A SQP-Method for Linearly Constrained Maximum Likelihood Problems

Christian Kredler

Institut für Angewandte Mathematik und Statistik
Technische Universität München

Dedicated to Prof. Klaus Ritter on his 60th birthday

Abstract. Newton-like methods are standard algorithms for unrestricted parameter estimation in a wide class of nonlinear regression models. The search directions of the here presented algorithm are solutions from a sequence of quadratic (sub-)problems (SQP) with linear constraints. The practically important generalized linear models with natural link functions, e.g. log-linear or logistic regression, lead to strictly convex optimization problems for which this easy to implement extension of Newton's method converges globally with quadratic rate. The numerical results are demonstrated at some ship damage data.

Key Words: Generalized linear models (GLM), natural link function, restricted maximum likelihood (ML-) estimation, SQP-method, global and quadratic convergence, uniform convexity.

1 Restricted ML-estimation in Generalized Linear Models

In a wide and important area of statistical data analysis we are faced with strictly convex optimization problems for which Newton-like optimization methods are more appropriate than the widely used standard algorithms, e.g. BFGS or the gradient based SQP-method of [Schittkowski (1981)]. The here proposed approach is a robust optimization tool, simply substitutable into statistical programs without using external libraries, provided a solver for quadratic programs is available; e.g. [Best & Ritter (1988)]. The convergence theorems illuminating the global and local behaviour of the suggested algorithm are proved using a technique similar to that presented in [Ritter (1982)].
The convergence properties of the investigated algorithm depend on strong convexity assumptions. Especially, these conditions are satisfied in statistical

regression problems where the canonical parameters $\boldsymbol{\theta} \in \mathbb{R}^q$ of some exponential families are involved. Consider a random variable $\boldsymbol{Y} \in \mathbb{R}^q$ having a density

$$h(\boldsymbol{y}, \boldsymbol{\theta}) = c(\boldsymbol{y})\, exp\,(\boldsymbol{y}^T \boldsymbol{\theta} - d(\boldsymbol{\theta})) \tag{1}$$

with respect to the Lebesgue or counting measure. The scalar valued c does not depend on $\boldsymbol{\theta}$. Note that normal, multinomial, Poisson and Gamma random variables belong to our exponential class. For the treatment of nuisance parameters like σ^2 in the normal case see e.g. [McCullagh & Nelder (1983)]. The partial derivatives $d_j, d_{jk}, ...$ of the analytic function d generate the cumulants of $\boldsymbol{Y}$ with

$$\begin{aligned} \boldsymbol{E}(\boldsymbol{Y}) &= (d_j) \\ \boldsymbol{cov}(\boldsymbol{Y}) = \boldsymbol{cov}(\boldsymbol{Y}_j, \boldsymbol{Y}_k) &= (d_{jk}) \quad etc. \end{aligned} \tag{2}$$

The third order cumulants of Y can be exploited to obtain explicit bounds for variable selection techniques based on certain quadratic approximations, cf. [Kredler (1986)].
The positive definiteness of (d_{jk}) ensures the convexity properties needed for the objective functions to be analyzed below.

To clarify presentation we now restrict ourselves to the univariate case $q = 1$, although all statements apply analogously to multivariate models from which the multinomial seems to be the most important, e.g. for contingency tables and logistic discriminant analysis (see [Fahrmeir & Kredler (1984)]). Generalized linear models (GLM), i.e. regression models for exponential type random variables, have first been proposed by [Nelder & Wedderburn (1972)]. Given independent observations and covariates

$$y_t \in \mathbb{R} \quad \text{and} \quad \boldsymbol{z}_t \in \mathbb{R}^p\,,\ t = 1, ..., T\,, \tag{3}$$

where the canonical or natural parameters θ_t vary in a linear subspace

$$\theta_t = \boldsymbol{z}_t^T \boldsymbol{\beta} \tag{4}$$

with an unknown parameter vector $\boldsymbol{\beta} \in \mathbb{R}^p$ to be estimated. Neglecting superfluous constants the log-likelihood function for all T observations is

$$l(\boldsymbol{\beta}) = \sum_{t=1}^{T} \left\{ y_t \cdot \boldsymbol{z}_t^T \boldsymbol{\beta} - d(\boldsymbol{z}_t^T \boldsymbol{\beta}) \right\}\,. \tag{5}$$

According to (2)

$$-\,\boldsymbol{\nabla}^2 \boldsymbol{l}(\boldsymbol{\beta}) = \sum_{t=1}^{T} d''(\boldsymbol{z}_t^T \boldsymbol{\beta})\, \boldsymbol{z}_t \boldsymbol{z}_t^T \qquad \text{is positive definite} \tag{6}$$

if the matrix $(\boldsymbol{z}_t)_{1\le t\le T}$ has full rank. Hence $-l$ is strictly convex for GLM's in the natural parameters θ_t. The existence of the ML-estimate $\hat{\boldsymbol{\beta}}$ with $l(\hat{\boldsymbol{\beta}}) \ge l(\boldsymbol{\beta})$ is not trivial, but can be ensured a priori by certain properties of the data y_t, $\boldsymbol{z}_t$, cf. [Wedderburn (1976)] and [Kaufmann (1988)]. [Albert & Anderson (1984)] guarantee especially for the binomial and multinomial case the existence of the ML-estimate $\hat{\boldsymbol{\beta}}$ under easy to verify linear separation properties of the data. For instance, $\hat{\boldsymbol{\beta}}$ cannot exist if there is a hyper-plane $H : \mathbf{0} = \boldsymbol{z}^T\tilde{\boldsymbol{\beta}}$ such that

$$\begin{aligned} \boldsymbol{z}_t^T\tilde{\boldsymbol{\beta}} &< 0 \quad \text{for all} \quad y_t = 0 \\ \boldsymbol{z}_t^T\tilde{\boldsymbol{\beta}} &> 0 \quad \text{for all} \quad y_t = 1\,. \end{aligned} \tag{7}$$

In section 3 a log-linear GLM for ship damage data with linear constraints for the parameter vector $\boldsymbol{\beta}$ is analyzed. This leads to

$$\max_{\boldsymbol{\beta}} \{\, l(\boldsymbol{\beta}) \,|\, \boldsymbol{\beta} \in \boldsymbol{R} \subseteq \mathbb{R}^p \,\} \tag{8}$$

where

$$\boldsymbol{R} \;=\; \{\, \boldsymbol{\beta} \,|\, \boldsymbol{A}\boldsymbol{\beta} \le \boldsymbol{b} \,\}\,, \quad \boldsymbol{A} \in \mathbb{R}^{m,p}\,,\; \boldsymbol{b} \in \mathbb{R}^m\,. \tag{9}$$

In the next section we discuss the convergence properties of an SQP-algorithm to compute a linearly restricted ML-estimate of a GLM with natural link function and existing $\hat{\boldsymbol{\beta}}$.

2 SQP-Algorithm, Convergence Properties

To go along with the usual notation in the optimization literature let in this section

$$\boldsymbol{x} := \boldsymbol{\beta} \in \mathbb{R}^p.$$

The problem to solve is then

$$\min_{\boldsymbol{x}} \{\, F(\boldsymbol{x}) \mid \boldsymbol{A}\,\boldsymbol{x} \le \boldsymbol{b} \,\} \tag{10}$$

where $F(\boldsymbol{x}) \equiv -l(\boldsymbol{\beta})$ is twice continuously differentiable and has additionally the convex properties stated below. A linear equality constraint $\boldsymbol{a}^T\boldsymbol{x} = b$ can be replaced with two linear inequalities

$$\boldsymbol{a}^T\boldsymbol{x} \le b \qquad \text{and} \qquad -\,\boldsymbol{a}^T\boldsymbol{x} \le -b\,. \tag{11}$$

Hence, for the sake of a brief notation, the restrictions in (10) cover this case, too. However, the numerical treatment of linear equalities is much easier than that of linear inequalities, and any good numerical algorithm handles

linear equality constraints in a separate way; see also formulation of corollary 2.12. Throughout we denote gradient and Hessian of F by

$$\boldsymbol{g}(\boldsymbol{x}) := \boldsymbol{\nabla} \boldsymbol{F}(\boldsymbol{x}) \quad \text{and} \quad \boldsymbol{G}(\boldsymbol{x}) := \boldsymbol{\nabla}^2 \boldsymbol{F}(\boldsymbol{x}) \,.$$

Def. 2.1 (Uniform Convexity)

$F \in C^2(\mathbb{R}^p)$ is said to be **uniformly convex** *on a convex set $\boldsymbol{D} \subseteq \mathbb{R}^p$, if and only if there are $0 < \mu \leq \eta$ such that*

$$\mu \, \|\boldsymbol{x}\|^2 \leq \boldsymbol{x}^T \, \boldsymbol{G}(\boldsymbol{u}) \, \boldsymbol{x} \;\; \leq \eta \, \|\boldsymbol{x}\|^2 \tag{12}$$

for all $\boldsymbol{x} \in \mathbb{R}^p$ and all $\boldsymbol{u} \in \boldsymbol{D}$.

Uniformly convex functions are strictly convex and attain their unique minimum. $F(x) = x^4$ shows that the reverse is not true.
However, for certain applications like natural link function GLM's the following theorem gives a concrete characterization depending on the existence of a minimizer $\bar{\boldsymbol{x}}$. As mentioned in the previous section this can be done, for instance in practically relevant GLM's, by an a priori inspection of the data involved in the nonlinear function F.

Theorem 2.2

If $F \in C^2(\mathbb{R}^p)$ has a positive definite Hessian for all $\boldsymbol{x} \in \mathbb{R}^p$ and attains its minimum at some $\bar{\boldsymbol{x}}$, then F is **uniformly convex** *on the level sets $\boldsymbol{N}_\circ := \{\, \boldsymbol{x} \in \mathbb{R}^p \mid F(\boldsymbol{x}) \leq F(\boldsymbol{x}_0) \,\}$ for all $\boldsymbol{x}_0 \in \mathbb{R}^p$.*

Proof:
We first show that the level sets are bounded. The compactness follows, because they are closed. Pick $\omega > 0$, let $\alpha := F(\bar{\boldsymbol{x}}) + \omega$, and consider $\boldsymbol{N}_\alpha := \{\, \boldsymbol{x} \mid F(\boldsymbol{x}) \leq \alpha \,\}$, which is convex since F is strictly convex. If it was not bounded there would exist a $\boldsymbol{s} \neq \boldsymbol{0} \in \mathbb{R}^p$ and $\varphi(\sigma) := F(\bar{\boldsymbol{x}} - \sigma \, \boldsymbol{s}) \leq \alpha$ for all $\sigma \geq 0$. φ is strictly convex, and hence $F(\bar{\boldsymbol{x}}) = \varphi(0) < \varphi(\sigma)$, for all $\sigma \neq 0$. With $\delta := \varphi(1) - \varphi(0) > 0$ we obtain for all $\sigma > 1$

$$\varphi(0) \, + \, \delta \, = \, \varphi(1) \, = \, \varphi\left(\frac{1}{\sigma} \sigma \, + \, (1 - \frac{1}{\sigma}) \, 0 \right) \, < \, \frac{1}{\sigma} \varphi(\sigma) \, + \, (1 \, - \, \frac{1}{\sigma}) \, \varphi(0) \,,$$

which implies $\varphi(0) \, + \, \delta \, < \, \frac{1}{\sigma} \Big(\varphi(\sigma) - \varphi(0) \Big) \, + \, \varphi(0)$, and further

$$0 \, < \, \delta \, < \, \frac{1}{\sigma} \, \Big(\varphi(\sigma) - \varphi(0) \Big) \, \leq \, \frac{1}{\sigma} \, \Big(\alpha \, - \, \varphi(0) \Big) = \frac{1}{\sigma} \, \Big(\omega \, + \, \varphi(0) \, - \, \varphi(0) \Big) = \frac{\omega}{\sigma}$$

for all $\sigma > 1$, which contradicts $\delta = \varphi(1) - \varphi(0) > 0$. Now, the extremal eigenvalues attain their minimum and maximum on the compact set $\boldsymbol{N}_\alpha$. □

The theorem covers most important generalized linear models including binomial, multinomial and Poisson variables. For the $\Gamma-$distribution and related cases the theorem can be extended, with some technical caution, to $F \in C^2(\boldsymbol{D})$ where $\boldsymbol{D}$ is an appropriate, convex subset of $\mathbb{R}^p$. In the proofs of the following statements we need only the uniform convexity of F on $\boldsymbol{N}_o \cap \boldsymbol{R}$, the intersection of the level set and the feasible region $\boldsymbol{R} = \{\, \boldsymbol{x} \,|\, \boldsymbol{A}\boldsymbol{x} \leq \boldsymbol{b}\}$. Trivially, this property can always be obtained by restricting F with positive definite Hessian to a box $\boldsymbol{v} \leq \boldsymbol{x} \leq \boldsymbol{w}$.

2.1 A General Descent Method

We consider an iterative algorithm which, starting from $\boldsymbol{x}_0 \in \mathbb{R}^p$, computes for $j = 0, 1, 2, ...$ a search direction $\boldsymbol{s}_j$ and a stepsize σ_j defining the new point

$$\boldsymbol{x}_{j+1} := \boldsymbol{x}_j - \sigma_j \, \boldsymbol{s}_j \,,$$

where $\boldsymbol{g}_j^T \boldsymbol{s}_j > 0$ for $\boldsymbol{g}_j := \boldsymbol{g}(\boldsymbol{x}_j)$.

Def. 2.3 (Armijo-Goldstein Rule)

Choose $0 < \delta < \frac{1}{2}$. *In* $\boldsymbol{x}_j$ *find the smallest integer* $\nu_j \geq 0$ *such that*

$$F(\boldsymbol{x}_j) - F(\boldsymbol{x}_j - (\tfrac{1}{2})^{\nu_j} \boldsymbol{s}_j) \geq \delta \cdot (\tfrac{1}{2})^{\nu_j} \cdot \boldsymbol{g}_j^T \boldsymbol{s}_j \,. \tag{13}$$

Finally, let $\sigma_j = (\frac{1}{2})^{\nu_j}$.

Def. 2.4 (Quadratic Approximation)

With $\boldsymbol{g}_j := \boldsymbol{g}(\boldsymbol{x}_j)$, $\boldsymbol{G}_j := \boldsymbol{G}(\boldsymbol{x}_j)$ *the quadratic Taylor approximation is* $F(\boldsymbol{x}_j - \boldsymbol{s}) \approx F(\boldsymbol{x}_j) + Q_j(\boldsymbol{s})$ *where*

$$Q_j(\boldsymbol{s}) := -\boldsymbol{g}_j^T \boldsymbol{s} + \tfrac{1}{2} \, \boldsymbol{s}^T \boldsymbol{G}_j \, \boldsymbol{s} \,. \tag{14}$$

Let throughout $\boldsymbol{\lambda} := (\lambda_1, ..., \lambda_m)^T \leq \boldsymbol{0}$ and

$$\boldsymbol{A} := \begin{pmatrix} \boldsymbol{a}_1^T \\ \cdots \\ \boldsymbol{a}_m^T \end{pmatrix} , \quad \boldsymbol{b} := \begin{pmatrix} b_1 \\ \cdots \\ b_m \end{pmatrix} .$$

Def. 2.5 (Kuhn-Tucker Conditions, Lagrange Multipliers)

The tupel $(\bar{\boldsymbol{x}}, \bar{\boldsymbol{\lambda}}) \in I\!R^{p+m}$, $\bar{\boldsymbol{\lambda}} \leq \mathbf{0}$, *is called* **Kuhn-Tucker pair** *of problem (10) if the following conditions hold*

$$\begin{aligned} \boldsymbol{g}(\bar{\boldsymbol{x}}) - \boldsymbol{A}^T \bar{\boldsymbol{\lambda}} &= \mathbf{0}\,, \quad \bar{\boldsymbol{\lambda}} \leq \mathbf{0}\,, && (15)\\ \boldsymbol{A}\,\bar{\boldsymbol{x}} &\leq \boldsymbol{b}\,, && (16)\\ \bar{\lambda}_k\,(\boldsymbol{a}_k^T\,\bar{\boldsymbol{x}} - b_k) &= 0\,, \quad 1 \leq k \leq m\,. && (17) \end{aligned}$$

$\bar{\boldsymbol{x}}$ *satisfying the above conditions is called* **stationary point** *and* $\bar{\boldsymbol{\lambda}}$ *is the vector of optimal* **Lagrange multipliers**.

In our case F is convex and differentiable, hence each stationary point $\bar{\boldsymbol{x}}$ is a global optimum of (10).

Algorithm SQP

Step 0: *Initialization*

Choose a feasible $\boldsymbol{x}_0$, i.e. $\boldsymbol{A}\,\boldsymbol{x}_0 \leq \boldsymbol{b}$, $0 < \delta < \frac{1}{2}$.
Let $j := 0$, and goto Step 1.

Step 1: *Search Direction, Stopping Criterion*

Compute $\quad \boldsymbol{g}_j := \boldsymbol{g}(\boldsymbol{x}_j), \quad \boldsymbol{G}_j := \boldsymbol{G}(\boldsymbol{x}_j) \quad$ and

$$\boldsymbol{s}_j := arg \min_{\boldsymbol{s}} \{\, Q_j(\boldsymbol{s}) \mid \boldsymbol{A}\,(\boldsymbol{x}_j - \boldsymbol{s}) \leq \boldsymbol{b} \,\} \qquad (18)$$

If $\boldsymbol{s}_j = \mathbf{0}$, then **STOP**, else goto Step 2.

Step 2: *Stepsize*

Choose σ_j according to Armijo-Goldstein.
Goto Step 3.

Step 3: *Update:*

Define

$$\boldsymbol{x}_{j+1} \;:=\; \boldsymbol{x}_j - \sigma_j \boldsymbol{s}_j\,,$$

replace j with $j+1$, and goto Step 1.

Theorem 2.6 *If F is uniformly convex on $\boldsymbol{N}_o \cap \boldsymbol{R}$, then algorithm SQP is well defined. The generated iterates $\{\boldsymbol{x}_j\}$ ensure $F(\boldsymbol{x}_{j+1}) < F(\boldsymbol{x}_j)$. If $\boldsymbol{s}_j = \boldsymbol{0}$, then $\boldsymbol{x}_j$ is a stationary point of (10).*

Proof:

The subproblem (18) is strictly convex, hence $\boldsymbol{s}_j$ is unique. We now have to show that $-\boldsymbol{s}_j$ is a descent direction, and that $\boldsymbol{s}_j = \boldsymbol{0}$ can only occur in a stationary point.
$\nabla \boldsymbol{Q}_j(\boldsymbol{s}) = -\boldsymbol{g}_j + \boldsymbol{G}_j\,\boldsymbol{s}$. The Kuhn-Tucker conditions for (18) imply

$$\begin{aligned} -\boldsymbol{g}_j + \boldsymbol{G}_j\,\boldsymbol{s}_j + \boldsymbol{A}^T\boldsymbol{\lambda}_j &= \boldsymbol{0}\,; \quad \boldsymbol{\lambda}_j \le \boldsymbol{0}\,, && (19)\\ \boldsymbol{A}\,(\boldsymbol{x}_j - \boldsymbol{s}_j) &\le \boldsymbol{b}\,, && (20)\\ \boldsymbol{\lambda}_j^T\,(\boldsymbol{A}\,(\boldsymbol{x}_j - \boldsymbol{s}_j) - \boldsymbol{b}) &= \boldsymbol{0}\,. && (21) \end{aligned}$$

(21) follows from the usual complementarity condition. (19) yields

$$\boldsymbol{s}_j = \boldsymbol{G}_j^{-1}(\boldsymbol{g}_j - \boldsymbol{A}^T\boldsymbol{\lambda}_j) \tag{22}$$

which is well defined since $\boldsymbol{G}_j$ is positive definite by assumption. With $\boldsymbol{\lambda}_j \le \boldsymbol{0}$ from (19) and $\boldsymbol{A}\,\boldsymbol{x}_j - \boldsymbol{b} \le \boldsymbol{0}$ we obtain by the complementarity condition (21)

$$\boldsymbol{\lambda}_j^T\,\boldsymbol{A}\,\boldsymbol{s}_j \ge 0\,. \tag{23}$$

Further, (19) multiplied by $\boldsymbol{s}_j$ gives

$$\boldsymbol{g}_j^T\,\boldsymbol{s}_j = \boldsymbol{s}_j^T\,\boldsymbol{G}_j\,\boldsymbol{s}_j + \boldsymbol{\lambda}_j^T\,\boldsymbol{A}\,\boldsymbol{s}_j\,. \tag{24}$$

(23) and the uniform convexity yield finally

$$\boldsymbol{g}_j^T\,\boldsymbol{s}_j \ge \mu\,\|\boldsymbol{s}_j\|^2 > 0\,. \tag{25}$$

This guarantees the descent property of the algorithm. We remark further that $\boldsymbol{x}_{j+1}$ stays within the feasible region $\boldsymbol{R} = \{\,\boldsymbol{x} \mid \boldsymbol{A}\boldsymbol{x} \le \boldsymbol{b}\}$ because $\boldsymbol{x}_j \in \boldsymbol{R}$, $\boldsymbol{x}_j - \boldsymbol{s}_j \in \boldsymbol{R}$ and $\sigma_j \in (0, 1]$. In case $\boldsymbol{s}_j = \boldsymbol{0}$ we conclude from (19) that $\boldsymbol{x}_j$ must be a stationary point of (10), and hence is optimal due to the convexity of F. □

Def. 2.7 (Lagrangian)

$$L(\boldsymbol{x}, \boldsymbol{\lambda}) := F(\boldsymbol{x}) - \boldsymbol{\lambda}^T\,(\boldsymbol{A}\,\boldsymbol{x} - \boldsymbol{b})\,, \quad \boldsymbol{\lambda} \le \boldsymbol{0}\,, \tag{26}$$

denotes the **Lagrangian function** *of problem (10).*

Theorem 2.8 (Global Convergence)

Provided F is uniformly convex on $\boldsymbol{N}_o \cap \boldsymbol{R}$, the intersection of level set and feasible region, the SQP-algorithm converges to a stationary point $\bar{\boldsymbol{x}}$ of (10) with

$$\lim_{j\to\infty} \boldsymbol{\nabla L}(\boldsymbol{x}_j) = \boldsymbol{0}\,. \tag{27}$$

Proof:

We follow the scheme of [Ritter (1982)]. To get a contradiction assume that $\|\boldsymbol{\nabla L}(\boldsymbol{x}_j)\| = \|\boldsymbol{g}_j - \boldsymbol{A}^T\boldsymbol{\lambda}_j\| > \varepsilon > 0$ for all $j \in J$ where J is an infinite subset of $\mathbb{N}$. Hence by (22)

$$\|\boldsymbol{s}_j\| \geq \frac{1}{\eta}\,\|\boldsymbol{g}_j - \boldsymbol{A}^T\boldsymbol{\lambda}_j\| \geq \frac{\varepsilon}{\eta}\,, \quad \text{for all} \quad j \in J\,. \tag{28}$$

Further, by (25)

$$\frac{\boldsymbol{g}_j^T\boldsymbol{s}_j}{\|\boldsymbol{s}_j\|} \geq \mu\|\boldsymbol{s}_j\| \geq \frac{\varepsilon\mu}{\eta}\,, \quad \text{for all} \quad j \in J\,. \tag{29}$$

According to Armijo-Goldstein we choose σ_j such that

$$a_j(\sigma_j) := \frac{F(\boldsymbol{x}_j) - F(\boldsymbol{x}_j - \sigma_j\boldsymbol{s}_j)}{\sigma_j\boldsymbol{g}_j^T\boldsymbol{s}_j} \geq \delta\,. \tag{30}$$

Taylor expansion with $\boldsymbol{u}_j \in [\boldsymbol{x}_j, \boldsymbol{x}_j - \boldsymbol{s}_j]$ gives

$$\begin{aligned} a_j(\sigma) &\geq 1 - \frac{\|\boldsymbol{g}(\boldsymbol{u}_j) - \boldsymbol{g}_j\| \cdot \|\boldsymbol{s}_j\|}{\boldsymbol{g}_j^T\boldsymbol{s}_j} \\ &= 1 - \frac{\|\boldsymbol{g}(\boldsymbol{u}_j) - \boldsymbol{g}_j\|}{\frac{\boldsymbol{g}_j^T\boldsymbol{s}_j}{\|\boldsymbol{s}_j\|}} \geq 1 - \frac{\eta}{\varepsilon\mu}\|\boldsymbol{g}(\boldsymbol{u}_j) - \boldsymbol{g}_j\|\,. \end{aligned} \tag{31}$$

As $\boldsymbol{g}(\cdot)$ is uniformly continuous on the compact set $\boldsymbol{N}_o$, there is a $\tau > 0$ such that $\|\boldsymbol{g}(\boldsymbol{u}_j) - \boldsymbol{g}_j\| \leq (\varepsilon\mu)/(2\eta)$ for $\|\sigma\boldsymbol{s}_j\| \leq \tau$. Then (31) with $\|\sigma\boldsymbol{s}_j\| \leq \tau$ yields

$$a_j(\sigma) \geq 1 - \frac{1}{2} = \frac{1}{2} > \delta\,, \quad \text{for all} \quad j \in J\,,$$

which means that (13) is satisfied. In the Armijo-Goldstein rule we choose the smallest ν_j (i.e. $\sigma_j = \frac{1}{2}^{\nu_j}$ as large as possible). Hence for $j \in J$ with $\sigma_j \leq 1$

$$\|\sigma_j\boldsymbol{s}_j\| = \|(\frac{1}{2})^{\nu_j}\boldsymbol{s}_j\| \geq \min\{\|\boldsymbol{s}_j\|, \frac{\tau}{2}\}\,,$$

and further for all $j \in J$

$$\begin{aligned} F(\boldsymbol{x}_j) - F(\boldsymbol{x}_j - \sigma_j \boldsymbol{s}_j) &\geq \delta \sigma_j \boldsymbol{g}_j^T \boldsymbol{s}_j = \delta \|\sigma_j \boldsymbol{s}_j\| \frac{\boldsymbol{g}_j^T \boldsymbol{s}_j}{\|\boldsymbol{s}_j\|} \geq \delta \min\{\|\boldsymbol{s}_j\|, \frac{\tau}{2}\} \frac{\varepsilon \mu}{\eta} \\ &\geq \delta \min\{\frac{\varepsilon}{\eta}, \frac{\tau}{2}\} \frac{\varepsilon \mu}{\eta} > 0 \,. \end{aligned} \tag{32}$$

Note that the bound is independent of j and $\boldsymbol{x}_j$. As J is infinite F cannot be bounded below, which contradicts the compactness of $\boldsymbol{N}_\circ$. The remaining Kuhn-Tucker conditions are obtained from those holding for subproblem (18).

F is uniformly, hence strictly convex. Therefore, each stationary point $\bar{\boldsymbol{x}}$ is a global minimum of the linearly constrained problem (10). Further, $\bar{\boldsymbol{x}}$ is unique such that the sequence $\{\boldsymbol{x}_j\}$ converges to $\bar{\boldsymbol{x}}$. □

2.2 Convergence Rate

A specific property of quadratically convergent Newton-like algorithms is that the stepsize σ_j converges to 1. This is also the case for our SQP-algorithm.

Lemma 2.9 (Stepsize $\sigma_j = 1$)

If F is uniformly convex on $\boldsymbol{N}_\circ \cap \boldsymbol{R}$ and $\{\boldsymbol{s}_j\}, \{\sigma_j\}$ are generated by algorithm SQP there is a $j_0 \in I\!N$ such that $\sigma_j = 1$ for $j \geq j_0$.

Proof:

According to Taylor there is some $\boldsymbol{u}_j \in [\boldsymbol{x}_j, \boldsymbol{x}_j - \boldsymbol{s}_j]$ such that

$$F(\boldsymbol{x}_j - \boldsymbol{s}_j) = F(\boldsymbol{x}_j) - \boldsymbol{g}_j^T \boldsymbol{s}_j + \tfrac{1}{2} \boldsymbol{s}_j^T \boldsymbol{G}(\boldsymbol{u}_j) \boldsymbol{s}_j \,. \tag{33}$$

We have to show $a_j(1) \geq \delta > 0$ for the Armijo-Goldstein function from (30).

$$\begin{aligned} a_j(1) &= \frac{F(\boldsymbol{x}_j) - F(\boldsymbol{x}_j - 1 \cdot \boldsymbol{s}_j)}{1 \cdot \boldsymbol{g}_j^T \boldsymbol{s}_j} \\ &= \frac{1}{\boldsymbol{g}_j^T \boldsymbol{s}_j} \big(\tfrac{1}{2} \boldsymbol{g}_j^T \boldsymbol{s}_j + \tfrac{1}{2} \boldsymbol{g}_j^T \boldsymbol{s}_j - \tfrac{1}{2} \boldsymbol{s}_j^T (\boldsymbol{G}_j + \boldsymbol{G}(\boldsymbol{u}_j) - \boldsymbol{G}_j) \boldsymbol{s}_j \big) \,. \end{aligned} \tag{34}$$

With $\boldsymbol{\lambda}_j^T \boldsymbol{A} \boldsymbol{s}_j \geq 0$ from (23) we obtain

$$a_j(1) \geq \frac{1}{2} + \frac{\boldsymbol{g}_j^T \boldsymbol{s}_j - \boldsymbol{s}_j^T \boldsymbol{G}_j \boldsymbol{s}_j - \boldsymbol{\lambda}_j^T \boldsymbol{A} \boldsymbol{s}_j}{2\, \boldsymbol{g}_j^T \boldsymbol{s}_j} - \frac{\boldsymbol{s}_j^T (\boldsymbol{G}(\boldsymbol{u}_j) - \boldsymbol{G}_j) \boldsymbol{s}_j}{2\, \boldsymbol{g}_j^T \boldsymbol{s}_j}. \tag{35}$$

According to (19) the second term on the right hand side vanishes. Hence with (25)

$$\begin{aligned} a_j(1) &\geq \frac{1}{2} - \frac{1}{2}\frac{\boldsymbol{s}_j^T(\boldsymbol{G}(\boldsymbol{u}_j) - \boldsymbol{G}_j)\boldsymbol{s}_j}{\boldsymbol{g}_j^T\boldsymbol{s}_j} \\ &\geq \frac{1}{2} - \frac{1}{2}\frac{\|\boldsymbol{G}(\boldsymbol{u}_j) - \boldsymbol{G}_j\| \cdot \|\boldsymbol{s}_j\|^2}{\mu\|\boldsymbol{s}_j\|^2} \\ &= \frac{1}{2} - \frac{1}{2}\frac{\|\boldsymbol{G}(\boldsymbol{u}_j) - \boldsymbol{G}_j\|}{\mu} \end{aligned} \tag{36}$$

which converges to $\frac{1}{2} > \delta$ since $\lim \boldsymbol{G}(\boldsymbol{u}_j) = \lim \boldsymbol{G}_j = \boldsymbol{G}(\bar{\boldsymbol{x}})$. □

So far the only used regularity condition was the uniform convexity of F. For technical reasons we now need assumptions on the linear independence of the gradients $\boldsymbol{a}_k$ corresponding to the active constraints. Denote by $\bar{K} := \{\, 1 \leq k \leq m \mid \boldsymbol{a}_k^T\bar{\boldsymbol{x}} = b_k \,\}$ the index set of active constraints at $\bar{\boldsymbol{x}}$, and by $\bar{\boldsymbol{\lambda}} := (\bar{\lambda}_1, ..., \bar{\lambda}_m)^T$ the optimal Lagrange multipliers.

Def. 2.10 (Assumption A)

The optimization problem (10) is said to satisfy **Assumption A** *if and only if F is twice continuously differentiable and if $(\bar{\boldsymbol{x}}, \bar{\boldsymbol{\lambda}})$, $\bar{\boldsymbol{\lambda}} \leq \boldsymbol{0}$, is a Kuhn-Tucker pair of (10) according to Def. 2.5 and*

$$\boldsymbol{s}^T\boldsymbol{G}(\bar{\boldsymbol{x}})\,\boldsymbol{s} > 0 \quad \text{for all} \quad \boldsymbol{s} \neq \boldsymbol{0} \quad \text{with} \quad \boldsymbol{a}_k^T\boldsymbol{s} = 0\,, \quad k \in \bar{K}\,, \tag{37}$$

and

$$\bar{\lambda}_k < 0 \quad \text{for} \quad k \in \bar{K}, \qquad (\textit{strict complementary slackness}) \tag{38}$$

and if

$$\bar{\boldsymbol{A}} := (\boldsymbol{a}_k)_{k\in\bar{K}}^T \in I\!R^{r,p} \quad \text{has full rank}\,. \tag{39}$$

Clearly, (37) follows for the special models of section 1 directly from the uniform convexity of F on the level set $\boldsymbol{N}_o$. The gradients of the active constraints in (39) can only be linearly independent if linear equalities are stated separately, and not in the way proposed for theoretical purposes in (11). Without loss of generality the corresponding Lagrange multiplier can be chosen positive; otherwise we redefine the equation by $-\boldsymbol{a}_k^T\boldsymbol{x} = -b_k$.

With stepsize $\sigma_j = 1$ our algorithm SQP is nothing else but the method of [Wilson (1963)] for linear constraints; see also [Gill et al. (1989)]. Among

others the quadratic convergence properties of Wilson's method are investigated in [Robinson (1974)]. Here, we only mention those aspects of the proof which can be exploited algorithmically. The subproblems (18) are viewed as perturbations of the original problem for which, under certain conditions, the solution converges to that of (10).

Theorem 2.11 (Active Constraints, Lagrange Multipliers)

Let the $\{\boldsymbol{x}_j\}$ *generated by algorithm SQP converge to* $\bar{\boldsymbol{x}}$ *and* $\boldsymbol{\lambda}_j$ *be the Lagrange multipliers of the quadratic subproblems (18). If (10) satisfies* **Assumption A** *at* $(\bar{\boldsymbol{x}}, \bar{\boldsymbol{\lambda}})$ *then the Lagrange multipliers* $\boldsymbol{\lambda}_j$ *of the quadratic subproblems (18) converge to the optimal multipliers* $\bar{\boldsymbol{\lambda}}$*. Further, there is an* j_0 *such that*

$$(\lambda_j)_k < 0 \qquad \text{if} \quad \bar{\lambda}_k < 0 \qquad \text{and} \tag{40}$$

$$\boldsymbol{a}_k^T \boldsymbol{x}_j < b_k \qquad \text{if} \quad \boldsymbol{a}_k^T \bar{\boldsymbol{x}} < b_k \tag{41}$$

for all $j \geq j_0$*.*

Proof: See theorem 2.1 of [Robinson (1974)], which is an application of the implicit function theorem. □

Hence we can be sure to pick ultimately the correct active constraints. In such a case, however, the solution of the subproblem (18) can be found very efficiently.

Corollary 2.12 (Newton Direction in Subspaces)

Choose $\boldsymbol{S} \in \mathbb{R}^{p-r,p}$ *with orthonormal columns and* $\bar{\boldsymbol{A}}^T \boldsymbol{S} = \boldsymbol{0}$*, where according to (39),* $r = |\bar{K}| = rank(\bar{\boldsymbol{A}})$*. Under the assumptions of theorem 2.11 there is an index* j_0 *such that the subproblem (18) can be replaced with*

$$\boldsymbol{w}_j := arg \min_{\boldsymbol{w}} \{ F(\boldsymbol{x}_j - \boldsymbol{S}\,\boldsymbol{w}) \mid \boldsymbol{w} \in \mathbb{R}^{p-r} \}\,. \tag{42}$$

$\boldsymbol{s}_j := \boldsymbol{S}\,\boldsymbol{w}_j$ *of (18) can explicitly be computed from*

$$\boldsymbol{S}^T \boldsymbol{G}_j \boldsymbol{S}\,\boldsymbol{w}_j = \boldsymbol{S}^T \boldsymbol{g}_j\,. \tag{43}$$

Theorem 2.13 (Quadratic Convergence)

Under the assumptions of theorem 2.11 the algorithm SQP converges quadratically to the optimal point $\bar{\boldsymbol{x}}$ *if* $\boldsymbol{\nabla}^2 \boldsymbol{F}$ *satisfies a Lipschitz condition in a neighbourhood of* $\bar{\boldsymbol{x}}$*.*

Proof:

Consider $\boldsymbol{x}_{j_0}$ with $\bar{\boldsymbol{A}}\,\boldsymbol{x}_{j_0} = (b_k)_{k\in\bar{K}}$ and

$$\bar{F}(\boldsymbol{w}) := F(\boldsymbol{x}_{j_0} - \boldsymbol{S}\,\boldsymbol{w})\,, \tag{44}$$

being uniformly convex with

$$\boldsymbol{\nabla}^2\bar{\boldsymbol{F}}(\boldsymbol{w}) = \boldsymbol{S}^T\boldsymbol{G}(\boldsymbol{x}_{j_0} - \boldsymbol{S}\,\boldsymbol{w})\,\boldsymbol{S}\,. \tag{45}$$

With $\boldsymbol{\nabla}^2\boldsymbol{F}$ also $\boldsymbol{\nabla}^2\bar{\boldsymbol{F}}$ satisfies a Lipschitz condition. Hence Newton's method converges quadratically. According to theorem 2.11 and corollary 2.12 in a neighbourhood close enough to $\bar{\boldsymbol{x}}$ the sequences generated by Newton's method for $\bar{F}$ and by algorithm SQP coincide. This completes the proof. □

Corollary 2.14 (Unconstrained Case)

All the stated global and local convergence properties remain valid for the case without linear restrictions. This fact is well known and has been exploited since ever to compute unconstrained ML-estimates in generalized linear models (GLM's) efficiently.

Remark:

1. The convergence results can be extended to problems with nonlinear constraints, cf. e.g. [Robinson (1974)]. From the data analysis point of view distribution results are hard to obtain even for linear constraints (see [McDonald & Diamond (1990)], [Piegorsch (1990)] and [Nyquist (1991)]). This may explain, why this paper is restricted to the linearly constrained case, where convergence results and algorithms can be formulated in quite a simple manner. Nevertheless, a review of the proofs shows that the assumptions can be weakened. The identities (19) - (25) play a crucial role throughout, and there may be scenarios where (23) and (25) hold on $\boldsymbol{N}_0 \cap \boldsymbol{R}$ and F is far from being convex.

2. General link functions for GLM's are of considerable practical interest. Like for unconstrained ML-estimates the Hessian should then be replaced with the positive (semi-)definite Fisher scoring matrix $\tilde{\boldsymbol{G}}_j$, cf. e.g. [Fahrmeir & Kredler (1984)], which is of the same approximation type as the Gauß-Newton-Matrix in Nonlinear Regression. The resulting modified SQP-algorithm can be expected to have satisfactory numerical behaviour. Weaker convergence properties than those stated here can be obtained under appropriate conditions, e.g. work with $\tilde{\boldsymbol{G}}_j + \kappa\,\boldsymbol{I}$ instead of $\tilde{\boldsymbol{G}}_j$, if necessary, to obtain good enough descent directions.

3 A GLM with Parameter Restrictions

Ship Damage Data (Lloyd's Register of Shipping)
cf. [McCullagh & Nelder (1983)], p. 137

Ship type	Year of construction	Period of operation	aggregate months service	number of damage incidents
A	1960-64	1960-74	127	0
A	1960-64	1975-79	63	0
A	1965-69	1960-74	1095	3
A	1965-69	1975-79	1095	4
A	1970-74	1960-74	1512	6
A	1970-74	1975-79	3353	18
A	1975-79	1960-74	0	0*
A	1975-79	1975-79	2244	11
B	1960-64	1960-74	44882	39
B	1960-64	1975-79	17176	29
B	1965-69	1960-74	28609	58
B	1965-69	1975-79	20370	53
B	1970-74	1960-74	7064	12
B	1970-74	1975-79	13099	44
B	1975-79	1960-74	0	0*
B	1975-79	1975-79	7117	18
C	1960-64	1960-74	1179	1
C	1960-64	1975-79	552	1
C	1965-69	1960-74	781	0
C	1965-69	1975-79	676	1
C	1970-74	1960-74	783	6
C	1970-74	1975-79	1948	2
C	1975-79	1960-74	0	0*
C	1975-79	1975-79	274	1
D	1960-64	1960-74	251	0
D	1960-64	1975-79	105	0
D	1965-69	1960-74	288	0
D	1965-69	1975-79	192	0
D	1970-74	1960-74	349	2
D	1970-74	1975-79	1208	11
D	1975-79	1960-74	0	0*
D	1975-79	1975-79	2051	4
E	1960-64	1960-74	45	0
E	1960-64	1975-79	0	0*
E	1965-69	1960-74	789	7
E	1965-69	1975-79	437	7
E	1970-74	1960-74	1157	5
E	1970-74	1975-79	2161	12
E	1975-79	1960-74	0	0*
E	1975-79	1975-79	542	1

* Necessarily empty cells for ships not (yet) in action.

The company has to balance financial risk and is interested in a statistical analysis that allows a distinction between categories of safe ships and others.

Damage Rates in $^0/_{00}$				
	year of construction			
ship type	60-64	65-69	70-74	75-79
A	0.0	3.2	4.9	4.9
B	1.1	2.3	2.3	2.5
C	1.2	0.7	2.9	3.6
D	0.0	0.0	8.3	2.0
E	0.0	11.4	5.1	1.8

Lowest risk is observed for ship types B and C and highest for type E. Unexpectedly, the oldest ships appear to be the safest, whereas those built in 1965-1974 seem to yield highest risk. To validate this conjecture we define an appropriate generalized linear model (GLM, see section 1) satisfying all assumptions made in the previous section. The ML-estimation of the model parameters was done with the SQP-algorithm described in the previous section. The reported numerical results can be found in [Klinger (1992)].

$T = 40$ design vectors

$$\boldsymbol{z} \;=\; (1, z_{\cdot 1}, \ldots, z_{\cdot 8}, z_{\cdot 9})^T$$

are defined with effect coding for the factors. For instance, reference category $60-64$ yields

$(z_{\cdot 1}, z_{\cdot 2}, z_{\cdot 3})$	year of construction
$(1,0,0)$	$60-64$
$(0,1,0)$	$65-69$
$(0,0,1)$	$70-74$
$(-1,-1,-1)$	$75-79$

Reference category ship type E defines $z_{\cdot 4} \cdots z_{\cdot 7}$ for the remaining 4 categories A,B,C,D. $z_{\cdot 8} = 1$ is for period of operation in 75-79, whereas $z_{\cdot 8} = -1$ means 60-74. Finally, $z_{\cdot 9}$ describes the monthly service hours of the individual ship.

We follow [McCullagh & Nelder (1983)] and view the number y_t of damage incidents as independent Poisson variables. This gives a log-linear GLM with parameter vector $\boldsymbol{\beta} \;=\; (\beta_0, \beta_1, \ldots, \beta_9)^T$

$$\ln(E[y_\cdot]) \;=\; \beta_0 \;+\; \sum_{j=1}^{9} z_{\cdot j}\beta_j \;, \tag{46}$$

which yields the log-likelihood function

$$l(\boldsymbol{\beta}) \;=\; \sum_{i=1}^{N} [\, y_i \boldsymbol{z}_i^T \boldsymbol{\beta} \;-\; \exp(\boldsymbol{z}_i^T \boldsymbol{\beta})] \,. \tag{47}$$

The analysis of interactions is omitted here. Unrestricted ML-estimation gives

$$\begin{array}{lccccc} \hat{\beta} = (\text{-}5.175, & \text{-}0.448, & 0.214, & 0.312, & \ldots, & 0.905) \\ \quad \beta_0 & \underbrace{60-64} & \underbrace{65-69} & \underbrace{70-74} & \ldots & \text{service hours} \end{array}$$

year of construction (75-79 is reference category)

Due to effect coding the reference category parameter becomes

$$\hat{\boldsymbol{\beta}}_{75-79} \;=\; -\sum_{j=1}^{3} \hat{\beta}_j \;=\; -0.078 \,. \tag{48}$$

Again, the conjecture that older ships are safer than new ones can be expressed in terms of the parameters of (46). Statistically spoken, we want to know whether the parameters corresponding to earlier construction periods are significantly smaller than the others. This can be formulated in the following hypotheses

$$H_0 : \beta_1 \;=\; \beta_2 \;=\; \beta_3$$

versus

$$H_1 : \beta_1 \;\leq\; \beta_2 \;\leq\; \beta_3 \,.$$

$\hat{\boldsymbol{\beta}}$ satisfies restriction H_1. Under the equality constraints of H_0 we get

$$\begin{array}{lccccc} \tilde{\beta} = (\text{-}3.940, & 0.084, & 0.084, & 0.084, & \ldots, & 0.759) \\ \quad \beta_0 & \underbrace{65-69} & \underbrace{70-74} & \underbrace{75-79} & \ldots & \text{service hours} \end{array}$$

year of construction

Following the theory of [Wollak (1987)], [McDonald & Diamond (1990)], and [Fahrmeir & Klinger (1994)] the corrected likelihood ratio statistic lq^* is distributed like a mixture of χ^2-variables with specific weights. [Klinger (1992)] reports $lq^* = 19.28$ with a "p-value" of 0.0001. Hence H_1 is accepted, and the year of construction has an highly significant influence on the number of damage incidents.

One possibility to check the significance of the negative influence of $\boldsymbol{\beta}_{75-79}$ is to recode the variables for the construction period with new reference category $60-64$. We obtain

$$\hat{\beta} = (\underset{\beta_0}{-5.175}, \underbrace{0.214}_{65-69}, \underbrace{0.312}_{70-74}, \underbrace{-0.078}_{75-79}, \ldots, \underset{\text{service hours}}{0.905})$$

year of construction (60-64 is reference category)

The statistical hypothesis to test is now

$$H_0 : \beta_1 \leq \beta_2 \leq \beta_3$$

versus

$$H_1 : \beta_1, \beta_2, \beta_3 \quad \text{arbitrary (not in } H_0).$$

With

$$\boldsymbol{A} = \begin{pmatrix} 0 & -1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & -1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix}, \boldsymbol{b} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$$

H_0 is equivalent with $\boldsymbol{A}\boldsymbol{\beta} \leq \boldsymbol{b}$. Under these restrictions we obtain with the algorithm of section 2 the ML-estimate

$$\tilde{\beta} = (\underset{\beta_0}{-5.366}, \underbrace{0.159}_{65-69}, \underbrace{0.180}_{70-74}, \underbrace{0.180}_{75-79}, \ldots, \underset{\text{service hours}}{0.935})$$

year of construction

Again, according to [Klinger (1992)], the corrected likelihood ratio statistic is $lq^* = 3.04$ with a p-value of 0.0804. So hypothesis H_0 is accepted, which finally confirms the conjecture that in the considered data set older ships are more robust than newer ones.

Examples with linear parameter restrictions for binomial credit scoring and endodontric risk data are discussed in [Fahrmeir & Klinger (1994)].

4 Conclusion

SQP-methods are appropriate for a wider class of parameter restricted estimation problems than the here described generalized linear models with canonical parameters. If analytic first derivatives are available the FORTRAN-algorithm of [Schittkowski (1981)] will usually produce good results.

Sometimes the exact model is not known a priori and alternatives are to be checked before the best model is found. In this case the derivative coding for each new model is not only time consuming but also a source of many errors. Numerical difference quotients instead of analytic derivatives are an

obvious but frequently not numerically stable alternative. In that situation automatic differentiation is a good way out. The PC-program PADMOS, cf. [Greiner, Kredler & Wagenpfeil (1992)], has been developed especially for such purposes. Text files contain the Pascal-like description of log-likelihood function and constraints. In the data file for each observation an extra line is reserved. A comfortable user interface with built-in editor enables a convenient input and modification of ML-problems. Models with some hundred observations and up to 15 parameters can be analyzed. Box, linear and even nonlinear type constraints are accepted. The execution time of PADMOS is about 10 times of that needed by analytic derivative procedures. But for any chosen model first **and** second derivatives are computed automatically, and hence the broad variety of powerful optimization algorithms, including methods with directions of negative curvature, is applicable, and ensures good convergence properties.

Acknowledgement

I wish to thank Joachim Klinger and Stefan Wagenpfeil for helpful comments and discussions.

References

[Albert & Anderson (1984)] Albert A. and Anderson J.A.: On the existence of maximum likelihood estimates in logistic regression models. Biometrica **71**, 1-10.

[Best & Ritter (1988)] Best M.J. and Ritter K.: A quadratic programming algorithm. ZOR **32**, 271-297.

[Fahrmeir & Klinger (1994)] Fahrmeir L. and Klinger J.: Estimating and testing generalized linear models under inequality restrictions. Statistical Papers **35**, 211-229.

[Fahrmeir & Kredler (1984)] Fahrmeir L. und Kredler Ch.: Verallgemeinerte lineare Modelle. In: Fahrmeir L. und Hamerle A. (Hrsg.): Multivariate statistische Verfahren. De Gruyter, Berlin.

[Gill et al. (1989)] Gill P.E., Murray W., Saunders M.A. and Wright M.H.: Constrained nonlinear programming. In: G.L. Nemhauser et al. (eds.): Handbooks in Operations Research and Management Science. North-Holland, Amsterdam.

[Greiner, Kredler & Wagenpfeil (1992)] Greiner M., Kredler Ch. and Wagenpfeil St.: Nonlinear optimization with PADMOS: User's guide and algorithms. Report **366**, DFG Schwerpunkt: Anwendungsbezogene Optimierung und Steuerung, TU München.

[Kaufmann (1988)] H.: On existence and uniqueness of maximum likelihood estimates in quantal and ordinal response models. Metrica **35**, 291-313.

[Klinger (1992)] J.: Tests für Ungleichungsrestriktionen in generalisierten linearen Modellen. Diplomarbeit (Referent. Prof. L. Fahrmeir), LMU München.

[Kredler (1986)] Ch.: Behaviour of third order terms in quadratic approximations of LR-statistics in multivariate generalized linear models. The Annals of Statistics **14**, 326-335.

[McCullagh & Nelder (1983)] McCullagh P. and Nelder J.A.: Generalized linear models. Chapman and Hall, London.

[McDonald & Diamond (1990)] McDonald P. and Diamond I.: On the fitting of generalized linear models with nonnegativity parameter constraints. Biometrics **46**, 201-206.

[Nelder & Wedderburn (1972)] Nelder J. and Wedderburn R.W.M.: Generalized linear models. J. Roy. Statist. Soc. **A 135**, 370-384.

[Nyquist (1991)] H.: Restricted estimation of generalized linear models. Applied Statistics, **40**, 133-141.

[Piegorsch (1990)] W.: One-side-significance tests for generalized linear models under dichotomous response. Biometrics **46**, 309-316.

[Ritter (1982)] K.: Numerical methods for nonlinear programming problems. In: Korte B. (ed.): MODERN APPLIED MATHEMATICS — Optimization and Operations Research. North-Holland Publ. Comp. Amsterdam.

[Robinson (1974)] S.M.: Perturbed Kuhn-Tucker points and rates of convergence for a class of nonlinear programming algorithms. Mathematical Programming **7**,1-16.

[Schittkowski (1981)] K.: The nonlinear programming method of Wilson, Han and Powell with an augmented Lagrangian type line search function. Part 1: Convergence analysis. Numer. Math. **38**, 83-114.

[Wedderburn (1976)] R.W.M.: On the existence and uniqueness of the maximum likelihood estimates for certain generalized linear models. Biometrika **63**, 27-32.

[Wilson (1963)] R.B.: A simplicial algorithm for concave programming. PhD thesis, Harvard University, Cambridge.

[Wollak (1987)] F.A.: An exakt test for multiple inequality and equality constraints in the linear model. JASA **82**, 782-793.