{G}\,M_i}\|g_i\| \tag{28}$$

by (26) and (25). Since $i \in N_2$, one has $\rho_i(d_i) \geq \underline{\rho}$, which together with (9), (13), (11), (28), (26) and (5)–(6) gives

$$\begin{aligned} F_i - F_{i+1} &\geq \frac{\underline{\rho}\,\underline{\sigma}}{2}\|g_i\|^2 \min\left(\frac{\underline{c}\,\underline{G}}{\sqrt{2}\,\overline{G}\,M_i}, \frac{1}{M_i}\right) = \frac{\underline{\rho}\,\underline{\sigma}\,\underline{c}\,\underline{G}}{2\sqrt{2}\,\overline{G}\,M_i}\|g_i\|^2 \\ &\geq \frac{\underline{\rho}\,\underline{\sigma}\,\underline{c}\,\underline{G}^3}{\sqrt{2}\,\overline{G}^2\,M_i}(F_i - F_*) \triangleq \frac{c}{M_i}(F_i - F_*), \end{aligned} \tag{29}$$

where

$$0 < \frac{c}{M_i} = \frac{\underline{\rho}\,\underline{\sigma}\,\underline{c}\,\underline{G}^3}{\sqrt{2}\,\overline{G}^2\,M_i} < \frac{\underline{\rho}\,\underline{\sigma}\,\underline{G}^3}{2\sqrt{2}\,\overline{G}^3} < 1,$$

since (26) and (25) imply

$$\underline{c} \leq \frac{\underline{\sigma}(1-\underline{\rho})\|B_1\|}{2(\overline{G}+\|B_1\|)} \leq \frac{\underline{\sigma}(1-\underline{\rho})M_i}{2\overline{G}} < \frac{M_i}{2\overline{G}}.$$

Let $N_2 = \{k_1, k_2, k_3, \dots\}$ (we assume, without loss of generality, that $k_1 = 1$). Using (29), we obtain

$$F_{k_{j+1}} - F_* \leq F_{k_j+1} - F_* \leq \left(1 - \frac{c}{M_{k_j}}\right)(F_{k_j} - F_*), \qquad j \in N, \tag{30}$$

since $F_{k_j+1} - F_* = F_{k_j} - F_* + (F_{k_j+1} - F_{k_j})$, and using the inequality $M_{k_j} \leq \overline{B}$, we can write

$$\sqrt{F_{k_{j+1}} - F_*} \leq q\sqrt{(F_{k_j} - F_*)}, \qquad q = \sqrt{1 - \frac{c}{\overline{B}}} < 1$$

for all $j \in N$, which together with (5) implies

$$\sum_{i \in N_2} \|x_i - x_*\| = \sum_{j=1}^{\infty} \|x_{k_j} - x_*\| \leq \sqrt{\frac{2}{\underline{G}}}\sum_{j=1}^{\infty}\sqrt{F_{k_j} - F_*} < \infty. \tag{31}$$

(b) We show that if

$$\|d_i\| \le \frac{\underline{\sigma}(1-\underline{\rho})}{\overline{G}+\overline{B}}\|g_i\|, \tag{32}$$

then $i \in N_2$. Using (9), the inequality $\rho_i(d_i) \ge \underline{\rho}$ can be expressed in the form

$$F_{i+1} - F_i - Q_i(d_i) \le (\underline{\rho}-1)Q_i(d_i) \tag{33}$$

(since $Q_i(d_i) \le 0$). Using the mean value theorem and (8), we obtain

$$F_{i+1} - F_i - Q_i(d_i) \le d_i^T g_i + \frac{1}{2}\overline{G}\|d_i\|^2 - d_i^T g_i + \frac{1}{2}\overline{B}\|d_i\|^2 = \frac{1}{2}(\overline{G}+\overline{B})\|d_i\|^2, \tag{34}$$

so using (13) and (32), we can write

$$(\underline{\rho}-1)Q_i(d_i) \ge \frac{\underline{\sigma}(1-\underline{\rho})}{2}\|g_i\| \min\left(\|d_i\|, \frac{\|g_i\|}{\overline{B}}\right) = \frac{\underline{\sigma}(1-\underline{\rho})}{2}\|g_i\|\|d_i\|. \tag{35}$$

If inequality (32) is satisfied, then (34) and (35) imply

$$F_{i+1} - F_i - Q_i(d_i) \le \frac{1}{2}(\overline{G}+\overline{B})\|d_i\|^2 \le \frac{\underline{\sigma}(1-\underline{\rho})}{2}\|g_i\|\|d_i\| \le (\underline{\rho}-1)Q_i(d_i),$$

so (33) holds.

(c) Let $k_j < i < k_{j+1}$. Then

$$\|d_i\| \le \Delta_i \le \overline{\beta}\|d_{i-1}\| \le \cdots \le \overline{\beta}^{i-k_j-1}\|d_{k_j+1}\| \le \overline{\beta}^{i-k_j-1}\Delta_{k_j+1} \le \frac{\overline{\gamma}}{\underline{\beta}}\overline{\beta}^{i-k_j}\|d_{k_j}\| \tag{36}$$

by (11) and (16)–(18). Since $k_j \in N_2$ and, therefore, $F_{k_j+1} = F(x_{k_j}+d_{k_j}) \le F_{k_j}$ by (15), Assumption A3 and (8) imply that

$$0 \ge F_{k_j+1} - F_{k_j} \ge d_{k_j}^T g_{k_j} + \frac{1}{2}\underline{G}\|d_{k_j}\|^2 \ge -\|d_{k_j}\|\|g_{k_j}\| + \frac{1}{2}\underline{G}\|d_{k_j}\|^2,$$

or

$$\|d_{k_j}\| \le \frac{2}{\underline{G}}\|g_{k_j}\|. \tag{37}$$

Using (36) and (37) one can write

$$\|d_i\| \le \frac{\overline{\gamma}}{\underline{\beta}}\overline{\beta}^{i-k_j}\|d_{k_j}\| \le \frac{2\overline{\gamma}}{\underline{G}\underline{\beta}}\overline{\beta}^{i-k_j}\|g_{k_j}\| = \frac{2\overline{\gamma}}{\underline{G}\underline{\beta}}\overline{\beta}^{i-k_j}\|g_i\| \tag{38}$$

since $x_i = x_{k_j}$ and, therefore, $\|x_i - x_*\| = \|x_{k_j} - x_*\|$ and $\|g_i\| = \|g_{k_j}\|$ by (15). Since $\overline{\beta} < 1$, there exists a minimum integer m such that

$$\frac{2\overline{\gamma}}{\underline{G}\underline{\beta}}\overline{\beta}^m \le \frac{(1-\underline{\rho})\underline{\sigma}}{\overline{G}+\overline{B}}. \tag{39}$$

Since $i \notin N_2$, then (32) cannot hold by (b) and, therefore, $i - k_j < m$ by (38) and (39) (if $i - k_j \ge m$ were satisfied, then (38) and (39) would imply (32)), so

$$\sum_{i=1}^{\infty}\|x_i - x_*\| < m\sum_{i\in N_2}\|x_i - x_*\| < \infty$$

by (31). □

Theorem 3. *(Local convergence) Let the point $x_* \in R^n$ satisfy Assumption A3. Consider the trust region method (T1)–(T3) such that $\|B_i\| \le C_i$, where $C_1 = \|B_1\|$ and $C_{i+1} \le C_i(1+O(\|e_i\|))$. Then there exists a number $\delta > 0$ such that $\|x_1 - x_*\| < \delta$ implies $\|B_i\| \le \overline{B} = 2\|B_1\|$, $i \in N$, $x_i \to x_*$ and $\sum_{i=1}^{\infty} \|x_i - x_*\| < \infty$.*

Proof By Assumption A3, there exists a number $\varepsilon > 0$, such that (5) and (6) hold whenever $x \in \mathcal{U}(x_*, \varepsilon)$.

(a) Assume that $x_i \in \mathcal{U}(x_*, \varepsilon)$, $i \in N$. Relation $C_{i+1} \le C_i(1 + O(\|e_i\|))$ and inequality (5) imply existence of a constant $C > 0$ such that

$$C_{i+1} \le C_i(1 + C\|e_i\|) \le C_i \left(1 + C\sqrt{\frac{2}{\underline{G}}}\sqrt{F_i - F_*}\right), \qquad i \in N,$$

and since

$$1 + C\sqrt{\frac{2}{\underline{G}}}\sqrt{F_i - F_*} \le \exp\left(C\sqrt{\frac{2}{\underline{G}}}\sqrt{F_i - F_*}\right)$$

we can write

$$C_{l+1} \le C_1 \prod_{i=1}^{l} \left(1 + C\sqrt{\frac{2}{\underline{G}}}\sqrt{F_i - F_*}\right) \le C_1 \exp\left(C\sqrt{\frac{2}{\underline{G}}} \sum_{i=1}^{l} \sqrt{F_i - F_*}\right) \tag{40}$$

for an arbitrary index $l \in N$. Let $N_2 = \{k_1, k_2, k_3, \dots\}$ (we assume, without loss of generality, that $k_1 = 1$) and $k_i \le l < k_{i+1}$. Since $F_k = F_{k_j}$ if $k_j \le k < k_{j+1}$, (40) can be written in the form

$$C_{l+1} \le C_1 \exp\left(C\sqrt{\frac{2}{\underline{G}}} \sum_{j=1}^{i} m_j \sqrt{F_{k_j} - F_*}\right) \le C_1 \exp\left(mC\sqrt{\frac{2}{\underline{G}}} \sum_{j=1}^{i} \sqrt{F_{k_j} - F_*}\right), \tag{41}$$

where $m_j = k_{j+1} - k_j$ for $1 \le j \le i$, since (as in the proof of Theorem 2) $m_j \le m$ for all $1 \le j \le i$, where m is the minimum integer such that

$$\frac{2\overline{\gamma}}{\underline{G}\overline{\beta}}\overline{\beta}^m \le \frac{(1-\underline{\rho})\underline{\sigma}}{\overline{G} + 2C_1}$$

(since either $k_{j+1} - 1 = k_j$, so $m_j = 1$, or $k_{j+1} - 1 \notin N_2$, so $m_j - 1 < m$).

(b) We show by induction that if

$$\sqrt{F_1 - F_*} \le \sqrt{\frac{\underline{G}}{2}} \min\left(\frac{c}{4\,m\,C\,C_1}, \varepsilon\right), \tag{42}$$

where $c > 0$ is the number defined in (29), then $C_i \le 2C_1$ for all $i \in N$. Since the sequence $F_i - F_*$, $i \in N$, is non-increasing, the inequality $\sqrt{F_i - F_*} \le \sqrt{\underline{G}/2}\,\varepsilon$ holds for all $i \in N$ by (42), so $x_i \in \mathcal{U}(x_*, \varepsilon)$ for all $i \in N$ by (5). Assume that $C_k \le 2C_1$ for $1 \le k \le l$, where $k_i \le l < k_{i+1}$ (it trivially holds for $l = k_1 = 1$). Using (30), we can write

$$\sqrt{F_{k_{j+1}} - F_*} \le \sqrt{1 - \frac{c}{M_{k_j}}}\sqrt{F_{k_j} - F_*} \le \sqrt{1 - \frac{c}{2C_1}}\sqrt{F_{k_j} - F_*} \le \left(1 - \frac{c}{4C_1}\right)\sqrt{F_{k_j} - F_*}$$

for $1 \le j \le i$, since $\sqrt{1-a} \le 1 - a/2$ for an arbitrary number $0 \le a \le 1$. Therefore

$$\sum_{j=1}^{i} \sqrt{F_{k_j} - F_*} \le \sqrt{F_1 - F_*} \sum_{j=1}^{i} \left(1 - \frac{c}{4\,C_1}\right)^{j-1} \le \sqrt{F_1 - F_*} \sum_{j=1}^{\infty} \left(1 - \frac{c}{4\,C_1}\right)^{j-1} = \frac{4\,C_1}{c}\sqrt{F_1 - F_*}.$$

Substituting this expression into (41), we obtain

$$C_{l+1} \le C_1 \exp\left(mC\sqrt{\frac{2}{\underline{G}}} \sum_{j=1}^{i} \sqrt{F_{k_j} - F_*}\right) \le C_1 \exp\left(mC\sqrt{\frac{2}{\underline{G}}}\frac{4C_1}{c}\sqrt{F_1 - F_*}\right). \tag{43}$$

Since (42) implies

$$\frac{4\,m\,C\,C_1}{c}\sqrt{\frac{2}{\underline{G}}}\sqrt{F_1-F_*}\le\frac{1}{2}$$

and $\exp(1/2)<2$, we obtain $C_{l+1}\le 2C_1$ by (43). Thus the induction step is finished and the assertion, that the choice of x_1 satisfying (42) implies inequalities $C_i\le 2C_1$, $i\in N$, is proved.

(c) Using (5) and (42), we can see that $x_i\in\mathcal{U}(x_*,\varepsilon)$ (i.e., $\|x_i-x^*\|<\varepsilon$) and $\|B_i\|\le C_i\le 2C_1=2\|B_1\|$ hold for all $i\in N$ if $x_1\in\mathcal{U}(x_*,\delta)$, where

$$\delta=\sqrt{\frac{\underline{G}}{\overline{G}}}\min\left(\frac{c}{4\,m\,C\,C_1},\varepsilon\right). \tag{44}$$

Then $x_i\to x_*$ and $\sum_{i=1}^{\infty}\|e_i\|<\infty$ by Theorem 2. □

Theorem 4. *(Superlinear convergence) Let the mapping $f:R^n\to R^m$ satisfy Assumption A1 and x_i, $i\in N$, be a sequence generated by the trust region method (T1)–(T3). Let $x_i\to x_*$, where the point $x_*\in R^n$ satisfies Assumption A3. Let $\|B_i\|\le\overline{B}$ for all $i\in N$ and*

$$\lim_{i\to\infty}\omega_i=0,\qquad\lim_{i\to\infty}\vartheta_i=0. \tag{45}$$

Then the sequence x_i, $i\in N$, converges Q-superlinearly to the point $x_*\in R^n$.

Proof (a) By Assumption A3, there exist a number $\varepsilon>0$ and constants $0<\underline{G}\le\overline{G}$ such that

$$d^TG(x)d\ge\underline{G}\|d\|^2,\qquad d^TG(x)d\le\overline{G}\|d\|^2, \tag{46}$$

whenever $\|x-x_*\|<\varepsilon$ and $d\in R^n$. Using (10), we can write

$$B_id_i=G_id_i+\vartheta_i\|d_i\|,$$

so

$$\|B_id_i\|\le(\overline{G}+\|\vartheta_i\|)\|d_i\|,\qquad d_i^TB_id_i\ge(\underline{G}-\|\vartheta_i\|)\|d_i\|^2$$

for $\|x_i-x_*\|\le\varepsilon$. Since $x_i\to x_*$, $\|\vartheta_i\|\to 0$, $\|\omega_i\|\to 0$, there exists an index $k_1\in N$ such that $\|x_i-x_*\|<\varepsilon$,

$$\|\vartheta_i\|\le\underline{G}/2\le\overline{G}/2,\qquad\|\omega_i\|\le 1/2 \tag{47}$$

for $i\ge k_1$, so $\|B_id_i\|\le(3\overline{G}/2)\|d_i\|$ and $d_i^TB_id_i\ge(\underline{G}/2)\|d_i\|^2$ for all $i\ge k_1$. The last inequality together with (8) and (13) imply

$$0\ge Q_i(d_i)=g_i^Td_i+\frac{1}{2}d_i^TB_id_i\ge\frac{1}{4}\underline{G}\|d_i\|^2-\|g_i\|\|d_i\|,$$

which for $i\ge k_1$ gives

$$\|g_i\|\ge\frac{1}{4}\underline{G}\|d_i\|. \tag{48}$$

Using (13) and (48), we obtain

$$-Q_i(d_i)\ge\frac{\sigma}{2}\|g_i\|\min\left(\|d_i\|,\frac{\|g_i\|}{\overline{B}}\right)\ge\frac{\sigma\,\underline{G}}{8}\min\left(1,\frac{\underline{G}}{4\overline{B}}\right)\|d_i\|^2\triangleq\frac{1}{2}\underline{C}\|d_i\|^2 \tag{49}$$

for all $i\ge k_1$. At the same time, one can write

$$G_id_i=(B_id_i+g_i)-(B_i-G_i)d_i-g_i=\omega_i\|g_i\|-\vartheta_i\|d_i\|-g_i,$$

so

$$(\overline{G}+\|\vartheta_i\|)\|d_i\|\ge\|G_id_i+\vartheta_i\|d_i\|\|=\|\omega_i\|g_i\|-g_i\|\ge(1-\|\omega_i\|)\|g_i\|$$

which together with (47) gives

$$\|g_i\| \leq 3\overline{G}\|d_i\| \tag{50}$$

for all $i \geq k_1$.

(b) We show that there exists an index $k_2 \geq k_1$ such that $i \in N_2$ for all $i \geq k_2$. Using Taylor expansion we can write

$$F(x_i + d_i) - F(x_i) = d_i^T g_i + \frac{1}{2} d_i^T G_i d_i + o(\|d_i\|^2) = Q_i(d_i) + \frac{1}{2} d_i^T (G_i - B_i) d_i + o(\|d_i\|^2),$$

so

$$\rho_i(d_i) = \frac{F(x_i + d_i) - F(x_i)}{Q_i(d_i)} = 1 + \frac{d_i^T (G_i - B_i) d_i + o(\|d_i\|^2)}{2Q_i(d_i)}$$

and (49) together with (10) imply

$$\left| \frac{d_i^T (G_i - B_i) d_i + o(\|d_i\|^2)}{2Q_i(d_i)} \right| \leq \frac{1}{\underline{C}} \frac{\|\vartheta_i\| \|d_i\|^2 + o(\|d_i\|^2)}{\|d_i\|^2} \to 0,$$

since $\|\vartheta_i\| \to 0$. Therefore, $\rho_i(d_i) \to 1$ and since $\underline{\rho} < 1$, there exists an index $k_2 \geq k_1$ such that $\rho_i(d_i) \geq \underline{\rho}$ for all $i \geq k_2$.

(c) We show that there exists an index $k \geq k_2$ such that $i \in N_1$ for all $i \geq k$. Note first that the set N_1 is infinite. If this set were finite, an index $k \geq k_2$ would exist such that $\|d_i\| \geq \Delta_i \geq \Delta_k$ (since $\rho_i \geq \underline{\rho}$) for all $i \geq k \geq k_2$. This is a contradiction since $\|g_i\| \to 0$ implies $\|d_i\| \to 0$ by (48). Using Taylor expansion, we can write

$$g_{i+1} = g(x_i + d_i) = g_i + G_i d_i + o(\|d_i\|) \tag{51}$$

for $i \geq k_2$ (since $i \in N_2$ for all $i \geq k_2$). Using (51) together with (10) and (48), we obtain

$$\|\tau_i\| = \frac{\|g_{i+1} - g_i - B_i d_i\|}{\|g_i\|} = \frac{\|(G_i - B_i) d_i + o(\|d_i\|)\|}{\|g_i\|} \leq \frac{\|\vartheta_i\| \|d_i\| + o(\|d_i\|)}{\|g_i\|} \leq \frac{4}{\underline{G}} \|\vartheta_i\| + o(1),$$

so $\tau_i \to 0$, and since $\omega_i \to 0$, there exists an index $k_3 \geq k_2$ such that

$$\|\tau_i\| < \frac{\underline{G}}{24\overline{G}}, \qquad \|\omega_i\| < \frac{\underline{G}}{24\overline{G}} \tag{52}$$

for $i \geq k_3$. Since the set N_1 is infinite, there exists an index $k \geq k_3$ such that $\|d_k\| < \Delta_k$. Using (48), (10), (50) and (52), we can write

$$\begin{aligned} \|d_{k+1}\| &\leq \frac{4}{\underline{G}} \|g_{k+1}\| \leq \frac{4}{\underline{G}} (\|g_{k+1} - g_k - B_k d_k\| + \|B_k d_k + g_k\|) \\ &= \frac{4}{\underline{G}} (\|\tau_k\| + \|\omega_k\|) \|g_k\| \leq \frac{12\overline{G}}{\underline{G}} (\|\tau_k\| + \|\omega_k\|) \|d_k\| < \left(\frac{1}{2} + \frac{1}{2} \right) \|d_k\| = \|d_k\|. \end{aligned}$$

Since $\rho_k \geq \underline{\rho}$ for $k \geq k_3 \geq k_2$, one has $\Delta_{k+1} \geq \Delta_k$, which gives

$$\|d_{k+1}\| < \|d_k\| \leq \Delta_k \leq \Delta_{k+1}.$$

Continuing this process, we deduce that $\|d_i\| < \Delta_i$ (so $i \in N_1$) for all $i \geq k$.

(d) Using (10), we obtain

$$\frac{\|g_{i+1}\|}{\|g_i\|} \leq \frac{\|g_{i+1} - g_i - B_i d_i\| + \|B_i d_i + g_i\|}{\|g_i\|} \leq \|\tau_i\| + \|\omega_i\|,$$

which together with $\|\tau_i\| \to 0$, $\|\omega_i\| \to 0$ and (6) gives

$$\lim_{i \to \infty} \frac{\|x_{i+1} - x_*\|}{\|x_i - x_*\|} \leq \frac{\overline{G}}{\underline{G}} \lim_{i \to \infty} \frac{\|g_{i+1}\|}{\|g_i\|} = 0.$$

□

3 Variable metric methods

Variable metric methods, intended for general unconstrained optimization, are usually realized as line search methods. Since they will be used for construction of hybrid methods for nonlinear least squares, we focus our attention on variable metric trust region methods. Unfortunately, these methods have a disadvantage consisting in the fact that the forbidden inequality $y_i^T s_i \leq 0$, which violates positive definiteness of the generated matrix, can hold for some $i \in N$. We use the notation

$$N_3 = \{i \in N : y_i^T s_i > c\|y_i\|^2\} \tag{53}$$

with $c > 0$ in the subsequent considerations. Variable metric trust region methods from the Broyden class generate matrices $B_i \approx G(x_i)$, $i \in N$, such that B_1 is positive definite (usually $B_1 = I$, where I is the unit matrix of order n) and

$$\begin{array}{lll} B_{i+1} = \mathcal{B}(B_i, y_i, s_i, \beta_i, \gamma_i) & \text{if} & i \in N_2 \cap N_3, \\ B_{i+1} = B_i & \text{if} & i \notin N_2 \cap N_3, \end{array} \tag{54}$$

where

$$\mathcal{B}(B, y, s, \beta, \gamma) = \frac{1}{\gamma}\left(B + \gamma\frac{yy^T}{y^T s} - \frac{Bs(Bs)^T}{s^T Bs} + \frac{\beta}{s^T Bs}\left(\frac{s^T Bs}{y^T s}y - Bs\right)^T\left(\frac{s^T Bs}{y^T s}y - Bs\right)\right). \tag{55}$$

In this formula, $s_i = x_{i+1} - x_i$, $y_i = g_{i+1} - g_i$, $1 \leq \gamma_i \leq \overline{\gamma}$ (usually $\gamma_i = 1$) and $\beta_i > \beta_i^*$ (usually $\beta_i = 0$), where $\beta_i^* < 0$ is a critical value for which matrix (55) is singular [42], [36] (matrix B_{i+1} is positive definite if B_i is positive definite, $y_i^T s_i > 0$ and $\beta_i > \beta_i^*$). The standard update (54) is sometimes replaced by an alternative strategy

$$\begin{array}{lll} B_{i+1} = \mathcal{B}(B_i, y_i, s_i, \beta_i, \gamma_i) & \text{if} & i \in N_3, \\ B_{i+1} = B_i & \text{if} & i \notin N_3, \end{array} \tag{56}$$

where $s_i = d_i$ and $y_i = g(x_i + d_i) - g_i$ if $i \notin N_2$ (matrices B_{i+1}, $i \in N_3$, are computed even if $i \notin N_2$). Note that (56) can be formally obtained by writing N instead of N_2 in (54). To simplify particular formulas derived from (55), we use the notation

$$a_i = y_i^T B_i^{-1} y_i, \qquad b_i = y_i^T s_i, \qquad c_i = s_i^T B_i s_i \tag{57}$$

in the subsequent considerations (the value a_i can be easily computed if the Choleski decomposition $B_i = L_i D_i L_i^T$ is known).

Remark 5. Values β_i, $i \in N$, determine individual variable metric methods. The most popular choice $\beta_i = \beta_i^{BFGS} = 0$ corresponds to the BFGS (Broyden [6], Fletcher [22], Goldfarb [27], Shanno [49]) method. The choice $\beta_i = \beta_i^{DFP} = 1$ corresponds to the DFP (Davidon [14], Fletcher and Powell [23]) method. The choice $\beta_i = \beta_i^H = \gamma_i b_i/(\gamma_i b_i + c_i)$ corresponds to the H (Hoshino [30]) method. The choice $\beta_i = \beta_i^{R1} = \gamma_i b_i/(\gamma_i b_i - c_i)$ corresponds to the R1 (rank-one [7]) method. The choice $\beta_i = \beta_i^{DW} = b_i/a_i$ corresponds to the DW (Dennis and Wolkowicz [19]) method. Values γ_i, $i \in N$, serve for scaling, which usually improve the efficiency of variable metric methods. Let $0 < \underline{\gamma} < \overline{\gamma}$. An efficient scaling technique is to use values

$$\gamma_i = b_i/a_i \quad \text{or} \quad \gamma_i = c_i/b_i \quad \text{or} \quad \gamma_i = \sqrt{c_i/a_i} \tag{58}$$

if $\underline{\gamma} \leq \gamma_i \leq \overline{\gamma}$. If the inequality $\underline{\gamma} \leq \gamma_i \leq \overline{\gamma}$ does not hold, we set $\gamma_i = 1$. Note that the choice $\underline{\gamma} = 1$ is required in convergence proofs. More details are given in [36].

Remark 6. If variable metric methods are used as parts of hybrid methods for nonlinear least squares, the matrix B in (55) can be singular (positive semidefinite). If in this case $Bs = 0$, we omit the last two terms in (55) (so we assume that $\beta = 0$). After this arrangement, divisions by zero cannot occur and variable metric methods (54) or (56) generate positive semidefinite matrices.

First we prove a global convergence theorem for variable metric trust region methods (54) (or (56)).

Theorem 5. *(Global convergence) Let the mapping $f : R^n \to R^m$ satisfy Assumption A1 and $x_i \in R^n$, $i \in N$, be a sequence generated by the trust region method (T1)–(T3), where the matrix B_1 is positive semidefinite and matrices B_{i+1}, $i \in N$, are computed by (54) (or (56)) with $1 \le \gamma_i \le \overline{\gamma}$ and $\beta_i^* < \beta_i \le \beta_i^K$, where*

$$\beta_i^* = \frac{b_i^2}{b_i^2 - a_i c_i}, \qquad \beta_i^K = \frac{b_i K}{b_i K + c_i}, \tag{59}$$

with $K > 0$. Then $\liminf_{i\to\infty} \|g(x_i)\| = 0$.

Proof Note that $\beta_i^* < 0$ is the critical value mentioned above, so matrices B_i, $i \in N$, are positive semidefinite by Remark 6. We show that there exists a constant C such that

$$Tr\, B_{i+1} \le Tr\, B_i + C \tag{60}$$

for all $i \in N$. Then setting $\overline{C} = \max(Tr\, B_1, C)$, we obtain $\|B_i\| \le Tr\, B_i \le i\overline{C}$ for all $i \in N$, so $M_i = \max_{1\le j\le i} \|B_j\| \le i\overline{C}$ and

$$\sum_{i=1}^{\infty} \frac{1}{M_i} \ge \sum_{i=1}^{\infty} \frac{1}{i\overline{C}} = \infty, \tag{61}$$

since the harmonic series is divergent. The considered assertion then follows from Theorem 1.

(a) If $i \notin N_2 \cap N_3$ (or $i \notin N_3$), then $B_{i+1} = B_i$, so $Tr\, B_{i+1} = Tr\, B_i$ and (60) holds with $C = 0$.

(b) If $i \in N_2 \cap N_3$ (or $i \in N_3$) and $\beta_i^* < \beta_i \le 0$, we can write

$$\begin{aligned} Tr\, B_{i+1} &= \frac{1}{\gamma_i} Tr\, B_i + \frac{y_i y_i^T}{y_i^T s_i} - \frac{(B_i s_i)^T B_i s_i}{\gamma_i c_i} + \frac{\beta_i}{\gamma_i c_i} \left(\frac{c_i}{b_i} y_i - B_i s_i\right)^T \left(\frac{c_i}{b_i} y_i - B_i s_i\right) \\ &\le Tr\, B_i + \frac{y_i^T y_i}{y_i^T s_i} \le Tr\, B_i + \frac{1}{c} \end{aligned} \tag{62}$$

by (55) and (53), since B_i is positive semidefinite and $\gamma_i \ge 1$. Setting $C = 1/c$, we obtain (60).

(c) By comparing, we easily find that the formula $B_{i+1} = \mathcal{B}(B_i, y_i, s_i, \beta_i, \gamma_i)$ (given by (55)) can be written in the form

$$\begin{aligned} B_{i+1} &= \frac{1}{\gamma_i} \left(B_i + \frac{1}{b_i} \left(\frac{\beta_i c_i}{(1-\beta_i) b_i} + \gamma_i \right) y_i y_i^T \right. \\ &- \left. \frac{1-\beta_i}{c_i} \left(\frac{\beta_i c_i}{(1-\beta_i) b_i} y_i + B_i s_i \right) \left(\frac{\beta_i c_i}{(1-\beta_i) b_i} y_i + B_i s_i \right)^T \right). \end{aligned} \tag{63}$$

If $i \in N_2 \cap N_3$ (or $i \in N_3$) and $0 < \beta_i \le \beta_i^K < 1$, the first part of (62) implies that $Tr\, B_{i+1} \le Tr\, B_{i+1}^K$, and since the last term in (63) has a negative trace if $\beta_i = \beta_i^K < 1$, we can write

$$\begin{aligned} Tr\, B_{i+1}^K &\le \frac{1}{\gamma_i} Tr\, B_i + \frac{1}{\gamma_i y_i^T s_i} \left(\frac{\beta_i^K c_i}{(1-\beta_i^K) b_i} + \gamma_i \right) y_i^T y_i \le Tr\, B_i + (K + \overline{\gamma}) \frac{y_i^T y_i}{y_i^T s_i} \\ &\le Tr\, B_i + \frac{K + \overline{\gamma}}{c}, \end{aligned} \tag{64}$$

since (59) implies (after substituting) that $\beta_i^K c_i / ((1-\beta_i^K) b_i) = K$. Setting $C = (K + \overline{\gamma})/c$, we obtain (60). □

Remark 7. Inequality (60) is satisfied for the Hoshino method since we can choose $K = \overline{\gamma}$ in this case. Inequality (60) is also satisfied for the modified rank-one method, which uses $\beta_i = \beta_i^{R1}$, if $\beta_i^* < \beta_i^{R1} < 0$, or $\beta_i = 0$, if $\beta_i^{R1} \le \beta_i^*$ or $\beta_i^{R1} \ge 0$, where

$$\beta_i^{R1} = \frac{\gamma_i b_i}{\gamma_i b_i - c_i} \quad \Rightarrow \quad B_{i+1}^{R1} = \frac{1}{\gamma_i} \left(B_i + \frac{(\gamma_i y_i - B_i s_i)(\gamma_i y_i - B_i s_i)^T}{s_i^T (\gamma_i y_i - B_i s_i)} \right)$$

by Remark 5 and (55). Evidently $\beta_i \le 0$, so we obtain (60) with $C = 1/c$ as in part (b) of Theorem 5.

Remark 8. The rank-one method can also be modified in such a way that

$$\begin{array}{lll} B_{i+1} = B_{i+1}^{R1} & \text{if} & i \in N_2 \cap N_4, \\ B_{i+1} = B_i & \text{if} & i \notin N_2 \cap N_4, \end{array} \tag{65}$$

where

$$N_4 = \{i \in N : \left| s_i^T \left(\gamma_i y_i - B_i s_i\right)\right| \geq c \left\|\gamma_i y_i - B_i s_i\right\|^2\} \tag{66}$$

with $c > 0$. In the first case $\|B_{i+1}\| \leq \|B_i\| + 1/c$ and in the second case $\|B_{i+1}\| = \|B_i\|$. Therefore $M_i \leq i\overline{C}$, where $\overline{C} = \max(\|B_1\|, 1/c)$, so

$$\sum_{i=1}^{\infty} \frac{1}{M_i} \geq \sum_{i=1}^{\infty} \frac{1}{i\overline{C}} = \infty$$

and the R1 method (65) is globally convergent by Theorem 1.

The convergence rate of variable metric methods is studied, e.g., in [9], [16], [28], [46]. The following lemma summarizes the results introduced in [28].

Lemma 1. *Let $x_i \in R^n$, $i \in N$, be a sequence generated by the trust region method (T1)–(T3), where matrices B_i, $i \in N$, (with B_1 positive definite) are updated by (54) with $\gamma_i = 1$ and $0 \leq \beta_i \leq 1$. Let $x_i \to x_*$, where the point $x_* \in R^n$ satisfies Assumption A3. Let $k \in N$ be an index such that $x_i \in \mathcal{U}(x_*, \varepsilon)$ for all $i \geq k$, where $\mathcal{U}(x_*, \varepsilon)$ is the neighborhood defined in Remark 2. Denote*

$$\begin{array}{lll} R_i = \tilde{G}_i^{-1/2} B_i \tilde{G}_i^{-1/2}, & R'_{i+1} = \tilde{G}_i^{-1/2} B_{i+1} \tilde{G}_i^{-1/2}, & z_i = \tilde{G}_i^{1/2} s_i = \tilde{G}_i^{-1/2} y_i, \\ R_i^* = G_*^{-1/2} B_i G_*^{-1/2}, & R_{i+1}^* = G_*^{-1/2} B_{i+1} G_*^{-1/2}, & \end{array}$$

where $\tilde{G}_i = \int_0^1 G(x_i + t s_i)\mathrm{d}t$ and $G_* = G(x_*)$.

(a) If $i \in N_2$, then

$$\begin{aligned} \|R'_{i+1} - I\|_F^2 \;=\; \|R_i - I\|_F^2 \;&-\; (1-\beta_i)\left(\left(1 - \frac{z_i^T R_i^2 z_i}{z_i^T R_i z_i}\right)^2 + 2\left(\frac{z_i^T R_i^3 z_i}{z_i^T R_i z_i} - \left(\frac{z_i^T R_i^2 z_i}{z_i^T R_i z_i}\right)^2\right)\right) \\ &-\; \beta_i\left(\left(1 - \frac{z_i^T R_i z_i}{z_i^T z_i}\right)^2 + 2\beta_i\left(\frac{z_i^T R_i^2 z_i}{z_i^T z_i} - \left(\frac{z_i^T R_i z_i}{z_i^T z_i}\right)^2\right)\right) \\ &-\; \beta_i(1-\beta_i)\left(\left(\frac{z_i^T R_i^2 z_i}{z_i^T R_i z_i}\right)^2 - \left(\frac{z_i^T R_i z_i}{z_i^T z_i}\right)^2\right), \end{aligned}$$

so $\|R'_{i+1} - I\|_F \leq \|R_i - I\|_F$ (since $0 \leq \beta_i \leq 1$).

(b) If $i \geq k$, then $\|R_{i+1} - I\|_F + 1 = (\|R_i - I\|_F + 1)(1 + O(\|e_i\|))$.

(c) If $i \geq k$, then $\max(1, \|R_{i+1}^*\|) \leq \max(1, \|R_i^*\|)(1 + O(\|e_i\|))$, so denoting $C_i = \overline{G}\max(1, \|R_i^*\|)$, we can write

$$\|B_i\| = \|G_*^{1/2} R_i^* G_*^{1/2}\| \leq \overline{G}\|R_i^*\| \leq \overline{G}\max(1, \|R_i^*\|) = C_i,$$

where $C_{i+1} = C_i(1 + O(\|e_i\|))$.

(d) If $\|s_i\| = O(\|e_i\|)$ for $i \geq k$ and $\sum_{i=1}^{\infty} \|e_i\| < \infty$, then

$$\lim_{i\to\infty} \frac{\|(B_i - G_i)s_i\|}{\|s_i\|} = 0.$$

Theorem 6. *(Local convergence) Let the point $x_* \in R^n$ satisfy Assumption A3. Consider the trust region method (T1)–(T3), where matrices B_i, $i \in N$, (with B_1 positive definite) are updated by (54) with $\gamma_i = 1$ and $0 \le \beta_i \le 1$. Then there exists a number $\delta > 0$ such that $\|x_1 - x_*\| < \delta$ implies $\|B_i\| \le \overline{B} = 2\|B_1\|$ for all $i \in N$, $x_i \to x_*$, $\sum_{i=1}^{\infty} \|x_i - x_*\| < \infty$ and $\vartheta_i \to 0$.*

Proof By Lemma 1 (c), there exists a constant $C > 0$ such that $\|B_i\| \le C_i$, where $C_1 = \|B_1\|$ and $C_{i+1} \le C_i(1+C\|e_i\|)$, for $x_i \in \mathcal{U}(x_*, \varepsilon)$. Thus choosing $\delta > 0$ as in (44), we obtain $x_i \in \mathcal{U}(x_*, \varepsilon)$, $\|B_i\| \le \overline{B} = 2\|B_1\|$ for all $i \in N$, $x_i \to x_*$ and $\sum_{i=1}^{\infty} \|e_i\| < \infty$ by Theorem 3. Since the sequence F_i, $i \in N$, is non-increasing, formula (5) implies $\|e_{i+1}\| = O(\|e_i\|)$ and, therefore, $\|s_i\| = \|e_{i+1} - e_i\| \le \|e_{i+1}\| + \|e_i\| = O(\|e_i\|)$, so Lemma 1 (d) implies $\vartheta_i \to 0$.

Theorem 7. *(Superlinear convergence) Let assumptions of Theorem 6 be satisfied and $x_i \to x_*$. If $\omega_i \to 0$, the rate of convergence is superlinear.*

Proof Since $x_i \to x_*$, there exists an index $k \in N$ such that $\|x_k - x_*\| \le \delta$, where δ is given by (44). Then $\|B_i\| \le \overline{B} = 2\|B_k\|$ for all $i \ge k$, $\sum_{i=k}^{\infty} \|x_i - x_*\| < \infty$ and $\vartheta_i \to 0$ by Theorem 5. If in addition $\omega_i \to 0$, the rate of convergence is superlinear by Theorem 4.

Remark 9. An advantage of variable metric methods consists in the fact that they can update factors of the Choleski decompositions $B_i = L_i D_i L_i^T$, which consumes $O(n^2)$ arithmetic operations per iteration [26] (computation of the Choleski decomposition consumes $O(n^3)$ arithmetic operations). If the variable metric methods are parts of hybrid methods for nonlinear least squares, then sometimes $B_i \approx J_i^T J_i$ where J_i does not have a full column rank, so $J_i^T J_i$ is singular (positive semidefinite). Then we can use the Gill-Murray decomposition $B_i = L_i D_i L_i^T = J_i^T J_i + E_i$, where E_i is a (small) diagonal positive semidefinite matrix obtained recursively in such a way that $B_i = L_i D_i L_i^T$ is positive definite [25]. Thus the important assumption, the positive definiteness of matrices B_i, $i \in N$, is always satisfied.

4 Newton method

The Newton trust region method uses matrices $B_i = G(x_i)$, $i \in N$, in (8) and (T1). If Assumption A1 is satisfied, then $\|B_i\| = \|G(x_i)\| \le \overline{G}$, $i \in N$, so this method is globally convergent by Theorem 1. The choice $B_i = G(x_i)$ implies equality $\vartheta_i = 0$ for all $i \in N$. If $x_i \to x_*$, $\omega_i \to 0$ and Assumption A3 is satisfied, then the rate of convergence is superlinear by Theorem 4 (if Assumption A4 is satisfied, then the rate of convergence is quadratic).

The main disadvantage of the Newton method is the necessity of computing second order derivatives. The second order derivatives of the objective function can be computed by numerical differentiation. In this case the Hessian matrix is determined inaccurately. The equality $B_i = G(x_i)$ does not hold, only the inequality $\|B_i - G_i\| \le \overline{\vartheta}$ is satisfied. The upper bound $\overline{\vartheta} > 0$ is given by the following theorem.

Theorem 8. *Let the Hessian matrix $G(x)$ be Lipschitz continuous on $\overline{\mathcal{D}}$, i.e., there exists a number $\overline{L} > 0$ such that*

$$\|G(x_2) - G(x_1)\| \le \overline{L}\|x_2 - x_1\|, \tag{67}$$

if $x_1 \in \overline{\mathcal{D}}$ and $x_2 \in \overline{\mathcal{D}}$, and let B be a matrix such that

$$Bv_j = \frac{g(x + \delta v_j) - g(x)}{\delta} \tag{68}$$

for $1 \le j \le n$, where v_j, $1 \le j \le n$, are columns of the unit matrix of order n and $\delta > 0$ is a small difference. Then

$$\|B - G(x)\| \le \frac{1}{2}\overline{L}\sqrt{n}\delta \triangleq \overline{\vartheta}. \tag{69}$$

Proof Using the mean value theorem, we obtain

$$g(x+\delta v_j) = g(x) + G(x)\delta v_j + \int_0^1 (G(x+\tau\delta v_j) - G(x))\delta v_j \mathrm{d}\tau,$$

so we can write

$$\begin{aligned} \|(B-G(x))v_j\| &= \left\| \frac{g(x+\delta v_j)-g(x)}{\delta} - G(x)v_j \right\| \le \frac{1}{\delta}\left\| \int_0^1 (G(x+\tau\delta v_j)-G(x))\delta v_j \mathrm{d}\tau \right\| \\ &\le \frac{1}{2\delta}\overline{L}\delta^2\|v_j\|^2 = \frac{1}{2}\overline{L}\delta. \end{aligned}$$

by (68) and (67). Let $w \in R^n$ be an arbitrary vector with the unit norm. Then

$$\begin{aligned} \|(B-G(x))w\| &= \left\| \sum_{j=1}^n (B-G(x))v_j v_j^T w \right\| \le \sum_{j=1}^n |v_j^T w| \|(B-G(x))v_j\| \le \frac{1}{2}\overline{L}\delta \sum_{j=1}^n |v_j^T w| \\ &\le \frac{1}{2}\overline{L}\sqrt{n}\delta\|w\| = \frac{1}{2}\overline{L}\sqrt{n}\delta \end{aligned}$$

and since

$$\|B-G(x)\| = \max_{\|w\|=1} \|(B-G(x))w\|,$$

we obtain (69). □

Remark 10. The discrete Newton method, based on numerical differentiation, is not recommended for problems with dense Hessian matrices, since the numerical approximation computed by (68) requires $n+1$ gradient evaluations. If the Hessian matrix is sparse, the number of gradient evaluations can be substantially reduced by a sophisticated choice of vectors v_j, $1 \le j \le \tilde{n}$, $\tilde{n} \ll n$ [10], [11]. Thus the discrete Newton method is really efficient for minimization of functions with sparse Hessian matrices.

5 Gauss–Newton method

The Gauss–Newton method is obtained from the Newton method by deleting the second order term $C(x_i)$ in (3), so

$$B_i = J^T(x_i)J(x_i) = \sum_{k=1}^m h_k(x_i)h_k^T(x_i),$$

where B_i, $i \in N$, are the matrices used in (8) and (T1).

Remark 11. There are two reasons for using such an approximation of the Hessian matrix:

(a) Zero residual problems. Let $F(x_*) = 0$. Then $x_i \to x_*$ implies $F(x_i) \to F(x_*) = 0$ and, therefore, $f_k(x_i) \to 0$, $1 \le k \le m$. If Assumption A1 is satisfied, then

$$\|C(x_i)\| = \left\| \sum_{k=1}^m f_k(x_i)H_k(x_i) \right\| \le \overline{G}\sum_{k=1}^m |f_k(x_i)| \to 0,$$

so $\|G(x_i) - B_i\| = \|C(x_i)\| \to 0$, which is a sufficient condition for the Q-superlinear rate of convergence (Theorem 4).

(b) Linearization. We can write

$$\begin{aligned} F(x_i+s) &= \frac{1}{2}f^T(x_i+s)f(x_i+s) \approx \frac{1}{2}(f(x_i)+J(x_i)s)^T(f(x_i)+J(x_i)s) = \\ &= \frac{1}{2}f^T(x_i)f(x_i) + f^T(x_i)J(x_i)s + \frac{1}{2}s^T J^T(x_i)J(x_i)s, \end{aligned}$$

so

$$F(x_i + s) - F(x_i) \approx g^T(x_i)s + \frac{1}{2}s^T B_i s,$$

which is a local quadratic approximation with the matrix $B_i = J_i^T J_i$.

Theorem 9. *If the mapping f satisfies Assumption A1, then the Gauss–Newton method, realized as the trust region method, is globally convergent. If in addition $x_i \to x_*$, where $F(x_*) = 0$, then the rate of convergence is Q-superlinear.*

Proof If the mapping f satisfies Assumption A1, then

$$\|B_i\| = \|J^T(x_i)J(x_i)\| \le m\overline{h}^2$$

by Remark 1, so the Gauss–Newton method is globally convergent by Theorem 1. If in addition $x_i \to x_*$ and $F(x_*) = 0$ (or $f(x_*) = 0$) holds, we obtain, as in Remark 11, relation

$$\frac{\|(G(x_i) - B_i)d_i\|}{\|d_i\|} \le \|C(x_i)\| \to 0,$$

which by Theorem 4 implies Q-superlinear rate of convergence. □

Remark 12. A direction vector corresponding to the Gauss–Newton method can be determined by several different ways [5]:

(a) Solution of the normal equation system. Substituting $B_i = J_i^T J_i$ and $g_i = J_i^T f_i$ into the formula $B_i d_i + g_i = 0$, we obtain a system of linear equations $J_i^T J_i d_i + J_i^T f_i = 0$, which is called the normal equation system.

(b) Solution of a linear least squares problem. We solve a linear over-determined system $J_i d_i + f_i \approx 0$ in the least squares sense (by minimization of $\|J_i d_i + f_i\|$. In this case, the stable QR-decomposition of the Jacobian matrix J_i can be used [2].

(c) Solution of the augmented system. Denote $r_i = -(J_i d_i + f_i)$. Since the direction vector has to satisfy the equation $J_i^T r_i = 0$, we can write

$$\begin{bmatrix} I & J_i \\ J_i^T & 0 \end{bmatrix} \begin{bmatrix} r_i \\ d_i \end{bmatrix} + \begin{bmatrix} f_i \\ 0 \end{bmatrix} = 0,$$

which is a system of $m+n$ linear equations with an indefinite matrix. This way is advantageous especially for sparse problems since sparsity of the Jacobian matrix J_i implies sparsity of the augmented system, while the normal equation matrix can be dense (e.g., if J_i has a dense row). The augmented system is also suitable for weighted least squares problems. If

$$F(x) = \frac{1}{2}f^T(x)W f(x),$$

where W is a weighting matrix, then the normal equation system has the form

$$J_i^T W J_i d_i + J_i^T W f_i = 0,$$

and denoting $r_i = -W(J_i d_i + f_i)$ we obtain

$$\begin{bmatrix} W^{-1} & J_i \\ J_i^T & 0 \end{bmatrix} \begin{bmatrix} r_i \\ d_i \end{bmatrix} + \begin{bmatrix} f_i \\ 0 \end{bmatrix} = 0.$$

Thus some weighting coefficients can acquire infinite values, which is useful for solving problems with equality constraints.

6 Simple hybrid methods with Hessian approximations

The Gauss–Newton method is very efficient for solving zero-residual problems, but it can fail for large-residual problems. Therefore, the following strategy seems to be reasonable:

(a) If $F_i \to F_* = 0$, we use the Gauss–Newton method.

(b) If $F_i \to F_* > 0$, we use some superlinearly convergent method (either the Newton method or a variable metric method).

The following theorem gives a reason for the choice of a suitable hybrid method [1].

Theorem 10. *Let $F_i \to F_* = 0$ Q-superlinearly. Then*

$$\lim_{i\to\infty} \frac{F_i - F_{i+1}}{F_i} = 1.$$

Let $F_i \to F_* > 0$. Then

$$\lim_{i\to\infty} \frac{F_i - F_{i+1}}{F_i} = 0.$$

Proof If $F_i \to F_* = 0$ Q-superlinearly, then

$$\lim_{i\to\infty} \frac{F_i - F_{i+1}}{F_i} = 1 - \lim_{i\to\infty} \frac{F_{i+1} - F_*}{F_i - F_*} = 1 - 0 = 1.$$

If $F_i \to F_* > 0$, then

$$\lim_{i\to\infty} \frac{F_i - F_{i+1}}{F_i} = \frac{1}{F_*} \lim_{i\to\infty} (F_i - F_{i+1}) = 0.$$

□

To describe a hybrid method based on Theorem 10, we choose a value $0 < \vartheta < 1$ and denote

$$N_5 = \{i \in N : (F_i - F_{i+1})/F_i \geq \vartheta\}.$$

Description 1. *(Simple hybrid method)* An efficient simple hybrid method arises as a combination of the Gauss–Newton method and a suitable variable metric method from the Broyden class [1]. Let $B_1 = J_1^T J_1$ and $\vartheta > 0$. Set

$$\begin{array}{lll} B_{i+1} = J_{i+1}^T J_{i+1} & \text{if} & i \in N_2, \quad i \in N_5, \\ B_{i+1} = \mathcal{B}(B_i, y_i, s_i, \beta_i, \gamma_i) & \text{if} & i \in N_2, \quad i \notin N_5, \quad i \in N_3, \\ B_{i+1} = B_i & \text{if} & i \in N_2, \quad i \notin N_5, \quad i \notin N_3, \\ B_{i+1} = B_i & \text{if} & i \notin N_2, \end{array} \tag{70}$$

where $\mathcal{B}$ is the mapping defined by (55), $1 \leq \gamma_i \leq \overline{\gamma}$, $0 \leq \beta_i \leq \beta_i^K$ and β_i^K is the number given by (59). Note that the simple hybrid method (70) passes to the Gauss-Newton method, if $\vartheta = 0$ (so $N_5 = N$), and to a variable metric method, if $\vartheta = \infty$ (so $N_5 = \emptyset$).

Remark 13. The procedure introduced in Description 1 will be considered as basic, since it is robust and efficient. Motivated by (56), we will consider here the following two strategies distinguished by the values of S1:

S1=1 – The basic strategy (70).
S1=2 – The strategy, where N_2 is replaced by N in (70) (so the last case in (70) is redundant).

Efficiency of these strategies is demonstrated in Table 2 in Section 10.

Theorem 11. *If the mapping f satisfies Assumption A1, then the simple hybrid (trust region) method described in Description 1 is globally convergent. If in addition $x_i \to x_*$, where the point $x_* \in R^n$ satisfies Assumption A3, and $\omega_i \to 0$, then the rate of convergence is Q-superlinear.*

Proof Clearly $i \in N_2 \cap N_5$ if and only if $B_i = J_i^T J_i$.

(a) If $i \in N_2 \cap N_5$, then $\|B_i\| = \|J_i\|^2 \leq m\overline{h}^2$ by Remark 1. If $i \notin N_2 \cap N_5$ and $k \in N$ is the maximum index such that $k < i$ and $k \in N_2 \cap N_5$, then $B_k = J_k^T J_k$ is positive semidefinite and also matrices B_j, $k < j \leq i$, are positive semidefinite by Remark 6 (since $y_j^T s_j > 0$ and $0 \leq \beta_j \leq \beta_j^K < 1$). Therefore, relation (60) implies that there exists a number $C > 0$ such that $Tr\, B_j \leq Tr\, B_{j-1} + C$, $k < j \leq i$. Thus we can write

$$\|B_i\| \leq Tr\, B_i \leq Tr\, B_k + (i-k)\, C \leq i\,(n\|B_k\| + C) \leq i\,(nm\overline{h}^2 + C) \leq i\overline{C},$$

where $\overline{C} = \max(\|B_1\|, nm\overline{h}^2 + C)$, so $M_i \leq i\overline{C}$, $i \in N$, and the global convergence follows from Theorem 1.

(b) Let $x_i \to x_*$ and $F(x_*) > 0$. Then $(F_{i-1} - F_i)/F_{i-1} \to 0$ by Theorem 10, so there exists an index $k \in N$ such that $(F_{i-1} - F_i)/F_{i-1} < \vartheta$ $\forall i \geq k$, so $i \notin N_2 \cap N_5$ $\forall i \geq k$ and the superlinear rate of convergence follows from Theorem 7.

(c) Let $x_i \to x_*$ and $F(x_*) = 0$. If the set $N_2 \cap N_5$ is finite, the superlinear rate of convergence follows from Theorem 7 (as in part (b)). If the set $N_2 \cap N_5$ is infinite, then $B_{i+1} \overset{N_2 \cap N_5}{\longrightarrow} G_{i+1}$ by Remark 11, which gives

$$\frac{\|(B_{i+1} - G_{i+1})d_{i+1}\|}{\|d_{i+1}\|} \overset{N_2 \cap N_5}{\longrightarrow} 0, \tag{71}$$

so, similarly as in parts (b) and (c) of the proof of Theorem 4, there exists an index $k_2 \in N_2 \cap N_5$ such that $i + 1 \in N_2$ for $i \in N_2 \cap N_5$, $i \geq k_2$, and

$$\tau_{i+1} = \frac{g_{i+2} - g_{i+1} - B_{i+1}d_{i+1}}{\|g_{i+1}\|} \overset{N_2 \cap N_5}{\longrightarrow} 0,$$

which together with $F(x_*) = 0$, $g(x_*) = 0$, (5) and (10) gives

$$\frac{F_{i+1} - F_{i+2}}{F_{i+1}} = 1 - \frac{F_{i+2}}{F_{i+1}} \geq 1 - \left(\frac{\overline{G}}{\underline{G}}\right)^2 \left(\frac{\|g_{i+2}\|}{\|g_{i+1}\|}\right)^2 = 1 - \left(\frac{\overline{G}}{\underline{G}}\right)^2 (\|\tau_{i+1}\| + \|\omega_{i+1}\|)^2 \overset{N_2 \cap N_5}{\longrightarrow} 1,$$

so there exists an index $k_3 \in N_2 \cap N_5$, $k_3 \geq k_2$ such that $(F_{i+1} - F_{i+2})/F_{i+1} \geq \vartheta$ for $i \in N_2 \cap N_5$, $i \geq k_3$. Thus $i + 1 \in N_5$ and since also $i + 1 \in N_2$, we can write $i + 1 \in N_2 \cap N_5$ for $i \in N_2 \cap N_5$, $i \geq k_3$. Therefore $i \in N_2 \cap N_5$ $\forall i \geq k_3$ holds by induction and the superlinear rate of convergence follows from (71) as in the proof of Theorem 4. □

7 Structured hybrid methods with Hessian approximations

In this section we will concentrate on further combinations of the Gauss-Newton and variable metric methods that are often called structured variable metric methods [16], [36]. To simplify the notation, we frequently omit index i and replace index $i + 1$ by symbol $+$. We will suppose that $B = J^T J + C$, where C is an approximation of $C(x)$, and we will look for a matrix C_+ such that the matrix $B_+ = J_+^T J_+ + C_+$ satisfies the quasi-Newton condition $B_+ s = y$, where again $s = x_+ - x$ and $y = g_+ - g = J_+^T f_+ - J^T f$. There exist two ways how to achieve this aim. The first one is based on using a transformed quasi-Newton condition

$$C_+ s = z = y - J_+^T J_+ s = J_+^T f_+ - J^T f - J_+^T J_+ s,$$

which immediately follows from the condition $B_+ s = J_+^T J_+ s + C_+ s = y$, and replacing matrix B with matrix C in (55). Thus, we will obtain an update $C_+ = \mathcal{B}(C, z, s, \beta, \gamma)$ (a Broyden class), where

$$\mathcal{B}(C, z, s, \beta, \gamma) = \frac{1}{\gamma}\left(C + \gamma\frac{zz^T}{s^T z} - \frac{Cs(Cs)^T}{s^T Cs} + \frac{\beta}{s^T Cs}\left(\frac{s^T Cs}{s^T z}z - Cs\right)\left(\frac{s^T Cs}{s^T z}z - Cs\right)^T\right) \tag{72}$$

(if $Cs = 0$, the last two terms will be cancelled). A disadvantage of this approach lies in that a number $s^T z$ need not be positive, which complicates using the BFGS method (with $\beta = 0$). In this connection, the rank-one method with an update $C_+ = C_+^{R1}$, where

$$C_+^{R1} = \frac{1}{\gamma}\left(C + \frac{(\gamma z - Cs)(\gamma z - Cs)^T}{s^T(\gamma z - Cs)}\right) \tag{73}$$

(the matrix C_+ need not be positive definite as it approximates the second order term which is added to a matrix $J_+^T J_+$), is mostly used. However, denominator in (73) may be zero, so this update is used only if $|s^T(\gamma z - Cs)| \geq c\|\gamma z - Cs\|^2$, where $c > 0$ is a suitably chosen small constant. It is also possible to use an update $C_+ = \mathcal{D}(C, z, s, v, \gamma)$ (Dennis class [15]), where

$$\mathcal{D}(C, z, s, v, \gamma) = \frac{1}{\gamma}\left(C + \frac{(\gamma z - Cs)v^T + v(\gamma z - Cs)^T}{s^T v} - \frac{(\gamma z - Cs)^T s}{s^T v}\frac{vv^T}{s^T v}\right) \tag{74}$$

and where $v = s$ for the PSB method, $v = z$ for the DFP method, $v = z + (z^T s/(\gamma s^T Cs))^{1/2} Cs$ (or $v = z$ if $Cs = 0$) for the BFGS method, and $v = z - (1/\gamma)Cs$ for the R1 method. Formula (74), which is in the cases of DFP, BFGS, and R1 methods equivalent to (72), can be derived by a variational approach [16]. In the case of the PSB (Powell symmetric Broyden [44]) method, a number $s^T v$ is positive and this method is (for least squares problems) relatively efficient.

Remark 14. The vectors y and z may be defined by a various way but it must hold $y = z + J_+^T J_+ s$. A standard choice

$$z = J_+^T f_+ - J^T f - J_+^T J_+ s \tag{75}$$

corresponds to a quasi-Newton condition $(J_+^T J_+ + C_+)s = y = J_+^T f_+ - J^T f$. A very efficient choice is based on an explicit form of the second order term [4]. Suppose that approximations B_k^+ of the Hessian matrices H_k satisfy quasi-Newton conditions $B_k^+ s = h_k^+ - h_k$, $1 \leq k \leq m$. Then we can write

$$z = C^+ s = \sum_{k=1}^{m} f_k^+ B_k^+ s = \sum_{k=1}^{m} f_k^+ (h_k^+ - h_k) = (J_+ - J)^T f_+, \tag{76}$$

so $y = (J_+ - J)^T f_+ + J_+^T J_+ s$. It is usually more advantageous to use vector (76) than the standard choice (75). However, two matrices J and J_+ appear in formula (76). It is not necessary to store these matrices simultaneously. A vector $J^T f_+$ can be determined recurrently. Let $p_0 = 0$. Since the gradients $h_k(x_+)$, $1 \leq k \leq m$, are computed stepwise, we first set $p_k = p_{k-1} + J^T e_k e_k^T f_+$ (the k-th column of J^T multiplied by the k-th component of f_+ is added to p_{k-1}) and only then we replace the k-th column of J^T with the gradient $h_k(x_+)$, thereby we get the k-th column of J_+^T. Finally, we set $J^T f_+ = p_m$.

In the subsequent considerations we use the notation

$$N_3' = \{i \in N : z_i^T s_i \geq c\,\|z_i\|^2\}, \qquad N_4' = \{i \in N : \left|s_i^T\,(\gamma_i z_i - C_i s_i)\right| \geq c\,\|\gamma_i z_i - C_i s_i\|^2\}. \tag{77}$$

Description 2. An efficient structured hybrid method arises as a combination of the Gauss-Newton method with the R1 variable metric method. Let C_1 be positive definite, $B_1 = J_1^T J_1$, $\vartheta > 0$, and $c > 0$. Set

$$\begin{array}{lll} C_{i+1} = C_i & \text{if} & i \in N_2, \quad i \in N_5, \\ C_{i+1} = \mathcal{B}(C_i, z_i, s_i, \beta_i, \gamma_i) & \text{if} & i \in N_2, \quad i \notin N_5, \quad i \in N', \\ C_{i+1} = C_i & \text{if} & i \in N_2, \quad i \notin N_5, \quad i \notin N', \\ C_{i+1} = C_i & \text{if} & i \notin N_2 \end{array} \tag{78}$$

and

$$\begin{array}{lll} B_{i+1} = J_{i+1}^T J_{i+1} & \text{if} & i \in N_2, \quad i \in N_5, \\ B_{i+1} = J_{i+1}^T J_{i+1} + C_{i+1} & \text{if} & i \in N_2, \quad i \notin N_5, \\ B_{i+1} = B_i & \text{if} & i \notin N_2, \end{array} \tag{79}$$

where $\mathcal{B}$ is the mapping defined by (72), $1 \le \gamma_i \le \overline{\gamma}$, $\beta_i = \beta_i^{R1}$ or $0 \le \beta_i \le \beta_i^K$, β_i^K is the value defined by (59) with $b_i = z_i^T s_i$, $c_i = s_i^T C_i s_i$, $s_i = x_{i+1} - x_i$ and either $z_i = J_{i+1}^T f_{i+1} - J_i^T f_i - J_{i+1}^T J_{i+1} s_i$ (formula (75)) or $z_i = (J_{i+1} - J_i)^T f_{i+1}$ (formula (76)). At the same time, either $N' = N_4'$ and $\beta_i = \beta_i^{R1}$ for all $i \in (N_2 \setminus N_5) \cap N_4'$ or $N' = N_3'$ and $0 \le \beta_i \le \beta_i^K$ for all $i \in (N_2 \setminus N_5) \cap N_3'$ (the chosen variant must be kept for all iterations).

Remark 15. The procedure introduced in Description 2 will be considered as basic since it is robust and efficient. Motivated by Remark 13, we will consider here the following eight strategies distinguished by values of S1, S2 and S3:

- S1 = 1 – Update (78) uses the set N_2.
- S1 = 2 – The set N_2 is replaced by N in (78) (as for strategy S1 = 2 in Remark 13), so the last case in (78) is redundant.
- S2 = 1 – Update (78) uses the set N_5.
- S2 = 2 – The set N_5 is replaced by the empty set in (78), so the first case in (78) is redundant (the matrix C_i is updated even if $i \in N_5$).
- S3 = 1 – Vector z is computed by (76).
- S3 = 2 – Vector z is computed by (75).

Efficiency of these strategies is demonstrated in Table 3 and Table 5 in Section 10.

Remark 16. The values $\gamma_i = b_i/a_i$ and $\gamma_i = \sqrt{c_i/b_i}$, introduced in (58), cannot be used for scaling structured hybrid methods because computation of a number $a_i = z_i^T C_i^{-1} z_i$ requires $O(n^3)$ arithmetic operations. The value $\gamma_i = c_i/b_i$, which is optimal for the DFP method, can be used with certain success for the BFGS and the Hoshino updates. However, this value is absolutely unsuitable for the R1 method, since in this case $\gamma_i z_i - C_i s_i = 0$. For scaling the R1 update, the value

$$\gamma_i = \frac{f_i^T f_i}{f_i^T f_{i+1}}, \tag{80}$$

proposed in [4], is recommended. This value was used for obtaining the results stated in Table 5 and Table 6 in Section 10.

Theorem 12. *If the mapping f satisfies Assumption A1, then the structured hybrid method introduced in Description 2 with either $N' = N_4'$ and $\beta_i = \beta_i^{R1}$ or $N' = N_3'$ and $0 \le \beta_i \le \beta_i^K$, realized as the trust region method, is globally convergent.*

Proof (a) Let $N' = N_4'$. If $i \in (N_2 \setminus N_5) \cap N_4'$ (so $\beta_i = \beta_i^{R1}$), we can write

$$\|C_{i+1}\| \le \|C_i\| + \frac{\|\gamma_i z_i - C_i s_i\|^2}{|s_i^T(\gamma_i z_i - C_i s_i)|} \le \|C_i\| + \frac{1}{c},$$

and if $i \notin (N_2 \setminus N_5) \cap N_4'$, then $\|C_{i+1}\| = \|C_i\|$. Thus for $i \in N_2$ we obtain

$$\|B_{i+1}\| \le \|J_{i+1}^T J_{i+1}\| + \|C_{i+1}\| \le m\overline{h}^2 + \|C_1\| + \frac{i}{c} \le (i+1)\overline{C}, \qquad \overline{C} = \max\left(m\overline{h}^2 + \|C_1\|, \frac{1}{c}\right)$$

and for $i \notin N_2$ we can write

$$\|B_{i+1}\| = \|B_{k+1}\| \le m\overline{h}^2 + \|C_1\| + \frac{k}{c} \le (k+1)\overline{C} \le (i+1)\overline{C},$$

where $k \in N$ is the largest index such that $k < i$ and $k \in N_2$. Thus $M_i \le i\overline{C}$ and the global convergence follows from Theorem 1.

(b) Let $N' = N_3'$. If $i \in (N_2 \setminus N_5) \cap N_3'$ (so $0 \le \beta_i \le \beta_i^K$), we can write

$$Tr\, C_{i+1} \le Tr\, C_i + (K + \overline{\gamma})\frac{z^T z}{z^T s} \le Tr\, C_i + \frac{K + \overline{\gamma}}{c}$$

(as in part (c) of the proof of Theorem 5) and if $i \notin (N_2 \setminus N_5) \cap N_3'$, then $Tr\, C_{i+1} = Tr\, C_i$. Thus for $i \in N_2$ we obtain

$$Tr\, B_{i+1} \leq Tr\, (J_{i+1}^T J_{i+1}) + Tr\, C_{i+1} \leq nm\overline{h}^2 + Tr\, C_1 + i\frac{K+\overline{\gamma}}{c} \leq (i+1)\overline{C},$$

where

$$\overline{C} = \max\left(nm\overline{h}^2 + Tr\, C_1, \frac{K+\overline{\gamma}}{c}\right),$$

and for $i \notin N_2$ we can write

$$Tr\, B_{i+1} = Tr\, B_{k+1} \leq nm\overline{h}^2 + Tr\, C_1 + k\frac{K+\overline{\gamma}}{c} \leq (k+1)\overline{C} \leq (i+1)\overline{C},$$

where $k \in N$ is the largest index such that $k < i$ and $k \in N_2$. Thus $M_i \leq i\overline{C}$ and the global convergence follows from Theorem 1. □

Remark 17. Update (78) can be replaced by the update

$$\begin{array}{lll} C_{i+1} = C_i & \text{if} \quad i \in N_2, \quad i \in N_5, & \\ C_{i+1} = \mathcal{D}(C_i, z_i, s_i, v_i, \gamma_i) & \text{if} \quad i \in N_2, \quad i \notin N_5, \quad i \in N', & \\ C_{i+1} = C_i & \text{if} \quad i \in N_2, \quad i \notin N_5, \quad i \notin N', & \\ C_{i+1} = C_i & \text{if} \quad i \notin N_2, & \end{array} \tag{81}$$

where $\mathcal{D}$ is the mapping defined by (74). In this case $N' = N_3'$ for both the BFGS and the DFP methods, $N' = N_4'$ for the R1 method and $N' = N$ for the PSB method.

A second way that has theoretical justification in the case $\gamma = 1$ is based on an update of the matrix $\bar{B} = J_+^T J_+ + C$ so that the matrix $B_+ = J_+^T J_+ + C_+$ satisfies the quasi-Newton condition $B_+ s = y$. In this case, a corresponding update can be written in the form $B_+ = \mathcal{B}(\bar{B}, y, s, \beta, 1)$. Subtracting the matrix $J_+^T J_+$ from both sides of this formula and taking account of $\gamma = 1$, we obtain

$$C_+ = C + \frac{yy^T}{s^T y} - \frac{\bar{B}s(\bar{B}s)^T}{s^T \bar{B}s} + \frac{\beta}{s^T \bar{B}s}\left(\frac{s^T \bar{B}s}{s^T y} y - \bar{B}s\right)\left(\frac{s^T \bar{B}s}{s^T y} y - \bar{B}s\right)^T, \tag{82}$$

where $\bar{B}s = J_+^T J_+ s + Cs$. Obviously, $y - \bar{B}s = y - J_+^T J_+ s - Cs = z - Cs$, so for the rank-one method it holds

$$C_+ = C + \frac{(y - \bar{B}s)(y - \bar{B}s)^T}{s^T(y - \bar{B}s)} = C + \frac{(z - Cs)(z - Cs)^T}{s^T(\gamma z - Cs)}. \tag{83}$$

Formula (83) is identical to formula (73), where $\gamma = 1$. It is also possible to use a formula

$$\begin{aligned} C_+ &= C + \frac{(y - \bar{B}s)v^T + v(y - \bar{B}s)^T}{s^T v} - \frac{(y - \bar{B}s)^T s}{s^T v}\frac{vv^T}{s^T v} \\ &= C + \frac{(z - Cs)v^T + v(z - Cs)^T}{s^T v} - \frac{(z - Cs)^T s}{s^T v}\frac{vv^T}{s^T v} \end{aligned} \tag{84}$$

(Dennis class [15]), where $v = s$ for the PSB method, $v = y$ for the DFP method, $v = y + (y^T s/(s^T \bar{B}s))^{1/2}\bar{B}s$ for the BFGS method, and $v = y - \bar{B}s = z - Cs$ for the R1 method. Formula (84) has form (74) (they are identical in the case of PSB and R1 methods). An advantage of this approach is the fact that the number $s^T v$ is always positive (except for the R1 method) if $s^T y > 0$ and $\bar{B}$ is positive definite. However, another problem, consisting in that the matrix $\bar{B} = J_+^T J_+ + C$ need not be positive definite even if the matrix $B = J^T J + C$ is positive definite, arises. Note that in (82)–(84), we can use scaling so that matrices C/γ and $\bar{B} = J_+^T J_+ + C/\gamma$ (thus $\bar{B}s = J_+^T J_+ s + Cs/\gamma$) are used instead of matrices C and $\bar{B}$. Note also that in this case we do not scale the matrix $\bar{B}$, as update $\mathcal{B}(\bar{B}, y, s, \beta, \gamma)$ requires, but the matrix C.

In the subsequent considerations we use the notation

$$N_3'' = N_3 \cap \{i \in N : s_i^T \bar{B}_i s_i = s_i^T (J_{i+1}^T J_{i+1} + C_i) s_i > 0\}, \qquad N_4'' = N_4'. \tag{85}$$

Description 3. An efficient structured hybrid method arises as a combination of the Gauss-Newton method with a suitable variable metric method which uses update (84). Let C_1 be positive definite, $B_1 = J_1^T J_1$, $\vartheta > 0$, and $c > 0$. Set

$$\begin{array}{lll} C_{i+1} = C_i & \text{if} & i \in N_2, \quad i \in N_5, \\ C_{i+1} = \mathcal{D}(C_i, z_i, s_i, v_i, 1) & \text{if} & i \in N_2, \quad i \notin N_5, \quad i \in N'', \\ C_{i+1} = C_i & \text{if} & i \in N_2, \quad i \notin N_5, \quad i \notin N'', \\ C_{i+1} = C_i & \text{if} & i \notin N_2 \end{array} \tag{86}$$

and

$$\begin{array}{lll} B_{i+1} = J_{i+1}^T J_{i+1} & \text{if} & i \in N_2, \quad i \in N_5, \\ B_{i+1} = J_{i+1}^T J_{i+1} + C_{i+1} & \text{if} & i \in N_2, \quad i \notin N_5, \\ B_{i+1} = B_i & \text{if} & i \notin N_2, \end{array} \tag{87}$$

where $\mathcal{D}$ is the mapping defined by (74). At the same time, $N'' = N_3''$ for both the BFGS and the DFP methods, $N'' = N_4''$ for the R1 method and $N'' = N$ for the PSB method. Even if $\gamma_i = 1$ in (86), these updates can be scaled in such a way that the matrix C_i is preliminary divided by a positive value γ_i as is mentioned above.

Remark 18. The procedure introduced in Description 3 will be considered as basic since it is robust and efficient. Motivated by Remark 15, we will consider here eight strategies distinguished by values of S1, S2 and S3, which concern update (86) and have the same meaning as strategies in Remark 15. Efficiency of these strategies is demonstrated in Table 4 and Table 5 in Section 10.

Remark 19. The update $C_{i+1} = \mathcal{D}(C_i, z_i, s_i, v_i, 1)$ can be replaced by (82) in (86). These updates are equivalent if $\beta_i = \beta_i^{R1}$ or $0 \le \beta_i \le \beta_i^K$.

Theorem 13. *If the mapping f satisfies Assumption A1, then the structured hybrid method introduced in Description 3, with $C_{i+1} = \mathcal{D}(C_i, z_i, s_i, v_i, 1)$ replaced by (82), such that either $N'' = N_4''$ and $\beta_i = \beta_i^{R1}$ or $N'' = N_3''$ and $0 \le \beta_i \le \beta_i^K$ (β_i^K is the value defined by (59) with $b_i = y_i^T s_i$ and $c_i = s_i^T \bar{B}_i s_i$), realized as the trust region method, is globally convergent.*

Proof The proof of this theorem is almost the same as the proof of Theorem 12. For all $i \in N$, vectors z_i are replaced by y_i and vectors $C_i s_i$ are replaced by $\bar{B}_i s_i$. □

Remark 20. From these considerations it follows that the methods using BFGS or R1 updates in (86) are globally convergent. The method using a PSB update in (86) is globally convergent by Theorem 18 of Section 12 because the PSB update is equivalent to the Toint update if the Hessian matrix is dense.

Totally structured variable metric methods mentioned in [31] offer a very interesting possibility of automatic scaling a matrix C. In this case, a matrix approximating the expression

$$T(x) = \sum_{k=1}^{m} \frac{f_k(x)}{\|f(x)\|} H_k(x)$$

is used and updated. Thus, we use model $B = J^T J + \|f\| T$ (so $C = \|f\| T$) and update the matrix T so that the matrix $B_+ = J_+^T J_+ + \|f_+\| T_+$ satisfied a quasi-Newton condition $B_+ s = (J_+^T J_+ + \|f_+\| T_+) s = y$, or $T_+ s = (y - J_+^T J_+ s)/\|f_+\| = z/\|f_+\| \triangleq \tilde{z}$ (the first way). Thus, we can write $T_+ = \mathcal{B}(T, \tilde{z}, s, \beta, 1)$, which by (55) gives

$$T_+ = T + \frac{\tilde{z}\tilde{z}^T}{s^T \tilde{z}} - \frac{Ts(Ts)^T}{s^T T s} + \frac{\beta}{s^T T s} \left(\frac{s^T T s}{s^T \tilde{z}} \tilde{z} - Ts \right) \left(\frac{s^T T s}{s^T \tilde{z}} \tilde{z} - Ts \right)^T. \tag{88}$$

For the rank-one method it holds

$$T_+ = T + \frac{(\tilde{z} - Ts)(\tilde{z} - Ts)^T}{s^T (\tilde{z} - Ts)}. \tag{89}$$

We can also set $T_+ = \mathcal{D}(T, \tilde{z}, s, v, 1)$, which by (74) gives

$$T_+ = T + \frac{(\tilde{z} - Ts)v^T + v(\tilde{z} - Ts)^T}{s^T v} - \frac{(\tilde{z} - Ts)^T s}{s^T v}\frac{vv^T}{s^T v}, \tag{90}$$

where $v = s$ for the PSB method, $v = \tilde{z}$ for the DFP method, $v = \tilde{z} + (s^T\tilde{z}/s^TTs)^{1/2}Ts$ (or $v = \tilde{z}$ if $Ts = 0$) for the BFGS method, and $v = z - Ts$ for the R1 method. Further, we can update the matrix $\tilde{B} = J_+^T J_+/\|f_+\| + T$ (the second way) so that the matrix $\tilde{B}_+ = B_+/\|f_+\| = J_+^T J_+/\|f_+\| + T_+$ satisfied quasi-Newton condition $\tilde{B}_+ s = y/\|f_+\| \stackrel{\Delta}{=} \tilde{y}$. In this case a corresponding update can be written in the form $\tilde{B}_+ = \mathcal{B}(\tilde{B}, \tilde{y}, s, \beta, 1)$. Subtracting a matrix $J_+^T J_+/\|f_+\|$ from both sides of this formula and taking account of $\gamma = 1$, we obtain

$$T_+ = T + \frac{\tilde{y}\tilde{y}^T}{s^T\tilde{y}} - \frac{\tilde{B}s(\tilde{B}s)^T}{s^T\tilde{B}s} + \frac{\beta}{s^T\tilde{B}s}\left(\frac{s^T\tilde{B}s}{s^T\tilde{y}}\tilde{y} - \tilde{B}s\right)\left(\frac{s^T\tilde{B}s}{s^T\tilde{y}}\tilde{y} - \tilde{B}s\right)^T, \tag{91}$$

where $\tilde{B}s = J_+^T J_+ s/\|f_+\| + Ts$. Obviously, $\tilde{y} - \tilde{B}s = \tilde{y} - J_+^T J_+ s/\|f_+\| - Ts = \tilde{z} - Ts$, so for the rank-one method it holds

$$T_+ = T + \frac{(\tilde{y} - \tilde{B}s)(\tilde{y} - \tilde{B}s)^T}{s^T(\tilde{y} - \tilde{B}s)} = T + \frac{(\tilde{z} - Ts)(\tilde{z} - Ts)^T}{s^T(\tilde{z} - Ts)}. \tag{92}$$

It is also possible to use a formula

$$\begin{aligned} T_+ &= T + \frac{(\tilde{y} - \tilde{B}s)v^T + v(\tilde{y} - \tilde{B}s)^T}{s^T v} - \frac{(\tilde{y} - \tilde{B}s)^T s}{s^T v}\frac{vv^T}{s^T v} \\ &= T + \frac{(\tilde{z} - Ts)v^T + v(\tilde{z} - Ts)^T}{s^T v} - \frac{(\tilde{z} - Ts)^T s}{s^T v}\frac{vv^T}{s^T v} \end{aligned} \tag{93}$$

(Dennis class [15]), where $v = s$ for the PSB update, $v = \tilde{y}$ for the DFP update, $v = \tilde{y} + (s^T\tilde{y}/s^T\tilde{B}s)^{1/2}\tilde{B}s$ for the BFGS update, and $v = \tilde{y} - \tilde{B}s$ for the R1 update. By Remark 14, one can use vectors $\tilde{z} = (J_+ - J)^T f_+/\|f_+\|$ and $\tilde{y} = J_+^T J_+ s/\|f_+\| + \tilde{z}$ instead of vectors $\tilde{z} = (y - J_+^T J_+ s)/\|f_+\|$ and $\tilde{y}$. In [31], a vector $\tilde{z} = (J_+ - J)^T f_+/\|f\|$ is used instead of a vector $\tilde{z} = (J_+ - J)^T f_+/\|f_+\|$ which is utilized in the proof of asymptotic rate of convergence.

Remark 21. Totally structured variable metric methods can be realized by the same way as standard structured variable metric methods. We again use updates introduced in Description 2 and Description 3, where matrices C_{i+1}, C_i and vectors y_i, z_i, $\bar{B}_i s_i$ are replaced with matrices T_{i+1}, T_i and vectors $\tilde{y}_i$, $\tilde{z}_i$, $\tilde{B}_i s_i$. Again, it is possible to use eight strategies stated in Remark 15. In order to present results of both classes of structured variable metric methods in common tables, we introduce the strategies:

S4 = 1 – Standard structured variable metric method is used.
S4 = 2 – Totally structured variable metric method is used.

Remark 22. Let the mapping f satisfy Assumption A1 and $x_i \in R^n$, $i \in N$, be a sequence generated by the totally structured hybrid method realized by Remark 21 such that $x_i \to x_*$ where $x_* \in R^n$ is a point in which function (1) attains its minimum and where the Hessian matrix $G(x) = \nabla^2 F(x)$ is positive definite. Then if $F(x_*) > 0$, the rate of convergence is superlinear and if $F(x_*) = 0$, the rate of convergence is quadratic [31].

8 Simple hybrid methods with Jacobian corrections

If the matrix B_i is ill-conditioned, then a more advantageous way is to use a full rank approximation A_i of the Jacobian matrix J_i and to replace the solution of the normal equation $d_i = -B_i^{-1} g_i$, where $B_i = A_i^T A_i$,

by the solution to the linear least-squares problem $d_i = -A_i^\dagger f_i$, where $A_i^\dagger$ is a Moore-Penrose pseudoinverse of A_i. This approach is not quite rigorous, since usually $A_i^\dagger f_i \neq B_i^{-1} g_i$. The equality $A_i^\dagger f_i = B_i^{-1} g_i$ is satisfied only if $A_i^T f_i = g_i = J_i^T f_i$. Using a variational principle, we derive an update which satisfies the quasi-Newton condition $A_{i+1}^T A_{i+1} s_i = y_i$ together with the condition $A_{i+1}^T f_{i+1} = g_{i+1} = J_{i+1}^T f_{i+1}$.

Let $B = A^T A$, where $A = J$ if the Gauss-Newton step is accepted, so $B = J^T J$ holds. To use the variational principle, we write the standard quasi-Newton condition $B_+ s = A_+^T A_+ s = y$ in the form

$$\sqrt{\gamma} A_+ s = \tilde{z}, \qquad \sqrt{\gamma} A_+^T \tilde{z} = \gamma y, \qquad \tilde{z}^T \tilde{z} = \gamma s^T y, \tag{94}$$

where $\gamma > 0$ is a scalar scaling parameter (as in (55)) and $\tilde{z} \in R^m$ is an arbitrary vector. Note that the last equality, which is a consequence of the first two equalities, is the only restriction on the choice of $\tilde{z}$.

Theorem 14. *Let W be an arbitrary symmetric positive definite matrix. Then the Frobenius norm $\|W^{-1/2}(\sqrt{\gamma} A_+ - A)^T\|_F$ is minimal on the set of all matrices satisfying quasi-Newton condition (94) if and only if*

$$\sqrt{\gamma} A_+^T = A^T - \frac{Ws}{s^T W s} \tilde{s}^T + \left(\gamma y - z + s^T z \frac{Ws}{s^T W s} \right) \frac{\tilde{z}^T}{\tilde{z}^T \tilde{z}}, \qquad \tilde{z}^T \tilde{z} = \gamma s^T y, \tag{95}$$

where $\tilde{s} = As$ and $z = A^T \tilde{z}$.

Proof This proof is similar to the proof of Theorem 3.1 proposed in [57]. Denote $X = \sqrt{\gamma} A_+^T$. Necessity will be proven using the Lagrangian function

$$\begin{aligned} \mathcal{L}(X, \tilde{u}, v) &= \frac{1}{2} \left\| W^{-1/2} \left(X - A^T \right) \right\|_F^2 + \tilde{u}^T \left(X^T s - \tilde{z} \right) + v^T \left(X \tilde{z} - \gamma y \right) \\ &= \sum_{i=1}^{m} \left[\frac{1}{2} \left(x_i - a_i \right)^T W^{-1} \left(x_i - a_i \right) + \tilde{u}_i s^T x_i + \tilde{z}_i v^T x_i \right] - \tilde{u}^T \tilde{z} - \gamma v^T y, \end{aligned}$$

where $A^T = [a_1, \ldots, a_m]$, $X = [x_1, \ldots, x_m]$ and where $\tilde{u}$, v are vectors of Lagrange multipliers. Differentiating the Lagrangian function we obtain

$$\frac{\partial \mathcal{L}(X, \tilde{u}, v)}{\partial x_i} = W^{-1} \left(x_i - a_i \right) + \tilde{u}_i s + \tilde{z}_i v.$$

Therefore, the conditions for stationarity of the Lagrangian function have the form $W^{-1}(x_i - a_i) + \tilde{u}_i s + \tilde{z}_i v = 0$, $1 \leq i \leq m$, or

$$X - A^T = -W s \tilde{u}^T - W v \tilde{z}^T.$$

Using the first condition from (94) we obtain

$$X^T s = As - s^T W s \tilde{u} - v^T W s \tilde{z} = \tilde{z} \quad \Rightarrow \quad \tilde{u} = \frac{1}{s^T W s} \left(As - (1 + v^T W s) \tilde{z} \right),$$

which after substitution to the previous equality gives

$$X - A^T = -\frac{Ws}{s^T W s} \tilde{s}^T + w \tilde{z}^T,$$

where $w \in R^n$ is an unknown vector (determined uniquely by the vector v). Using the second condition from (94) we obtain

$$X \tilde{z} = A^T \tilde{z} - s^T A^T \tilde{z} \frac{Ws}{s^T W s} + \tilde{z}^T \tilde{z} w = \gamma y \quad \Rightarrow \quad w = \frac{1}{\tilde{z}^T \tilde{z}} \left(\gamma y - z + s^T z \frac{Ws}{s^T W s} \right),$$

which after substitution to the previous equality (with using relation $X = \sqrt{\gamma} A_+^T$) gives (95). Sufficiency follows from the convexity of the Frobenius norm. □

Update (95) contains two vector parameters $w = Ws/s^T Ws$ and $\tilde{z}$. These parameters should be chosen in such a way to guarantee the condition $A_+^T f_+ = g_+$.

Lemma 2. *Equalities*

$$\sqrt{\gamma}A_+s = \tilde{z}, \qquad \sqrt{\gamma}A_+^T\tilde{z} = \gamma y, \qquad A_+^T f_+ = g_+ \tag{96}$$

can be satisfied simultaneously only if

$$f_+^T f_+ s^T y \geq (s^T g_+)^2. \tag{97}$$

Proof From the first two equalities in (96), the relation $\tilde{z}^T\tilde{z} = \gamma s^T y$ follows, which determines the norm of vector $\tilde{z}$. The first and the third equalities imply $f_+^T\tilde{z} = \sqrt{\gamma}f_+^T A_+ s = \sqrt{\gamma}s^T g_+$. Since the distance of the hyperplane $f_+^T\tilde{z} = \sqrt{\gamma}s^T g_+$ from the origin is equal to $\sqrt{\gamma}|s^T g_+|/\|f_+\|$, the norm of vector $\tilde{z}$ cannot be smaller than this number, which together with equality $\|\tilde{z}\| = \sqrt{\gamma s^T y}$ gives $\sqrt{\gamma}|s^T g_+|/\|f_+\| \leq \sqrt{\gamma s^T y}$, or $f_+^T f_+ s^T y \geq (s^T g_+)^2$. □

Remark 23. If a perfect line search method is used, i.e., if equality $s_i^T g_{i+1} = 0$ holds in every iteration, then $s_i^T y_i = s_i^T g_{i+1} - s_i^T g_i = -s_i^T g_i > 0$ and condition (97) is satisfied. If the strong Wolfe condition is satisfied [42], then $|s_i^T g_{i+1}| \leq \varepsilon_2 |s_i^T g_i|$ holds in every iteration, so $s_i^T y_i = s_i^T g_{i+1} - s_i^T g_i \geq (1-\varepsilon_2)|s_i^T g_i|$, and condition (97) is satisfied whenever

$$f_{i+1}^T f_{i+1} \geq \frac{\varepsilon_2^2}{1-\varepsilon_2}|s_i^T g_i|. \tag{98}$$

If $x_i \to x_*$ (so $g_i \to 0$ and $s_i \to 0$) and $F(x_*) > 0$, there exists an index $k \in N$ such that condition (98) (and therefore also condition (97)) is satisfied for all $i \geq k$. In our numerical experiments, condition (97) was always satisfied if $F_i - F_{i+1} \leq \vartheta F_i$ with $\vartheta = 0.0005$.

Theorem 15. *Let vectors f_+ and As be linearly independent and assume that inequality (97) holds. If we use the vectors*

$$\tilde{z} = \sqrt{\gamma}(\lambda_1 f_+ + \lambda_2 As), \tag{99}$$

where

$$\lambda_2^2 = \frac{s^T y f_+^T f_+ - (s^T g_+)^2}{f_+^T f_+ s^T A^T A s - (s^T A^T f_+)^2}, \qquad \lambda_1 = \frac{s^T g_+ - \lambda_2 s^T A^T f_+}{f_+^T f_+}, \tag{100}$$

and

$$\frac{Ws}{s^T W s} = \frac{\gamma s^T y(A^T f_+ - \sqrt{\gamma}g_+) + \sqrt{\gamma}s^T g_+(\gamma y - A^T\tilde{z})}{\gamma s^T y s^T A^T f_+ - \sqrt{\gamma}s^T g_+ s^T A^T\tilde{z}}, \tag{101}$$

in formula (95), then equalities (96) hold.

Proof Vector $\tilde{z}$ has to satisfy equalities $f_+^T\tilde{z} = \sqrt{\gamma}s^T g_+$ and $\tilde{z}^T\tilde{z} = \gamma s^T y$. Setting $\tilde{z} = \sqrt{\gamma}(\lambda_1 f_+ + \lambda_2 As)$, we obtain the system of equations

$$\begin{aligned} \lambda_1 f_+^T f_+ + \lambda_2 s^T A^T f_+ &= \sqrt{\gamma}s^T g_+, \\ \lambda_1^2 f_+^T f_+ + 2\lambda_1\lambda_2 s^T A^T f_+ + \lambda_2^2 s^T A^T A s &= \gamma s^T y \end{aligned}$$

for unknowns λ_1 and λ_2. Since the vectors f_+ and As are linearly independent, these equations have the unique solution given by (100). Update (95) satisfies the first two equalities in (96) (Theorem 14). Using the third equality, we obtain

$$\begin{aligned} \sqrt{\gamma}g_+ &= A^T f_+ - \frac{Ws}{s^T W s}s^T A^T f_+ + \left(\gamma y - A^T\tilde{z} + s^T A^T\tilde{z}\frac{Ws}{s^T W s}\right)\frac{\tilde{z}^T f_+}{\tilde{z}^T\tilde{z}} \\ &= A^T f_+ - \left(s^T A^T f_+ - s^T A^T\tilde{z}\frac{\sqrt{\gamma}s^T g_+}{\gamma s^T y}\right)w + \left(\gamma y - A^T\tilde{z}\right)\frac{\sqrt{\gamma}s^T g_+}{\gamma s^T y}, \end{aligned}$$

where $w = Ws/s^T Ws$. This relation implies that

$$w = \lambda\left(A^T f_+ - \sqrt{\gamma}g_+ + \left(\gamma y - A^T\tilde{z}\right)\frac{\sqrt{\gamma}s^T g_+}{\gamma s^T y}\right), \tag{102}$$

where λ is an unknown multiplier, and since $s^T w = s^T W s / s^T W s = 1$, one can write

$$\lambda \left(\gamma s^T y s^T A^T f_+ - \sqrt{\gamma} s^T g_+ s^T A^T \tilde{z} \right) = \gamma s^T y.$$

Substituting this value λ into (102), we obtain (101). □

Description 4. Theorem 15 is a basis for an efficient simple hybrid method with Jacobian corrections. Let $A_1 = J_1$ and $\vartheta > 0$. Denote

$$N_5' = \{i \in N : (F_i - F_{i+1})/F_i \geq \vartheta_i\},$$

where

$$\vartheta_i = \min \left(\vartheta, 1 - \frac{(s_i^T g_{i+1})^2}{s_i^T y_i f_i^T f_i} \right),$$

and set

$$\begin{array}{lll} A_{i+1} = J_{i+1} & \text{if} \quad i \in N_2, \quad i \in N_5', \\ A_{i+1} = \mathcal{A}(A_i, y_i, s_i, \tilde{z}_i, w_i, \gamma_i) & \text{if} \quad i \in N_2, \quad i \notin N_5', \\ A_{i+1} = A_i & \text{if} \quad i \notin N_2, \end{array} \tag{103}$$

where $\mathcal{A}(A, y, s, \tilde{z}, w, \gamma)$ is (transposed) update (95) with $\tilde{z}$ and $w = Ws/s^T Ws$ given by (99)–(101). If $i \notin N_5'$, then

$$f_{i+1}^T f_{i+1} \geq (1 - \vartheta_i) f_i^T f_i \geq \frac{(s_i^T g_{i+1})^2}{s_i^T y_i f_i^T f_i} f_i^T f_i = \frac{(s_i^T g_{i+1})^2}{s_i^T y_i},$$

so condition (97) is satisfied.

Remark 24. The approximation A_+ of the Jacobian matrix J_+ satisfying condition $A_+^T f_+ = J_+^T f_+ = g_+$ can be also determined by residual adjoint quasi-Newton updates [48] (e.g., by the two-sided adjoint quasi-Newton update [39]). Let $A_1 = J_1$. If $(F_i - F_{i+1})/F_i > \vartheta_i$, we set

$$A_{i+1} = J_{i+1}.$$

If $(F_i - F_{i+1})/F_i \leq \vartheta_i$, we set

$$A_{i+1} = A_i + \frac{(f_{i+1} - f_i - As)(g_{i+1} - A_i f_{i+1})^T}{f_{i+1}^T (f_{i+1} - f_i - A_i s_i)}.$$

However, numerical experiments show that hybrid methods with adjoint quasi-Newton updates are less efficient then hybrid methods described in Description 4, so we do not recommend them.

9 Structured hybrid methods with Jacobian corrections

In this section we will concentrate on structured variable metric methods that utilize knowledge of the Jacobian matrix [58]. In order to express these methods in multiplicative form, we set $A = J + L$, $A_+ = J_+ + L_+$ and update the matrix L in such a way that

$$B_+ s = A_+^T A_+ s = (J_+ + L_+)^T (J_+ + L_+) s = y.$$

Since in the case of a sum of squares, only the BFGS method can be efficiently realized in multiplicative form, we will restrict ourselves to a subclass of variable metric methods that contains the BFGS method and for which deriving multiplicative form is much simpler than in the general case. For deriving a multiplicative form we will use a variational principle [58]. In order to use it, we write a quasi-Newton condition in the form

$$(J_+ + L_+)^T \tilde{z} = y, \qquad (J_+ + L_+) s = \tilde{z}, \qquad \tilde{z}^T \tilde{z} = s^T y, \tag{104}$$

where $\tilde{z} \in R^m$ is an optional vector (parameter). Note that the last equality, which is a consequence of the first two equalities, is the only restriction put on the choice of the vector $\tilde{z}$. The following theorem is a modification of a similar theorem proposed in [58].

Theorem 16. *Let W be a symmetric positive definite matrix. Then the Frobenius norm $\|W^{-1/2}(L_+ - L)\|_F$ is minimal on the set of all matrices meeting the equality $(J_+ + L_+)^T \tilde{z} = y$ if and only if*

$$L_+ = L + \frac{W\tilde{z}(y - \bar{A}^T\tilde{z})^T}{\tilde{z}^T W\tilde{z}}, \tag{105}$$

where $\bar{A} = J_+ + L$. Quasi-Newton condition (104) is satisfied in this case if and only if the vector $W\tilde{z}$ is parallel with the vector $\tilde{z} - \bar{A}s$, where $\tilde{z}^T\tilde{z} = y^T s$, so

$$L_+ = L + \frac{(\tilde{z} - \bar{A}s)(y - \bar{A}^T\tilde{z})^T}{\tilde{z}^T(\tilde{z} - \bar{A}s)}. \tag{106}$$

Proof (a) Necessity of the first part of the assertion will be proved using the Lagrangian function

$$\begin{aligned}\mathcal{L}(L_+, u) &= \frac{1}{2}\left\|W^{-1/2}(L_+ - L)\right\|_F^2 + u^T\left((J_+ + L_+)^T\tilde{z} - y\right) \\ &= \sum_{i=1}^{n}\left[\frac{1}{2}\left(l_i^+ - l_i\right)^T W^{-1}\left(l_i^+ - l_i\right) + u_i\tilde{z}^T l_i^+\right] + u^T(J_+^T\tilde{z} - y),\end{aligned}$$

where $L_+ = [l_1^+, \ldots, l_n^+]$ and $L = [l_1, \ldots, l_n]$. Sufficiency is then an immediate consequence of the convexity of the Frobenius norm. Differentiating the Lagrangian function we obtain

$$\frac{\partial\mathcal{L}(L_+, u)}{\partial l_i^+} = W^{-1}\left(l_i^+ - l_i\right) + u_i\tilde{z}, \quad 1 \le i \le n.$$

Thus, the condition for stationarity of the Lagrangian function has the form $W^{-1}(l_i^+ - l_i) + u_i\tilde{z} = 0$, $1 \le i \le n$, or

$$A_+ - \bar{A} = L_+ - L = -W\tilde{z}u^T.$$

From equation $A_+^T\tilde{z} = y$ we obtain $(A_+ - \bar{A})^T\tilde{z} = -\tilde{z}^T W\tilde{z}u = y - \bar{A}^T\tilde{z}$, so

$$u = -\frac{y - \bar{A}^T\tilde{z}}{\tilde{z}^T W\tilde{z}},$$

which after substitution into the previous equation gives

$$A_+ - \bar{A} = L_+ - L = \frac{W\tilde{z}(y - \bar{A}^T\tilde{z})^T}{\tilde{z}^T W\tilde{z}}.$$

(b) Suppose that quasi-Newton condition (104) is satisfied, so $(A_+ - \bar{A})s = \tilde{z} - \bar{A}s$. Then

$$\frac{W\tilde{z}(y - \bar{A}^T\tilde{z})^T s}{\tilde{z}^T W\tilde{z}} = \tilde{z} - \bar{A}s. \tag{107}$$

From this expression it is obvious that the vector $W\tilde{z}$ is parallel with the vector $\tilde{z} - \bar{A}s$. Since the matrix W can be multiplied by an arbitrary number without changing the fraction on the left-hand side, we can set $W\tilde{z} = \tilde{z} - \bar{A}s$ (it can be performed only if $W \neq I$, the case $W = I$ is investigated in Remark 25). On the other hand, let $W\tilde{z} = \tilde{z} - \bar{A}s$ and $\tilde{z}^T\tilde{z} = s^T y$. Then (106) holds and

$$A_+ s = \bar{A}s + (\tilde{z} - \bar{A}s)\frac{(y - \bar{A}^T\tilde{z})^T s}{\tilde{z}^T(\tilde{z} - \bar{A}s)} = \bar{A}s + (\tilde{z} - \bar{A}s) = \tilde{z},$$

so the second condition in (104) is satisfied as well. □

Remark 25. Letting $W = I$ we obtain the BFGS method. Then, by (107), the vector $\tilde{z}$ is parallel with the vector $\bar{A}s$, or $W\tilde{z} = \tilde{z} = \lambda\bar{A}s$. From the last condition in (104), the equality $\tilde{z}^T\tilde{z} = \lambda^2 s^T\bar{A}^T\bar{A}s = s^T y$ follows, so $\lambda^2 = s^T y/\bar{s}^T\bar{s}$, where $\bar{s} = \bar{A}s$. Substituting vector $W\tilde{z} = \tilde{z} = \lambda\bar{s}$ into (105) we obtain

$$L_+ = L + \frac{\lambda\bar{s}(y - \lambda\bar{A}^T\bar{s})^T}{s^T y} = L - \frac{\bar{s}}{\bar{s}^T\bar{s}}\left(\bar{A}^T\bar{s} \pm \sqrt{\frac{\bar{s}^T\bar{s}}{s^T y}}y\right)^T. \tag{108}$$

A certain disadvantage of update (108) is that a solution to a linear least squares problem $(J+L)s+f \approx 0$ is not a solution to a normal system of equations $(J+L)^T(J+L)s = -g = -J^T f$, which is used for computation of a direction vector. Thus, neither efficient methods based on QR decomposition nor the LSQR method [43] can be used. This disadvantage can be removed by choosing a matrix L such that $(J+L)^T f = J^T f$, or $L^T f = 0$. Thus, it is advantageous to add a constraint $L_+^T f_+ = 0$ into a variational problem defining the BFGS method [50]. If $L_+^T f_+ = 0$, then minimization of the Frobenius norm $\|L_+ - L\|_F$ is equivalent to minimization of the Frobenius norm $\|P(L_+ - L)\|_F$, where $P = I - f_+ f_+^T/f_+^T f_+$ is an orthogonal projection matrix (remind that $P^2 = P$). From this it follows that $PL_+ = L_+$, so

$$\begin{aligned}(L_+ - L)^T P(L_+ - L) &= L_+^T PL_+ - L^T PL_+ - L_+^T PL + L^T PL \\ &= L_+^T L_+ - L^T L_+ - L_+^T L + L^T PL \\ &= (L_+ - L)^T(L_+ - L) + L^T(P - I)L,\end{aligned}$$

where the last term is independent of L_+. The following theorem is introduced in [50].

Theorem 17. *The Frobenius norm $\|P(L_+ - L)\|_F$ is minimal on the set of all matrices meeting quasi-Newton condition (104) and the constraint $L_+^T f_+ = 0$ if and only if*

$$L_+ = PL - \frac{\tilde{s}}{\tilde{s}^T\tilde{s}}\left(\bar{A}^T\tilde{s} \pm \sqrt{\frac{\tilde{s}^T\tilde{s}}{s^T\tilde{y}}}\tilde{y}\right)^T, \tag{109}$$

where $\bar{A} = J_+ + L$, $\bar{s} = \bar{A}s$, and

$$\tilde{s} = P\bar{s} = \bar{s} - \frac{f_+^T\bar{s}}{f_+^T f_+}f_+, \qquad \tilde{y} = y - \frac{J_+^T f_+(J_+^T f_+)^T s}{f_+^T f_+} = y - \frac{g_+^T s}{f_+^T f_+}g_+. \tag{110}$$

Proof (a) First, we show that if $(J_+ + L_+)s = \tilde{z}$ and $L_+^T f_+ = 0$, then the condition $(J_+ + L_+)^T\tilde{z} = y$ is equivalent to the condition $(J_+ + L_+)^T P\tilde{z} = \tilde{y}$. Actually, $f_+^T J_+ s = f_+^T\tilde{z}$ follows from $(J_+ + L_+)s = \tilde{z}$ and $L_+^T f_+ = 0$, so

$$\begin{aligned}(J_+ + L_+)^T P\tilde{z} - \tilde{y} &= J_+^T\tilde{z} - \frac{J_+^T f_+ f_+^T J_+ s}{f_+^T f_+} + L_+^T P\tilde{z} - y + \frac{J_+^T f_+ f_+^T J_+ s}{f_+^T f_+} \\ &= J_+^T\tilde{z} + L_+^T P\tilde{z} - y = (J_+ + L_+)^T\tilde{z} - y.\end{aligned}$$

Note that the relation $\tilde{z}^T P\tilde{z} = s^T\tilde{y}$ follows from equations $(J_+ + L_+)s = \tilde{z}$ and $(J_+ + L_+)^T P\tilde{z} = \tilde{y}$.
(b) Necessity will be proved using the Lagrangian function

$$\begin{aligned}\mathcal{L}(L_+, u) &= \frac{1}{2}\|P(L_+ - L)\|_F^2 + u^T\left((J_+ + L_+)^T P\tilde{z} - \tilde{y}\right) \\ &= \sum_{i=1}^{n}\left[\frac{1}{2}\left(l_i^+ - l_i\right)^T P\left(l_i^+ - l_i\right) + u_i\tilde{z}^T Pl_i^+\right] + u^T(J_+^T P\tilde{z} - \tilde{y}),\end{aligned}$$

where $L_+ = [l_1^+, \ldots, l_n^+]$ and $L = [l_1, \ldots, l_n]$. Sufficiency is then an immediate consequence of the convexity of the Frobenius norm. Differentiating the Lagrangian function we obtain

$$\frac{\partial \mathcal{L}(L_+, u)}{\partial l_i^+} = P\left(l_i^+ - l_i\right) + u_i P\tilde{z}.$$

Thus, the condition for stationarity of the Lagrangian function has the form $P(l_i^+ - l_i) + u_i P\tilde{z} = 0$, $1 \leq i \leq n$, or

$$P(L_+ - L) = -P\tilde{z}u^T.$$

From equation $(J_+ + L_+)^T P\tilde{z} = \tilde{y}$ we obtain $(L_+ - L)^T P\tilde{z} = -\tilde{z}^T P\tilde{z}u = \tilde{y} - P\bar{A}^T\tilde{z}$, so

$$u = -\frac{\tilde{y} - P\bar{A}^T\tilde{z}}{\tilde{z}^T P\tilde{z}},$$

which after substitution into the previous equation gives

$$P(L_+ - L) = \frac{P\tilde{z}(\tilde{y} - \tilde{A}^T P\tilde{z})^T}{\tilde{z}^T P\tilde{z}}. \tag{111}$$

If we use the second condition in (104), we obtain $P(L_+ - L)s = P\tilde{z} - P\bar{A}s$, so we can write

$$\frac{P\tilde{z}(\tilde{y} - \bar{A}^T P\tilde{z})^T s}{\tilde{z}^T P\tilde{z}} = P\tilde{z} - P\bar{A}s = P\tilde{z} - \tilde{s}.$$

From the last expression it is obvious that the vector $P\tilde{z}$ is parallel with the vector $\tilde{s}$, or $P\tilde{z} = \lambda\tilde{s}$. Using the relation $\tilde{z}^T P\tilde{z} = s^T\tilde{y}$, proved in (a), we can write

$$\lambda^2\tilde{s}^T\tilde{s} = \tilde{z}^T P\tilde{z} = s^T\tilde{y} \quad \Rightarrow \quad \lambda = \pm\sqrt{\frac{s^T\tilde{y}}{\tilde{s}^T\tilde{s}}},$$

which after substitution into $P\tilde{z} = \lambda\tilde{s}$ and then into (111) proves the assertion of theorem. □

Remark 26. Quasi-Newton methods based on Jacobian approximations described in [6], [39] work surprisingly well for solving nonlinear least squares problem as it is demonstrated in Table 7. Matrices $A_i \approx J_i$ are generated using quasi-Newton updates and a direction vector is determined by solving a linear system $A_i d_i + f_i \approx 0$. These methods usually require more (up to twice the amount of) function evaluations compared to the methods stated in this section. If we use orthogonal decomposition updates, as described in [13], then the total computational time is usually shorter, see the results of numerical experiments presented in Table 7 in Section 10.

10 Numerical comparison of methods for least squares problems

All methods described in this report are implemented in the software system for Universal Functional Optimization (UFO) [37] together with collections of problems for their testing. We have used collection TEST24 [35] for our computational experiments presented in this section. Collection TEST24 contains 80 different problems with optional dimensions. We have used the first 30 problems with 200 variables ($n = 200$) which are a mixture of both zero and nonzero residual problems. The remaining 50 problems are all zero residual, so they are not suitable for testing hybrid methods.

The tested methods are distinguished by the following codes:

- `VM` - variable metric methods,
- `GN` - the Gauss-Newton method,
- `GB` - simple hybrid methods,
- `GS` - structured hybrid methods,
- `GL` - simple hybrid method with Jacobian corrections,
- `QB` - the good Broyden quasi-Newton method [6],
- `QL` - the quasi-Newton method proposed in [39].

Individual dog-leg realizations of these methods are distinguished by the following codes:

- `GM1` - update of an approximation of the Hessian matrix followed by the LDL^T decomposition,
- `GG1` - update of the LDL^T (Choleski) decomposition,
- `GR1` - update of the R^TR decomposition obtained by orthogonal transformations,
- `GA1` - update of an approximation of the Jacobian matrix followed by the QR decomposition,
- `GQ1` - update of the QR (orthogonal) decomposition.

The strategy `GM1` is used in connection with the Gauss-Newton method or with structured hybrid methods. In this case, the matrix B is factorized in every iteration by the standard Choleski decomposition, if B is positive definite, or by the Gill-Murray decomposition [25], if B is not positive definite (after the R1 update). The `GG1` realization is intended for variable metric or simple hybrid methods. In this case, the matrix $B = J^TJ$, obtained using the Gauss-Newton method, is factorized in the form $B = LDL^T$. The matrices L (lower triangular) and D (diagonal) are stored and updated by the method proposed in [26] if the variable metric method is used. In the `GR1` strategy, which is advantageous for the Gauss-Newton method or for simple hybrid methods, the factorization $B = R^TR$ is obtained recursively using gradients of functions $f_k(x)$, $1 \le k \le m$, by the method proposed in [3] and introduced also in [32]. The `GA1` realization is intended for methods with Jacobian corrections. In this case, the complete orthogonal decomposition of the matrix A is performed in every iteration by the method described in [5] and used also in [2]. The strategy `GQ1` is used in connection with the quasi-Newton methods. In this case, the matrix $A = J$, obtained using the Gauss-Newton method, is factorized in the form $A = QR$. The matrices Q (orthogonal) and R (upper triangular) are stored and updated by the method proposed in [13] if the quasi-Newton method is used. In the implementation of tested methods, we have used a simplified trust region method described in Remark 3 and the values $\underline{\rho} = 0.1$, $\overline{\rho} = 0.9$, $\overline{\Delta} = 1000$ (this value was decreased for some problems), $\underline{\gamma} = 0.7$ (this value gave better results than $\underline{\gamma} = 1$), $\overline{\gamma} = 6.0$, $\vartheta = 0.0005$, $c = 10^{-32}$ (the square of machine precision) and $K = 10^{32}$.

The test results are presented in Tables 1–7 whose columns mean the tested method with the chosen strategy S1, S2, S3, S4 (values 1 or 2), the scaling used (Y - yes, N - no), the total number of function evaluations `NFV` (which is equal to the total number of iterations of the trust region method), the total number of gradient evaluations `NFG`, the total number of matrix decompositions `NDC`, the total computation time and sometimes the number of failures (the number of problems that were not solved). The asterisk in some tables indicates that at least one problem was not solved (the maximum number of function evaluations was exceeded). Note that the problems in the TEST24 collection are essentially sparse, so the total computational time is primarily determined by the number of arithmetic operations in matrix computations. When we need to solve problems where computation of function values consumes most of the machine time, then the values `NFV` and `NFG` are crucial (the computational time is then proportional to these values).

Table 1 contains the results of comparison of particular variable metric methods defined in Remark 5. These methods use updates of Choleski decomposition, so there is no need to perform this decomposition.

Update	S1	Scaling	NFV	NFG	NDC	Time	Fail
BFGS	1	N	12683	11102	-	4.85	-
BFGS	1	Y	9139	8340	-	3.82	-
BFGS	2	N	12448	12434	-	5.19	-
BFGS	2	Y	8172	8160	-	3.58	-
DFP	1	N	82844	81670	-	33.50	14
DFP	1	Y	22001	20457	-	9.53	2
DFP	2	N	107081	107079	-	44.55	20
DFP	2	Y	21304	21284	-	9.62	2
H	1	N	11585	10465	-	4.46	-
H	1	Y	8510	7933	-	3.52	-
H	2	N	13194	13184	-	5.36	-
H	2	Y	8541	8527	-	3.58	-
DW	1	N	11377	10121	-	5.10	-
DW	1	Y	8366	7817	-	3.83	-
DW	2	N	12729	12717	-	6.17	-
DW	2	Y	8389	8378	-	3.99	-

Table 1: Variable metric methods (VM)

			Choleski decomposition				Recursive QR decomposition			
Update	S1	Scaling	NFV	NFG	NDC	Time	NFV	NFG	NDC	Time
BFGS	1	N	2295	2061	1547	14.27	4227	3382	2146	7.39
BFGS	1	Y	2280	2023	1543	14.07	4201	3356	2159	7.72
BFGS	2	N	2353	2345	1574	14.71	4963	4953	2535	11.26
BFGS	2	Y	2371	2362	1578	15.14	4993	4978	2543	11.45
DFP	1	N	2738	2512	1472	14.11	3739	2898	2079	7.23
DFP	1	Y	2427	2118	1451	13.75	3467	2741	2043	6.89
DFP	2	N	2610	2601	1496	14.43	9225	9512	2228	*9.40
DFP	2	Y	2351	2344	1483	14.07	3799	3787	2083	6.08
H	1	N	2090	1881	1428	13.05	3609	2906	2213	6.55
H	1	Y	2096	1837	1428	13.33	3592	2872	2221	6.58
H	2	N	2289	2279	1560	14.74	3332	3322	2078	5.50
H	2	Y	2204	2196	1563	14.47	3239	3224	2059	5.49
DW	1	N	2054	1838	1426	13.35	3340	2659	2047	6.34
DW	1	Y	2051	1836	1429	13.07	3342	2628	2035	6.31
DW	2	N	2229	2218	1556	14.52	3329	3316	2109	5.56
DW	2	Y	2194	2186	1554	14.55	3330	3315	2155	5.56
GN	-	-	3714	3323	3207	29.59	4480	3936	3936	*9.26

Table 2: Simple hybrid methods (GB)

In Table 2 we can see comparison of simple hybrid methods (with BFGS, DFP, H, DW updates) using different strategies (S1 and scaling) namely in two versions. In the first version, a matrix $J^T J$ is constructed and its Choleski decomposition $J^T J = LDL^T$ is computed. In the second version, decomposition $J^T J = R^T R$ is directly recursively determined using orthogonal transformations. In both cases, variable metric methods update a triangular decomposition, so arithmetic operations are saved. In Table 2, the results obtained by the Gauss-Newton method are also shown.

The next two tables contain comparison of different structured hybrid methods (Table 3 the first way and Table 4 the second way) for S1=1 and S2=1, with different strategies S3, S4 and without scaling

realized using updates from the Broyden class (formulas (72), (88), (82), (91)) or from the Dennis class (formulas (74), (90), (84), (93)). Here H/PS means that the results of both the Hoshino method from the Broyden class and the PSB method from the Dennis class are presented in the same row.

			Broyden class (72) or (88)				Dennis class (74) or (90)			
Update	S3	S4	NFV	NFG	NDC	Time	NFV	NFG	NDC	Time
BFGS	1	1	2523	2314	2313	21.18	2522	2312	2312	20.85
BFGS	1	2	2882	2618	2609	24.07	2878	2619	2614	24.05
BFGS	2	1	2865	2621	2609	23.55	2885	2636	2620	23.68
BFGS	2	2	3191	2896	2884	26.50	3190	2895	2884	26.44
DFP	1	1	3185	2928	2927	26.50	2904	3216	2929	26.50
DFP	1	2	4111	3780	3795	*34.88	4141	3814	3824	*35.20
DFP	2	1	2624	2383	2380	21.54	2548	2317	2310	20.88
DFP	2	2	4738	4499	4491	*43.88	4739	4480	4475	*43.68
H/PS	1	1	2845	2565	2557	23.19	2253	2068	2083	18.56
H/PS	1	2	2947	2680	2669	24.69	2319	2185	2180	19.70
H/PS	2	1	3200	2980	2973	*26.80	2378	2118	2167	19.03
H/PS	2	2	3470	3211	3204	*29.37	2389	2205	2209	19.82
R1	1	1	2241	1995	2014	17.83	2205	1997	2018	17.90
R1	1	2	2141	1975	1963	17.72	2201	1990	1982	17.86
R1	2	1	2221	1993	2005	17.81	2188	1995	2005	17.79
R1	2	2	2337	2028	2029	18.19	2228	2007	2001	17.95

Table 3: Structured hybrid methods (GS) - the first way

			Broyden class (82) or (91)				Dennis class (84) or (93)			
Update	S3	S4	NFV	NFG	NDC	Time	NFV	NFG	NDC	Time
BFGS	1	1	2215	2001	2010	18.15	2220	1996	2005	18.14
BFGS	1	2	2254	2054	2046	18.75	2266	2064	2058	18.84
BFGS	2	1	2222	1997	2014	17.51	2255	2016	2042	17.73
BFGS	2	2	3314	3012	2025	*25.00	3289	3017	2038	*25.06
DFP	1	1	2180	2039	2045	18.51	2183	2045	2060	18.61
DFP	1	2	3494	3187	3201	*30.29	3453	3183	3187	*30.26
DFP	2	1	2441	2094	2181	18.56	2386	2075	2123	18.33
DFP	2	2	3860	3433	3574	*30.95	2756	2464	2478	22.13
H/PS	1	1	2198	1993	1985	18.08	2253	2068	2083	18.82
H/PS	1	2	2369	3131	2128	19.53	2319	2185	2180	19.99
H/PS	2	1	3202	2913	2068	24.25	2378	2118	2167	18.72
H/PS	2	2	2313	2083	2117	18.45	2389	2205	2209	19.51
R1	1	1	2240	1995	2009	18.14	2192	1990	2005	18.09
R1	1	2	2212	1992	1992	18.12	2154	1987	1979	18.04
R1	2	1	2202	1982	1995	17.36	2214	1987	1996	17.42
R1	2	2	2315	2021	2027	17.80	2311	2029	2031	17.87

Table 4: Structured hybrid methods (GS) - the second way

Table 5 contains more detailed comparison of different structured hybrid methods with BFGS and R1 updates, realized using formulas (72) and (93), depending on the choice of strategies S1, S2 and on scaling (the value (80) is used as a scaling parameter). In Table 5, the choice S3=1 is used.

				Broyden class (72)				Dennis class (93)			
Method	S1	S2	Scaling	NFV	NFG	NDC	Time	NFV	NFV	NDC	Time
BFGS	1	1	N	2523	2314	2313	21.18	2220	1996	2005	18.14
BFGS	1	1	Y	2514	2312	2310	21.06	2166	1992	1992	18.09
BFGS	1	2	N	2622	2391	2778	21.55	4136	3929	1994	*32.13
BFGS	1	2	Y	2946	2593	2642	23.43	4932	4740	1883	*41.82
BFGS	2	1	N	2523	2511	2246	22.54	6998	6995	1874	*54.83
BFGS	2	1	Y	2590	2575	2269	23.22	7078	7072	1992	*55.49
BFGS	2	2	N	3172	3160	2898	*28.06	7916	7907	1798	*63.18
BFGS	2	2	Y	3603	3582	2236	*30.12	7854	7851	1720	*64.30
R1	1	1	N	2141	1975	1963	17.72	2154	1987	1979	18.04
R1	1	1	Y	2127	1987	1984	17.98	2135	1989	1978	18.09
R1	1	2	N	2279	2013	2046	18.03	2258	2017	2042	18.22
R1	1	2	Y	2324	2008	2035	17.97	2260	2011	2037	18.15
R1	2	1	N	2490	2484	2231	21.66	2496	2484	2142	21.94
R1	2	1	Y	2518	2512	2226	21.91	2453	2441	2122	21.68
R1	2	2	N	2698	2683	2292	23.93	2761	2743	2283	24.57
R1	2	2	Y	2583	2574	2296	22.54	2692	2677	2315	24.05

Table 5: Structured hybrid methods (GS) - various strategies

The next two tables contain comparison of selected methods using Choleski decomposition or recursive QR decomposition. The results stated in these tables are also demonstrated in more detail using graphs showing the performance profiles [20].

Method	Update	S1	S2	S3	S4	Scaling	NFV	NFG	NDC	Time	Fail
GN	-	-	-	-	-	-	3714	3323	3207	29.59	-
VM	H	1	-	-	-	Y	8510	7933	-	3.52	-
GB	DW	1	-	-	-	Y	2051	1836	1429	13.07	-
GS	R1	1	1	1	2	N	2141	1975	1963	17.72	-

Table 6a: Methods with the Choleski decomposition

Method	Update	S1	S2	S3	S4	Scaling	NFV	NFG	NDC	Time	Fail
GN	-	-	-	-	-	-	4480	3936	3936	9.26	1
VM	H	2	-	-	-	Y	8541	8527	-	3.63	-
GB	H	2	-	-	-	Y	3239	3224	2059	5.49	-

Table 6b: Methods with the recursive QR decomposition

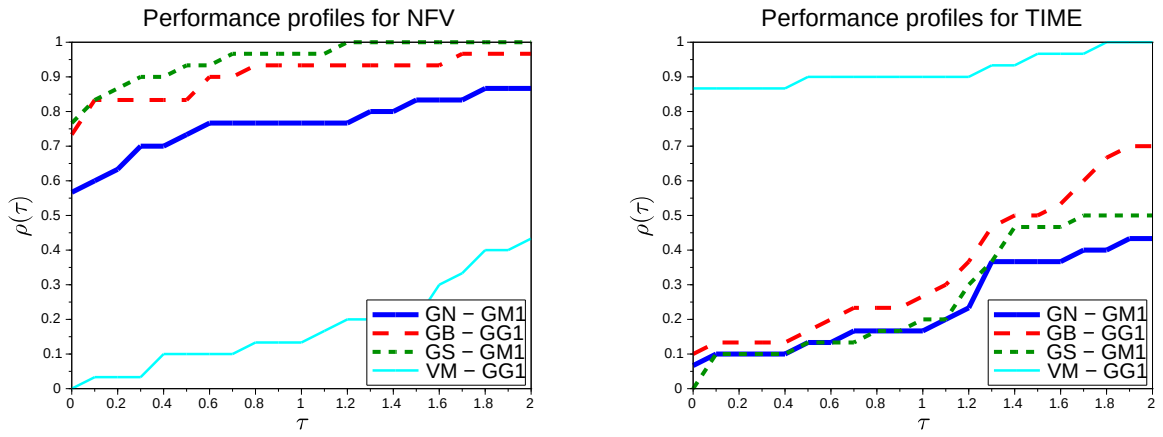

Figure 1a: Methods with Choleski decomposition

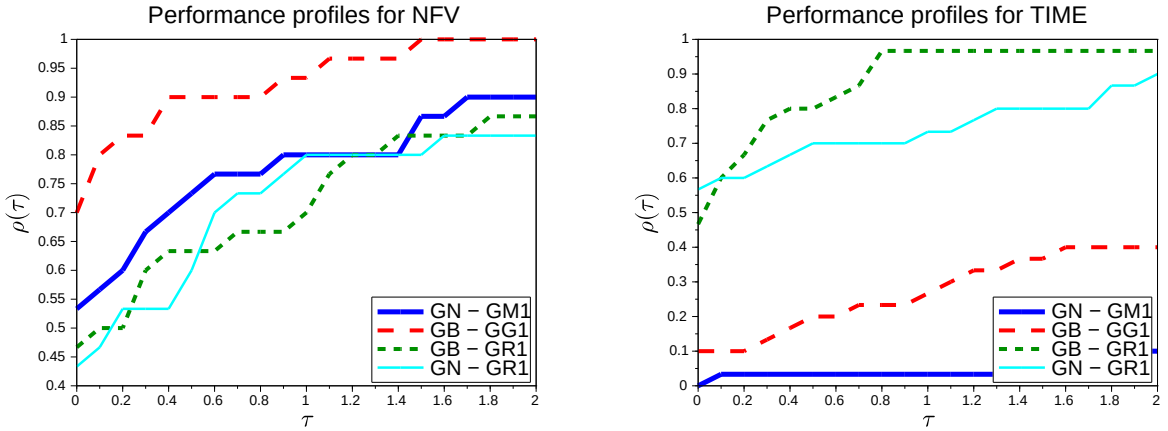

Figure 1b: Methods with recursive QR decomposition

Method	S2	S1	NFV	NFG	NDC	Time	Fail
GL	-	1	1875	1741	1648	22.15	-
GL	-	2	1963	1959	1813	24.02	-
GS	1	1	1935	1746	1663	21.93	-
GS	1	2	1763	1760	1634	21.43	-
GS	2	1	1875	1752	1690	23.15	-
GS	2	2	1799	1796	1668	22.83	-
QB	-	-	7636	1026	1019	22.35	1
QL	-	-	6858	7593	760	21.10	1
GN-C	-	-	2683	2494	2391	33.11	-
GN-R	-	-	2751	2571	2460	141.60	-

Table 7: Hybrid methods with Jacobian corrections

The last table contains the results obtained using the methods with Jacobian corrections. Besides the methods described in Section 8 and Section 9, using strategies S1 and S2, the results obtained using the quasi-Newton methods QB [6] and QL [39] are also stated. These methods, originally developed for solving systems of nonlinear equations, can be also used for minimization of the sum of squares. However,

a generalized update of orthogonal decomposition described in [13] must be used. In Table 7, the results obtained by the Gauss-Newton method using the QR decomposition of the Jacobian matrix stored column-wise (GN-C) or row-wise (GN-R) are also stated.

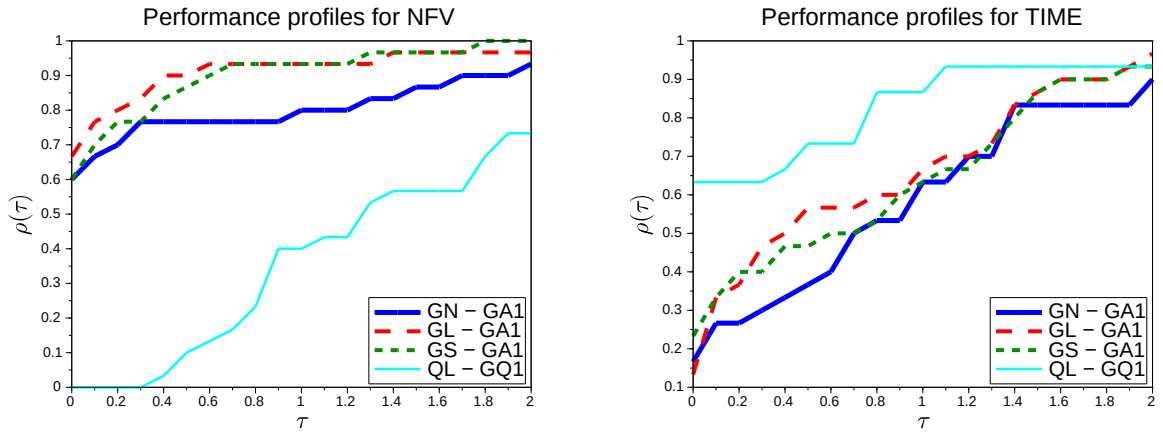

Figure 2: Hybrid methods with Jacobian corrections

From the results in this section we can draw several conclusions.

- Variable metric methods are relatively reliable and their advantage is that matrix decompositions need not be performed. We set $B_1 = I$ in the first iteration step, so from $B_1 = L_1 D_1 L_1^T$ it follows $L_1 = I$ and $D_1 = I$. In the next iteration steps, matrices L_i and D_i are updated using $O(n^2)$ arithmetic operations. However, variable metric methods converge slowly, so they require a higher number of function (and gradient) evaluations which can be disadvantageous if the minimized function is computationally complicated.
- The Gauss-Newton method is very efficient for solving problems with zero residua. Nevertheless, convergence can substantially slow down if the problems have large residua. Moreover, the Gauss-Newton method requires solving of a system of linear equations (using Choleski decomposition) or linear least squares problem (using orthogonal decomposition) which requires $O(n^3)$ arithmetic operations.
- The Gauss-Newton method can be substantially improved by combination with variable metric methods namely using a simple combination (Description 1) or using structured updates (Description 2 and Description 3). An advantage of simple hybrid methods consists in that variable metric methods use the Choleski decomposition update which requires only $O(n^2)$ arithmetic operations (structured hybrid methods determine the Choleski decomposition in each iteration step).
- It is advantageous to use H and DW updates in simple hybrid methods which are more efficient than BFGS updates and much more efficient than DFP updates. Using the R1 update is the most advantageous for realization of structured methods. Here, indefiniteness of matrices C or T is no problem (the PSB update from the Dennis class is also good for the same reason). There is no significant difference between structured methods (S4=1) and totally structured methods (S4=2). The choice S3=1 (formula (76)) is usually more advantageous than the choice S3=2 (formula (75)).
- It is becoming apparent that the standard choice S1=1 and S2=1 is the most advantageous strategy. The choice S1=2 leads in some cases to a lower number of evaluations (function values and gradients of the minimized functions) but a higher number of updates slows down the computation. The choice S1=2 and S2=2 is very unsuitable in the case of structured methods with BFGS, DFP, and H updates.
- The methods based on stable orthogonal decomposition of the Jacobi matrix or its approximation converge a bit faster than the methods using Choleski decomposition of the matrix $J^T J$ or its approximation (they require a smaller number of function and gradient evaluations of the minimized function). However, more demanding orthogonal transformations slow down computation.

- Hybrid methods with Jacobian corrections are again more efficient than the Gauss-Newton method whose version based on the orthogonal decomposition of the Jacobi matrix is more robust than the standard version using the Choleski decomposition of the matrix $J^T J$. Since determination of matrices Q and R in the orthogonal decomposition $J = QR$, needed for performing updates described in [13], is time consuming, it is not worth performing these updates. Thus, the simple hybrid methods lose their main advantage and structured hybrid methods are a bit more efficient.
- Quasi-Newton methods, developed for solving systems of nonlinear equations, are surprisingly efficient when using the updating technique described in [13]. Although they require up to three times more function and gradient evaluations of the minimized function than the Gauss-Newton method, their lower complexity causes that their consumed computational time to be comparable with the computational time of hybrid methods.
- The methods based on orthogonal decomposition of the Jacobi matrix require this matrix to be saved column-wise. Row-wise storage substantially slows down computation.

11 Methods for sparse least-squares problems

If the objective function $F(x)$ has the form (1), the expression (3) implies that the Hessian matrix $G(x)$ has the same sparsity pattern as the matrix $J^T(x)J(x)$, so the Hessian matrix $G(x)$ is sparse only if the Jacobian matrix $J(x)$ has sparse rows (if the vector h_k has n_k nonzero elements, then the matrix $h_k h_k^T$ has n_k^2 nonzero elements). In this case, every partial function $f_k : R^n \to R$, $1 \le k \le m$, depends only on $n_k = O(1)$ variables. Therefore, we obtain a special case of a partially separable problem. Partially separable problems can be solved by the separable Newton method or by separable variable metric methods [29]. However, the special form of the objective function allows us to use separable modifications of the Gauss-Newton method, which are more efficient.

The definition domains of the functions f_k, $1 \le k \le m$, lie in subspaces $R_k^n \subset R^n$ of dimension $n_k \ll n$. Since the subspace R_k^n is isomorphic to the subspace R^{n_k}, it is advantageous to introduce reduced quantities. Then the structurally zero elements of the Hessian matrix need not be considered.

Definition 2. *Let I_k, $1 \le k \le m$, be the sets of indices of variables defining the subspaces R_k^n (containing the definition domains of the functions $f_k(x)$), and let $Z_k \in R^{n \times n_k}$ be matrices whose columns are orthogonal bases of R_k^n (i.e. columns of the unit matrix of order n with indices from I_k). Then the vectors $\hat{x}_k = Z_k^T x$ of dimension n_k are called reduced vectors of variables, the functions $\hat{f}_k : R^{n_k} \to R$, for which $\hat{f}_k(\hat{x}_k) = f_k(x)$, are called reduced functions, the vectors $\hat{h}_k(\hat{x}_k) = Z_k^T h_k(x)$ of dimension n_k are called reduced gradients of functions $f_k(x)$ and the symmetric matrices $\hat{H}_k(\hat{x}_k) = Z_k^T H_k(x) Z_k$ of order n_k are called reduced Hessian matrices of functions $f_k(x)$.*

Remark 27. From the practical point of view, we assume that $n_k > 0$ (i.e. $I_k \neq \emptyset$), $1 \le k \le m$, and $I_1 \cup \cdots \cup I_m = \{1, \ldots, n\}$. Therefore, all matrices Z_k, $1 \le k \le m$, are nonempty (they have at least one column) and contain all columns of the unit matrix of order n.

The reduced gradients $\hat{h}_k(\hat{x}_k)$ and the reduced Hessian matrices $\hat{H}_k(\hat{x}_k)$, $1 \le k \le m$, determine explicitly the gradient $g(x)$ and the sparse Hessian matrix $G(x)$ of function (1). We can write

$$F(x) = \sum_{k=1}^{m} \hat{f}_k^2(\hat{x}_k), \qquad g(x) = \sum_{k=1}^{m} \hat{f}_k(\hat{x}_k) Z_k \hat{h}_k(\hat{x}_k), \tag{112}$$

$$G(x) = \sum_{k=1}^{m} Z_k \hat{h}_k(\hat{x}_k)(Z_k \hat{h}_k(\hat{x}_k))^T + \sum_{k=1}^{m} \hat{f}_k(\hat{x}_k) Z_k \hat{H}_k(\hat{x}_k) Z_k^T. \tag{113}$$

The trust region Newton method uses matrices

$$B_i = G_i = \sum_{k=1}^{m} Z_k \hat{h}_i^k (Z_k \hat{h}_i^k)^T + \sum_{k=1}^{m} \hat{f}_i^k Z_k \hat{H}_i^k Z_k^T$$

in (8) and (T1), where $\hat{f}_i^k = \hat{f}_k(x_i)$, $\hat{h}_i^k = \hat{h}_k(x_i)$, $\hat{H}_i^k = \hat{H}_k(x_i)$, $i \in N$. There are two ways of computing an approximation of the Hessian matrix $B \approx G(x)$ by numerical differentiation. The first way, intended for general sparse problems, uses formulas (68), where vectors v_j, $1 \leq j \leq l$, are not columns of the unit matrix, but they contain more unit elements chosen in a way that allows us to compute more elements of B in one differentiation. The choice of vectors v_j, $1 \leq j \leq l$, (which corresponds to grouping of columns in the Hessian matrix) is a difficult combinatorial problem, solved e.g. in [11] (the resulting algorithm is presented in [10]). The second way, intended for partially separable problems, consists in approximation of elements of $\hat{B}_k \approx \hat{H}_k(\hat{x}_k)$, $1 \leq k \leq m$, by formulas

$$\hat{B}_k v_k^j = \frac{\hat{h}(\hat{x}_k + \delta v_k^j) - \hat{h}_k(\hat{x_k})}{\delta}. \tag{114}$$

where v_k^j, $1 \leq j \leq n_k$, are columns of the unit matrix of order n_k. This quite straightforward way is competitive with the method based on grouping columns of the Hessian matrix described in [11].

The trust region Gauss-Newton method uses matrices

$$B_i = J_i^T J_i = \sum_{k=1}^{m} Z_k \hat{h}_i^k (Z_k \hat{h}_i^k)^T$$

in (8) and (T1), where $\hat{h}_i^k = \hat{h}_k(x_i)$, $i \in N$. If the matrix $B_i = J_i^T J_i$ is sparse, then the best way for computing a direction vector d_i is to use the sparse Choleski decomposition [5] of B_i for computing d_i^N by (22). If the matrix J_i is sparse but it has dense columns, then an efficient possibility for obtaining d_i^N is to use the sparse Bunch-Parlett decomposition [21] of the augmented system matrix (Remark 12 (c)).

The sparse Gauss-Newton method can be improved by using an approximation of the second order terms $\hat{f}_i^k \hat{H}_i^k$, $1 \leq k \leq m$. The following method is proposed in [33].

Description 5. (Combined method for partially separable problems) An efficient method arises as a combination of the partitioned Gauss-Newton method and a difference version of the partitioned Newton method. Let $B_1 = J_1^T J_1$ and $0 < \vartheta < 1$. Set

$$\begin{array}{lll} B_{i+1} = J_{i+1}^T J_{i+1} & \text{if} & i \in N_5, \\ B_{i+1} = J_{i+1}^T J_{i+1} + \sum_{k=1}^{m} \hat{f}_{i+1}^k Z_k \hat{B}_{i+1}^k Z_k^T & \text{if} & i \notin N_5, \end{array} \tag{115}$$

where $J_{i+1} = J(x_{i+1})$, $\hat{f}_{i+1}^k = \hat{f}_k(\hat{x}_{i+1}^k)$ and $\hat{B}_{i+1}^k \approx \hat{H}_k(\hat{x}_{i+1}^k)$ for $1 \leq k \leq m$, ($\hat{B}_{i+1}^k$ is a difference approximation of $\hat{H}_k(\hat{x}_{i+1}^k)$ determined by the formula (114)).

Global and superlinear convergence of the combined method for partially separable problems (with $\hat{B}_{i+1}^k = \hat{H}_k(\hat{x}_{i+1}^k)$) follow from Theorem 1, Theorem 4 and Theorem 9.

12 Variable metric methods for sparse problems

Variable metric updates for sparse problems should preserve symmetry and the sparsity pattern of the Hessian matrix and satisfy a quasi-Newton condition. Denote

$$\begin{aligned} \mathcal{V}_Q &= \{B \in R^{n \times n} : Bs = y\}, \\ \mathcal{V}_S &= \{B \in R^{n \times n} : B^T = B\}, \\ \mathcal{V}_G &= \{B \in R^{n \times n} : B_{ij} = 0, \text{ if } G_{ij} = 0\}. \end{aligned}$$

Here $G_{ij} = 0$ denotes structural zeroes, i.e. elements such that $G_{ij}(x) = 0$ for all $x \in R^n$. Since the Hessian matrix is symmetric, one has $G_{ij} = 0 \Leftrightarrow G_{ji} = 0$. Since the Hessian matrix should be positive definite, we assume that $G_{ii} \neq 0$, $1 \leq i \leq n$ (diagonal elements of the Hessian matrix are structurally nonzero). Clearly, $\mathcal{V}_Q \subset R^{n \times n}$, $\mathcal{V}_S \subset R^{n \times n}$, $\mathcal{V}_G \subset R^{n \times n}$ are linear manifolds ($\mathcal{V}_S$ and $\mathcal{V}_G$ are linear subspaces) in

$R^{n\times n}$. Since the Frobenius norm of a matrix is an Euclidean norm in $R^{n\times n}$ (if matrices are considered as vectors of dimension $n \times n$), we can define orthogonal projection operators $\mathcal{P}_Q$, $\mathcal{P}_S$, $\mathcal{P}_G$ into $\mathcal{V}_Q$, $\mathcal{V}_S$, $\mathcal{V}_G$ by expressions

$$\begin{aligned}\mathcal{P}_Q B &= \arg\min_{\tilde{B}\in\mathcal{V}_Q} \|\tilde{B} - B\|_F,\\ \mathcal{P}_S B &= \arg\min_{\tilde{B}\in\mathcal{V}_S} \|\tilde{B} - B\|_F,\\ \mathcal{P}_G B &= \arg\min_{\tilde{B}\in\mathcal{V}_G} \|\tilde{B} - B\|_F.\end{aligned}$$

Similarly, we can define orthogonal projection operators $\mathcal{P}_{QS}$, $\mathcal{P}_{QG}$, $\mathcal{P}_{SG}$ and $\mathcal{P}_{QSG}$ into linear manifolds $\mathcal{V}_Q \cap \mathcal{V}_S$, $\mathcal{V}_Q \cap \mathcal{V}_G$, $\mathcal{V}_S \cap \mathcal{V}_G$ and $\mathcal{V}_Q \cap \mathcal{V}_S \cap \mathcal{V}_G$. Clearly, $\mathcal{V}_Q \cap \mathcal{V}_S \cap \mathcal{V}_G \neq \emptyset$, since $\tilde{G} \in \mathcal{V}_Q \cap \mathcal{V}_S \cap \mathcal{V}_G$, where $\tilde{G} = \int_0^1 G(x + ts)\mathrm{d}t$.

It is obvious that the requirements for the sparse variable metric update are satisfied by the Toint [55] update

$$B_+ = \mathcal{P}_{QSG} B. \tag{116}$$

This update is relatively difficult to realize (it requires solving an additional linear system) and its efficiency is not excellent since the generated matrices may be indefinite. Therefore, additional updates, which in some sense violate the quasi-Newton condition, were proposed. We are concerned with the Marwil [40] update

$$B_+ = \mathcal{P}_S \mathcal{P}_{QG} B, \tag{117}$$

the Powell [45] update

$$B_+ = \mathcal{P}_G \mathcal{P}_{QS} B \tag{118}$$

and the Steihaug [51] update

$$B_+ = \mathcal{P}_{SG} \mathcal{P}_Q B. \tag{119}$$

Formulas (116)–(119) can be written in the form $B_+ = \mathcal{P}_B \mathcal{P}_A B$, where $\mathcal{P}_A$, $\mathcal{P}_B$ are orthogonal projection operators into $\mathcal{V}_A$, $\mathcal{V}_B$, where $\mathcal{V}_A \subset \mathcal{V}_Q$ and $\mathcal{V}_A \cap \mathcal{V}_B = \mathcal{V}_Q \cap \mathcal{V}_S \cap \mathcal{V}_G$ ($\mathcal{P}_B$ is an identical operator in (116)). Thus the trust region variable metric methods for sparse problem use updates

$$\begin{aligned} &B_{i+1} = \mathcal{P}_B \mathcal{P}_A B_i \quad \text{if} \quad i \in N_2,\\ &B_{i+1} = B_i \qquad\quad \text{if} \quad i \notin N_2.\end{aligned} \tag{120}$$

The update (120) is basic. However, we can use two strategies distinguished by the value S1 in the same way as in Remark 13.

The following theorems, valid for trust region methods, are variants of theorems proved in [51].

Theorem 18. *Let $x_i \in R^n$, $i \in N$, be a sequence of points generated by the trust region method (T1)–(T3), where matrix B_1 is positive definite and matrices B_{i+1}, $i \in N$, are computed by (120) ($\mathcal{P}_B \mathcal{P}_A B_i$ is one of sparse updates (116)–(119)). If the mapping $f : R^n \to R^m$ satisfies Assumption A1 and the second derivatives are Lipschitz continuous (so (69) holds) on $\overline{\mathcal{D}}$, then matrices B_i, $i \in N$, have bounded norms and*

$$\liminf_{i\to\infty} \|g_i\| = 0.$$

Theorem 19. *Let the assumptions of Theorem 18 be satisfied and $x_i \to x^*$, where the point $x_* \in R^n$ satisfies Assumption A3. If $\omega_i \to 0$, then the rate of convergence is Q-superlinear.*

Variable metric methods for sparse problems are less efficient than corresponding methods for dense problems as demonstrated in Table 8. The most sophisticated is the Toint method with the update (116), but the best results are usually obtained by the Marwil method with the update (117).

Remark 28. Let B be a symmetric matrix. Then the Toint update (116), proposed in [55], can be expressed in the form

$$\mathcal{P}_{QSG}B = \mathcal{P}_G(B + us^T + su^T), \tag{121}$$

where $u \in R^n$ is a solution of the linear system $Qu = y - (\mathcal{P}_G B)s$ with the symmetric positive semidefinite matrix

$$Q = \mathcal{P}_G(ss^T) + \sum_{i=1}^{n} \|s^i\|^2 e_i e_i^T.$$

Vectors $s^i \in R^n$, $1 \le i \le n$, are defined by formulas

$$\begin{array}{lll} e_j^T s^i = e_j^T s & \text{if} & G_{ij} \neq 0, \\ e_j^T s^i = 0 & \text{if} & G_{ij} = 0, \end{array}$$

where e_j, $1 \le j \le n$, are columns of the unit matrix of order n, and

$$\begin{array}{lll} (\mathcal{P}_G M)_{ij} = M_{ij} & \text{if} & G_{ij} \neq 0, \\ (\mathcal{P}_G M)_{ij} = 0 & \text{if} & G_{ij} = 0 \end{array}$$

for an arbitrary matrix M of order n.

Remark 29. Let B be a symmetric matrix. Then the Marwil update (117), proposed in [40] and introduced also in [51], can be expressed in the form

$$\mathcal{P}_S \mathcal{P}_{QG} B = \mathcal{P}_G(B + (us^T + su^T)/2), \tag{122}$$

where $u \in R^n$ is a solution of the linear system $Qu = y - (\mathcal{P}_G B)s$ with the diagonal matrix

$$Q = \sum_{i=1}^{n} \|s^i\|^2 e_i e_i^T.$$

13 Variable metric methods for partially separable problems

In the subsequent considerations, we use the notation

$$N_3^k = \{i \in N : (\hat{y}_i^k)^T \hat{s}_i^k > c\|\hat{y}_i^k\|^2\}, \qquad N_4^k = \left\{ i \in N : \left| (\hat{s}_i^k)^T \left(\gamma_i^k \hat{y}_i^k - \hat{B}_i^k \hat{s}_i^k \right) \right| \ge c \left\| \gamma_i^k \hat{y}_i^k - \hat{B}_i^k \hat{s}_i^k \right\|^2 \right\}$$

for $1 \le k \le m$, where $\hat{y}_i^k = \hat{f}_{i+1}^k \hat{h}_{i+1}^k - \hat{f}_i^k \hat{h}_i^k = Z_k^T y_i$, $\hat{s}_i^k = \hat{x}_{i+1}^k - \hat{x}_i^k = Z_k^T s_i$ are vectors of dimension n_k. Line search variable metric methods for partially separable problems (partitioned variable metric methods) were proposed in [29]. Trust region partitioned variable metric methods generate matrices $\hat{B}_i^k \approx \hat{h}_i^k (\hat{h}_i^k)^T + \hat{f}_i^k \hat{H}_i^k$, $1 \le k \le m$, which serve for construction of the matrix

$$B_i = \sum_{k=1}^{m} Z_k \hat{B}_i^k Z_k^T \tag{123}$$

used in (8) and (T1). The matrices $\hat{B}_i^k$, $1 \le k \le m$, are generated in such a way that $\hat{B}_1^k$, $1 \le k \le m$, are positive definite and

$$\begin{array}{lll} \hat{B}_{i+1}^k = \mathcal{B}(\hat{B}_i^k, \hat{y}_i^k, \hat{s}_i^k, \beta_i^k, \gamma_i^k) & \text{if} & i \in N_2, \quad i \in N^k, \\ \hat{B}_{i+1}^k = \hat{B}_i^k & \text{if} & i \in N_2, \quad i \notin N^k, \\ \hat{B}_{i+1}^k = \hat{B}_i^k & \text{if} & i \notin N_2 \end{array} \tag{124}$$

for $1 \le k \le m$, where $\mathcal{B}$ is the mapping defined by (55) and β_i^k, γ_i^k are free parameters. At the same time, $N^k = N_3^k$ for a general update and $N^k = N_4^k$ for the R1 update. Update (124) will be considered as basic. However, we can use two strategies distinguished by the value S1 in the same way as in Remark 13.

Remark 30. A disadvantage of update (124) consists in the fact that $i \notin N_3^k$ can hold for majority of indices $i \in N$ for some $1 \le k \le m$, so the matrix $\hat{B}_i^k$ is mostly unchanged if a general update is used. To overcome this difficulty, we use the following three strategies distinguished by the value S5:

S5 = 1 – Basic strategy (124) is used.

S5 = 2 – If $i \notin N_3^k$ for some $1 \le k \le m$ and $i \in N$, we use (for this k) the R1 update in the current iteration and in all subsequent iterations, i.e., a general update of $\hat{B}_i^k$, where k is given, is switched to the R1 update in the current iteration and in all subsequent iterations (with N_3^k replaced by N_4^k).

S5 = 3 – If $i \notin N_3^k$ for at least $m/2$ indices $1 \le k \le m$, we use the R1 update for all matrices $\hat{B}_i^k$, $1 \le k \le m$, in the current iteration and in all subsequent iterations, i.e., a general update of $\hat{B}_i^k$, where k is arbitrary, is switched to the R1 update in the current iteration and in all subsequent iterations (with N_3^k replaced by N_4^k).

Theorem 20. *(Global convergence) Let the mapping $f : R^n \to R^m$ satisfy Assumption A1 and $x_i \in R^n$, $i \in N$, be a sequence generated by the trust region method (T1)–(T3) with the matrices B_i given by (123), where the matrices $\hat{B}_1^k$, $1 \le k \le m$, are positive definite and the matrices $\hat{B}_{i+1}^k$, $1 \le k \le m$, are computed by (124) with $1 \le \gamma_i^k \le \overline{\gamma}$ and either $\beta_i^k = \gamma_i^k b_i^k/(\gamma_i^k b_i^k - c_i^k)$ (R1 method) or $0 \le \beta_i^k \le \overline{\beta}_i^k$, where*

$$\overline{\beta}_i^k = \frac{K}{K + c_i^k/b_i^k}, \tag{125}$$

with $b_i^k = (\hat{y}_i^k)^T \hat{s}_i^k$, $c_i^k = (\hat{s}_i^k)^T \hat{B}_i^k \hat{s}_i^k$ and $K > 0$. Then $\liminf_{i\to\infty} \|g(x_i)\| = 0$.

Proof Formulas (124) for reduced matrices $\hat{B}_i^k$, $1 \le k \le m$, are the same as formulas (54) (or (65) if the R1 update is used) for matrices B_i. Therefore, there exist constants C_k, $1 \le k \le m$, such that

$$Tr\, \hat{B}_{i+1}^k \le Tr\, \hat{B}_i^k + C_k, \qquad 1 \le k \le m. \tag{126}$$

Since the matrices Z_k, $1 \le k \le m$, contain columns of the unit matrix, we can write

$$Tr\, B_i = Tr \sum_{k=1}^{m} Z_k \hat{B}_i^k Z_k^T = \sum_{k=1}^{m} Tr\, Z_k \hat{B}_i^k Z_k^T = \sum_{k=1}^{m} Tr\, \hat{B}_i^k, \tag{127}$$

so

$$Tr\, B_{i+1} = \sum_{k=1}^{m} Tr\, \hat{B}_{i+1}^k \le \sum_{k=1}^{m} (Tr\, \hat{B}_i^k + C_k) = Tr\, B_i + C, \qquad C = \sum_{k=1}^{m} C_k, \tag{128}$$

and global convergence is proved as in the proof of Theorem 5. □

If the objective function has the form (1), superlinear convergence cannot be proved by the approach used in [28]. The functions $(f_k(x_i))^2$, $1 \le k \le m$, are usually not all convex (if $f_k(x)$ is convex and $f_k(x) < 0$, then function $(f_k(x_i))^2$ may not be convex). Therefore, $i \notin N_3^k$ can hold for some k even if x_i is close to x_*. Nevertheless, the globally convergent trust region variable metric method, described in this section, is relatively efficient for solving partially separable least squares problems, so it is a suitable choice for the construction of hybrid methods.

14 Simple hybrid methods for sparse least squares problems

There are two classes of simple hybrid methods for sparse least squares problems [33]. The first class combines the Gauss-Newton method with variable metric methods for sparse problems described in Section 12 [33].

Description 6. (Simple hybrid method for sparse problems) An efficient hybrid method arises as a combination of the sparse Gauss-Newton method and a suitable variable metric method for sparse problems. Let $B_1 = J_1^T J_1$ and $0 < \vartheta < 1$. Set

$$\begin{array}{lll} B_{i+1} = J_{i+1}^T J_{i+1} & \text{if} & i \in N_2, \quad i \in N_5, \\ B_{i+1} = \mathcal{P}_B \mathcal{P}_A B_i & \text{if} & i \in N_2, \quad i \notin N_5, \\ B_{i+1} = B_i & \text{if} & i \notin N_2, \end{array} \tag{129}$$

where $\mathcal{P}_B \mathcal{P}_A B_i$ is one of the updates (116)–(119). Update (129) will be considered as basic. We will consider two strategies here distinguished by the value S1 in the same way as in Remark 13.

Theorem 21. *If the mapping f satisfies Assumption A1 and its second derivatives are Lipschitz continuous (so (67) holds), then the simple hybrid (trust region) method introduced in Description 6 is globally convergent. If in addition $x_i \to x_*$, where the point $x_* \in R^n$ satisfies Assumption A3, and $\omega_i \to 0$, then the rate of convergence is Q-superlinear.*

Proof (a) The matrices B_i, $i \in N$, have bounded norms by Remark 1 and Theorem 18, so $g(x_i) \to 0$ by Theorem 1.

(b) Let $x_i \to x_*$ and $F(x_*) > 0$. Then $(F_{i-1} - F_i)/F_{i-1} \to 0$ by Theorem 10, so there exists an index $l \in N$ such that $(F_{i-1} - F_i)/F_{i-1} < \vartheta \ \forall i \geq l$, so $i \notin N_2 \cap N_5 \ \forall i \geq l$ and the superlinear rate of convergence follows from Theorem 19.

(c) Let $x_i \to x_*$ and $F(x_*) = 0$. If the set $N_2 \cap N_5$ is finite, the superlinear rate of convergence follows from Theorem 19 (as in part (b)). If the set $N_2 \cap N_5$ is infinite, then $B_i \overset{N_2 \cap N_5}{\longrightarrow} G_i$ by Remark 11, which gives (71) and, similarly as in the proof of Theorem 4, there exists an index $l_2 \in N_2 \cap N_5$ such that (48)–(50) hold for $i \in N_2 \cap N_5$, $i \geq l_2$ and $i \in N_2$, so

$$\tau_i = \frac{g_{i+1} - g_i - B_i d_i}{\|g_i\|} \overset{N_2 \cap N_5}{\longrightarrow} 0.$$

Let $i \in N_2 \cap N_5$, $i \geq l_2$. Since $i \in N_2$, we can write

$$g_{i+1} = g(x_i + d_i) = g_i + G_i d_i + o(\|d_i\|),$$

which together with $F(x_*) = 0$ and $g(x_*) = 0$ gives

$$\frac{F_i - F_{i+1}}{F_i} = 1 - \frac{F_{i+1}}{F_i} \geq 1 - \left(\frac{\overline{G}}{\underline{G}}\right)^2 \left(\frac{\|g_{i+1}\|}{\|g_i\|}\right)^2 = 1 - \left(\frac{\overline{G}}{\underline{G}}\right)^2 (\|\tau_i\| + \|\omega_i\|)^2 \overset{N_2 \cap N_5}{\longrightarrow} 1,$$

so there exists an index $l_3 \geq l_2$ such that $(F_i - F_{i+1})/F_i \geq \vartheta$, or $i + 1 \in N_2 \cap N_5$, for $i \in N_2 \cap N_5$, $i \geq l_3$. Therefore we obtain $i \in N_2 \cap N_5 \ \forall i \geq l_3$ by induction and the superlinear rate of convergence follows from (71) (Theorem 4). □

The second class uses partitioned variable metric methods described in Section 13.

Description 7. (Simple hybrid method for partially separable problems) An efficient hybrid method arises as a combination of the partitioned Gauss-Newton method and a suitable partitioned variable metric method. Let $\hat{B}_1^k = \hat{h}_1^k (\hat{h}_1^k)^T$, $1 \leq k \leq m$, and $0 < \vartheta < 1$. Set

$$\begin{array}{lll} \hat{B}_{i+1}^k = \hat{h}_{i+1}^k (\hat{h}_{i+1}^k)^T & \text{if} & i \in N_2, \quad i \in N_5, \\ \hat{B}_{i+1}^k = \mathcal{B}(\hat{B}_i^k, \hat{y}_i^k, \hat{s}_i^k, \beta_i^k, \gamma_i^k) & \text{if} & i \in N_2, \quad i \notin N_5, \quad i \in N_3^k, \\ \hat{B}_{i+1}^k = \hat{B}_i^k & \text{if} & i \in N_2, \quad i \notin N_5, \quad i \notin N_3^k, \\ \hat{B}_{i+1}^k = \hat{B}_i^k & \text{if} & i \notin N_2 \end{array} \tag{130}$$

for $1 \leq k \leq m$, where $\mathcal{B}$ is the mapping defined by (55), $\hat{y}_i^k = \hat{f}_{i+1}^k \hat{h}_{i+1}^k - \hat{f}_i^k \hat{h}_i^k = Z_k^T y_i$, $\hat{s}_i^k = \hat{x}_{i+1}^k - \hat{x}_i^k = Z_k^T s_i$ are vectors of dimension n_k, $1 \leq \gamma_i^k \leq \overline{\gamma}$ and $0 \leq \beta_i^k \leq \overline{\beta}_i^k$, where $\overline{\beta}_i^k$ is the value determined by (125). Update (130) will be considered as basic. We will consider here two strategies distinguished by the value S1 in the same way as in Remark 13.

Theorem 22. *If the mapping f satisfies Assumption A1, then the simple hybrid (trust region) method introduced in Description 7 is globally convergent.*

Proof If $i \in N_2 \cap N_5$, then $Tr\, \hat{B}_{i+1}^k = (\hat{h}_{i+1}^k)^T \hat{h}_{i+1}^k = \|\hat{h}_{i+1}^k\|^2 = \|h_{i+1}^k\|^2 \leq \overline{h}^2$, $1 \leq k \leq m$, by Assumption A1. If $i \notin N_2 \cap N_5$, then there exist numbers C_k, $1 \leq k \leq m$, such that $Tr\, \hat{B}_{i+1}^k \leq Tr\, \hat{B}_i^k + C_k$ by (126). Since matrices Z_k, $1 \leq k \leq m$, contain columns of the unit matrix, we can use an analogy of (128), so $Tr\, B_{i+1} \leq \overline{h}^2 + iC \leq (i+1)\overline{C}$, $\overline{C} = \max(\overline{h}^2, C)$, so $M_i \leq i\overline{C}$, $i \in N$, and the global convergence follows from Theorem 5. □

15 Structured hybrid methods for sparse least squares problems

As in Section 7, we will investigate two classes of structured hybrid methods for sparse least squares problems [54], [33]. Methods from the first class use approximations $\hat{T}_i^k \approx \hat{H}_k(\hat{x}_i)$, $1 \leq k \leq m$, so the matrices $\hat{T}_{i+1}^k$ should satisfy quasi-Newton conditions $\hat{T}_{i+1}^k \hat{s}_i^k = \hat{h}_{i+1}^k - \hat{h}_i^k \triangleq \tilde{z}_i^k$, $1 \leq k \leq m$ [56]. These quasi-Newton conditions are satisfied if matrices $\hat{T}_{i+1}^k$ are generated by the variable metric methods from the Broyden class as shown in the following description, where we use the notation

$$N_3^{'k} = \{i \in N : (\tilde{z}_i^k)^T \hat{s}_i^k \geq c\|\tilde{z}_i^k\|^2\}, \qquad N_4^{'k} = \left\{i \in N : \left|(\hat{s}_i^k)^T \left(\gamma_i^k \tilde{z}_i^k - \hat{T}_i^k \hat{s}_i^k\right)\right| \geq c \left\|\gamma_i^k \tilde{z}_i^k - \hat{T}_i^k \hat{s}_i^k\right\|^2\right\}.$$

Description 8. (Totally structured hybrid method for partially separable problems) An efficient hybrid method arises as a combination of the partitioned Gauss-Newton method and a suitable partitioned variable metric method. Let $\hat{T}_1^k = 0$, $\hat{B}_1^k = \hat{h}_1^k (\hat{h}_1^k)^T$, $1 \leq k \leq m$, and $0 < \vartheta < 1$. Set

$$\begin{array}{lll} \hat{T}_{i+1}^k = \hat{T}_i^k & \text{if} & i \in N_2, \quad i \in N_5, \\ \hat{T}_{i+1}^k = \mathcal{B}(\hat{T}_i^k, \tilde{z}_i^k, \hat{s}_i^k, \beta_i^k, \gamma_i^k) & \text{if} & i \in N_2 \quad i \notin N_5, \quad i \in N^{'k}, \\ \hat{T}_{i+1}^k = \hat{T}_i^k & \text{if} & i \in N_2, \quad i \notin N_5, \quad i \notin N^{'k}, \\ \hat{T}_{i+1}^k = \hat{T}_i^k & \text{if} & i \notin N_2 \end{array} \tag{131}$$

and

$$\begin{array}{lll} \hat{B}_{i+1}^k = \hat{h}_{i+1}^k (\hat{h}_{i+1}^k)^T & \text{if} & i \in N_2, \quad i \in N_5, \\ \hat{B}_{i+1}^k = \hat{h}_{i+1}^k (\hat{h}_{i+1}^k)^T + \hat{f}_{i+1}^k \hat{T}_{i+1}^k & \text{if} & i \in N_2, \quad i \notin N_5, \\ \hat{B}_{i+1}^k = \hat{B}_i^k & \text{if} & i \notin N_2 \end{array} \tag{132}$$

for $1 \leq k \leq m$, where $\mathcal{B}$ is the mapping defined by (55), $\tilde{z}_i^k = \hat{h}_{i+1}^k - \hat{h}_i^k$, $\hat{s}_i^k = \hat{x}_{i+1}^k - \hat{x}_i^k$ are vectors of dimension n_k, $1 \leq \gamma_i^k \leq \overline{\gamma}$ and either $\beta_i^k = \gamma_i^k b_i^k / (\gamma_i^k b_i^k - c_i^k)$ (R1 method) or $0 \leq \beta_i^k \leq \overline{\beta}_i^k$, where $\overline{\beta}_i^k$ is the value determined by (125) with $b_i^k = (\tilde{z}_i^k)^T \hat{s}_i^k$, $c_i^k = (\hat{s}_i^k)^T \hat{B}_i^k \hat{s}_i^k$ and $K > 0$. At the same time, $N^{'k} = N_4^{'k}$ for the R1 method or $N^{'k} = N_3^{'k}$ if $0 \leq \beta_i^k \leq \overline{\beta}_i^K$.

Theorem 23. *If the mapping f satisfies Assumption A1, then the structured hybrid (trust region) method introduced in Description 8 is globally convergent.*

Proof We use the same idea as in the proof of Theorem 12.
(a) If $N^{'k} = N_4^{'k}$, then (similarly as in part (a) of the proof of Theorem 12) we obtain

$$\|\hat{T}_{i+1}^k\| \leq \|\hat{T}_i^k\| + \frac{\|\gamma_i^k \tilde{z}_i^k - \hat{T}_i^k \hat{s}_i^k\|^2}{(\hat{s}_i^k)^T (\gamma_i^k \tilde{z}_i^k - \hat{T}_i^k \hat{s}_i^k)} \leq \|\hat{T}_i^k\| + \frac{1}{c},$$

so

$$\|\hat{B}_{i+1}^k\| = \|\hat{B}_{l+1}^k\| \leq \|\hat{h}_{l+1}^k\|^2 + \|\hat{T}_{l+1}^k\| \leq \overline{h}^2 + \|\hat{T}_1^k\| + \frac{l}{c} \leq (l+1)C_k \leq (i+1)C_k,$$

$$C_k = \max\left(\overline{h}^2 + \|\hat{T}_1^k\|, \frac{1}{c}\right),$$

where $l \in N$ is a maximum index such that $l \leq i$ and $l \in N_2$. Thus (123) implies

$$\|B_{i+1}\| \leq \sum_{k=1}^{m} \|Z_k \hat{B}_{i+1}^k Z_k^T\| \leq \sum_{k=1}^{m} \|\hat{B}_{i+1}^k\| \leq (i+1) \sum_{k=1}^{m} C_k \triangleq \overline{C},$$

so $M_i \leq i\overline{C}$, $i \in N$, and the global convergence follows from Theorem 1.

(b) If $N'^k = N_3'^k$, then (similarly as in part (b) of the proof of Theorem 12) we obtain

$$Tr\, \hat{T}_{i+1}^k \leq Tr\, \hat{T}_i^k + (K + \overline{\gamma}) \frac{(\tilde{z}_i^k)^T \tilde{z}_i^k}{(\tilde{z}_i^k)^T \hat{s}_i^k} \leq Tr\, \hat{T}_i^k + \frac{K + \overline{\gamma}}{c}, \tag{133}$$

so

$$Tr\, \hat{B}_{i+1}^k = Tr\, \hat{B}_{l+1}^k \leq \|\hat{h}_{l+1}^k\|^2 + Tr\, \hat{T}_{l+1}^k \leq \overline{h}^2 + Tr\, \hat{T}_1^k + l \frac{K + \overline{\gamma}}{c} \leq (l+1) C_k \leq (i+1) C_k,$$

$$C_k = \max\left(\overline{h}^2 + Tr\, \hat{T}_1^k, \frac{K + \overline{\gamma}}{c}\right),$$

where $l \in N$ is a maximum index such that $l \leq i$ and $l \in N_2$. Thus (123) implies

$$Tr\, B_{i+1} \leq \sum_{k=1}^{m} Tr\, (Z_k \hat{B}_{i+1}^k Z_k^T) \leq \sum_{k=1}^{m} Tr\, \hat{B}_{i+1}^k \leq (i+1) \sum_{k=1}^{m} C_k \triangleq \overline{C},$$

so $M_i \leq i\overline{C}$, $i \in N$, and the global convergence follows from Theorem 1. □

Methods from the second class use approximations $\hat{C}_i^k \approx \hat{f}_k(\hat{x}_i) \hat{H}_k(\hat{x}_i) \approx \hat{f}_i^k \hat{T}_i^k$, $1 \leq k \leq m$. Since we assume that $\hat{T}_{i+1}^k \hat{s}_i^k = \hat{h}_{i+1}^k - \hat{h}_i^k$, quasi-Newton conditions for $\hat{C}_i^k$ have the form $\hat{C}_{i+1}^k \hat{s}_i^k = \hat{f}_{i+1}^k (\hat{h}_{i+1}^k - \hat{h}_i^k) \triangleq \hat{z}_i^k$, $1 \leq k \leq m$. These quasi-Newton conditions are satisfied if matrices $\hat{C}_{i+1}^k$ are generated by the variable metric methods from the Broyden class as shown in the following description, where we use the notation

$$N_3''^k = \{i \in N : (\hat{z}_i^k)^T \hat{s}_i^k \geq c \|\hat{z}_i^k\|^2\}, \qquad N_4''^k = \left\{i \in N : \left|(\hat{s}_i^k)^T \left(\gamma_i^k \hat{z}_i^k - \hat{B}_i^k \hat{s}_i^k\right)\right| \geq c \left\|\gamma_i^k \hat{z}_i^k - \hat{B}_i^k \hat{s}_i^k\right\|^2\right\}.$$

Description 9. (Structured hybrid method for partially separable problems) An efficient hybrid method arises as a combination of the partially separable Gauss-Newton method and a suitable variable metric method for partially separable problems. Let $\hat{C}_1^k = 0$, $\hat{B}_1^k = \hat{h}_1^k (\hat{h}_1^k)^T$, $1 \leq k \leq m$, and $0 < \vartheta < 1$. Set

$$\begin{array}{lll} \hat{C}_{i+1}^k = \hat{C}_i^k & \text{if} & i \in N_2, \quad i \in N_5, \\ \hat{C}_{i+1}^k = \mathcal{B}(\hat{C}_i^k, \hat{z}_i^k, \hat{s}_i^k, \beta_i^k, \gamma_i^k) & \text{if} & i \in N_2 \quad i \notin N_5, \quad i \in N''^k, \\ \hat{C}_{i+1}^k = \hat{C}_i^k & \text{if} & i \in N_2, \quad i \notin N_5, \quad i \notin N''^k, \\ \hat{C}_{i+1}^k = \hat{C}_i^k & \text{if} & i \notin N_2 \end{array} \tag{134}$$

and

$$\begin{array}{lll} \hat{B}_{i+1}^k = \hat{h}_{i+1}^k (\hat{h}_{i+1}^k)^T & \text{if} & i \in N_2, \quad i \in N_5, \\ \hat{B}_{i+1}^k = \hat{h}_{i+1}^k (\hat{h}_{i+1}^k)^T + \hat{f}_{i+1}^k \hat{C}_{i+1}^k & \text{if} & i \in N_2, \quad i \notin N_5, \\ \hat{B}_{i+1}^k = \hat{B}_i^k & \text{if} & i \notin N_2 \end{array} \tag{135}$$

for $1 \leq k \leq m$, where $\mathcal{B}$ is the mapping defined by (55), $\hat{z}_i^k = \hat{f}_{i+1}^k (\hat{h}_{i+1}^k - \hat{h}_i^k)$, $\hat{s}_i^k = \hat{x}_{i+1}^k - \hat{x}_i^k$ are vectors of dimension n_k, $1 \leq \gamma_i^k \leq \overline{\gamma}$ and either $\beta_i^k = \gamma_i^k b_i^k / (\gamma_i^k b_i^k - c_i^k)$ (R1 method) or $0 \leq \beta_i^k \leq \overline{\beta}_i^k$, where $\overline{\beta}_i^k$ is the value determined by (125) with $b_i^k = (\hat{z}_i^k)^T \hat{s}_i^k$, $c_i^k = (\hat{s}_i^k)^T \hat{B}_i^k \hat{s}_i^k$ and $K > 0$. At the same time, $N''^k = N_4''^k$ for the R1 method or $N''^k = N_3''^k$ if $0 \leq \beta_i^k \leq \overline{\beta}_i^K$.

Theorem 24. *If the mapping f satisfies Assumption A1, then the structured hybrid (trust region) method introduced in Description 9 is globally convergent.*

Proof The proof of this theorem is almost the same as the proof of Theorem 23. Only inequality (133) is replaced by the inequality

$$Tr\,\hat{C}_{i+1}^k = Tr\,\hat{C}_i^k + (K+\overline{\gamma})\frac{(\hat{z}_i^k)^T\hat{z}_i^k}{(\hat{z}_i^k)^T\hat{s}_i^k} \leq Tr\,\hat{C}_i^k + \frac{K+\overline{\gamma}}{c}. \tag{136}$$

□

16 Numerical comparison of methods for sparse least squares problems

We have used the collection TEST26 [35] for our computational experiments described in this section. Collection TEST26 contains 60 different sparse problems with optional dimensions. We used the first 30 problems with 1000 variables ($n = 1000$) which are a mixture of both zero and nonzero residual problems. The remaining 30 problems are all zero residual, so they are not suitable for testing hybrid methods.

The tested methods are distinguished by the following codes:

- `VM` - the Marwil sparse variable metric method,
- `VT` - the Toint sparse variable metric method,
- `VB` - partitioned variable metric methods,
- `MN` - the partitioned Newton method with numerical differentiation,
- `GN` - the partitioned Gauss-Newton method,
- `MG` - combined method introduced in Description 5,
- `GM` - simple hybrid method introduced in Description 6 with the Marwil update,
- `GT` - simple hybrid method introduced in Description 6 with the Toint update,
- `GB` - simple hybrid methods introduced in Description 7,
- `GD` - structured hybrid methods introduced in Description 8,
- `GS` - structured hybrid methods introduced in Description 9.

The code `GM1` has the same meaning as in Section 10. We used the sparse Choleski (or Gill-Murray) decomposition described in [24]. The tested methods were implemented in the same way as the methods investigated in Section 10, i.e., we used the values $\underline{\rho} = 0.1$, $\overline{\rho} = 0.9$, $\overline{\Delta} = 1000$, $\underline{\gamma} = 0.7$, $\overline{\gamma} = 6.0$, $\vartheta = 0.0005$, $c = 10^{-32}$ and $K = 10^{32}$.

The test results are presented in several following tables whose columns mean the tested method with chosen strategy S1, S2 (values 1 or 2), the scaling used (Y - yes, N - no), the total number of function evaluations `NFV` (which is equal to the total number of iterations of the trust region method), the total number of gradient evaluations `NFG`, the total number of matrix decompositions `NDC`, the total computational time and sometimes the number of failures (the number of problems that were not solved). The asterisk in some tables indicates that at least one problem was not solved (the maximum number of function evaluations was exceeded).

Table 8 contains the results of comparing the particular sparse variable metric methods `VM` and `VT`.

Update	S1	NFV	NFG	NDC	Time	Fail
VM	1	28476	24719	23887	11.57	1
VM	2	30538	30535	29251	14.07	1
VT	1	30565	25879	24972	18.36	2
VT	2	41347	41342	39864	29.20	2

Table 8: Sparse variable metric methods

Table 9 contains the results of comparing the particular partitioned variable metric updates and various strategies S1 and S5.

Update	S5	S1	NFV	NFG	NDC	Time	Fail
BFGS	1	1	28939	27025	26782	20.21	-
BFGS	2	1	28953	23865	20394	13.22	1
BFGS	3	1	19209	16828	16216	12.04	-
BFGS	1	2	30126	30102	24948	21.38	1
BFGS	2	2	35214	35209	17344	17.30	2
BFGS	3	2	35292	35287	17355	17.42	2
H	1	1	53217	52272	52027	47.47	3
H	2	1	29562	24353	20758	13.55	1
H	3	1	37409	35878	35268	30.77	2
H	1	2	52873	52831	49682	43.33	1
H	2	2	35481	35476	17394	17.45	2
H	3	2	30386	30344	27076	19.14	1
R1	1	1	30138	28436	28196	21.66	-
R1	2	1	29261	24119	20537	13.69	1
R1	3	1	28939	27025	26782	20.21	-
R1	1	2	32770	32766	31798	23.93	-
R1	2	2	35514	35504	17095	17.20	2
R1	3	2	30533	30524	27966	20.46	-

Table 9: Partitioned variable metric methods (VB)

In Table 10 we can see a comparison of simple hybrid methods with sparse updates GM and GT using different choices of S1.

Method	S1	NFV	NFG	NDC	Time	Fail
GM	1	6480	6302	6191	3.10	-
GM	2	6572	6568	6380	3.55	-
GT	1	7209	6912	6800	4.30	-
GT	2	7826	7824	7634	4.30	-

Table 10: Simple hybrid methods with sparse updates

In Table 11 we can see a comparison of simple hybrid methods with partitioned updates GB using different choices of S1 and scaling.

Update	S1	Scaling	NFV	NFG	NDC	Time	Fail
BFGS	1	N	6816	6614	6510	3.79	-
BFGS	1	Y	6797	6606	6503	3.77	-
BFGS	2	N	9108	9104	7938	5.64	-
BFGS	2	Y	7754	7747	7070	4.71	-
DFP	1	N	7898	7647	7509	5.10	-
DFP	1	Y	7506	7244	7120	4.50	-
DFP	2	N	12488	12483	11395	10.56	1
DFP	2	Y	13405	13403	12862	13.57	1
H	1	N	6882	6745	6636	3.93	-
H	1	Y	8219	7821	7711	4.99	-
H	2	N	9260	9252	8208	6.22	-
H	2	Y	8156	8148	7599	5.63	-
R1	1	N	7752	7490	7385	4.33	-
R1	1	Y	7002	6855	6754	3.80	-
R1	2	N	7693	7692	6893	4.85	-
R1	2	Y	9236	9238	7903	6.06	-

Table 11: Simple hybrid methods (GB) with partitioned updates

Tables 12 contains a comparison of different structured hybrid methods GS and GD with different strategies S1, S2 and scaling.

				Methods GS				Methods GD			
Update	S2	S1	Scaling	NFV	NFG	NDC	Time	NFV	NFG	NDC	Time
BFGS	1	1	N	7253	7083	7062	4.35	7778	7529	7491	4.74
BFGS	1	1	Y	6902	6711	6681	3.72	8100	7818	7815	5.10
BFGS	1	2	N	12768	12765	7198	*11.85	8786	8774	7671	5.27
BFGS	1	2	Y	12522	12519	6926	*11.66	10678	10669	10080	7.97
BFGS	2	1	N	11939	11458	7496	*11.31	7957	7635	7671	5.27
BFGS	2	1	Y	11929	11761	6712	*11.32	10396	10175	10194	9.26
BFGS	2	2	N	13024	13021	6760	*12.87	10078	10069	9440	8.06
BFGS	2	2	Y	13946	13943	6673	*14.04	10606	10591	9960	7.20
R1	1	1	N	6488	6368	6354	3.46	6768	6615	6630	3.65
R1	1	1	Y	7262	6998	7129	4.15	7368	7134	7237	4.04
R1	1	2	N	8280	8280	7379	5.38	7652	7651	6976	4.71
R1	1	2	Y	6743	6743	6349	4.13	7549	7549	6905	4.64
R1	2	1	N	6483	6364	6355	4.27	7640	7353	7500	5.16
R1	2	1	Y	7202	6955	7068	4.97	7461	7205	7321	5.13
R1	2	2	N	6323	6587	6587	4.74	7034	7034	6585	4.99
R1	2	2	Y	8453	8449	7418	6.63	8348	8344	7383	6.38

Table 12: Structured hybrid methods

The last Table 13 contains the results obtained by various methods for sparse nonlinear least squares problems studied in this report. The results stated in this table are also demonstrated in more detail using graphs showing performance profiles [20].

Method	NFV	NFG	NDC	Time	Fail
VM	28476	24719	23887	11.57	1
VT	30565	25879	24972	18.36	2
VB	19209	16828	16216	12.04	-
MN	37793	37365	6440	8.26	-
GN	7525	7265	7139	4.16	-
MG	7236	7592	6802	4.10	-
GM	6480	6302	6191	3.10	-
GT	7209	6912	6800	3.72	-
GB	6816	6614	6510	3.79	-
GS	6488	6368	6354	3.46	-
GD	6768	6615	6630	3.65	-

Table 13: Comparison of various methods

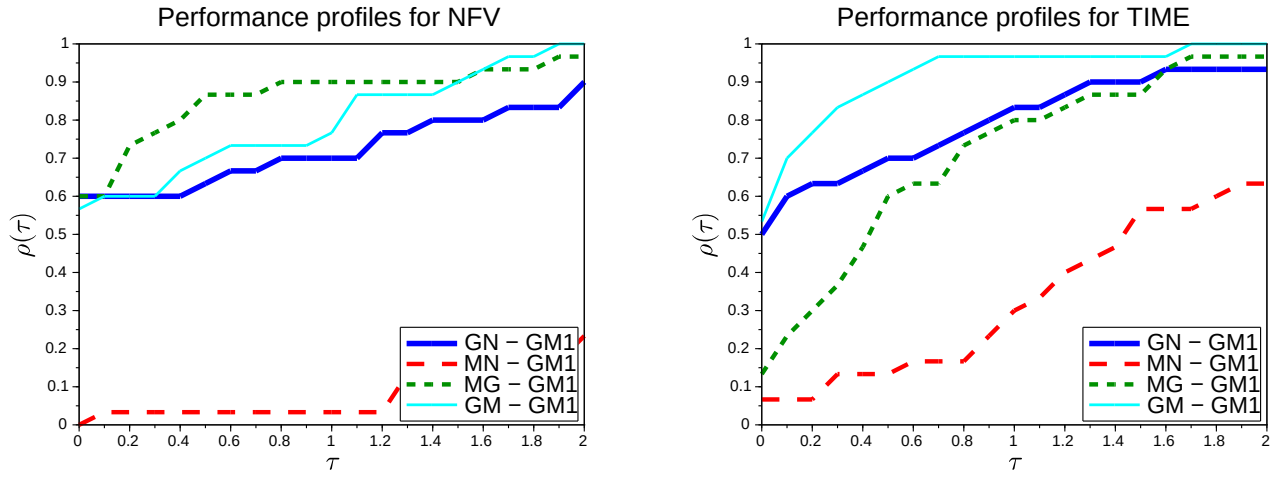

Figure 3: Methods for sparse problems

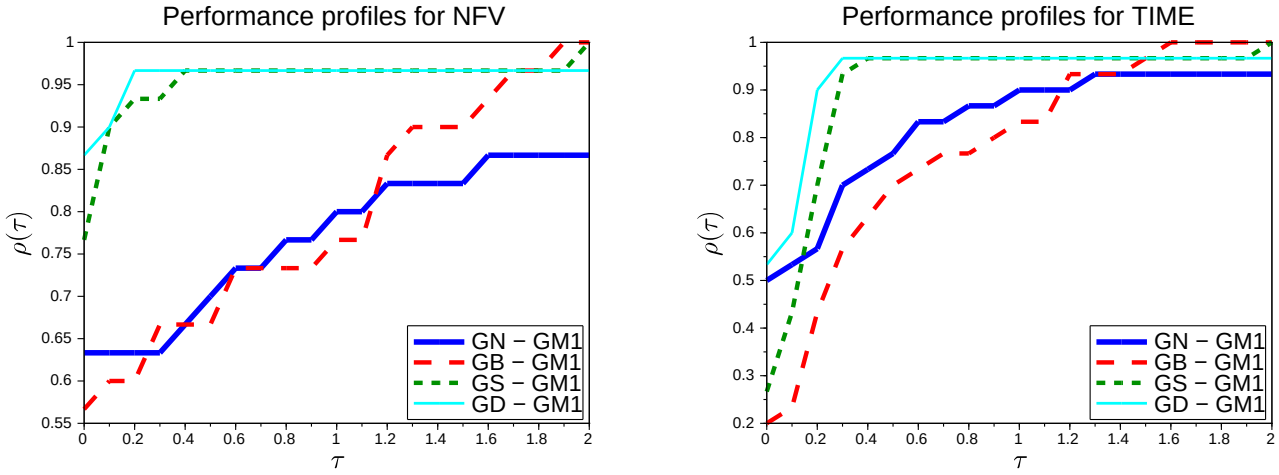

Figure 4: Methods for partially separable problems

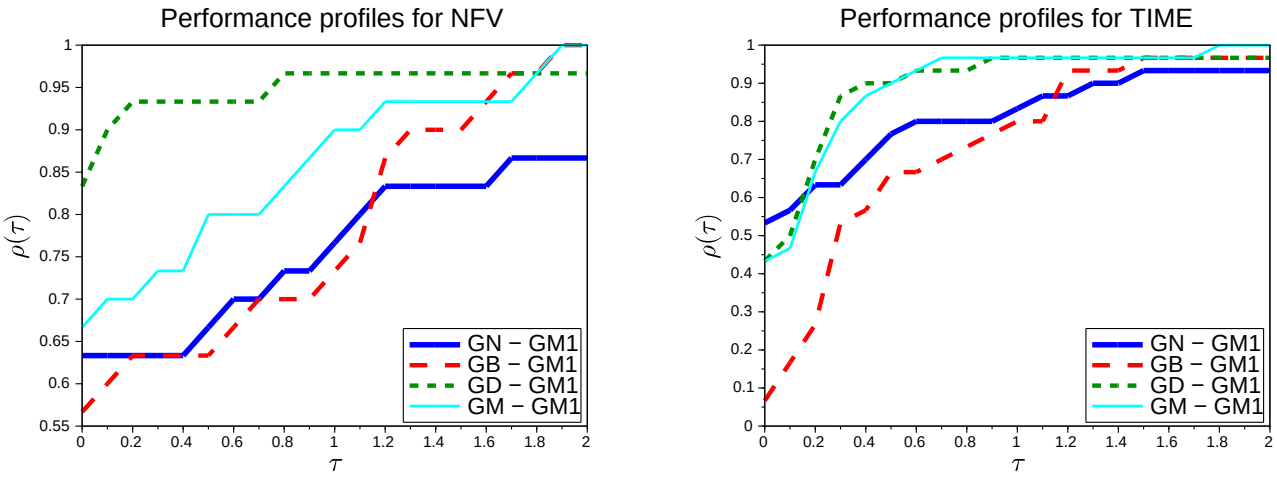

Figure 5: Comparison of efficient methods

From the results in this section we can draw several conclusions.

- Variable metric methods for sparse least squares problems are not too efficient because they lack the main advantage of classical variable metric methods (triangular decomposition update). It is necessary to decompose matrices B_i, $i \in N$, at each iteration step which requires a lot of arithmetic operations and slower convergence increases the number of function and gradient evaluations of the minimized function. Both prolong computational time.
- A difference version of the Newton method is relatively efficient (concerning the number of matrix decompositions which corresponds to the number of iterations) but a higher number of function and gradient evaluations of the minimized function needed for numerical computation of the Hessian matrix prolongs computation time, so this method is not competitive with the Gauss-Newton method.
- The Gauss-Newton method is rather efficient for solving sparse least squares problems (at least those contained in our collection TEST26).
- The combination of the Gauss-Newton method and the Newton method is very efficient for solving sparse least squares problems. A higher number of function and gradient evaluations of the minimized function needed for numerical computation of the Hessian matrix in corrected steps somewhat prolongs computational time.
- A simple hybrid method using a Marwil sparse update is surprisingly efficient even though it uses a variable metric method that gives bad results when used separately. A Marwil update is more efficient than a Toint update.
- A simple hybrid method using partitioned BFGS and R1 updates is less efficient because partitioned updates require a higher number of numerical operations than sparse updates and, furthermore, there is no possibility of a Choleski decomposition update which saves arithmetic operations.
- Structured hybrid methods `GD` and `GS` (together with the `GM` method) seem to be the most efficient for solving sparse (partially separable) least squares problems.

References

[1] M.Al-Baali, R.Fletcher: Variational methods for nonlinear least squares. J. Optimization Theory and Applications 36 (1985) 405-421.

[2] V.A.Barker, J.Dongarra, J.D.Croz, S.Hammarling, M.Marinova, J.Wasniewski, P.Yalamov: LAPACK95 Users' Guide. SIAM, Philadelphia, 2001.

[3] R.Bartels, L.Kaufman: Cholesky factor updating techniques for rank 2 matrix modifications. SIAM J. Matrix Analysis and Applications 10 (1989) 557-592.

[4] M.C.Biggs: The estimation of the Hessian matrix in nonlinear least squares problems with non-zero residuals. Mathematical Programming 12 (1977) 67-80.

[5] A.Björck: Numerical Methods in Matrix Computations. Springer, Heidelberg, 2015.

[6] C.G.Broyden: A class of methods for solving nonlinear simultaneous equations. Mathematics of Computation 19 (1965) 577-593.

[7] C.G.Broyden: Quasi-Newton methods and their application to function minimization. Mathematics of Computation 21 (1967) 368-381.

[8] C.G.Broyden: The convergence of a class of double rank minimization algorithms. Part 1 – general considerations. Part 2 – the new algorithm. J. Institute of Mathematics and its Applications 6 (1970) 76-90, 222-231.

[9] R.H.Byrd, D.C.Liu, J.Nocedal: On the behavior of Broyden's class of quasi-Newton methods. SIAM J. Optimization 2 (1992) 533-557.

[10] T.F.Coleman, B.S.Garbow, J.J.Moré: Algorithm 636. Fortran subroutines for estimating sparse Hessian matrices. ACM Transactions on Mathematical Software 11 (1985) 378-378.

[11] T.F.Coleman, J.J.Moré: Estimation of sparse Hessian matrices and graph coloring problems. Mathematical Programming 28 (1984) 243-270.

[12] A.R.Conn, N.I.M.Gould, P.L.Toint: Trust-Region Methods. SIAM, Philadelphia, 2000.

[13] J.W.Daniel, W.B.Gragg, L.Kaufman, G.W.Stewart: Reorthogonalization and Stable Algorithms for Updating the Gram-Schmidt QR Factorization. Mathematics of Computation 30 (1976) 772-795.

[14] W.C.Davidon: Variable metric method for minimization. SIAM J. Optimization 1 (1991) 1-17.

[15] J.E.Dennis: On some methods based on Broyden's secant approximation to the Hessian. In: Numerical Methods for Nonlinear Optimization (L.A.Lootsma ed.) Academic Press, New York, 1972.

[16] J.E.Dennis, H.J.Martinez, R.A.Tapia: Convergence theory for the structured BFGS secant method with an application to nonlinear least squares. J. Optimizaton Theory and Applications 61 (1989) 161-178.

[17] J.E.Dennis, H.H.W.Mei: An unconstrained optimization algorithm which uses function and gradient values. Report No. TR-75-246. Dept. of Computer Science, Cornell University, 1975.

[18] J.E.Dennis, J.J.Moré: A characterization of superlinear convergence and its application to quasi-Newton methods. Mathematics of Computation 28 (1974) 549-560.

[19] J.E.Dennis, H Wolkowicz: Sizing and least-change secant methods. SIAM J. on Numerical Analysis 30 (1993) 1291-1314.

[20] E.D.Dolan, J.J.Moré: Benchmarking optimization software with performance profiles. Mathematical Programming 91 (2002) 201-213.

[21] I.S.Duff, N.I.M.Gould, J.K.Reid, K.Turner: The factorization of sparse symmetric indefinite matrices. IMA Journal of Numerical Analysis 11 (1991) 181-204.

[22] R.Fletcher: A new approach to variable metric algorithms. Computer J. 13 (1970) 317-322.

[23] R.Fletcher, M.J.D.Powell: A rapidly convergent descent method for minimization. Computer J. 6 (1963) 163-168.

[24] A.George, W.H.Liu: Computer Solution of Large Sparse Positive Definite Systems. Prentice Hall, Englewood Cliffs, New Jersey 1981.

[25] P.E.Gill, W.Murray: Newton type methods for unconstrained and linearly constrained optimization. Mathematical Programming 7 (1974) 311-350.

[26] P.E.Gill, W.Murray, M.A.Saunders: Methods for computing and modifying LDV factors of a matrix. Mathematics of Computation 29 (1975) 1051-1077.

[27] D.Goldfarb: A family of variable metric algorithms derived by variational means. Math Comput. 24 (1970) 23-26.

[28] A.Griewank, P.L.Toint: Local convergence analysis for partitioned quasi-Newton updates. Numerische Mathematik 39 (1982) 429-448.

[29] A.Griewank, P.L.Toint: Partitioned variable metric updates for large-scale structured optimization problems. Numerische Mathematik 39 (1982) 119-137.

[30] S.Hoshino: A formulation of variable metric methods. J. Inst. Math. Appl. 10 (1972) 394-403.

[31] J.Huschens: On the use of product structure in secant methods for nonlinear least squares. SIAM J. Optimization 4 (1994) 108-129.

[32] C.L.Lawson, R.J.Hanson: Solving Least Squares Problems. SIAM, Philadelphia, 1995.

[33] L.Lukšan: Hybrid methods for large sparse nonlinear least squares. J. Optimization Theory and Applications 89 (1996) 575-595.

[34] L.Lukšan, C.Matonoha, J.Vlček: Sparse Test Problems for Nonlinear Least Squares. Technical Report V-1258. ICS AS CR, Prague, 2018.

[35] L.Lukšan, C.Matonoha, J.Vlček: Problems for Nonlinear Least Squares and Nonlinear Equations. Technical Report V-1259. ICS AS CR, Prague, 2018.

[36] L.Lukšan, E.Spedicato: Variable metric methods for unconstrained optimization and nonlinear least squares. Journal of Computational and Applied Mathematics 124 (2000) 61-93.

[37] L.Lukšan, M.Tůma, C.Matonoha, J.Vlček, N.Ramešová, M.Šiška, J.Hartman: UFO 2017. Interactive System for Universal Functional Optimization. Technical Report V-1252. ICS AS CR, Prague, 2017 (http://www.cs.cas.cz/luksan/ufo.pdf).

[38] L.Lukšan, J.Vlček: Computational experience with globally convergent descent methods for large sparse systems of nonlinear equations. Optimization Methods and Software 8 (1998) 201-223.

[39] L.Lukšan, J.Vlček: New quasi-Newton method for solving systems of nonlinear equations. Applications of Mathematics 62 (2017) 121-134.

[40] E.S.Marwil: Exploiting sparsity in Newton-like methods. Ph.D. Thesis, Cornell University, Ithaca 1978.

[41] J.J.Moré, D.C.Sorensen: Computing a trust region step. SIAM J. on Scientific and Statistical Computing 4 (1983) 553-572.

[42] J.Nocedal, S.J.Wright: Numerical Optimization. Springer, New York, 2006.

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

[44] M.J.D.Powell: A new algorithm for unconstrained optimization. In: Nonlinear Programming (J.B.Rosen O.L.Mangasarian, K.Ritter eds.) Academic Press, London, 1970.

[45] M.J.D. Powell: A note on quasi-Newton formulae for sparse second derivative matrices. Mathematical Programming 20 (1981) 144-151.

[46] M.J.D.Powell: Convergence properties of a class of minimization algorithms. In: Nonlinear Programming 2 (O.L. Mangasarian, R.R.Meyer, S.M.Robinson, eds.) Academic Press, London, 1975.

[47] M.J.D.Powell: On the global convergence of trust region algorithms for unconstrained minimization. Mathematical Programming 29 (1984) 297-303.

[48] S.Schlenkrich, A.Walther: Global convergence of quasi-Newton methods based on adjoint Broyden updates. Applied Numerical Mathematics 59 (2009) 1120-1136.

[49] D.F.Shanno: Conditioning of quasi-Newton methods for function minimization. Mathematics of Computation 24 (1970) 647-656.

[50] S.B.Sheng, Z.H.Zou: A new secant method for nonlinear least squares problems. Numerical Mathematics, A Journal of Chinese Universities 2 (1993) 125-137.

[51] T.Steihaug: Local and superlinear convergence for truncated iterated projections methods. Mathematical Programming 27 (1983) 176-190.

[52] T.Steihaug: The conjugate gradient method and trust regions in large-scale optimization. SIAM J. Numerical Analysis 20 (1983) 626-637.

[53] S.W.Thomas: Sequential estimation techniques for quasi-Newton algorithms. Thesis, Cornell University, Ithaca, New York, 1975.

[54] P.L.Toint: On large scale nonlinear least squares calculations. SIAM J. on Scientific and Statistical Computations 8 (1987) 416-435.

[55] P.L.Toint: On sparse and symmetric matrix updating subject to a linear equation. Mathematics of Computation 31 (1977) 954-961.

[56] P.L.Toint: Towards an efficient sparsity exploiting Newton method for minimization. In: Sparse Matrices and Their Uses (I.S.Duff, ed.), Academic Press, London, 1981, 57-88.

[57] J.Vlček, L.Lukšan: Shifted limited-memory variable metric methods for large-scale unconstrained minimization. J. of Computational and Applied Mathematics 186 (2006) 365-390.

[58] H.Yabe, T.Takahashi: Factorized quasi-Newton methods for nonlinear least squares problems. Mathematical Programming 51 (1991) 75-100.