A Quadratically Constrained Minimization Problem Arising from PDE of Monge-Ampère Type *

D.C. SORENSEN † and ROLAND GLOWINSKI‡

e-mail: sorensen@rice.edu, roland@math.uh.edu

July 10, 2008

Abstract

This note develops theory and a solution technique for a quadratically constrained eigenvalue minimization problem. This class of problems arises in the numerical solution of fully-nonlinear boundary value problems of Monge-Ampère type. Though it is most important in the three dimensional case, the solution method is directly applicable to systems of arbitrary dimension.

The focus here is on solving the minimization subproblem which is part of a method to numerically solve a Monge-Ampère type equation. These subproblems must be evaluated many times in this numerical solution technique and thus efficiency is of utmost importance.

A novelty of this minimization algorithm is that it is finite, of complexity $\mathcal{O}(n^3)$, with the exception of solving a very simple rational function of one variable. This function is essentially the same for any dimension. This result is quite surprising given the nature of the constrained minimization problem.

1 Introduction

1.1 Generalities

This note is a contribution to the numerical solution of the following fully-nonlinear three-dimensional boundary value problem which is an equation of Monge-Ampère type:

Find ψ such that

$$\lambda_1\lambda_2 + \lambda_2\lambda_3 + \lambda_3\lambda_1 = f \text{ in } \Omega,\ \psi = g \text{ on } \partial\Omega, \tag{1.1}$$

where (i) Ω is a bounded domain of $\mathbb{R}^3$, (ii) $\{\lambda_1, \lambda_2, \lambda_3\}$ is the spectrum of the Hessian $\mathbf{D}^2\psi = \left(\frac{\partial^2\psi}{\partial x_i, \partial x_j}\right)_{1\leq i,j\leq 3}$ of the unknown function ψ, (iii) f and g are two given functions with $f > 0$.

Fully nonlinear elliptic equations of the Monge-Ampère type are encountered in Differential Geometry, Fluid Mechanics, Finance and Stochastic Processes, Shape Design and many others. An excellent synopsis of these applications may be found in the report [6] by Chang , Guan and Yang describing a 2003 BIRS workshop on Monge-Ampère type equations and applications.

Problem (1.1) is known as the Dirichlet problem for the σ_2-operator. The first equation in (1.1) can be rewritten as:

$$|\nabla^2\psi|^2 - \mathbf{D}^2\psi : \mathbf{D}^2\psi = 2f \text{ in } \Omega, \tag{1.2}$$

*This work was supported in part by the NSF through Grants DMS-0412267, DMS-9972591, CCR-9988393 and ACI-0082645.

†Department of Computational and Applied Mathematics, MS 134, Rice University, Houston, Texas 77251-1892.

‡University of Houston, Department of Mathematics, 651 P. G. Hoffman Hall, Houston, Texas 77204-3008.

where

$$\mathbf{A} : \mathbf{B} \equiv \sum_{1 \le i,j \le d} a_{ij} b_{ij} = \mathrm{trace}\mathbf{A}^T\mathbf{B} \tag{1.3}$$

is the Frobenius scalar product of the two matrices $\mathbf{A}, \mathbf{B}$. It follows from (1.3) that the fully nonlinear partial differential equation in (1.2) can be rewritten as

$$[\mathrm{trace}\{\mathbf{D}^2\psi\}]^2 - \mathrm{trace}\{(\mathbf{D}^2\psi)^2\} = 2f. \tag{1.4}$$

Suppose that $\Omega = (0,1)^3$, $f = 1$ and $g = 0$. It is clear that despite the smoothness of f and g, the above problem can not have smooth solutions in $\bar{\Omega}$. *Viscosity solutions* provide a classical generalization to handle those situations where problem (1.1) has no classical solutions. An alternative to the viscosity solution approach is provided by the least squares solution. This least-squares methodology has been quite successful at solving two-dimensional fully nonlinear elliptic equations such as *Monge-Ampère's* and *Pucci's* (these least-squares solution methods are detailed in [1, 2, 3]).

Our goal here is to apply variants of these methods to the solution of the nonlinear Dirichlet problem when the σ_2-operator is *elliptic*. If the σ_2-operator is linearized in the neighborhood of ψ, we obtain the following linear second-order operator:

$$\phi \to 2[\nabla^2\psi\nabla^2\phi - \mathbf{D}^2\psi : \mathbf{D}^2\phi].$$

The coefficient matrix associated with the above linear operator is

$$2[\nabla^2\psi\mathbf{I} - \mathbf{D}^2\psi], \tag{1.5}$$

and the σ_2-operator will be elliptic in the neighborhood of ψ if and only if the matrix in (1.5) is either positive-definite or negative-definite, a.e. on Ω, that is:

$$(\lambda_1 + \lambda_2)(\lambda_2 + \lambda_3) > 0, (\lambda_1 + \lambda_3)(\lambda_1 + \lambda_2) > 0.$$

The remainder of this paper addresses a numerical solution of a restricted form of this problem. We shall require that solutions must satisfy the more restrictive constraints:

$$\lambda_1 + \lambda_2 > 0, \;\; \lambda_2 + \lambda_3 > 0, \;\; \lambda_1 + \lambda_3 > 0. \tag{1.6}$$

In the following section we shall discuss the least-squares solution of the σ_2-problem (1.1) assuming that the inequalities (1.6) hold. The solutions will be feasible for the original problem but not optimal since the feasible set of the modified problem is a proper subset of the original feasible set.

1.2 On the least-squares solution of the σ_2-problem

Hilbert spaces provide a natural framework for the least-squares solution of linear and nonlinear partial differential equation problems. The σ_2-problem is most naturally posed in the *Sobolev space* $\mathcal{H}^2(\Omega)$. This supposes necessarily that the data f and g satisfy:

$$(f, g) \in \mathcal{L}^1(\Omega) \times \mathcal{H}^{3/2}(\partial\Omega). \tag{1.7}$$

If (1.7) holds, the following space $\mathcal{V}_g$ and the set $\mathbf{Q}_f$

$$\begin{aligned} \mathcal{V}_g &= \{\phi | \phi \in \mathcal{H}^2(\Omega), \phi = g \text{ on } \partial\Omega\}, \\ \mathbf{Q}_f &= \{\mathbf{G} | \mathbf{G} \in (\mathcal{L}^2(\Omega))^{3\times 3}, \mathbf{G} = \mathbf{G}^T, \\ &\quad \mu_1\mu_2 + \mu_2\mu_3 + \mu_3\mu_1 = f, \;\; \mu_1 + \mu_2 > 0, \mu_2 + \mu_3 > 0, \mu_3 + \mu_1 > 0\} \end{aligned}$$

are non-empty (above, $\{\mu_1, \mu_2, \mu_3\}$ is the spectrum of matrix $\mathbf{G}$). The σ_2-problem (1.1) has a solution in $\mathcal{V}_g$ satisfying (1.6) if and only if

$$\mathbf{D}^2\mathcal{V}_g \cap \mathbf{Q}_f \neq \emptyset. \tag{1.8}$$

In order to address simultaneously those situations where either (1.8) or $\mathbf{D}^2\mathcal{V}_g \cap \mathbf{Q}_f = \emptyset$ hold, we consider, as in [1, 2, 3], the following least squares problem:

Find $(\psi, \mathbf{P}) \in \mathcal{V}_g \times \mathbf{Q}_f$ such that

$$J(\psi, \mathbf{P}) \leq J(\phi, \mathbf{G}), \forall (\phi, \mathbf{G}) \in \mathcal{V}_g \times \mathbf{Q}_f, \tag{1.9}$$

where

$$J(\phi, \mathbf{G}) = \frac{1}{2}\int_\Omega (\mathbf{D}^2\phi - \mathbf{G}) : (\mathbf{D}^2\phi - \mathbf{G})\mathrm{d}x,$$

with $\mathrm{d}x = \mathrm{d}x_1\mathrm{d}x_2\mathrm{d}x_3$.

In order to solve the minimization problem (1.9), we advocate the following block relaxation algorithm

$$\text{Given } \psi^0 \in \mathcal{V}_g,$$

for $k = 0, 1, 2, \ldots$

$$\mathbf{P}^{k+1} = argmin_{\mathbf{G}\in\mathbf{Q}_f} J(\psi^k, \mathbf{G}), \tag{1.10}$$

$$\psi^{k+1/2} = argmin_{\phi\in\mathcal{V}_g} J(\phi, \mathbf{P}^{k+1}), \tag{1.11}$$

$$\psi^{k+1} = \psi^k + \omega(\psi^{k+1/2} - \psi^k), \tag{1.12}$$

with ω a relaxation parameter (typically, $0 < \omega < 2$). The solution of the sub-problems (1.10) and (1.11) will be discussed in the following sections. To initialize the iteration, we take ψ^0 to be the unique solution in $\mathcal{V}_g$ to the Dirichlet problem

$$\nabla^2\psi^0 = \sqrt{3f} \text{ in } \Omega, \ \psi^0 \text{on } \partial\Omega. \tag{1.13}$$

Note that ψ^0 has the $\mathcal{H}^2(\Omega)$-regularity if $\partial\Omega$ is 'sufficiently' smooth or if Ω is convex.

1.3 Solution of the minimization sub-problems

Convexity and differentiability arguments will easily show that the minimization problem in (1.11) has a unique solution characterized by

$$\psi^{k+1/2} \in \mathcal{V}_g \text{ and } \int_\Omega \mathbf{D}^2\psi^{k+1/2} : \mathbf{D}^2\phi\mathrm{d}x = \int_\Omega \mathbf{P}^{k+1} : \mathbf{D}^2\phi\mathrm{d}x, \ \forall\phi \in \mathcal{H}^2(\Omega) \cap \mathcal{H}^1_0(\Omega) \tag{1.14}$$

which amounts to the Euler-Lagrange equation for problem (1.11), written in variational form. Following refs. [1, 2, 3], for the solution of (1.14), we advocate a *conjugate gradient algorithm* operating in the spaces $\mathcal{V}_g$ and $\mathcal{H}^2(\Omega) \cap \mathcal{H}^1_0(\Omega)$, both equipped with the scalar product

$$\langle v, w\rangle = \int_\Omega \nabla^2 v \nabla^2 w \mathrm{d}x$$

and the associated norm.

The minimization problem in (1.10) reads as follows:

Find $\mathbf{P}^{k+1} \in \mathbf{Q}_f$ such that

$$J_k(\mathbf{P}^{k+1}) \leq J_k(\mathbf{G}),\ \forall \mathbf{G} \in \mathbf{Q}_f, \tag{1.15}$$

where

$$J_k(\mathbf{G}) = \frac{1}{2}\int_\Omega \mathbf{G} : \mathbf{G} - \int_\Omega \mathbf{D}^2\psi^k : \mathbf{G}\mathrm{d}x,\ \forall \mathbf{G} \in \mathbf{Q},$$

with $\mathbf{Q} = \{\mathbf{G} | \mathbf{G} \in (\mathcal{L}^2(\Omega))^{3\times 3}, \mathbf{G} = \mathbf{G}^T\}$.

Problem (1.15) can be solved point-wise (in practice at the vertices of a finite element or finite difference mesh). We have to solve, for a.e. $x \in \Omega$, the following minimization problem:

$$\text{find } \mathbf{P}^{k+1}(x) \in \mathbf{E}(x),\ \text{ suchthat } \mathbf{j}_k(\mathbf{P}^{k+1}(x); x) \leq \mathbf{j}_k(\mathbf{A}; x),\ \ \forall \mathbf{A} \in \mathbf{E}(x), \tag{1.16}$$

where

$$\mathbf{E}(x) = \{\mathbf{A} | \mathbf{A} \in \mathbb{R}^{3\times 3},\ \ \mathbf{A} = \mathbf{A}^T,\ \ \mu_1\mu_2 + \mu_2\mu_3 + \mu_3\mu_1 = f(x),\ \ \mu_1 + \mu_3 > 0, \mu_2 + \mu_3 > 0, \mu_3 + \mu_1 > 0\}$$

and

$$\mathbf{j}_k(\mathbf{A}; x) = \frac{1}{2}\mathbf{A} : \mathbf{A} - \mathbf{D}^2\psi^k(x) : \mathbf{A},$$

with $\{\mu_1, \mu_2, \mu_3\}$ being the spectrum of $\mathbf{A}$.

The minimization problem (1.16) can be normalized by dividing the objective functional $\mathbf{j}_k(\mathbf{A}; x)$ with $f(x)$. Let $\mathbf{B} = \frac{1}{\sqrt{f(x)}}\mathbf{D}^2\psi^k(x)$. Then solving

$$\min_{\mathbf{A}\in\mathbf{E}_1} \mathrm{trace}[\mathbf{A}^T(\mathbf{A} - 2\mathbf{B})] = \min_{\mathbf{A}\in\mathbf{E}_1} \mathrm{trace}[\mathbf{A}^2 - 2\mathbf{A}\mathbf{B})] \tag{1.17}$$

where

$$\mathbf{E}_1 = \{\mathbf{A} | \mathbf{A} \in \mathbb{R}^{3\times 3}, \mathbf{A} = \mathbf{A}^T, \mu_1\mu_2 + \mu_2\mu_3 + \mu_3\mu_1 = 1,\ \ \mu_1 + \mu_3 > 0, \mu_2 + \mu_3 > 0, \mu_3 + \mu_1 > 0\}$$

(with $\{\mu_1, \mu_2, \mu_3\}$ being the spectrum of $\mathbf{A}$) and then replacing the solution $\mathbf{A} \leftarrow \sqrt{f(x)}\mathbf{A}$ will give the solution to (1.16). The remainder of this note will be devoted to the numerical solution of problem (1.17).

It is convenient for the sequel to express Problem (1.17) for general dimension n with the constraints formulated in matrix form:

Problem Qmin

$$\begin{aligned} &\min \mathrm{trace}\{\mathbf{A}\mathbf{A} - 2\mathbf{B}\mathbf{A}\} \\ &\text{s.t.} \\ &\ell^T\mathbf{M}\ell = 2 \\ &\mathbf{M}\ell \geq \mathbf{0} \end{aligned}$$

where

$$\mathbf{M} = \mathbf{e}\mathbf{e}^T - \mathbf{I},\ \text{ with } \mathbf{e}^T = (1, 1, \ldots, 1),$$

and

$$\begin{aligned} \mathbf{B} &= \mathbf{B}^T \text{ is specified,} \\ \mathbf{A} &= \mathbf{A}^T = \mathbf{Q}\Lambda\mathbf{Q}^T, \\ \Lambda &= \mathrm{diag}(\ell). \end{aligned}$$

While $n = 3$ is perhaps most relevant, the Monge-Ampère equation has been studied in arbitrary dimensions [4, 5]. Moreover, the algorithm we shall develop is completely general and applies to arbitrary dimensions.

We wish to point out that $\mathrm{trace}(\mathbf{B}) \neq 0$ at iterate $k = 0$, if algorithm (1.10)-(1.12) is initialized as specified in (1.13). We note also that close to a solution of Problem 1.1, $\mathrm{trace}(\mathbf{B}) \neq 0$ is implied by (1.4). The condition $\mathrm{trace}(\mathbf{B}) \neq 0$ will have implications for the numerical method we shall develop. This point will be discussed further subsequent to the development of our method.

2 Formulation and Analysis of a Solution Method for Problem Qmin

Note that Problem Qmin is equivalent to

$$\min \mathrm{trace}\{\Lambda\Lambda - 2\hat{\mathbf{B}}\Lambda\} = \min \ell^T\ell - 2\mathbf{b}^T\ell \tag{2.1}$$

$$\text{s.t.} \tag{2.2}$$

$$\ell^T\mathbf{M}\ell = 2 \tag{2.3}$$

$$\mathbf{M}\ell \geq \mathbf{0} \tag{2.4}$$

where

$$\hat{\mathbf{B}} = \mathbf{Q}^T\mathbf{B}\mathbf{Q} \text{ and } \mathbf{b} = \mathrm{diag}(\hat{\mathbf{B}}).$$

We shall first consider (2.1) with the orthogonal matrix $\mathbf{Q}$ considered to be fixed. To begin, we shall show that there is no finite solution to (2.1) with an active inequality constraint. This will then lead to an unconstrained minimization problem on the manifold specified by the equality constraint.

Lemma 2.1 *If the vector $\ell \in \mathbb{R}^n$ is finite and feasible, then none of the inequality constraints can be active. In other words,*

$$\ell^T\mathbf{M}\ell = 2 \;\Rightarrow\; \mathbf{e}_j^T\mathbf{M}\ell > 0, \text{ for } j = 1, 2, \ldots, n.$$

Proof: It is sufficient to consider any one of the inequality constraints since the same argument can be applied to all of the others simply by re-labeling the variables. Let $\ell^T = (\lambda_1, \lambda_2, \ldots, \lambda_n)$. If the first inequality constraint $\mathbf{e}_1^T\mathbf{M}\ell$ is active then

$$\lambda_2 + \lambda_3 + \ldots + \lambda_n = 0. \tag{2.5}$$

The equality constraint $\ell^T\mathbf{M}\ell = 2$ provides

$$\begin{aligned} \lambda_1(\lambda_2 + \lambda_3 + \ldots + \lambda_n) = 2 \quad &- \quad \lambda_2(\lambda_3 + \lambda_4 + \ldots \lambda_n) \\ &- \quad \lambda_3(\lambda_2 + \lambda_4 + \lambda_5 + \ldots \lambda_n) \\ &\ldots \\ &- \quad \lambda_n(\lambda_2 + \lambda_3 + \ldots + \lambda_{n-1}). \end{aligned}$$

From (2.5) it follows that

$$0 = \lambda_1(\lambda_2 + \lambda_3 + \ldots + \lambda_n) = 2 + \lambda_2^2 + \lambda_3^2 \ldots + \lambda_n^2 \geq 2$$

which is a contradiction. ■

Since no inequality can be active, the minimization problem can be solved by a Lagrange multiplier argument involving only the equality constraint. The Lagrangian will be

$$\mathcal{L}(\ell, \mu) = \ell^T\ell - 2\mathbf{b}^T\ell + \mu(\ell^T\mathbf{M}\ell - 2).$$

Setting the gradient of the Lagrangian to zero gives the linear equation

$$(\mathbf{I} + \mu\mathbf{M})\ell = \mathbf{b}. \tag{2.6}$$

If $1/\mu$ is not an eigenvalue of $-\mathbf{M}$ then the equality constraint becomes

$$\mathbf{b}^T(\mathbf{I} + \mu\mathbf{M})^{-1}\mathbf{M}(\mathbf{I} + \mu\mathbf{M})^{-1}\mathbf{b} = 2 \tag{2.7}$$

To understand this equation, it is helpful to know the eigensystem of $\mathbf{M}$.

Lemma 2.2 *The matrix* $\mathbf{M} = \mathbf{e}\mathbf{e}^T - \mathbf{I}$ *has an eigenvalue* $\omega_1 = n - 1$ *corresponding to the eigenvector* $\mathbf{e}$ *and another eigenvalue* $\omega_2 = -1$ *of multiplicity* $n - 1$ *corresponding to the eigenspace* $\mathrm{span}\{\mathbf{e}\}^\perp$. *An eigenvector matrix for* $\mathbf{M}$ *may be constructed as a Householder transformation*

$$\mathbf{U} \equiv (\mathbf{I} - 2\mathbf{w}\mathbf{w}^T), \text{ with } \mathbf{w} = (\mathbf{e} + \sqrt{n}\mathbf{e}_1)/\|\mathbf{e} + \sqrt{n}\mathbf{e}_1\|,$$

where $\mathbf{e}_1^T = (1, 0, \ldots 0)$.

Proof: The eigensystem of $\mathbf{M}$ follows immediately from the eigensystem of $\mathbf{e}\mathbf{e}^T$ which obviously has $\mathbf{e}$ as an eigenvector corresponding to eigenvalue n along with an $n - 1$ dimensional null space that must be orthogonal to $\mathrm{span}\{\mathbf{e}\}$. The Householder transformation $\mathbf{U}$ is an orthogonal matrix constructed such that

$$\mathbf{U}\mathbf{e} = -\mathbf{e}_1\sqrt{n},$$

and since

$$\mathbf{U}\mathbf{e}_1 = \mathbf{U}^T\mathbf{e}_1 = -\mathbf{e}/\sqrt{n},$$

the first column of $\mathbf{U}$ is a normalized eigenvector corresponding to the eigenvalue $\omega_1 = n-1$. Since the remaining columns of $\mathbf{U}$ are orthogonal to the first column, they must form an orthogonal basis for the eigenvalue $\omega_2 = -1$.

This is easily checked since

$$\mathbf{U}^T\mathbf{M}\mathbf{U} = \mathbf{U}\mathbf{M}\mathbf{U}^T = \mathbf{U}\mathbf{e}\mathbf{e}^T\mathbf{U}^T - \mathbf{I} = n\mathbf{e}_1\mathbf{e}_1^T - \mathbf{I}.$$

■

With this simple result, we can greatly simplify equation (2.7). Let $\Omega = n\mathbf{e}_1\mathbf{e}_1^T - \mathbf{I}$. Then substituting $\mathbf{M} = \mathbf{U}\Omega\mathbf{U}^T$ into (2.7) gives

$$\frac{\beta_1^2\omega_1}{(1 + \mu\omega_1)^2} = 2 + \frac{\beta_2^2}{(1 - \mu)^2} \tag{2.8}$$

where $(\beta_1, \mathbf{b}_2^T) = \mathbf{b}^T\mathbf{U}$ and $\beta_2^2 = \mathbf{b}_2^T\mathbf{b}_2$. Note further that $\beta_1 = \mathbf{e}^T\mathbf{b}/\sqrt{n}$ is *invariant* since $\mathbf{e}^T\mathbf{b} = \mathrm{trace}\{\mathbf{B}\}$ regardless of the choice of the orthogonal matrix $\mathbf{Q}$. Figure 1 shows a graph of the secular equation (2.8).

This secular equation has precisely two solutions unless $\mathrm{trace}\{\mathbf{B}\} = 0$. If the trace is nonzero, this simple rational equation in the scalar variable μ may be solved numerically using a simple safeguarded Newton method. However, as the trace nears 0, this equation becomes increasingly difficult to solve numerically. A better approach is to solve the "reciprocal square root" equivalent form of this equation:

$$\pm(1 + \mu\omega_1) = \frac{(1 - \mu)|\beta_1|\sqrt{\omega_1}}{\sqrt{2(1 - \mu)^2 + \beta_2^2}}. \tag{2.9}$$

In Figure 1 a graph of the left hand side linear function (solid blue for plus sign, dashed blue for minus sign) is shown in the right hand graph. The right hand side of (2.9) is the solid black line. To solve this equation, a starting point of $\mu = -1/\omega_1$ is taken in either case and Newton method is applied. This does not need safeguarding

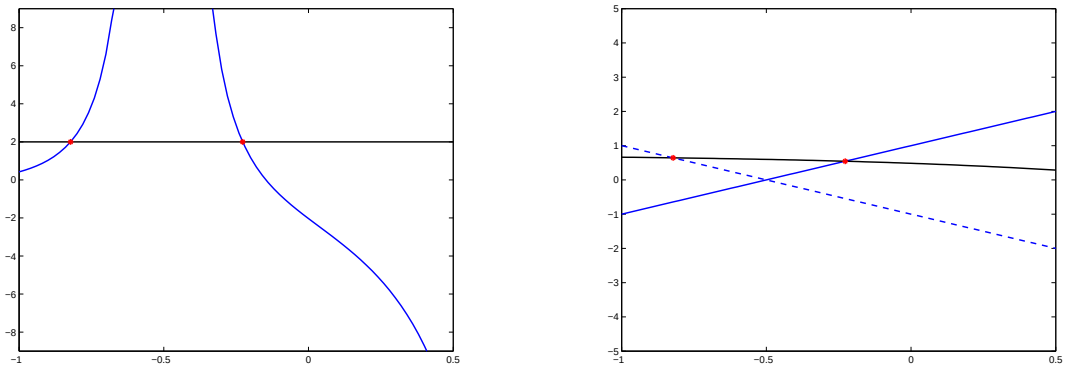

Figure 1: The Secular Equation for Multiplier μ (left) and the Reciprocal Square Root Secular Equation (right)

and typically converges in 3-4 steps. Moreover, this equivalent reformulation does not suffer from numerical breakdown when $\beta_1 \approx 0$. In fact, when $\beta_1 = 0$, i.e., when $\mathrm{trace}\{\mathbf{B}\} = 0$, the two solutions coalesce to the double root $\mu_1 = \mu_2 = -1/\omega_1$.

Choosing the correct sign: The correct choice of sign in Equation 2.9 and hence the correct root μ can be determined from $\mathbf{e}^T\mathbf{b} = \mathrm{trace}(\mathbf{B})$. Since the right hand side of Equation 2.9 must be positive, it follows that the choice of sign in $\pm(1 + \mu\omega_1)$ must be " + " if $1 + \mu\omega_1 > 0$ and "$-$" otherwise. From Equation 2.6 we have

$$(1 + \mu\omega_1)\mathbf{e}^T\ell = \mathbf{e}^T(\mathbf{I} + \mu\mathbf{M})\ell = \mathbf{e}^T\mathbf{b}.$$

and this shows that $\mathrm{sign}(1 + \mu\omega_1) = \mathrm{sign}(\mathbf{e}^T\mathbf{b})$ since $\omega_1\mathbf{e}^T\ell = \mathbf{e}^T\mathbf{M}\ell > 0$.

Assuming it is solved for the scalar μ, the solution ℓ is then constructed via

$$\ell = \mathbf{U}\mathbf{c} = \mathbf{c} - \mathbf{w}(2\mathbf{w}^T\mathbf{c}), \quad \text{with } \mathbf{c}^T = (\beta_1/(1 + \mu\omega_1), \frac{1}{1-\mu}\mathbf{b}_2^T).$$

Comment: The quadratic objective function in Problem Qmin and its transformation (2.1) must have bounded and hence compact level sets due to the dominance of the quadratic term over the linear term. This property is preserved when one of the variables is explicitly eliminated via the equality constraint so the problem must have a feasible minimum.

Now, we return to the full problem and find a very surprising result. Suppose that $\mathbf{Q}$ is chosen to be an eigenvector matrix for $\mathbf{B}$ so that $\hat{\mathbf{B}} = \mathbf{Q}^T\mathbf{B}\mathbf{Q} = \mathrm{diag}(\mathbf{b})$ is diagonal with the entries of the vector $\mathbf{b}$ being the eigenvalues of the original $\mathbf{B}$. Then, construct μ and the corresponding feasible solution $\ell = (\mathbf{I} + \mu\mathbf{M})^{-1}\mathbf{b}$. With this construction, the following lemma holds:

Lemma 2.3 *Let $\Lambda = \mathrm{diag}(\ell)$ where μ and ℓ are constructed as described above to solve (2.1). Then $\mathbf{A_Q} = \mathbf{Q}\Lambda\mathbf{Q}^T$ solves*

$$\min \mathrm{trace}\{\mathbf{AA} - 2\mathbf{BA}\}$$

over all $\mathbf{A} = \mathbf{W}\Lambda\mathbf{W}^T$ with $\mathbf{W}^T\mathbf{W} = \mathbf{I}$.

Proof:

Any orthogonal matrix $\mathbf{W}$ must be of the form $\mathbf{W} = \mathbf{Q}\hat{\mathbf{Q}}$ where $\hat{\mathbf{Q}} = \mathbf{Q}^T\mathbf{W}$ is orthogonal. Hence $\mathbf{W}^T\mathbf{B}\mathbf{W} = \hat{\mathbf{Q}}^T\hat{\mathbf{B}}\hat{\mathbf{Q}}$. Let $\hat{\mathbf{b}} = \mathrm{diag}(\hat{\mathbf{Q}}^T\hat{\mathbf{B}}\hat{\mathbf{Q}})$. It is easily seen that the j-th entry of $\hat{\mathbf{b}}$ is

$$\hat{\beta}_j \equiv \hat{\mathbf{b}}(j) = \mathbf{q}_j^T\hat{\mathbf{B}}\mathbf{q}_j = \sum_{i=1}^{n} \beta_i\gamma_{ij}^2,$$

where $\beta_j = \mathbf{b}(j)$, γ_{ij} is the i,j-th entry of the orthogonal matrix $\hat{\mathbf{Q}}$ and $\mathbf{q}_j$ is its j-th column. From this it follows that

$$\hat{\mathbf{b}} = \mathbf{G}^T\mathbf{b}, \text{ where the } i,j-\text{th entry of } \mathbf{G} \text{ is } \gamma_{ij}^2.$$

Note that the orthogonality of $\hat{\mathbf{Q}}$ implies that **G**is doubly stochastic: $\mathbf{Ge} = \mathbf{G}^T\mathbf{e} = \mathbf{e}$.

The result will be established by demonstrating that

$$\ell^T\ell - 2\mathbf{b}^T\ell \leq \ell^T\ell - 2\hat{\mathbf{b}}^T\ell.$$

Of course, it is sufficient to show that

$$\ell^T\hat{\mathbf{b}} - \ell^T\mathbf{b} = \ell^T(\hat{\mathbf{b}} - \mathbf{b}) \leq 0.$$

Therefore, let us consider the vector $\hat{\mathbf{b}} - \mathbf{b}$.

$$\begin{aligned}
\hat{\mathbf{b}} - \mathbf{b} &= (\mathbf{G}^T - \mathbf{I})\mathbf{b} && (2.10)\\
&= (\mathbf{G}^T - \mathbf{I})(\mathbf{I} + \mu\mathbf{M})\ell && (2.11)\\
&= (\mathbf{G}^T - \mathbf{I})((1-\mu)\mathbf{I} + \mu\mathbf{e}\mathbf{e}^T)\ell && (2.12)\\
&= (1-\mu)(\mathbf{G}^T - \mathbf{I})\ell, && (2.13)
\end{aligned}$$

since $(\mathbf{G}^T - \mathbf{I})\mathbf{e} = \mathbf{0}$. Therefore,

$$\begin{aligned}
\ell^T(\hat{\mathbf{b}} - \mathbf{b}) &= (1-\mu)\ell^T(\mathbf{G}^T - \mathbf{I})\ell\\
&= (1-\mu)\frac{1}{2}(\ell^T(\mathbf{G}^T - \mathbf{I})\ell + \ell^T(\mathbf{G} - \mathbf{I})\ell)
\end{aligned}$$

The i-th diagonal element of both $\mathbf{G}^T - \mathbf{I}$ and $\mathbf{G} - \mathbf{I}$ is

$$\gamma_{ii}^2 - 1 = -\sum_{j\neq i}\gamma_{ij}^2 = -\sum_{j\neq i}\gamma_{ji}^2.$$

Hence, the j-th entry of $\mathbf{x} = (\mathbf{G} - \mathbf{I})\ell$ is

$$\mathbf{x}(j) = \sum_{i\neq j}\gamma_{ij}^2(\lambda_i - \lambda_j),$$

while the i-th entry of $\mathbf{y} = (\mathbf{G}^T - \mathbf{I})\ell$ is

$$\mathbf{y}(i) = \sum_{j\neq i}\gamma_{ij}^2(\lambda_j - \lambda_i).$$

From this, it follows that in the terms of the innerproduct

$$\ell^T(\hat{\mathbf{b}} - \mathbf{b}) = \ell^T\mathbf{y} + \ell^T\mathbf{x}$$

we shall find the term $\gamma_{ij}^2\lambda_i(\lambda_j - \lambda_i)$ resulting from $\lambda_i\mathbf{y}(i)$ along with the term $\gamma_{ij}^2\lambda_j(\lambda_i - \lambda_j)$ resulting from $\lambda_j\mathbf{x}(j)$. There is exactly one such pair of terms for each index pair i, j with $i \neq j$.

Adding these two terms together gives

$$\gamma_{ij}^2\lambda_i(\lambda_j - \lambda_i) + \gamma_{ij}^2\lambda_j(\lambda_i - \lambda_j) = \gamma_{ij}^2(\lambda_j - \lambda_i)(\lambda_i - \lambda_j) = -\gamma_{ij}^2(\lambda_j - \lambda_i)^2.$$

It follows that

$$\ell^T(\hat{\mathbf{b}} - \mathbf{b}) = -(1-\mu)\sum_{i\neq j}\gamma_{ij}^2(\lambda_j - \lambda_i)^2 \leq 0,$$

since $(1 - \mu) > 0$. This concludes the proof. ■

We have experimented numerically with the alternating iteration

1. $\min_\ell \ell^T\ell - 2\mathbf{b}^T\ell$ subject to constraints
2. $\min_{\mathbf{W}} \ell^T\ell - 2\mathbf{b}^T\ell$ with $\mathbf{b} = diag(\mathbf{W}^T\mathbf{B}\mathbf{W})$

and found this converged in 2 steps with $\mathbf{W}^T\mathbf{B}\mathbf{W}$ diagonal. This result explains why that happened and indicates that if $\mathbf{B}$ is diagonalized at the outset then the alternating iteration terminates in one step with the fixed point.

The following lemma provides the converse.

Lemma 2.4 *Suppose $\mathbf{A} = \mathbf{W}\Lambda\mathbf{W}^T$ solves Problem Qmin. Then there is an orthogonal $\hat{\mathbf{Q}}$ such that $\mathbf{Q} = \mathbf{W}\hat{\mathbf{Q}}$ diagonalizes $\mathbf{B}$ and $\mathbf{A} = \mathbf{Q}\Lambda\mathbf{Q}^T$. In other words, if $\mathbf{A}$ solves Problem Qmin, then $\mathbf{A} = \mathbf{Q}\Lambda\mathbf{Q}^T$ with $\mathbf{Q}^T\mathbf{B}\mathbf{Q}$ diagonal.*

Proof: Suppose $\mathbf{A} = \mathbf{W}\Lambda\mathbf{W}^T$ solves Problem Qmin. Let $\Lambda = diag(\lambda_1, \lambda_2, \ldots, \lambda_n)$. Let the elements of $\mathbf{W}^T\mathbf{B}\mathbf{W}$ be denoted by β_{ij}.

Claim: If $\beta_{ij} \neq 0$ for some off diagonal element ($i \neq j$), then $\lambda_i = \lambda_j$ must hold. To establish the claim, suppose $\lambda_i \neq \lambda_j$. Without loss of generality, assume $i = 1, j = 2$ (which can be arranged by a symmetric permutation). Consider the leading 2×2 leading principal submatrix $\begin{pmatrix} \beta_{11} & \beta_{12} \\ \beta_{12} & \beta_{22} \end{pmatrix}$ and let $\gamma^2 + \sigma^2 = 1$ so that

$$\begin{pmatrix} \hat{\beta}_{11} & \hat{\beta}_{12} \\ \hat{\beta}_{12} & \hat{\beta}_{22} \end{pmatrix} = \begin{pmatrix} \gamma & \sigma \\ -\sigma & \gamma \end{pmatrix}\begin{pmatrix} \beta_{11} & \beta_{12} \\ \beta_{12} & \beta_{22} \end{pmatrix}\begin{pmatrix} \gamma & -\sigma \\ \sigma & \gamma \end{pmatrix}$$

is an orthogonal similarity transformation. A little computation will provide

$$\begin{aligned} \lambda_1\hat{\beta}_{11} + \lambda_2\hat{\beta}_{22} &= \lambda_1(\beta_{11}\gamma^2 + 2\beta_{12}\gamma\sigma + \beta_{22}\sigma^2) + \lambda_2(\beta_{11}\sigma^2 - 2\beta_{12}\gamma\sigma + \beta_{22}\gamma^2) \\ &= \lambda_1\beta_{11} + \lambda_2\beta_{22} + \sigma^2(\lambda_1 - \lambda_2)\beta_{12}\left[2\frac{\gamma}{\sigma} + \frac{\beta_{22} - \beta_{11}}{\beta_{12}}\right], \end{aligned}$$

where $\gamma^2 = 1 - \sigma^2$ has been substituted and terms collected to arrive at the final equality. We are free to choose the sign of $\gamma = \pm\sqrt{1 - \sigma^2}$ and then by choosing σ small enough it will be possible to make the term

$$\theta \equiv \sigma^2(\lambda_1 - \lambda_2)\beta_{12}\left[2\frac{\gamma}{\sigma} + \frac{\beta_{22} - \beta_{11}}{\beta_{12}}\right]$$

positive. No other diagonal elements of $\mathbf{W}^T\mathbf{B}\mathbf{W}$ are modified. If we replace $\hat{\mathbf{W}} \leftarrow \mathbf{W}\mathbf{H}$ where $\mathbf{H}$ is an $n \times n$ plane rotation in the (1,2) plane defined by (γ, σ) and put $\hat{\mathbf{A}} = \hat{\mathbf{W}}\Lambda\hat{\mathbf{W}}^T$ then

$$\mathrm{trace}\hat{\mathbf{A}}\hat{\mathbf{A}} - 2\hat{\mathbf{A}}\mathbf{B} = \mathrm{trace}\mathbf{A}\mathbf{A} - 2\mathbf{A}\mathbf{B} - 2\theta < \mathrm{trace}\mathbf{A}\mathbf{A} - 2\mathbf{A}\mathbf{B}$$

contradicting the optimality of $\mathbf{A}$.

Therefore, there is a permutation $\mathbf{P}$ such that $\hat{\Lambda} \equiv \mathbf{P}\Lambda\mathbf{P}^T = diag(\Lambda_1, \Lambda_2, \ldots, \Lambda_m)$ with each $\Lambda_j = \lambda_j \mathbf{I}_{k_j}$ where the λ_j are now the distinct eigenvalues of $\mathbf{A}$ and k_j denotes multiplicity. Applying the permutation to $\mathbf{W}^T\mathbf{B}\mathbf{W}$ must produce a block diagonal matrix

$$\mathbf{P}\mathbf{W}^T\mathbf{B}\mathbf{W}\mathbf{P}^T = diag(\mathbf{B}_1, \mathbf{B}_2, \ldots, \mathbf{B}_m) \text{ with } k_j = \mathrm{order}(\mathbf{B}_j),$$

since any nonzero in an off diagonal block would violate the condition of the above claim. Now, let

$$\tilde{\mathbf{Q}} = diag(\mathbf{Q}_1, \mathbf{Q}_2, \ldots, \mathbf{Q}_m) \text{ with } \mathbf{Q}_j^T\mathbf{B}_j\mathbf{Q}_j \text{ diagonal for } j = 1, 2, \ldots, m.$$

Put, $\mathbf{Q} = \mathbf{P}^T\tilde{\mathbf{Q}}$. Note that $\hat{\Lambda} = \mathbf{Q}^T\Lambda\mathbf{Q}$ and that

$$\mathbf{A} = \mathbf{W}\Lambda\mathbf{W}^T = (\mathbf{W}\hat{\mathbf{Q}})(\hat{\mathbf{Q}}^T\Lambda\hat{\mathbf{Q}})(\hat{\mathbf{Q}}^T\mathbf{W}^T).$$

Therefore,

$$\mathbf{A} = \mathbf{Q}\hat{\Lambda}\mathbf{Q}^T \text{ with } \mathbf{Q}^T\mathbf{B}\mathbf{Q} \text{ diagonal.}$$

This concludes the proof. ■

3 Computational Results

We did a considerable number of test problems for the case $n = 3$. Figure 2 shows surface and contour plots of the objective function subject to the constraints. The red * gives the co-ordinates of the computed minimizer.

The Matlab function `randn(n)` was used to generate the 3×3 matrices $\mathbf{B}$. This function returns a matrix containing pseudo-random entries drawn from a normal distribution with mean zero and standard deviation. Each randomly generated matrix $\mathbf{B}$ was symmetrized via $\mathbf{B} \leftarrow \frac{1}{2}(\mathbf{B} + \mathbf{B}^T)$.

The average number of iterations for the inverse square root iteration was 4.26 taken over 100,000 random symmetric matrices $\mathbf{B}$. The elapsed time (using Matlab's `tic, toc` commands) was 23 seconds to solve the 100K quadratically constrained minimization problems. There were no failures. The computations were done on a Laptop with an Intel Duo Core processor in Matlab under Windows XP.

A two dimensional version of this minimization algorithm has been applied to the solution of the Dirichlet problem for the two-dimensional Monge-Ampère equation, namely

$$\det\mathbf{D}^2\psi = f(> 0) \text{ in } \Omega, \;\; \psi = g \text{ on } \partial\Omega.$$

The corresponding numerical results coincide (to working precision) with those reported in [1] and [3], including the case where $\Omega = (0, 1) \times (0, 1)$, $f = 1$ and $g = 0$, a situation for which the Monge-Ampère equation has no classical solutions. This latter case justifies our recourse to least-squares solutions (see [1] and [3] for details) as a better approach to preserving the boundary condition than the well-known viscosity solution alternative.

The minimization algorithm presented and analyzed here is valid for arbitrary n and we intend to construct a 3-D code based upon this approach.

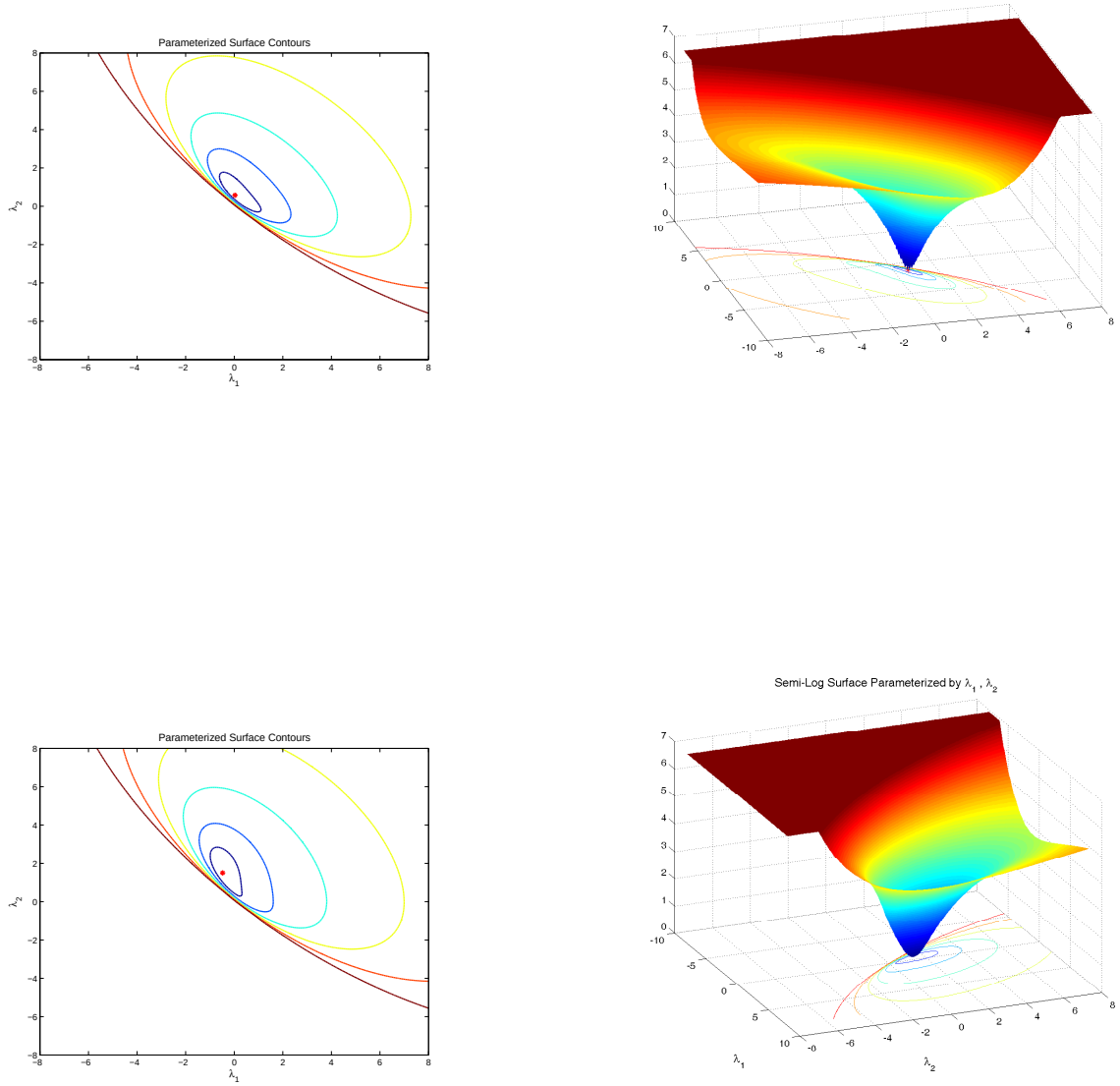

Figure 2: Contour (left) and surface (right) plots of the surface $\ell^T\ell - 2\mathbf{b}^t\ell$ subject to the constraints $\ell^T\mathbf{M}\ell = 2,\ \mathbf{M}\ell \leq 0$.

4 Acknowledgements

The authors would like to thank Prof. L.A. Caffarelli for encouraging us to investigate the numerical solution of the σ_2 problem, and Prof. Yin Zhang for discussions on the nature of the quadratically constrained minimization problem. We would also like to thank Prof. Ed Dean for pointing out an earlier coding error and for implementing a Fortran version and testing it on the $n = 2$ case within the context of the algorithm for the Monge-Ampère problem described in [1].

References

[1] E.J. Dean and R. Glowinski, Numerical solution of the two-dimensional Monge-Ampère equation with Dirichlet boundary conditions: a least-squares approach, *C. R. Acad. Sci. Paris, Ser. I*, **339**(12), 887-892, (2004).

[2] E.J. Dean and R. Glowinski, On the numerical solution of a two-dimensional Pucci's equation with Dirichlet boundary conditions: a least-squares approach, *C. R. Acad. Sci. Paris, Serie I*, **341**, 375-380, (2005).

[3] R. Glowinski, E.J. Dean, G. Guidoboni, L.H. Juarez and T.W. Pan, Applications of operator-splitting methods to the direct numerical simulation of particulate and free-surface flows and to the numerical solution of the two- dimensional Monge-Ampère equation, *Jap. J. Industr. Appl. Math.*, **25**, 1- 63, (2008).

[4] D.B.Fairlie and A.N. Leznov, General solutions of the Monge-Ampère equation in n-dimensional space, *Journal of Geometry and Physics*, **16** , 385-390, (1995).

[5] D.B.Fairlie and A.N. Leznov,The General Solution of the Complex Monge-Ampère Equation in two dimensional space, preprint August (1999)
http://arxiv.org/abs/solv-int/9909014/ (accessed 27 May 08).

[6] A. Chang , P. Guan and P. Yang , Monge-Ampère Type Equations and Applications, BIRS Workshop Report, August, (2003)
http://www.birs.ca/workshops/2003/03w5067/report03w5067.pdf (accessed 27 May 08).

5 Appendix:

Algorithm 1:

```
function [A,lam,err] = QconMinPos(B);
%
%  This routine solves  min trace(A'A - 2B'A)
%                        A
%
%           s.t.  lam'M lam = 2,  M lam > 0,
%
%           with  M = ones(n) - eye(n).
%
%  Input:   B    - a symmetric matrix of order n.
%
%  Output:  A    - a symmetric matrix, the minimizer.
%
%           lam  - the eigenvalues of A.
%
%           err  - scalar indicating faults:
%                  err = 0 , successful solution
%                  err = 1 , max iters exceeded
%                  err = 2 , B is not symmetric
%
%-----------------------------------------------------------------------
%  D.C. Sorensen
%  2 July 08
%

   if (norm(B - B') > 100*eps), err = 2; return, end  % error: B is not symmetric

   [Q,D] = eig(B);
   b = diag(D);
   n = length(b);

   [mu,lam,err] = minpsi_inv(b);

   A = Q*diag(lam)*Q';


function [mu,lam,err] = minpsi_inv(b);
%
%  Input:  b     an n-vector
%
%  Output: mu    a  scalar  solution to
%                secular equation
%
%          lam   an n-vector solution to
%                constrained min problem
%
%          err   a scalar indicating error (err = 1)
%                if too many iters (default maxit = 30);
%
%
```

```
%  Newton's method is applied to the reciprocal
%  square root equations (equivalent to secular
%  equation) to find the multiplier mu (see calling code).
%
%  The given vector b  is diag(B) from
%  the calling code, with components expressed
%  in the eigenbasis of M (see calling code)
%
%-----------------------------------------------------------------------
%  D.C. Sorensen
%  30 June 08
%
   maxit = 30; err = 0;
   n = length(b);
   traceB = sum(b);
%
%  Express components of b  in the eigenvector
%  basis of M = e*e' - I
%
   z = ones(n,1); z(1) = z(1) + sqrt(n);
   z = z/norm(z);
   bo = b - z*(2*z'*b);
%
%  Compute weights of terms in secular equation
%
   b1 = bo(1)*bo(1); b2 = bo(2:n)'*bo(2:n);
   w = n-1;
   mu = -1/w;
   swb1 = sqrt(w)*abs(bo(1));

   iter = 0;
%
%  Iterative solution of inverse square root equation
%
   f1 = 1; f2 = 0;
   if (traceB > 0), sgn = 1; else sgn = -1; end
      df1 = sgn*w;
      while (abs(f1-f2) > 1000*eps & iter < maxit),
          iter = iter + 1;
          d1 = 1 + w*mu;
          d2 = 1 - mu;
          f1 = sgn*d1;
          df2 = b2 + 2*d2*d2;
          f2 = swb1*d2/sqrt(df2);
          df2 = -swb1*b2/(sqrt(df2)*df2);
          mu = mu + (f2 - f1)/(df1 - df2);
      end
      lam = (b - ones(n,1)*(traceB*mu/d1))/d2;

   if (iter >= maxit), err = 1; end % Error: max iters exceeded
```