AN EXACT PENALTY METHOD FOR SMOOTH EQUALITY CONSTRAINED OPTIMIZATION WITH APPLICATION TO MAXIMUM LIKELIHOOD ESTIMATION [1]

Bergsma, W.[2] and Rapcsák, T.[3]

Abstract. A new exact penalty function is presented which turns a smooth constrained nonlinear optimization problem into an unconstrained one. The advantage of the proposed penalty method is that arbitrary positive penalty parameters ensure local optimality, avoiding this way the possible ill-conditioning of the problem.

In a statistical example, the method was successfully applied to maximum likelihood estimation of a class of marginal models for categorical data, involving a large number of unknown parameters. Some theoretical results are given for general maximum likelihood problems with constraints.

1. Introduction

The Lagrange multiplier rule was introduced in 1762, and later, in "Lagrange, J.L., Méchanique analytique I-II., Paris, 1788" for minimizing a function subject to equality constraints. Based on the Lagrange multiplier rule, penalty methods were developed in order to eliminate some or all of the constraints and add to the objective function a penalty term which prescribes a high cost to infeasible points. A good survey on Lagrange multiplier methods can be found in Bertsekas (1982).

In 1943, Courant introduced the quadratic penalty method where the penalty term is the squared Euclidean norm of the constraint violations. In 1970, Fletcher studied the Lagrange function depending only on the variables, then, gave the theoretical justification of a class of exact penalty methods for solving smooth equality constrained nonlinear optimization problems. Exact penalty methods were intensively investigated and a well-prepared survey was published by Di Pillo (1994). A new smooth exact penalty function was suggested by Christianson (1995).

The Lagrange multiplier rule was further developed by Rapcsák (1991, 1997) who combined the optimization theory with Riemannian geometry in order to describe the geometric structure of smooth nonlinear optimization problems by tensors and to extend the local

[1] Research partially supported by the Hungarian Scientific Research Fund, Grant Nos. OTKA-T043276, OTKA-T043241 and Netherlands Organization for Scientific Research Grant 400-20-001.

[2] EURANDOM, P.O.Box 513, 5600 MB Eindhoven, The Netherlands.

[3] Computer and Automation Research Institute, Hungarian Academy of Sciences, Budapest, Hungary.

results of Lagrange to global ones. In [21], the idea of Fletcher (1970) to define smooth exact penalty functions and that of Courant (1943) to use a quadratic penalty term were reconsidered and developed further by the global version of the global Lagrange multiplier rule clarifying the geometric meaning as well.

In the paper, a new exact penalty function is presented for solving smooth equality constrained nonlinear optimization problems the advantage of which is that arbitrary positive penalty parameters ensure local optimality. This exact penalty function has numerical conditioning similar to the problem functions from which it is constructed. A penalty function like this can be used to establish termination properties for algorithms which avoid ill-conditioned steps.

In Section 2, the constrained optimization problem is formulated and the necessary and sufficient optimality conditions are presented. In Section 3, the Lagrange multiplier functions depending on the variables are considered, and the derivatives of these multiplier functions at an optimal point are obtained. In Section 4, a new exact penalty function is introduced. This function has a local minimum at an optimal point of the constrained problem for the arbitrary positive values of the penalty parameter. In Section 5, a quasi-Newton algorithm is proposed which achieves superlinear convergence, and each iterative step is no more difficult to be performed than the evaluation of the penalty function itself. In Section 6, a statistical example is presented which leads to a smooth optimization problem. This example concerns a class of statistical models whose theoretical properties were described in detail in Bergsma and Rudas (2002) and whose application was described in Croon, Bergsma and Hagenaars (2000) and references therein. In Section 7, numerical experience related to the class of statistical models introduced in the preceding part are reported, in Section 8, the statistical interpretation of the Lagrange function and of the exact penalty function are investigated, and in Section 9, some conclusions are drawn.

2. Formulation of the problem

For $f : \mathbf{R}^n \to \mathbf{R}$ and $\mathbf{h} : \mathbf{R}^n \to \mathbf{R}^{\,n-k}, k \geq 0$, consider the following nonlinear optimization problem (NOP):

$$\begin{gathered} \text{minimize } f(\mathbf{x}) \\ \text{subject to } \mathbf{x} \in M = \{\mathbf{x} \in R^n \mid \mathbf{h}(\mathbf{x}) = 0\}\,, \end{gathered} \tag{2.1}$$

and the Lagrangian function

$$L(\mathbf{x}, \boldsymbol{\lambda}) = f(\mathbf{x}) - \boldsymbol{\lambda}^T \mathbf{h}(\mathbf{x}), \qquad \mathbf{x} \in R^n, \qquad \boldsymbol{\lambda} \in R^{n-k}, \tag{2.2}$$

with Lagrange multiplier vector $\boldsymbol{\lambda}$. The first and second derivatives of L with respect to $\mathbf{x}$

are

$$\nabla L(\mathbf{x}, \boldsymbol{\lambda}) = \nabla f(\mathbf{x}) - \sum_{i=1}^{n-k} \lambda_i \nabla h_i(\mathbf{x}) = \nabla f(\mathbf{x}) - \boldsymbol{\lambda}^T J\mathbf{h}(\mathbf{x}), \qquad \mathbf{x} \in R^n, \tag{2.3}$$

$$HL(\mathbf{x}, \boldsymbol{\lambda}) = Hf(\mathbf{x}) - \sum_{i=1}^{n-k} \lambda_i Hh_i(\mathbf{x}) = Hf(\mathbf{x}) - \boldsymbol{\lambda}^T H\mathbf{h}(\mathbf{x}), \qquad \mathbf{x} \in R^n, \tag{2.4}$$

where the symbol ∇ denotes gradient vectors which are row vectors, the symbol H Hessian matrices, the $(n-k)$-tuple $H\mathbf{h}^T = (Hh_1, \ldots, Hh_{n-k})$, and $J\mathbf{h}$ the $(k \times n)$ Jacobian matrix of the mapping $\mathbf{h}$.

Let us assume that the regularity condition

$$r\big(J\mathbf{h}(\mathbf{x})\big) = n - k, \qquad \mathbf{x} \in M, \tag{2.5}$$

holds where $r(J\mathbf{h}(\mathbf{x}))$ denotes the rank of the Jacobian matrix at the point $\mathbf{x}$. Then, the constraint set M is a differentiable manifold (see, e.g., Rapcsák, 1997). The tangent spaces of M are given by

$$TM_{\mathbf{x}} = \{\mathbf{v} \in R^n \mid J\mathbf{h}(\mathbf{x})\mathbf{v} = 0\}, \qquad \mathbf{x} \in M,$$

and the projection matrix onto the space spanned by the columns of $J\mathbf{h}(\mathbf{x})$ in the form of

$$P(\mathbf{x}) = J\mathbf{h}(\mathbf{x})^T \big(J\mathbf{h}(\mathbf{x}) J\mathbf{h}(\mathbf{x})^T\big)^{-1} J\mathbf{h}(\mathbf{x}), \qquad \mathbf{x} \in M. \tag{2.6}$$

The first-order necessary optimality condition of problem (2.1) at an optimal point $\mathbf{x}^* \in M$ is the existence of a Lagrange multiplier vector $\boldsymbol{\lambda}^* \in R^{n-k}$ such that

$$\nabla L(\mathbf{x}^*, \boldsymbol{\lambda}^*) = \mathbf{0}. \tag{2.7}$$

The second-order necessary condition of problem (2.1) for optimality is to fulfil the first-order necessary condition (2.7), and that the matrix

$$\big(HL(\mathbf{x}^*, \boldsymbol{\lambda}^*)\big)_{|TM} = \big(I - P(\mathbf{x}^*)\big) HL(\mathbf{x}^*, \boldsymbol{\lambda}^*) \big(I - P(\mathbf{x}^*)\big) \tag{2.8}$$

should be positive semidefinite where the symbol $_{|TM}$ denotes the restriction of a matrix to the tangent space of M. A second-order sufficient condition of local minimality is to fulfil the first-order necessary condition (2.7) and the positive definiteness of $(HL(\mathbf{x}^*, \boldsymbol{\lambda}^*))_{|TM}$, where the latter is equivalent to the positive definiteness of the second covariant derivatives of the function f on M with respect to the induced Riemannian metric (Rapcsák, 1997).

3. Lagrange multiplier functions

The pseudoinverses of the full-rank matrices $J\mathbf{h}(\mathbf{x})$, $\mathbf{x} \in M$, can be written as

$$J\mathbf{h}(\mathbf{x})^{+} = J\mathbf{h}(\mathbf{x})^{T}\big(J\mathbf{h}(\mathbf{x})J\mathbf{h}(\mathbf{x})^{T}\big)^{-1}, \qquad \mathbf{x} \in M. \tag{3.1}$$

From the first-order optimality condition, the Lagrange multipliers can be extended in the form of

$$\boldsymbol{\lambda}(\mathbf{x})^{T} = \nabla f(\mathbf{x}) J\mathbf{h}(\mathbf{x})^{+}, \qquad \mathbf{x} \in M, \tag{3.2}$$

from which

$$\nabla L\big(\mathbf{x}, \boldsymbol{\lambda}(\mathbf{x})\big) = \nabla f(\mathbf{x})\big(I - P(\mathbf{x})\big), \qquad \mathbf{x} \in M. \tag{3.3}$$

Hence, $\nabla L(\mathbf{x}, \boldsymbol{\lambda}(\mathbf{x}))$, $\mathbf{x} \in M$, are the orthogonal projections of the gradients $\nabla f(\mathbf{x})$, $\mathbf{x} \in M$, with respect to the Euclidean metric onto the tangent spaces of M, which are the first covariant derivatives of the function f on M with respect to the induced Riemannian metric (Rapcsák, 1997).

The following lemma will turn out to be important:

Lemma 3.1.

(1) *If* $\mathbf{x} \in M$ *satisfies* $\nabla L\big(\mathbf{x}, \boldsymbol{\lambda}(\mathbf{x})\big) = \mathbf{0}$, *then,*

$$J\boldsymbol{\lambda}(\mathbf{x})^{T} = HL\big(\mathbf{x}, \boldsymbol{\lambda}(\mathbf{x})\big) J\mathbf{h}(\mathbf{x})^{+}. \tag{3.4}$$

(2)

$$J\boldsymbol{\lambda}(\mathbf{x})^{T} = HL\big(\mathbf{x},\boldsymbol{\lambda}(\mathbf{x})\big) J\mathbf{h}(\mathbf{x})^{+} + \nabla L\big(\mathbf{x}, \boldsymbol{\lambda}(\mathbf{x})\big) J\Big(J\mathbf{h}(\mathbf{x})^{T}\Big)\Big(J\mathbf{h}(\mathbf{x})J\mathbf{h}(\mathbf{x})^{T}\Big)^{-1}, \quad \mathbf{x} \in M, \tag{3.5}$$

where the (i,j,k)*th entry of the 3-dimensional matrix* $J\big(J\mathbf{h}(\mathbf{x})^{T}\big)$ *is* $\dfrac{\partial^2 h_j}{\partial x_i \partial x_k}, \forall (i,j,k)$.

Proof. By (3.2) ,

$$\mathbf{0} = \Big(\nabla f(\mathbf{x}) - \boldsymbol{\lambda}(\mathbf{x})^{T} J\mathbf{h}(\mathbf{x})\Big) J\mathbf{h}(\mathbf{x})^{T}, \qquad \mathbf{x} \in M. \tag{3.6}$$

By differentiating equality (3.6) , we have the matrix equation

$$\mathbf{0} = J\Big[\Big(\nabla f(\mathbf{x}) - \boldsymbol{\lambda}(\mathbf{x})^{T} J\mathbf{h}(\mathbf{x})\Big) J\mathbf{h}(\mathbf{x})^{T}\Big] =$$
$$\Big(HL\big(\mathbf{x},\boldsymbol{\lambda}(\mathbf{x})\big) - J\boldsymbol{\lambda}(\mathbf{x})^{T} J\mathbf{h}(\mathbf{x})\Big) J\mathbf{h}(\mathbf{x})^{T} + \Big(\nabla f(\mathbf{x}) - \boldsymbol{\lambda}(\mathbf{x})^{T} J\mathbf{h}(\mathbf{x})\Big) J\Big(J\mathbf{h}(\mathbf{x})^{T}\Big), \quad \mathbf{x} \in M. \tag{3.7}$$

By using the assumption $\nabla L\big(\mathbf{x},\boldsymbol{\lambda}(\mathbf{x})\big)=\mathbf{0}$, we obtain that

$$J\boldsymbol{\lambda}(\mathbf{x})^T J\mathbf{h}(\mathbf{x}) J\mathbf{h}(\mathbf{x})^T = HL\big(\mathbf{x},\boldsymbol{\lambda}(\mathbf{x})\big) J\mathbf{h}(\mathbf{x})^T, \qquad \mathbf{x}\in M,$$

from which

$$J\boldsymbol{\lambda}(\mathbf{x})^T = HL\big(\mathbf{x},\boldsymbol{\lambda}(\mathbf{x})\big) J\mathbf{h}(\mathbf{x})^T \Big(J\mathbf{h}(\mathbf{x}) J\mathbf{h}(\mathbf{x})^T\Big)^{-1}, \qquad \mathbf{x}\in M,$$

which is equivalent to (3.4).

From (3.7), we obtained (3.5). ■

A similar analysis can be found in Luenberger (1972).

4. An exact penalty function

Consider the penalty function

$$\mathcal{P}(\mathbf{x};q) = L\big(\mathbf{x},\boldsymbol{\lambda}(\mathbf{x})\big) + \frac{1}{2}\mathbf{h}(\mathbf{x})^T W(\mathbf{x})\mathbf{h}(\mathbf{x}) + \frac{1}{2}q\,\mathbf{h}(\mathbf{x})^T V(\mathbf{x})\mathbf{h}(\mathbf{x}), \qquad \mathbf{x}\in\mathcal{M}, \tag{4.1}$$

where $\mathcal{M}=\{\mathbf{x}\in R^n | J\mathbf{h}(\mathbf{x}) \text{ is of full rank}\}$,

$$W(\mathbf{x}) = J\mathbf{h}(\mathbf{x})^{+T} HL\big(\mathbf{x},\boldsymbol{\lambda}(\mathbf{x})\big) J\mathbf{h}(\mathbf{x})^{+}, \qquad \mathbf{x}\in\mathcal{M}, \tag{4.2}$$

and $V(\mathbf{x})$, $\mathbf{x}\in\mathcal{M}$, is any $(n-k)\times(n-k)$ positive definite matrix function, for example, the identity matrix or

$$V(\mathbf{x}) = \big(J\mathbf{h}(\mathbf{x}) J\mathbf{h}(\mathbf{x})^T\big)^{-1}, \qquad \mathbf{x}\in\mathcal{M}, \tag{4.3}$$

and $q>0$ is a constant. The term $\frac{1}{2}q\,\mathbf{h}(\mathbf{x})^T V(\mathbf{x})\mathbf{h}(\mathbf{x})$ is referred to as a penalty term. The advantage of the use of the function $\mathcal{P}$ is that it has a strict local minimum value at the optimal point of problem (2.1) for arbitrary positive parameters.

Theorem 4.1. *Let $\mathbf{x}^*$ be a local optimal point of NOP* (2.1) *and suppose $V(\mathbf{x}^*)$ is positive definite. Then, the function $\mathcal{P}$ has a strict local minimum at $\mathbf{x}^*$ for all $q>0$.*

Proof. It is sufficient to show that the gradient of $\mathcal{P}$ is equal to zero at $\mathbf{x}^*$ and the second derivative matrix function of $\mathcal{P}$ is positive definite at the local optimal point $\mathbf{x}^*$. The first derivative is

$$\begin{aligned}\nabla\mathcal{P}(\mathbf{x};q) = {} & \nabla L\big(\mathbf{x},\boldsymbol{\lambda}(\mathbf{x})\big) - \mathbf{h}(\mathbf{x})^T J\boldsymbol{\lambda}(\mathbf{x}) + \mathbf{h}(\mathbf{x})^T W(\mathbf{x}) J\mathbf{h}(\mathbf{x}) \\ & + q\mathbf{h}(\mathbf{x})^T V(\mathbf{x}) J\mathbf{h}(\mathbf{x}) + \mathbf{r}(\mathbf{x})^T, \qquad \mathbf{x}\in\mathcal{M},\end{aligned} \tag{4.4}$$

where the components of $\mathbf{r}$ are quadratic in $\mathbf{h}$. Thus, the first-order necessary optimality condition holds at $\mathbf{x}^*$. Note that, by using Lemma 3.1, if $\nabla L\big(\mathbf{x}, \boldsymbol{\lambda}(\mathbf{x})\big) = 0$ for some $\mathbf{x} \in M$, then,

$$J\boldsymbol{\lambda}(\mathbf{x})^T J\mathbf{h}(\mathbf{x}) = HL\big(\mathbf{x}, \boldsymbol{\lambda}(\mathbf{x})\big)P(\mathbf{x}).$$

Hence, the second derivative evaluated at a stationary point reduces to

$$\begin{aligned} H\mathcal{P}(\mathbf{x}; q) =& HL\big(\mathbf{x}, \boldsymbol{\lambda}(\mathbf{x})\big) - J\boldsymbol{\lambda}(\mathbf{x})^T J\mathbf{h}(\mathbf{x}) - J\mathbf{h}(\mathbf{x})^T J\boldsymbol{\lambda}(\mathbf{x}) \\ &+ J\mathbf{h}(\mathbf{x})^T W(\mathbf{x}) J\mathbf{h}(\mathbf{x}) + qJ\mathbf{h}(\mathbf{x})^T V(\mathbf{x}) J\mathbf{h}(\mathbf{x}) \\ =& HL(\mathbf{x}, \boldsymbol{\lambda}(\mathbf{x})) - HL(\mathbf{x}, \boldsymbol{\lambda}(\mathbf{x}))P(\mathbf{x}) - P(\mathbf{x})HL(\mathbf{x}, \boldsymbol{\lambda}(\mathbf{x})) \\ &+ P(\mathbf{x})HL\big(\mathbf{x}, \boldsymbol{\lambda}(\mathbf{x})\big)P(\mathbf{x}) + qJ\mathbf{h}(\mathbf{x})V(\mathbf{x})J\mathbf{h}(\mathbf{x})^T \\ =& \big(I - P(\mathbf{x})\big)HL\big(\mathbf{x}, \boldsymbol{\lambda}(\mathbf{x})\big)\big(I - P(\mathbf{x})\big) + qJ\mathbf{h}(\mathbf{x})^T V(\mathbf{x}) J\mathbf{h}(\mathbf{x}) \\ =& \Big(HL\big(\mathbf{x}, \boldsymbol{\lambda}(\mathbf{x})\big)\Big)_{|TM} + qJ\mathbf{h}(\mathbf{x})V(\mathbf{x})J\mathbf{h}(\mathbf{x})^T, \qquad q > 0. \end{aligned} \tag{4.5}$$

This matrix is positive definite at any local optimal point for arbitrary positive definite matrix $V(\mathbf{x})$, since $\Big(HL\big(\mathbf{x}, \boldsymbol{\lambda}(\mathbf{x})\big)\Big)_{|TM}$ is nonnegative definite at a local optimal point and the second term is positive definite. ■

Example 4.1.

Let us solve the problem

$$\min x_2$$
$$x_1^2 + x_2^2 = 1, \;\; (x_1, x_2) \in \mathbf{R}^2.$$

Let us introduce the notation

$$M = \left\{(x_1, x_2) \in \mathbf{R}^2 \mid x_1^2 + x_2^2 = 1\right\}.$$

As $f(\mathbf{x}) = x_2,\ h(\mathbf{x}) = \frac{1}{2}x_1^2 + \frac{1}{2}x_2^2 - \frac{1}{2}, \;\; \mathbf{x} \in \mathbf{R}^2,$

$\nabla f(\mathbf{x}) = (0,\, 1)\,,$
$\nabla h(\mathbf{x}) = (x_1,\, x_2)\,,$
$\|\nabla h(\mathbf{x})\|^2 = x_1^2 + x_2^2 = 1,$
$\lambda(\mathbf{x}) = \nabla f(\mathbf{x})\ \nabla h(\mathbf{x})^T = x_2,$

$$Hf(\mathbf{x}) = \begin{pmatrix} 0 & 0 \\ 0 & 0 \end{pmatrix},$$

$$Hh(\mathbf{x}) = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix},$$

$$P(\mathbf{x}) = \nabla h(\mathbf{x})^T \ \nabla h(\mathbf{x}) = \begin{pmatrix} x_1^2 & x_1x_2 \\ x_1x_2 & x_2^2 \end{pmatrix},$$

$$I - P(\mathbf{x}) = \begin{pmatrix} 1-x_1^2 & x_1x_2 \\ x_1x_2 & 1-x_2^2 \end{pmatrix} = \begin{pmatrix} x_2^2 & x_1x_2 \\ x_1x_2 & x_1^2 \end{pmatrix}, \qquad \mathbf{x} \in M.$$

In order to determine the stationary points of the problem, the following system should be solved:

$$x_1x_2 = 0,$$

$$x_2^2 = 1,$$

$$x_1^2 + x_2^2 = 1.$$

The solutions are $x_2 = \pm 1, \quad x_1 = 0,$ *thus, the stationary points are*

$$(0,-1) \quad \textit{and} \quad (0,+1)\,,$$

and the global minimum point is $\mathbf{x}^* = (0,-1)$.

By calculating the second covariant derivatives on M*, we obtain that*

$$D^2f(\mathbf{x}) = \Big(Hf(\mathbf{x}) - \lambda(\mathbf{x})\, Hh(\mathbf{x})\Big)_{|TM} = -x_2\Big(I - P(\mathbf{x})\Big)\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}\Big(I - P(\mathbf{x})\Big) =$$

$$-x_2\Big(I - P(\mathbf{x})\Big) = -x_2\begin{pmatrix} x_2^2 & x_1x_2 \\ x_1x_2 & x_1^2 \end{pmatrix}, \qquad \mathbf{x} \in M.$$

Since the matrices $I - P(\mathbf{x})$, $\mathbf{x} \in M$*, are positive semidefinite,* D^2f *is positive semidefinite iff* $x_2 \leq 0$*. It follows that the objective function is geodesic convex with respect to the induced Riemannian metric in this domain.*

Let us consider the case when the function

$$L\Big(\mathbf{x},\lambda(\mathbf{x})\Big) = x_2 - \frac{x_2h(\mathbf{x})}{x_1^2+x_2^2},\ \mathbf{x} \in \mathcal{M}, \quad \mathcal{M} = \big\{\mathbf{x} \in \mathbf{R}^2 |\ \nabla h(\mathbf{x}) \neq 0\big\} = \big\{\mathbf{x} \in \mathbf{R}^2 \setminus \{\mathbf{0}\}\big\}\,,$$

has to be minimized. Then,

$$\nabla L\Big(\mathbf{x},\lambda(\mathbf{x})\Big) = \frac{1}{x_1^2+x_2^2}\left(\frac{2h(\mathbf{x})x_1x_2}{x_1^2+x_2^2} - x_1x_2,\ x_1^2 + \frac{2h(\mathbf{x})x_2^2}{x_1^2+x_2^2} - h(\mathbf{x})\right), \qquad \mathbf{x} \in \mathcal{M},$$

$$HL\Big(\mathbf{x},\lambda(\mathbf{x})\Big) = \frac{x_1^2+x_2^2-2h(\mathbf{x})}{(x_1^2+x_2^2)^3}\begin{pmatrix} -x_2(-3x_1^2+x_2^2) & -x_1(x_1^2-3x_2^2) \\ -x_1(x_1^2-3x_2^2) & -3x_2(-3x_1^2+x_2^2) \end{pmatrix}, \qquad \mathbf{x} \in \mathcal{M}.$$

It follows that the function $L(\mathbf{x},\lambda(\mathbf{x}))$, $\mathbf{x} \in\mathcal{M}$, *is neither convex, nor concave.*

Let us determine the stationary points by solving the following system:

$$\frac{2h(\mathbf{x})x_1x_2}{x_1^2+x_2^2} - x_1x_2 = 0, \qquad x_1^2 + \frac{2h(\mathbf{x})x_2^2}{x_1^2+x_2^2} - h(\mathbf{x}) = 0, \qquad (x_1x_2) \in \mathcal{M}.$$

The function $L\left(\mathbf{x},\lambda(\mathbf{x})\right)$, $\mathbf{x} \in\mathcal{M}$, *has the following two stationary points:*

$$(0, -1)\,, \quad (0, +1)\,,$$

which coincide with the preceding ones. It follows that the global minimum point of the function $L\Big(\mathbf{x}, \lambda(\mathbf{x})\Big), \mathbf{x} \in \mathcal{M}$, *is* $\mathbf{x}^* = (0, -1)$.

Let us consider the exact penalty function given by formulas (4.1),(4.2), (4.3) as follows:

$$\mathcal{P}(\mathbf{x}; q) = x_2\left(1 - \frac{h(\mathbf{x})}{(x_1^2+x_2^2)} - \frac{h(\mathbf{x})^2}{2(x_1^2+x_2^2)^2} + q\frac{h(\mathbf{x})^2}{2(x_1^2+x_2^2)^2}\right), \qquad \mathbf{x} \in\mathcal{M},$$

where $q > 0$. *For general* q, *there are at most seven stationary points. The first two are* $(0, -1)$ *and* $(0, 1)$ *not depending on* q. *Then,* $(0, k)$, *where* k *is a real root of the equation*

$$2qk^3 + 3k^2 + 2qk - 3 = 0$$

if such a root exists. Finally, if

$$-\frac{3}{2}3^{\frac{1}{4}} \leq q \leq \frac{3}{2}3^{\frac{1}{4}}\,,$$

there are stationary points at (x_1, x_2) *with*

$$x_1 = \pm\sqrt{3}\sqrt{q^2(-4+2\sqrt{3}) - 27 + 18\sqrt{3}}\,.$$

and

$$x_2 = \frac{1}{9}(3-\sqrt{3})q\,.$$

In the case of $\mathbf{x}^* = (0, -1)$,

$$H\mathcal{P}(\mathbf{x}^*; q) = \begin{pmatrix} 1 & 0 \\ 0 & q \end{pmatrix}, \qquad q > 0.$$

It follows that a stationary point of $\mathcal{P}$ need not be a local minimum or a stationary point of the original constrained optimization problem.

5. A quasi-Newton algorithm

The exact penalty function $\mathcal{P}$ to be minimized involves the second derivatives of the original objective function f and of the constraint functions $h_i, \forall i$. An implementation of the classical Newton method to find the minimum would need derivatives up to the fourth-order, which would often be prohibitively expensive. However, it is important to use an algorithm that takes the special structure of these functions into account. The salient feature of this structure is that the gradients of the smooth exact penalty functions at a stationary point involve the second derivatives of the objective and constraint functions. If these second derivatives are unavailable or are difficult to compute, they can be suitably approximated by using first derivatives.

If second derivatives can be computed relatively easily, then there arises the possibility of using a Newton-like scheme for unconstrained minimization. The difficulty with this is that the Hessian matrix of the smooth penalty functions involves the third derivatives of the problem functions. It turns out, however, that at a Kuhn-Tucker point, the term of the third derivative vanishes, so they can be neglected in a Newton-like algorithm without loss of the superlinear convergence property.

Below, we describe a quasi-Newton method with iterative steps no more difficult to be performed than to evaluate $\mathcal{P}$. This method is of superlinear convergence. In the iterative process, we replace $V(\mathbf{x})$ and $W(\mathbf{x})$ depending on $\mathbf{x}$ in $\mathcal{P}(\mathbf{x}; q)$ by their fixed current estimate so that the derivative of $\mathcal{P}$ simplifies. Sufficiently close to the optimal point this yields a feasible direction to the optimal point. The simplified first derivative is

$$\nabla \mathcal{P}(\mathbf{x} \mid \mathbf{x}_0; q) = \nabla L\big(\mathbf{x}, \boldsymbol{\lambda}(\mathbf{x})\big) - \mathbf{h}(\mathbf{x})^T J\boldsymbol{\lambda}(\mathbf{x}) + \mathbf{h}(\mathbf{x})^T W(\mathbf{x}_0) J\mathbf{h}(\mathbf{x}) + q\mathbf{h}(\mathbf{x})^T V(\mathbf{x}_0) J\mathbf{h}(\mathbf{x}),$$
$$\mathbf{x} \in \mathcal{M},$$

where

$$J\boldsymbol{\lambda}(\mathbf{x}) = J\mathbf{h}(\mathbf{x})^{+^T} HL\big(\mathbf{x}, \boldsymbol{\lambda}(\mathbf{x})\big) + \Big[J\mathbf{h}(\mathbf{x}) J\mathbf{h}(\mathbf{x})^T\Big]^{-1} R(\mathbf{x}), \qquad \mathbf{x} \in \mathcal{M},$$

with

$$R(\mathbf{x}) = \nabla L\big(\mathbf{x}, \boldsymbol{\lambda}(\mathbf{x})\big) \Big[J\big(J\mathbf{h}(\mathbf{x})^T\big)\Big]^T = \sum_{i=1}^{n} \frac{\partial L\big(\mathbf{x}, \boldsymbol{\lambda}(\mathbf{x})\big)}{\partial x_i} \left(\frac{\partial^2 h_j(\mathbf{x})}{\partial x_k \partial x_i}\right), \qquad \mathbf{x} \in \mathcal{M},$$

and $\mathbf{x}_0$ is the fixed current estimate. Note that $R(\mathbf{x}) = 0$ at a local optimal point. Furthermore, the second derivative evaluated at a local optimal point is given as

$$\begin{aligned} H\mathcal{P}(\mathbf{x}; q) &= (I - P(\mathbf{x})) HL(\mathbf{x}, \boldsymbol{\lambda}(\mathbf{x}))(I - P(\mathbf{x})) + qJ\mathbf{h}(\mathbf{x}) V(\mathbf{x}) J\mathbf{h}(\mathbf{x})^T \\ &= \Big(HL\big(\mathbf{x}, \boldsymbol{\lambda}(\mathbf{x})\big)\Big)_{|TM} + qJ\mathbf{h}(\mathbf{x}) V(\mathbf{x}) J\mathbf{h}(\mathbf{x})^T. \end{aligned}$$

For $k = 0, 1, 2, \ldots$, the quasi-Newton algorithm with superlinear convergence is

$$\mathbf{x}^{(k+1)} = \mathbf{x}^{(k)} - step_k \left[H\mathcal{P}(\mathbf{x}^{(k)}; q) \right]^{-1} \nabla \mathcal{P}(\mathbf{x}^{(k)}; q)$$

for an appropriate step size $step_k$ and an appropriate starting point $\mathbf{x}^{(0)}$, which ensures a decrease in $\mathcal{P}(\mathbf{x}; q)$. The starting point depends on the problem at hand and might be an initial guess of the optimal point. An appropriate step size can be found by trying the value 1 first, and if the new estimate does not give a decrease in the value of the objective function, repeat halving the value of the step size until it does.

Example 5.1. *Rosenbrock's parcel problem (1960) is to minimize*

$$f(\mathbf{x}) = x_1 x_2 x_3$$

subject to

$$h(\mathbf{x}) = x_1 + 2x_2 + 2x_3 - 72 = 0, \qquad \mathbf{x} \in R^3.$$

With $V(\mathbf{x})$ and $W(\mathbf{x})$ as in (4.2) and (4.3), the exact penalty function evaluates to

$$\mathcal{P}(\mathbf{x}; q) = x_1 x_2 x_3 - \frac{1}{9} h(\mathbf{x})(2x_1x_2 + x_2x_3 + 2x_1x_3) +$$
$$\frac{2}{81} h(\mathbf{x})^2 (2x_1 + x_2 + x_3) - \frac{q}{18} h(\mathbf{x})^2, \qquad \mathbf{x} \in R^3.$$

Application of the above quasi-Newton algorithm with $q = 1$ and the starting point $(0, 0, 0)$ yielded convergence to the optimal point $\mathbf{x}^* = (24, 12, 12)$ in six iterations.

6. Maximum likelihood estimation subject to constraints

Now, we should like to apply the above exact penalty method to a special class of nonlinear optimization problems.

Let $\mathbf{y}^T = (y_1, \ldots, y_n)$ be a vector of Poisson random variables y_i with expectation vector $\boldsymbol{\mu}^T = (\mu_1, \ldots, \mu_n)$, where $\mu_i > 0$, $i = 1 \ldots, n$. The likelihood function is given as

$$P(\mathbf{y}|\boldsymbol{\mu}) = \prod_{i=1}^{n} \frac{e^{-\mu_i} \mu_i^{y_i}}{y_i!}. \tag{6.1}$$

Note that often a multinomial distribution is assumed rather than an independent Poisson, but both give the same inferential results (see, e.g., Bergsma, 1997). The expectation vector is unknown, and it is desirable to estimate it from the data where the expectation vector may be subject to certain constraints. The set of values of the expectation vector satisfying a set of constraints is called a **statistical model**. A commonly used estimator of the expectation vector is the **maximum likelihood estimate** defined as the value of the expectation vector maximizing the probability of observing the frequency vector subject to the model constraints.

An important class of models is of the form

$$K = \left\{ \boldsymbol{\mu} \in \mathbf{R}^n_+ \mid B \log(A\boldsymbol{\mu}) = \mathbf{0} \right\}$$

for appropriate matrices A and B, where the "log" function is taken coordinatewise. Typically, the number of variables n may become very large, while the number of constraints will remain relatively small. This class of models has received considerable attention recently. If A is the identity matrix, then K is called a **loglinear model**, which is probably the most commonly used model for categorical data analysis (see, e.g., Hagenaars, 1990, and Agresti, 2002). In general, it is awkward or impossible to smoothly parameterize K, and it is preferable to leave the original problem formulation intact, i.e., to use a constrained optimization technique.

To illustrate the matrix notation for the constraints let $\boldsymbol{\mu} = (\mu_{11}, \mu_{12}, \mu_{21}, \mu_{22})$ and consider the restriction

$$\frac{(\mu_{11} + \mu_{12})(\mu_{12} + \mu_{22})}{(\mu_{21} + \mu_{22})(\mu_{11} + \mu_{21})} = 1.$$

This is one way to formulate the model of equality of the marginal distributions in a 2 × 2 table. Taking logarithms, the constraint can be given in matrix notation as

$$\begin{pmatrix} 1 & -1 & -1 & 1 \end{pmatrix} \log \left[\begin{pmatrix} 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 \\ 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 \end{pmatrix} \begin{pmatrix} \mu_{11} \\ \mu_{12} \\ \mu_{21} \\ \mu_{22} \end{pmatrix} \right] = 0, \quad \mathbf{x} \in \mathbf{R}^4,$$

which is of the form $B \log(A\boldsymbol{\mu}) = \mathbf{0}$, $\mathbf{x} \in \mathbf{R}^n$.

Further details on the matrix notation and the application of models of the form $B \log(A \exp(\mathbf{x})) = \mathbf{0}$, $\mathbf{x} \in R^n$, are given in Croon, Bergsma and Hagenaars (2000). Extensions to even more complex types of constraints are given in Bergsma and Croon (2004). Bergsma and Rudas (2004) described a subclass of statistical models which can be written in the form (6.3) and which are differentiable manifolds. Section 7 describes two models like this. The structure of the matrices A and B is implicit in their paper. In general, manifolds of the form (6.3) can have a complicated structure.

We now formulate the maximum likelihood problem of maximizing the probability of observing the frequency vector $P(\mathbf{y}|\boldsymbol{\mu})$ as a function of the expectation vector $\boldsymbol{\mu}$ subject to model constraints, in a mathematically convenient way. The kernel of the log likelihood, i.e., the logarithm of $P(\mathbf{y}|\boldsymbol{\mu}) = P(\mathbf{y}|\exp(\mathbf{x}))$ with the constant term removed, is:

$$\sum_{i=1}^{n}(y_i \log \mu_i - \mu_i), \qquad \boldsymbol{\mu} \in \mathbf{R}^n.$$

To avoid the positivity restriction on the $\boldsymbol{\mu}$'s, we reparameterize them by using

$$x_i = \log \mu_i, \quad i = 1, \dots, n.$$

Then the kernel of the log likelihood becomes

$$f_{\mathbf{y}}(\mathbf{x}) = \sum_{i=1}^{n}(y_i x_i - \exp(x_i)) = \mathbf{y}^T\mathbf{x} - \mathbf{1}^T \exp(\mathbf{x}), \quad \mathbf{x} \in \mathbf{R}^n,$$

where the function "exp" is applied coordinatewise. The model constraints in matrix notation can now be given in the form

$$\mathbf{h}(\mathbf{x}) = B \log(A \exp(\mathbf{x})), \quad \mathbf{x} \in \mathbf{R}^n,$$

The optimization problem now is

$$\begin{array}{c} \max\ \mathbf{y}^T\mathbf{x} - \mathbf{1}^T \exp(\mathbf{x}) \\ \text{subject to}\ \ \mathbf{x} \in M, \end{array} \tag{6.2}$$

where

$$M = \{\mathbf{x} \in \mathbf{R}^n |\, \mathbf{h}(\mathbf{x}) = B \log(A \exp(\mathbf{x})) = \mathbf{0}\}. \tag{6.3}$$

The derivatives of the objective function of problem (6.2) are as follows:

$$\nabla f_{\mathbf{y}}(\mathbf{x}) = \mathbf{y}^T - \exp(\mathbf{x})^T, \quad \mathbf{x} \in M, \tag{6.4}$$

$$H f_{\mathbf{y}}(\mathbf{x}) = -D(\exp(\mathbf{x})), \quad \mathbf{x} \in M, \tag{6.5}$$

where $D(.)$ represents the diagonal matrix with the argument on the main diagonal. Let $\mathbf{b}_i$ be the i-th row of B. The derivatives of $\mathbf{h}$ are

$$J\mathbf{h}(\mathbf{x}) = BD^{-1}(A\exp(\mathbf{x}))AD(\exp(\mathbf{x})), \quad \mathbf{x} \in M, \tag{6.6}$$

$$\begin{aligned} Hh_i(\mathbf{x}) = {} & D(A^T D^{-1}(A\exp(\mathbf{x}))\mathbf{b}_i)D(\exp(\mathbf{x})) - \\ & D(\exp(\mathbf{x}))A^T D^{-2}(A\exp(\mathbf{x}))D(\mathbf{b}_i)AD(\exp(\mathbf{x})), \quad \mathbf{x} \in M. \end{aligned} \tag{6.7}$$

Now it will be shown that the suggested exact penalty method fits well to the maximum likelihood problems subject to marginal loglinear constraints. The following theorem gives a sufficient condition for M to be a C^∞ differentiable manifold.

Theorem 6.1. *Let B be an $(n-k_1)\times(n-k_2)$-dimensional matrix* $(0 \le k_1, k_2 \le n-1)$ *and A an $(n-k_2)\times n$-dimensional positive matrix. Then, the set*

$$M = \big\{\mathbf{x} \in R^n \,|\, \mathbf{h}(\mathbf{x}) = B\log\big(A\exp(\mathbf{x})\big) = 0\big\}$$

is an $r(B)$-dimensional C^∞ differentiable manifold.

Proof. The Jacobian matrices of $\mathbf{h}$ are given by (6.6) where $D^{-1}\big(A\exp(\mathbf{x})\big)$ is an $(n-k_2)\times(n-k_2)$ and $\big(D\exp(\mathbf{x})\big)$ is an $n\times n$ diagonal matrix for every $\mathbf{x}\in M$, respectively.

Now, it will be shown that

$$r\big(J\mathbf{h}(\mathbf{x})\big) = r(B), \qquad \mathbf{x}\in M.$$

The rows of a product matrix are linear combinations of the rows of the second matrix, therefore,

$$r\left(\left(BD^{-1}\big(A\exp(\mathbf{x})\big)A\right)D\big(\exp(\mathbf{x})\big)\right) \le r\left(BD^{-1}\big(A\exp(\mathbf{x})\big)A\right), \qquad \mathbf{x}\in M. \tag{6.9}$$

Since $D\big(\exp(\mathbf{x})\big)$ is invertible,

$$\begin{aligned} r\left(BD^{-1}\big(A\exp(\mathbf{x})\big)A\right) = r\left(BD^{-1}\big(A\exp(\mathbf{x})\big)AD\big(\exp(\mathbf{x})\big)D^{-1}\big(\exp(\mathbf{x})\big)\right) \le \\ r\left(BD^{-1}\big(A\exp(\mathbf{x})\big)AD\big(\exp(\mathbf{x})\big)\right), \qquad \mathbf{x}\in M. \end{aligned} \tag{6.10}$$

From inequalities (6.9) and (6.10),

$$r\left(BD^{-1}\big(A\exp(\mathbf{x})\big)A\right) = r\left(BD^{-1}\big(A\exp(\mathbf{x})\big)AD\big(\exp(\mathbf{x})\big)\right), \qquad \mathbf{x}\in M. \tag{6.11}$$

Consider the matrices $BD^{-1}\big(A\exp(\mathbf{x})\big)A$, $\mathbf{x}\in M$, and use the same reasoning for the matrix A and a right inverse A^- $(AA^- = I)$, then, for the matrices $D^{-1}\big(A\exp(\mathbf{x})\big)$, $\mathbf{x}\in M$, respectively, thus, we obtain that

$$r\big(J\mathbf{h}(\mathbf{x})\big) = r(B), \qquad \mathbf{x}\in M,$$

from which the statement follows. ■

7. Numerical experience

Consider the data in Table 1 obtained from a national sample of the Dutch electorate interviewed in February 1977 and March 1977. Let A denote vote intention in February, B vote intention in March, C preference for the prime minister in February and D preference for the prime minister in March. By y_{ijkl} we denote the number of people who responded in category i of variable A, in category j of B, in category k of C and in category l of D. These observed frequencies are given in Table 1. Thus, for example, $y_{1111} = 293$, $y_{1112} = 1$, $y_{1113} = 6$, and so on. The observed marginal distributions pertaining to the turnover in vote intention and preference for the prime minister are given in Table 2 where the meanings of the variables are given as well. Note that the marginal observed frequencies are given as

$$\sum_{k=1}^{3}\sum_{l=1}^{3} y_{ijkl} \quad \textit{and} \quad \sum_{i=1}^{3}\sum_{j=1}^{3} y_{ijkl}.$$

The data have previously been studied by Hagenaars (1990) and Bergsma (1997). A reasonable assumption is that the y_{ijkl} have independent Poisson distributions with unknown means $Ey_{ijkl} = \mu_{ijkl} > 0$.

		C	1	1	1	2	2	2	3	3	3	
		D	1	2	3	1	2	3	1	2	3	*Total*
A	B											
1	1		293	1	6	4	2	0	22	1	21	350
1	2		2	1	2	1	1	0	0	1	1	9
1	3		8	1	7	0	1	0	0	0	9	26
2	1		8	0	1	1	0	0	2	2	3	17
2	2		13	6	7	9	84	23	8	24	68	242
2	3		2	0	3	1	3	2	3	2	9	25
3	1		31	0	0	1	0	1	9	2	7	51
3	2		5	4	0	1	6	1	1	9	16	43
3	3		48	3	23	1	14	15	21	12	200	337
Total			410	16	49	19	111	42	66	53	334	1100

Table 1: Vote Intention and Preference Prime Minister in The Netherlands
(source: Hagenaars, 1990)

We note that the symbols A, B, C, and D have the same meaning in Table 1 and Table 2.

(a) Vote Intention: February 1977 - March 1977				
	B. March			
	Left	Chr. Dem.	Other	*Total*
A. February				
1. Left Wing	350	9	26	385
2. Christ. Dem.	17	242	25	284
3. Other	51	43	337	431
Total	418	294	388	1100
(b) Preference for Prime Minister: February 1977 - March 1977				
	D. March			
	Left	Chr. Dem.	Other	*Total*
C. February				
1. Left Wing (Den Uyl)	410	16	49	475
2. Christ. Dem. (Van Agt)	19	111	42	172
3. Other	66	53	334	453
Total	495	180	425	1100

Table 2: Turnover in Political Preference in The Netherlands: February 1977 - March 1977 (source: Hagenaars, 1990)

In Political Science, various hypotheses concerning the marginal population distributions are of interest. A first hypothesis is that they are equal, i.e.,

$$\sum_{k,l=1}^{3} \mu_{ijkl} = \sum_{k,l=1}^{3} \mu_{klij}, \qquad i,j = 1,2,3, \tag{7.1}$$

where the number of independent constraints is 8 and the number of variables $n = 81$. This is called the **marginal homogeneity model**. Note, however, that a conspicuous difference between the two observed (two-way) marginal tables is that their one-way marginal distributions are different: both in February and March, the number of people who preferred the socialist candidate Den Uyl is larger than the number of people who preferred the left-wing party, whereas, the opposite is true for the Christian Democratic candidate, Van Agt and the Christian Democratic Party. This reflects the "empirical regularity" that, in general, the Prime Minister in office, Den Uyl at that time, was more popular than his party, even among those who do not support the Prime Minister's politics (Hagenaars, 1990, page 172). A weaker hypothesis asserts that only the association, as measured by odds ratios, is equal in the two tables, and is written as follows:

$$\frac{\left(\sum_{k,l} \mu_{ijkl}\right)\left(\sum_{k,l} \mu_{i+1,j+1,kl}\right)}{\left(\sum_{k,l} \mu_{i+1,jkl}\right)\left(\sum_{k,l} \mu_{i,j+1,kl}\right)} = \frac{\left(\sum_{k,l} \mu_{klij}\right)\left(\sum_{k,l} \mu_{kl,i+1,j+1}\right)}{\left(\sum_{k,l} \mu_{kl,i+1,j}\right)\left(\sum_{k,l} \mu_{kl,i,j+1}\right)}, \qquad i,j = 1,2,3. \tag{7.2}$$

An interpretation of this model is that the two-way marginal distributions are equal, except for differences in their one-way marginals. It follows from Theorem 6.1 as well as Bergsma and Rudas (2002) that M is a differentiable manifold, because $J\mathbf{h}$ is of full rank on M, and from Bergsma and Rudas (2002) that M are connected for both models, (7.1) and (7.2).

Using V as given in (4.3) and $q = 1$, the quasi-Newton algorithm described in the previous section was used for finding $\mathbf{x}^*$ for problem (6.2) with model constraints (7.1) and (7.2), respectively. A potential problem with finding $\mathbf{x}^*$ is that certain x_i may go to minus infinity, corresponding to the μ_i going to zero. It is easy to verify that this is only possible if the corresponding $y_i = 0$. This problem was solved by adding a small constant, namely 10^{-50}, to those $y_i = 0$. This constant negligably affects the value of $f_{\mathbf{y}}$ and the modified vector $\mathbf{y}$ was also used as a starting value. As a convergence criterion, the algorithm was stopped at iteration $k > 1$ if

$$\max_i |x_i^{(k)} - x_i^{(k-1)}| < 10^{-10}.$$

At each iteration, we started with a step-size equal to 1, and if the resulting new estimate did not lead to a higher value of $\mathcal{P}(\mathbf{x}; q)$, the step-size was repeatedly halved until a higher value of $\mathcal{P}(\mathbf{x}; q)$ was obtained. The algorithm was implemented in *Mathematica 5.0.* The search direction was calculated by using function *LinearSolve* and, to avoid numerical difficulties the option *Method→Multifrontal* was used.

The estimated marginal frequencies are given in Tables 4 and 5. The maxima of $f_{\mathbf{y}}$ are 3380.66 and 3431.37, respectively. A summary of the maximization procedure is given in Table 3.

	Model (7.1)	Model (7.2)
# variables	81	81
# constraints	8	4
# iterations	75	6
time taken	8.1 seconds	0.7 seconds

Table 3: Summary of the maximum likelihood fitting by using a Pentium IV, 2.4 MHz

For model (7.2), the algorithm has converged very fast and automatically. However, for model (7.1), we were unable to obtain automatic convergence, due to an ill-conditioning of the problem close to the boundary of the parameter space and far from the optimal value. After some experimentation, we let the algorithm run with a step-size of 0.01 for 50 iterations, a step-size of 0.1 for 20 iterations, and finally, we were able to obtain convergence by letting the algorithm run with step-size 1 for another five iterations.

If $\mathbf{x}$ is in M, the likelihood ratio test statistic, defined as

$$G^2(\mathbf{x}^*, \mathbf{y}) = -2 \log \frac{P(\mathbf{y}|\mathbf{y})}{P(\mathbf{y}|\mathbf{x}^*)} = -2(f_{\mathbf{y}}(\mathbf{y}) - f_{\mathbf{y}}(\mathbf{x}^*))$$

has a large sample chi-square distribution with degrees of freedom equal to the dimension of M. We find $G^2(\mathbf{x}^*, \mathbf{y}) = 132.29$ with 8 degrees of freedom for model (7.1) and

$G^2(\mathbf{x}^*, \mathbf{y}) = 30.86$ with 4 degrees of freedom for model (7.2). The p-value for assessing the fit of a model is then defined as the probability that a random variable with a chi-square distribution of the appropriate number of degrees of freedom exceeds $G^2(\mathbf{x}^*, \mathbf{y})$. We obtain p-values of less than 10^{-5} for both models, which gives a very strong evidence that $\mathbf{x} \notin M$ in both cases.

(a) Vote Intention: February 1977 - March 1977				
	B. March			
	Left	Chr. Dem.	Other	*Total*
A. February				
1. Left Wing	387.43	12.48	39.25	439.16
2. Christ. Dem.	16.17	179.66	31.35	227.18
3. Other	55.95	47.44	330.27	433.66
Total	459.55	239.56	400.87	1100
(b) Preference for Prime Minister: February 1977 - March 1977				
	D. March			
	Left	Chr. Dem.	Other	*Total*
C. February				
1. Left Wing	387.43	12.48	39.25	439.16
2. Christ. Dem.	16.17	179.66	31.25	227.18
3. Other	55.95	47.44	330.27	433.66
Total	459.55	239.56	400.87	1100

Table 4: Maximum likelihood estimates for the marginal homogeneity model (7.1) based on the data of Table 1

(a) Vote Intention: February 1977 - March 1977				
	B. March			
	Left	Chr. Dem.	Other	*Total*
A. February				
1. Left Wing	343.08	12.77	33.09	388.94
2. Christ. Dem.	20.81	224.02	36.49	281.32
3. Other	57.94	53.55	318.25	429.74
Total	421.83	290.35	387.82	1100
(b) Preference for Prime Minister: February 1977 - March 1977				
	D. March			
	Left	Chr. Dem.	Other	*Total*
C. February				
1. Left Wing	421.49	12.14	43.86	477.49
2. Christ. Dem.	15.09	125.58	28.54	169.21
3. Other	59.32	42.41	251.57	453.30
Total	495.90	180.13	423.97	1100

Table 5: Maximum likelihood estimates for the model (7.2) based on the data of Table 1

Other algorithms have been described in the literature for the maximum likelihood fitting of models of the form (6.3). For more general maximum likelihood problems where parameters are subject to equality constraints, Aitchison and Silvey (1958) proposed a modified Newton method, based on the first and second derivatives of the Lagrangian (2.2), so that the search is in $(2n - k)$-dimensional space. This modification is based on replacing the second derivative matrix by its expected value, presumably to obtain a stabler algorithm, but at the loss of superlinear convergence. This method was applied by Lang (1996) to models of the form (6.3). Bergsma modified the algorithm to obtain a dimension reduction of the search to n-dimensional space. Numerical experience by the first author has indicated that the latter's algorithm appears to improve on the Aitchison and Silvey algorithm, and it was possible to obtain automatic convergence for a wide range of problems. However, why that algorithm works well is not understood. A drawback of that algorithm is that it has only linear convergence, and convergence could be quite slow.

Certain models of the form (6.3) are parameterizable (see Bergsma and Rudas, 2002), and hence the maximum likelihood estimate can also be found by unconstrained optimization. However, since the parameterization is implicit, an approach like this would require "iteration within iteration," which can be prohibitively expensive.

8. Large sample behavior of parameter estimates and interpretation of asymptotic covariance matrix

In a setting more general than in the previous section, a smoothly constrained maximum likelihood estimation problem can be formulated as follows. Let S be the sample space, $N \geq 1$ the sample size, $\mathbf{y} \in S^N$ a vector of observations, and let $\mathbf{x} \in \mathbf{R}^n$ be a vector of unknown parameters. Denote the log-likelihood as a function of the unknown parameters $\mathbf{x}$ by $f_{\mathbf{y}}(\mathbf{x})$ and suppose $\mathbf{x} \in M = \{\mathbf{x} \in \mathbf{R}^n \mid \mathbf{h}(\mathbf{x}) = \mathbf{0}\}$ for some constraint vector $\mathbf{h}$.

Below, without further statement, we assume that all necessary regularity conditions, given by Aitchison and Silvey (1958), are satisfied. They proved that $\mathbf{x}^* \to \mathbf{x}$ in probability as $N \to \infty$. Then, since $\nabla f(\mathbf{x}^*) \to \mathbf{0}$, we also obtain $\boldsymbol{\lambda}(\mathbf{x}^*) \to \mathbf{0}$ in probability. Similarly, it follows that $L(\mathbf{x}^*, \boldsymbol{\lambda}(\mathbf{x}^*)) - f_{\mathbf{y}}(\mathbf{x}^*) \to 0$ and $\mathcal{P}(\mathbf{x}^*; q) - f_{\mathbf{y}}(\mathbf{x}^*) \to 0$. This implies that for large N, both the Lagrangian function $L(\mathbf{x}, \boldsymbol{\lambda}(\mathbf{x}))$ and the penalty function $\mathcal{P}(\mathbf{x}; q)$ are similar to the objective function $f_{\mathbf{y}}(\mathbf{x})$ close to the optimal point $\mathbf{x}^*$.

Now, let $\tilde{\mathbf{x}}$ be the unconstrained maximum of $f_{\mathbf{y}}$. Then, it can be shown that $Hf_{\mathbf{y}}(\tilde{\mathbf{x}})$ is positive definite with probability going to one. Since additionally $|\tilde{\mathbf{x}} - \mathbf{x}^*| \to 0$ in probability, we obtain that $Hf_{\mathbf{y}}(\mathbf{x}^*)$ is positive definite with probability going to one as $N \to \infty$. The following theorem follows:

Theorem 8.1. *Suppose $\mathbf{x} \in M$. Then as $N \to \infty$, the probability that the maximum likelihood estimate $\mathbf{x}^*$ is a local maximum of $L(\mathbf{x}, \boldsymbol{\lambda}(\mathbf{x}^*))$ goes to one.*

By the central limit theorem, both $\sqrt{N}(\mathbf{x} - \mathbf{x}^*)$ and $\sqrt{N}(\mathbf{x} - \tilde{\mathbf{x}})$ have an asymptotic normal distribution with mean zero and covariance matrix, say, Σ_0 and Σ_1, respectively. Now, it is well know that $\Sigma_1 = E(Hf_{\mathbf{y}}(\mathbf{x}))^{-1}$. The asymptotic covariance matrix of Σ_0 was given by Aitchison and Silvey. Interestingly, Σ_0 can be expressed as Σ_1 restricted to the tangent space of M. We next introduce some notations in order to do this.

Let $R_{|TM,Q}$ denote the restriction of the matrix R to the tangent space of M by using the metric induced by the nonsingular matrix Q, i.e., with inner product $\langle \mathbf{x}, \mathbf{y} \rangle = \mathbf{x}^T Q \mathbf{y}$. Specifically,

$$R_{|TM,Q} = (I - P_Q) R (I - P_Q),$$

where

$$P_Q = Q^{-1} J\mathbf{h}(\mathbf{x})^T (J\mathbf{h}(\mathbf{x}) Q^{-1} J\mathbf{h}(\mathbf{x})^T)^{-1} J\mathbf{h}(\mathbf{x}), \qquad \mathbf{x} \in M,$$

is the projection matrix onto the space spanned by the columns of $Q^{-1}\ J\mathbf{h}^T$ with respect to the metric induced by Q. With this notation, it can be shown that

$$\Sigma_0 = (\Sigma_1)_{TM|\Sigma_1^{-1}}$$

That is, Σ_0 equals Σ_1 restricted to the tangent space of M using the metric induced by Σ_1^{-1}.

9. Conclusion

A new exact penalty function has been introduced which converts a smooth constrained nonlinear optimization problem to an unconstrained optimization problem, and a quasi-Newton method to find the optimal point with superlinear convergence was proposed. An advantage with respect to "classical" exact penalty function approaches is that it is easy to find an appropriate penalty parameter, and the ill-conditioning of the problem is more easily avoided. An application of the method to Rosenbrock's parcel problem yielded convergence in 6 iterations. A potential drawback of the proposed approach is that the second derivatives of both the object function and the constraint functions are needed. However, numerical values for the penalty function and its derivatives can be efficiently calculated by using automatic differentiation techniques.

The method was tested on maximum likelihood estimation subject to constraints. The problem concerned multivariate categorical data with loglinear constraints on the marginal distributions. Since for constraints like this, it is (numerically) awkward to parameterize the constraint surface, it is best to use constrained optimization techniques. For two problems with 81 unknown parameters, the quasi-Newton algorithm converged quickly, although for one of the problems some experimentation with the step size was necessary to be done.

Finally, for general constrained maximum likelihood problems, the large sample behavior of parameter estimates was considered. It was shown that the probability that the maximum likelihood estimate is an unconstrained maximum of the Lagrangian function goes to one if the statistical model is true and the sample size goes to infinity. Furthermore, an interpretation of the asymptotic covariance matrix of parameter estimates was given in terms of the tangent space of the constraint surface.

References

[1] Aitchison, J. and Silvey, S.D., Maximum-likelihood estimation of parameters subject to restraints, Annals of Mathematical Statistics 29 (1958) 813-828.

[2] Agresti, A., Categorical data analysis, *Wiley*, New York, 2002. (2nd edition)

[3] Bergsma, W.P., Marginal models for categorical data, *Tilburg University Press*, Tilburg, 1997.

[4] Bergsma, W.P. and Croon, M.A., Analyzing categorical date by marginal models, in: New developments in categorical data analysis: the social and behavioral sciences, L.A. van der Ark et al. (eds.), *Mahwah NJ*, Erlbaum (2004) 83-101.

[5] Bergsma, W.P. and Rudas, T., Marginal models for categorical data, *Annals of Statistics* 30 (2002) 140-159.

[6] Bertsekas, D.P., Constrained optimization and Lagrange multiplier methods, *Academic Press*, New York, London, 1982.

[7] Christianson, B., Automatic Hessians by reverse accumulation, *IMA Journal of Numerical Analysis* 12 (1992) 135-150.

[8] Christianson, B., Geometric approach to Fletcher's ideal penalty function, *Journal of Optimization Theory and Applications* 84 (1995) 433-441.

[9] Croon, M.A., Bergsma, W.P. and Hagenaars, J.A., Analyzing change in categorical variables by generalized log-linear models, *Sociological Methods and Research* 29(2) (2000) 195-229.

[10] Courant, R., Variational methods for the solution of problems of equilibrium and vibrations, *Bulletin of the American Mathematical Society* 49 (1943) 1-23.

[11] Di Pillo, G., Exact penalty methods, in: Algorithms for continuous optimization: the state-of-the-art, E. Spedicato (ed.), *Kluwer Academic Publishers*, Boston (1994) 1-45.

[12] Di Pillo, G. and Grippo, L., Exact penalty functions in constrained optimization, *SIAM Journal on Control and Optimization* 27 (1989) 1333-1360.

[13] Fletcher, R., A class of methods for nonlinear programming with termination and convergence properties, in: Integer and nonlinear programming, J. Abadie (ed.), *North-Holland Publishing Company*, Amsterdam, London (1970) 157-175.

[14] Hagenaars, J.A., Categorical longitudinal data, *Sage*, Newbury Park, 1990.

[15] Lagrange, J.L., Essai sur une nouvelle méthode pour determiner les maxima et minima des formules intégrales indefinies, in: Miscellanea Taurinensia II. (1762) 173-195.

[16] Lagrange, J.L., Méchanique analytique I-II., Paris, 1788.

[17] Lang, J.B., Maximum likelihood methods for a generalized class of loglinear models, *Annals of Statistics* 24 (1996) 726-752.

[18] Luenberger, D.G., The gradient projection method along geodesics, *Management Science* 18 (1972) 620-631.

[19] Rapcsák, T., Geodesic convexity in nonlinear programming, *Journal of Optimization Theory and Applications* 69 (1991) 169-183.

[20] Rapcsák, T., Smooth nonlinear optimization in R^n, *Kluwer Academic Publishers* (1997). (374 p)

[21] Rapcsák, T., Global Lagrange multiplier rule and smooth exact penalty functions for equality constraints, in: Nonlinear optimization and related topics, eds.: G. Di Pillo and F. Giannessi, *Kluwer Academic Publishers* (2000) 351-368.

[22] Rosenbrock, H. H., An automatic method for finding the greatest value of a function, *Computer Journal* 3 (1960) 175-184.