Forschungsschwerpunkt S92

Industrial Geometry

http://www.ig.jku.at

INDUSTRIAL GEOMETRY

FSP Report No. 54

Distance Regression by Gauss-Newton-type Methods and Iteratively Re-weighted Least-Squares

M. Aigner, B. Jüttler

July 2007

FWF

Der Wissenschaftsfonds.

Distance Regression by Gauss-Newton-type Methods and Iteratively Re-weighted Least-Squares

M. Aigner, B. Jüttler
Institute of Applied Geometry, Johannes Kepler University, Linz, Austria

We discuss the problem of fitting a curve or surface to given measurement data. In many situations, the usual least–squares approach (minimization of the sum of squared residuals) is not appropriate, as it implicitly assumes a Gaussian distribution of the measurement errors. In those cases, it is more appropriate to minimize other functions of the residuals, which we will call norm–like functions. We show that the technique of iteratively re-weighted least-squares (IRLS), which originated in statistics (Huber, 1981) is a Gauss-Newton-type method for minimizing a sum of such functions of the residuals.

Keywords: Gauss-Newton method, iteratively re-weighted least-squares (IRLS), curve fitting, surface fitting

1. Introduction

The problem of fitting a given mathematical model to a certain set of measurement data appears in the context of various applications. For instance, during the process of reverse engineering of geometric models, one is interested in fitting a parametric curve or surface (often represented by NURBS: non–uniform rational B-splines) to a given point cloud (see e.g. Rogers & Fog, 1989; Sarkar & Menq, 1991; Hoschek & Lasser, 1993; Watson, 1999; Alhanaty & Bercovier, 2001; Varady & Martin, 2002; Atieg & Watson, 2003; Pottmann *et al.*, 2005). In statistics, regression lines or other geometric primitives need to be identified in order to analyze the relations between two or more variables (Huber, 1981). In biomedicine, image segmentation or the automatic recognition of biological structures is an important issue (Mahadevan *et al.*, 2004).

Most existing techniques rely on least-squares approximation, minimizing the sum of the squared residuals (i.e. the deviations between the model and the measured data). Equivalently, this can be seen as minimization of the ℓ_2 norm of the vector of (scalar–valued) residuals. In the case of (parametric or implicitly defined) curves and surfaces, the residuals are often chosen to be the distances between the data and the associated closest points. This leads to the special case of orthogonal distance regression (Rogers & Fog, 1989; Sarkar & Menq, 1991; Speer*et al.*, 1998; Blake & Isard, 1998; Alhanaty & Bercovier, 2001; Atieg & Watson, 2003; Pottmann & Leopoldseder, 2003; Wang *et al.*, 2006; Aigner *et al.*, 2007).

If the residuals depend linearly on the parameters governing the model, then the solution can be found by solving a system of linear equations. In the case of a non-linear dependence between residuals and parameters, Gauss-Newton-type methods are widely used, as they avoid the evaluation of second derivatives (see Atieg & Watson, 2003; Wang *et al.*, 2006). In particular, the nonlinearity is present in the case of orthogonal distance regression. See Yamamoto (2000) for a historical overview of Newton and Newton-like methods.

In many situations, considering the ℓ_2 norm of the residuals is not the appropriate approach.

On the one hand, so–called outliers (data with large error) may destroy the quality of the approximation, since their influence grows quadratically with the distance to the curve. On the other hand, if the data are very precise, then it is more appropriate to minimize the maximum deviation between the model and the data. It is clearly important to adjust the norm carefully to the problem (see Watson, 1977; Jüttler, 1998; Hermey & Watson, 1999; Watson, 2000, 2002; Al-Subaihi & Watson, 2004; Atieg & Watson, 2004).

In the field of statistics (Huber, 1981), this is achieved with the help of the concept of M-estimators. An M-estimator is a function $\rho(r)$ of the scalar residual r that grows sub-quadratically with r and that is monotonically nondecreasing with increasing r. For instance, one one may use ℓ_p norms of the vector of residuals for different values of p. This leads to more meaningful results, while increasing the computational costs.

This paper is organized as follows. In the next two sections we introduce the generalized fitting problem and discuss iterative techniques for solving it. Section 4 analyzes the connection between Gauss-Newton and the method of iteratively re-weighted least-squares (IRLS). Section 5 presents several examples. Finally, we conclude this paper.

2. The generalized fitting problem

Given a set of data points $\{\mathbf{P}_j\}_{j=1..N}$ in $\mathbb{R}^d$, we are interested in a surface of dimension $k < d$ that approximates these points in the best way. In this paper we restrict ourselves to curves ($k = 1$). The results which are presented in this paper can be extended to higher dimensions and to other settings, such as the problem of fitting a parametric surface ($k = 2$).

We denote a parametric curve by $\mathbf{c}_{\mathbf{s}}(t)$, with the curve parameter $t \in I = [a, b] \subset \mathbb{R}$. Similarly to the framework described in (Aigner & Jüttler, 2007; Aigner *et al.*, 2007), we assume that the curve is described by a vector of shape parameters $\mathbf{s} \in \mathbb{R}^n$.

Example 2.1 Consider a spline curve

$$\mathbf{c}_{\mathbf{s}}(t) = \sum_{i=0}^{m} \mathbf{d}_i \psi_i(t) \tag{1}$$

where $\psi_i(t)$ are the B-splines of a certain degree k, defined over a suitable fixed knot vector (see Hoschek & Lasser, 1993). The vector of shape parameters is obtained by concatenation of all control points $\mathbf{d}_i$. In addition it may also contain the knots of the curve.

Example 2.2 An ellipse in the plane can be parameterized using trigonometric functions as

$$\mathbf{c}_{\mathbf{s}}(t) = \begin{pmatrix} \cos\phi & -\sin\phi \\ \sin\phi & \cos\phi \end{pmatrix} \begin{pmatrix} x_0 + a\sin(t) \\ y_0 + b\cos(t) \end{pmatrix}, \quad t \in I = [0, 2\pi). \tag{2}$$

The shape parameters are the coordinates of the center (x_0, y_0) along with the semimajor a and the semiminor axes b and the rotation angle ϕ of the ellipse.

For a fixed data point $\mathbf{P}_j$ and associated point $\mathbf{c}_{\mathbf{s}}(t_j(\mathbf{s}))$ on the curve, the residual vector is given by

$$\mathbf{R}_j(\mathbf{s}, t_j(\mathbf{s})) = \mathbf{P}_j - \mathbf{c}_{\mathbf{s}}(t_j(\mathbf{s})). \tag{3}$$

Its Euclidean norm $||\mathbf{R}_j||$ is the scalar residual associated with $\mathbf{P}_j$. In order to fit the curve to the data, one minimizes simultaneously the residuals for all data points. Depending on the

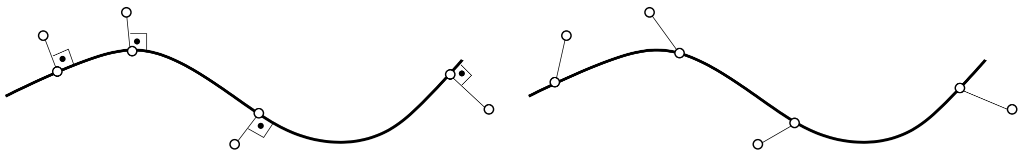


FIG. 1. Orthogonal distance regression (a) and distance regression (b).

choice of $t_j(\mathbf{s})$, two basic fitting problems can be formulated (see also Watson, 2000; Atieg & Watson, 2004):

(i) *Orthogonal distance regression* (ODR). For any given vector of shape parameters, the parameter values t_j are chosen such that the points $\mathbf{c}_\mathbf{s}(t_j(\mathbf{s}))$ are as close as possible to the given data,

$$t_j(\mathbf{s}) = \arg\min_{t\in I} \|\mathbf{c}_\mathbf{s}(t(\mathbf{s})) - \mathbf{P}_j\|. \tag{4}$$

In this case, the residual vectors are perpendicular to the curve, see Figure 1(a). For a regular curve point, i.e. $\mathbf{c}'_\mathbf{s}(t_j) \neq (0,0)^\top$, this orthogonality can be formulated as

$$\mathbf{c}'_\mathbf{s}(t_j(\mathbf{s}))^\top(\mathbf{c}_\mathbf{s}(t_j(\mathbf{s})) - \mathbf{P}_j) = 0, \tag{5}$$

where $'$ denotes the derivative with respect to the curve parameter. Clearly, one has to identify both the curve that provides the minimal distance to the data points, and also the curve parameters t_j where these distances are realized. Due to (4), the curve parameters t_j depend highly on the shape parameters $\mathbf{s}$.

(ii) *Distance regression*. In this simpler case, the parameter values t_j are fixed a priori using a suitable parameterization method, such as chordal, centripetal or uniform parameterization (see e.g. Hoschek & Lasser, 1993). The residuals are given by $\mathbf{R}_j(\mathbf{s}, t_j) = \mathbf{P}_j - \mathbf{c}_\mathbf{s}(t_j)$, see Figure 1(b). In this case, the distances from the data point to the curve are not necessarily the shortest possible ones.

We consider the following *generalized fitting problem:*

$$F(\mathbf{s}) = \sum_{j=1}^{M} N(\|\mathbf{R}_j(\mathbf{s})\|) \to \min_{\mathbf{s}}. \tag{6}$$

It generalizes the usual fitting problems by replacing the ℓ_p norms (mostly with $p = 2$) of the vector of scalar residuals $||\mathbf{R}_j||$ with a sum of a certain functions $N(.)$ applied to the residuals. We choose these functions from the class of norm-like functions in the following sense.

Definition 2.3 A $\mathcal{C}^2$ function $N(x) : \mathbb{R}^+ \to \mathbb{R}^+$ is said to be *norm-like* if there exists $\epsilon \in \mathbb{R}^+$ such that the derivative satisfies

$$N'(x) = xw(x) \quad \text{for} \quad x \in (0, \epsilon] \tag{7}$$

where the associated *weight function* $w(x)$ is positive. If the weight function $w(x)$ can smoothly be extended such that $w(x)) : [0, \varepsilon] \to [c, C]$ with $c, C \in \mathbb{R}^+$, then we will call it *positive* and *bounded*.

EXAMPLE 2.4 1. The norm–like function $N(x) = x^2$, which corresponds to the ℓ_2 norm of the vector of residuals, has the weight function $w(x) = 2$ which is positive and bounded.

2. The norm–like function $N(x) = \exp(x^2) - 1$ has the weight function $w(x) = 2\exp x^2$. Again, the weight function is positive and bounded.

3. The norm–like functions $N(x) = x^p$ for $p \neq 2$, which correspond to the ℓ_p norm of the vector of residuals, have the weight functions $w(x) = px^{p-2}$. The weight functions are positive, but not bounded, for $1 < p < 2$, and bounded, but not positive, for $p > 2$.

3. Iterative techniques for solving the generalized fitting problem

The generalized fitting problem (6) defines a non-linear optimization problem. We consider iterative techniques for computing an approximate solution. In each step, a better approximation $\mathbf{s}^+$ of the exact solution is found by updating the current solution $\mathbf{s}^c$ via $\mathbf{s}^+ = \mathbf{s}^c + h\Delta\mathbf{s}$, where $h \leqslant 1$ is a certain step size.

3.1 *Gradient and Hessian*

For later reference we provide the gradient and the Hessian of the objective function $F(\mathbf{s})$, see (6). The gradient, which is the row vector of the first partial derivatives with respect to the shape parameters s_i, is

$$\nabla F = \sum_{j=1}^{M} N'(\|\mathbf{R}_j\|)\frac{\mathbf{R}_j^\top}{\|\mathbf{R}_j\|}\nabla\mathbf{R}_j = \sum_{j=1}^{M} w(\|\mathbf{R}_j\|)\mathbf{R}_j^\top\nabla\mathbf{R}_j \tag{8}$$

where $\mathbf{R}_j$ denotes the residual at the current position $\mathbf{s}^c$. The Hessian is the $n \times n$–matrix

$$H_F = \nabla(\nabla F^\top) = \sum_{j=1}^{M} \frac{w_j'}{\|\mathbf{R}_j\|}\nabla\mathbf{R}_j^\top\mathbf{R}_j\mathbf{R}_j^\top\nabla\mathbf{R}_j + w_j\nabla\mathbf{R}_j^\top\nabla\mathbf{R}_j + w_j\nabla(\nabla\mathbf{R}_j^\top)\circ\mathbf{R}_j, \tag{9}$$

where we use the abbreviations $w_j = w(\|\mathbf{R}_j\|)$ and $w_j' = w'(\|\mathbf{R}_j\|)$. The second order derivative $\nabla(\nabla\mathbf{R}_j^\top)$ is a tensor, and is to be interpreted in the following way:

$$\left[\nabla(\nabla\mathbf{R}_j^\top)\circ\mathbf{R}_j\right]_{l,k} = \sum_{i=1}^{d}\left[\frac{\partial}{\partial s_l}\frac{\partial}{\partial s_k}[\mathbf{R}_j]_i\right][\mathbf{R}_j]_i. \tag{10}$$

Here, $[\mathbf{V}]_i$ denotes the i–th entry of the vector $\mathbf{V}$ and $[\mathbf{M}]_{i,j}$ denotes the j–th entry of the i–th row of the matrix $\mathbf{M}$.

3.2 *Linearization of* $\mathbf{R}_j$

The classical Gauss-Newton-type methods are based on the linearization of the residuals. For the generalized fitting problem (6) this gives

$$F^*(\Delta\mathbf{s}) = \sum_{j=1}^{M} N(\|\mathbf{R}_j\| + \nabla\|\mathbf{R}_j\|\Delta\mathbf{s}) \to \min_{\Delta\mathbf{s}}. \tag{11}$$

If $N(x) = x^2$, then this can be solved efficiently for the update vector $\Delta\mathbf{s}$. However, in the case of general norm-like functions $N(x)$ the equations obtained from the direct linearization (11) are non-linear in the unknowns $\Delta\mathbf{s}$,

$$\nabla F^{*\top}(\Delta\mathbf{s}) = \sum_{j=1}^{M} N'(\|\mathbf{R}_j\| + \nabla\|\mathbf{R}_j\|\Delta\mathbf{s})(\nabla\|\mathbf{R}_j\|)^\top = 0.$$

Therefore the direct linearization (11) of the residuals should be avoided. In the particular case of the ℓ_1 norm, the problem becomes even non-differentiable in the limit. Watson (2002) shows that the problem can be reformulated such that it becomes differentiable. Unfortunately, one has to know a-priori which data points are interpolated, hence which residuals vanish in the limit.

3.3 *From Newton to Gauss-Newton*

Clearly, a local minimum of the objective function F fulfills the condition

$$\nabla F^\top = 0.$$

In general, this equation is non-linear in the vector of unknowns $\mathbf{s}$. Using Newton's method one obtains the linearized system

$$\nabla F^\top + H_F \Delta\mathbf{s} = 0 \tag{12}$$

for the update step $\Delta\mathbf{s}$ where H_F denotes the Hessian (9) of the objective function F. Under certain assumptions, Newton's method exhibits quadratic convergence (see Yamamoto, 2000, and the references cited therein).

Newton's method is especially well suited if the residuals are linear in the unknown coefficients $\mathbf{s}$. In this case, the second order derivatives in the Hessian (9) vanish identically and the problem gets particularly simple. In other cases the exact Hessian may be unknown or very costly to evaluate. In such a situation an approximation of the exact Hessian can be used. This leads to Gauss-Newton-type methods.

In the following we consider the approximate Hessian

$$H_F^* = \sum_{j=1}^{M} w_j \nabla\mathbf{R}_j^\top \nabla\mathbf{R}_j \tag{13}$$

which is obtained by omitting the first and the last part in (9). Note that the latter contains the second order derivatives with respect to the shape parameters.

In the case of vanishing residuals we get the following result.

LEMMA 3.1 Let $\bar{\mathbf{s}}$ be the minimizer of (6) such that $\|\mathbf{R}_j(\bar{\mathbf{s}})\| = 0$ for all j and let $N(x)$ have a positive and bounded weight function. Then

$$\lim_{\mathbf{s}\to\bar{\mathbf{s}}} H_F = \lim_{\mathbf{s}\to\bar{\mathbf{s}}} H_F^*. \tag{14}$$

Proof. We have to show that all but the second term of the expansion of the Hessian (9) vanish. For the last term this can be seen immediately. The first part of the Hessian is the sum of

$$T_j = w_j' \|\mathbf{R}_j\| \nabla\mathbf{R}_j^\top \frac{\mathbf{R}_j}{\|\mathbf{R}_j\|} \frac{\mathbf{R}_j^\top}{\|\mathbf{R}_j\|} \nabla\mathbf{R}_j, \quad j = 1, \ldots, M. \tag{15}$$

We consider the spectral norm of T_j,

$$\|T_j\|_2^2 \leqslant \left\| w_j' \|\mathbf{R}_j\| \nabla \mathbf{R}_j^\top \frac{\mathbf{R}_j}{\|\mathbf{R}_j\|} \frac{\mathbf{R}_j^\top}{\|\mathbf{R}_j\|} \nabla \mathbf{R}_j \right\|_2^2 \leqslant |w_j'|^2 \ \|\mathbf{R}_j\|^2 \ \left\| \nabla \mathbf{R}_j^\top \frac{\mathbf{R}_j}{\|\mathbf{R}_j\|} \frac{\mathbf{R}_j^\top}{\|\mathbf{R}_j\|} \nabla \mathbf{R}_j \right\|_2^2$$

Exploiting the Cauchy-Schwarz inequality we obtain

$$\|T_j\|_2^2 \leqslant |w_j'|^2 \ \|\mathbf{R}_j\|^2 \ \left\| \nabla \mathbf{R}_j^\top \frac{\mathbf{R}_j}{\|\mathbf{R}_j\|} \right\|_2^2 \left\| \frac{\mathbf{R}_j^\top}{\|\mathbf{R}_j\|} \nabla \mathbf{R}_j \right\|_2^2 .$$

Finally we get

$$\|T_j\|_2^2 \leqslant |w_j'|^2 \ \|\mathbf{R}_j\|^2 \ \left\| \nabla \mathbf{R}_j^\top \right\|_2^2 \left\| \frac{\mathbf{R}_j}{\|\mathbf{R}_j\|} \right\|_2^2 \left\| \frac{\mathbf{R}_j}{\|\mathbf{R}_j\|} \right\|_2^2 \left\| \nabla \mathbf{R}_j^\top \right\|_2^2 , \tag{16}$$

as the spectral norm for matrices and the ℓ_2 norm for vectors are consistent. If the shape parameters tend to the zero residual case $\mathbf{s} \to \bar{\mathbf{s}}$, then $\|\mathbf{R}_j\| \to 0$, hence $\|T_j\|_2 \to 0$. □

REMARK 3.1 This Lemma excludes the case of ℓ_p norms for $p > 2$ since they do not fulfill the assumption that $w(0) \neq 0$ in a zero-residual case. Nevertheless, the conclusion remains valid, since $H_F = H_F^* = \mathbf{0}$ in the limit. For the case $1 \leqslant p < 2$ the optimization problem is not differentiable.

Now we use the approximate Hessian H_F^* in order to define a Gauss-Newton-type method for the generalized fitting problem (6),

$$H_F^* \Delta \mathbf{s} + \nabla F^\top = 0. \tag{17}$$

With the help of (8) and (13), we are led to formulate the following definition.

Definition 3.1 In each step of the *Gauss-Newton method for the generalized fitting problem* (6) we solve the system of linear equations

$$\sum_{j=1}^{M} w_j(\|\mathbf{R}_j(\mathbf{s}^c)\|) \nabla \mathbf{R}_j^\top(\mathbf{s}^c) \nabla \mathbf{R}_j(\mathbf{s}^c) \Delta \mathbf{s} + \sum_{j=1}^{M} w_j(\|\mathbf{R}_j(\mathbf{s}^c)\|) \nabla \mathbf{R}_j^\top(\mathbf{s}^c) \mathbf{R}_j(\mathbf{s}^c) = 0 \tag{18}$$

for the update step $\Delta \mathbf{s}$ and compute the update $\mathbf{s}^+ = \mathbf{s}^c + h\Delta\mathbf{s}$, where h is the step size $h \leqslant 1$.

3.4 *Iteratively re-weighted least-squares (IRLS)*

IRLS generalizes the standard least-squares technique by assigning an individual weight to each residual. More precisely, we reformulate the general minimization problem (6) in the following way

$$\sum_{j=1}^{M} N(\|\mathbf{R}_j\|) = \sum_{j=1}^{M} v_j \|\mathbf{R}_j\|^2. \tag{19}$$

By assigning constant values to the weights v_j, the non-linearity of the norm–like function is eliminated. Unfortunately, the weights v_j that correspond to the exact solution are a-priori unknown. This problem can be addressed by an iterative process. The weights v_j are updated in each step. They are computed from the previous approximate solution.

In addition, the residuals may depend in a non–linear way on the shape parameters. By linearizing the residuals as in the usual Gauss-Newton approach we are led to the following definition.

Definition 3.2 One step of *iteratively re-weighted least-squares (IRLS)* solves

$$\sum_{i=1}^{M} v_j(\|\mathbf{R}_j(\mathbf{s}^c)\|)\|\mathbf{R}_j(\mathbf{s}^c) + \nabla\mathbf{R}_j(\mathbf{s}^c)\Delta\mathbf{s}\|^2 \to \min_{\Delta\mathbf{s}} \tag{20}$$

for the unknown update vector $\Delta\mathbf{s}$. Then we compute the update $\mathbf{s}^+ = \mathbf{s}^c + h\Delta\mathbf{s}$, where h is the step size. The next step then uses the new weights $v_j(\|\mathbf{R}_j(\mathbf{s}^+)\|)$.

Example 3.3 In the case of ℓ_p minimization, $N(x) = x^p$, the objective function (19) can be rewritten as

$$\sum_{j=1}^{M} \|\mathbf{R}_j\|^p = \sum_{j=1}^{M} v_j\|\mathbf{R}_j\|^2 \tag{21}$$

which suggests the weights $v_j = \|\mathbf{R}_j\|^{p-2}$. Note that potential problems may occur for $1 \leqslant p < 2$. The weights diverge as the residuals tend to zero and even a single vanishing residual destroys the regularity of the problem. For this reason, different approximations for the ℓ_1 weights are used. Especially in the statistics community (see Huber, 1981), a wealth of different weights is being used.

4. Discussion

We show that IRLS and the Gauss-Newton method in the sense of (18) are equivalent.

4.1 *Equivalence of the methods*

In the case of scalar residuals which depend linearly on the unknowns, Watson (1977) establishes a connection between IRLS and Newton's method for ℓ_p approximation. O'Leary (1990) gives an overview of some minimization methods for a general norm–like function of a linear residual $N(r_i)$. Mahadevan *et al.* (2004) show that a Gauss-Newton-type method for a general norm–like function of non-linear, scalar residuals leads to an IRLS problem. We extend this result to the case of vector-valued residuals.

Proposition 4.1 One step of the Gauss-Newton-type method for solving the generalized fitting problem (6) in the sense of Definition 3.1 is equivalent to one step of IRLS in the sense of Definition 3.2, where the weights are chosen as $v_j = w(\|\mathbf{R}_j\|)$.

Proof. The minimizer $\Delta\mathbf{s}$ of the IRLS problem

$$\sum_{j=1}^{M} w(\|\mathbf{R}_j\|)\|\mathbf{R}_j + \nabla\mathbf{R}_j\Delta\mathbf{s}\|^2 \to \min_{\Delta\mathbf{s}} \tag{22}$$

satisfies

$$\sum_{j=1}^{M} w(\|\mathbf{R}_j\|)\nabla\mathbf{R}_j^\top \nabla\mathbf{R}_j \Delta\mathbf{s} + \sum_{j=1}^{M} w(\|\mathbf{R}_j\|)\nabla\mathbf{R}_j^\top \mathbf{R}_j = 0. \tag{23}$$

This is equivalent to the linear system (18). □

Consequently, any generalized fitting problem (6) can be solved via IRLS.

REMARK 4.1 Note that Eq. (21), which suggests the choice $v_j = \|\mathbf{R}_j\|^{p-2}$ of the weights, is valid only in the case of ℓ_p norms. In this case, the proposition gives the weights $v_j = p\|\mathbf{R}_j\|^{p-2}$. The factor p, which is shared by all weights, does not have any influence. In the case of general norm–like functions, the analogue of Eq. (21) is not satisfied.

4.2 *Convergence results*

In the ℓ_2 case, hence $N(x) = x^2$, standard results from optimization theory provide informations about the convergence behavior of certain iteration methods. Under certain conditions the standard Gauss-Newton method converges quadratically in the zero-residual case. We extend this result to the Gauss-Newton method in the sense of (18) for a general $N(x)$.

PROPOSITION 4.2 If the Gauss-Newton update (18) is defined, then $\Delta\mathbf{s}$ is a direction of descent.

Proof. The update step is given by

$$\Delta\mathbf{s} = -\left(\sum_{j=1}^{M} w_j \nabla\mathbf{R}_j^\top \nabla\mathbf{R}_j\right)^{-1} \sum_{j=1}^{M} w_j \nabla\mathbf{R}_j^\top \mathbf{R}_j. \tag{24}$$

Recall that the gradient is given by (8). We show that the inner product

$$\langle \nabla F^\top, \left(\sum_{j=1}^{M} w_j \nabla\mathbf{R}_j^\top \nabla\mathbf{R}_j\right)^{-1} \nabla F^\top \rangle$$

of the negative gradient of the objective function and the update vector $\Delta\mathbf{s}$ is positive. This is the case if all singular values of $\sum_{j=1}^{M} w_j \nabla\mathbf{R}_j^\top \nabla\mathbf{R}_j$ are positive. First we know that the matrix is regular, since we assumed that the Gauss-Newton update is defined. Second, from $w_j \geqslant 0$ and

$$\langle x, \sum_{j=1}^{M} w_j \nabla\mathbf{R}_j^\top \nabla\mathbf{R}_j x\rangle = \sum_{j=1}^{M} w_j \langle x^\top \nabla\mathbf{R}_j^\top, \nabla\mathbf{R}_j x\rangle > 0 \quad \forall x \neq \vec{0} \tag{25}$$

we conclude that all singular values are positive. □

As an important consequence of this observation, even for points where $N(x)$ is not $\mathcal{C}^2$, the Gauss-Newton method provides a direction of descent. The next theorem analyzes the convergence rate of the Gauss-Newton method, and therefore also of IRLS. It extends the known properties of the usual Gauss-Newton method (cf. Kelley, 1999).

THEOREM 4.3 Let $\bar{\mathbf{s}}$ be a minimizer of

$$\sum_{j=1}^{M} N(\|\mathbf{R}_j(\mathbf{s})\|) \to \min_{\mathbf{s}}$$

with a norm-like function $N(x)$ and $w(x)$ be a positive and bounded weight function. The residuals $\mathbf{R}_j$ shall be Lipschitz continuously differentiable and $\sum_{j=1}^{M} \nabla\mathbf{R}_j^\top w_j \nabla\mathbf{R}_j$ be regular in $\mathcal{B}_\delta(\overline{\mathbf{s}}) = \{\mathbf{s} : \|\mathbf{s}-\overline{\mathbf{s}}\| < \delta\}$ with $\delta > 0$. Then there exists for the Gauss-Newton method (18) some constant $K \in \mathbb{R}$ such that

$$\|\mathbf{s}^+ - \overline{\mathbf{s}}\| \leqslant K(\|\mathbf{s}^c - \overline{\mathbf{s}}\|^2 + \|\sum_{j=1}^{M} \overline{\mathbf{R}}_j\| \, \|\mathbf{s}^c - \overline{\mathbf{s}}\|)$$

where $\mathbf{R}_j = \mathbf{R}_j(\mathbf{s}^c)$, $\overline{\mathbf{R}}_j = \mathbf{R}_j(\overline{\mathbf{s}})$ and $\nabla\overline{\mathbf{R}}_j = \nabla\mathbf{R}_j(\overline{\mathbf{s}})$.

Proof. With $\Delta\mathbf{s}$ from Eq. (24), the error after one update step is given by

$$\begin{aligned}
\mathbf{s}^+ - \overline{\mathbf{s}} &= (\mathbf{s}^c - \overline{\mathbf{s}}) - [\sum_{j=1}^{M} \nabla\mathbf{R}_j^\top w_j \nabla\mathbf{R}_j]^{-1} \sum_{j=1}^{M} \nabla\mathbf{R}_j^\top w_j \mathbf{R}_j \\
&= [\sum_{j=1}^{M} \nabla\mathbf{R}_j^\top w_j \nabla\mathbf{R}_j]^{-1} \sum_{j=1}^{M} \nabla\mathbf{R}_j^\top w_j [\nabla\mathbf{R}_j(\mathbf{s}^c - \overline{\mathbf{s}}) - \mathbf{R}_j].
\end{aligned}$$

The last factor $[\ldots]$ can be rewritten as

$$\nabla\mathbf{R}_j(\mathbf{s}^c - \overline{\mathbf{s}}) - \mathbf{R}_j = \nabla\mathbf{R}_j(\mathbf{s}^c - \overline{\mathbf{s}}) - \overline{\mathbf{R}}_j + \overline{\mathbf{R}}_j - \mathbf{R}_j = -\overline{\mathbf{R}}_j + \big(\nabla\mathbf{R}_j(\mathbf{s}^c - \overline{\mathbf{s}}) + \overline{\mathbf{R}}_j - \mathbf{R}_j\big)$$

hence

$$\|\mathbf{s}^+ - \overline{\mathbf{s}}\| \leqslant \|[\sum_{j=1}^{M} \nabla\mathbf{R}_j^\top w_j \nabla\mathbf{R}_j]^{-1}\| \cdot (\|\sum_{j=1}^{M} \nabla\mathbf{R}_j^\top w_j \overline{\mathbf{R}}_j\| + \|\sum_{j=1}^{M} \nabla\mathbf{R}_j^\top w_j\| \frac{\alpha}{2} \|\mathbf{s}^c - \overline{\mathbf{s}}\|^2)$$

with the Lipschitz constant α, where we used that

$$\|\nabla\mathbf{R}_j(\mathbf{s}^c - \overline{\mathbf{s}}) + \overline{\mathbf{R}}_j - \mathbf{R}_j\| \leqslant \frac{\alpha}{2} \|\mathbf{s}^c - \overline{\mathbf{s}}\|^2.$$

Finally,

$$\|\sum_{j=1}^{M} \nabla\mathbf{R}_j^\top w_j \overline{\mathbf{R}}_j\| = \|\sum_{j=1}^{M} (\overline{w}_j \nabla\overline{\mathbf{R}}_j - w_j \nabla\mathbf{R}_j)^\top \overline{\mathbf{R}}_j\| \leqslant \alpha(C+c) \|\sum_{j=1}^{M} \overline{\mathbf{R}}_j\| \, \|\mathbf{s}^c - \overline{\mathbf{s}}\|$$

since $\sum_{j=1}^{M} \overline{w}_j (\nabla\overline{\mathbf{R}}_j)^\top \overline{\mathbf{R}}_j = 0$ and due to the Lipschitz continuity of $\nabla\mathbf{R}_j$. Summing up,

$$\begin{aligned}
\|\mathbf{s}^+ - \overline{\mathbf{s}}\| &\leqslant \|[\sum_{j=1}^{M} \nabla\mathbf{R}_j^\top w_j \nabla\mathbf{R}_j]^{-1}\| \, \alpha \, [\, (C+c) \|\sum_{j=1}^{M} \overline{\mathbf{R}}_j\| \, \|\mathbf{s}^c - \overline{\mathbf{s}}\| + \frac{1}{2} \|\sum_{j=1}^{M} \nabla\mathbf{R}_j^\top w_j\| \, \|\mathbf{s}^c - \overline{\mathbf{s}}\|^2 \,] \\
&\leqslant K \, [\, \|\sum_{j=1}^{M} \overline{\mathbf{R}}_j\| \, \|\mathbf{s}^c - \overline{\mathbf{s}}\| + \|\mathbf{s}^c - \overline{\mathbf{s}}\|^2 \,]
\end{aligned}$$

with the constant

$$K = \alpha \max_{\mathbf{s} \in \mathcal{B}_\delta(\overline{\mathbf{s}})} \|[\sum_{j=1}^{M} \nabla\mathbf{R}_j^\top w_j \nabla\mathbf{R}_j]^{-1}\| \, [\, (C+c) \|\sum_{j=1}^{M} \overline{\mathbf{R}}_j\| + \frac{1}{2} \|\sum_{j=1}^{M} \nabla\mathbf{R}_j^\top w_j\| \,].$$

This concludes the proof. □

According to this result, the Gauss-Newton-type method (Definition 3.1) and the equivalent IRLS method perform with quadratic convergence rate in the zero-residual case. However, the assumptions restrict the choice of the weight function $w(x)$, and therefore of the norm–like function $N(x)$. More precisely, the weight function $w(x)$ must be positive and has to be bounded for $x \in [0, \epsilon]$.

The weights of the ℓ_p norms for $p \neq 2$, where $w(x) = px^{p-2}$ do not fulfill these conditions. First we consider the case $p > 2$. In a zero–residual case, all residuals vanish simultaneously which causes a singular Hessian. In the general situation, consider only the weights that vanish in the limit. This has the same effect as if the points with the associated weights would be excluded from the data set. Since we assume that $M \gg n$ this will not influence the regularity compared to the ℓ_2 case too much. More difficult is the case $p < 2$, where even a single vanishing residual destroys the regularity of the Hessian.

4.3 *Remarks on orthogonal distance regression*

In the case of orthogonal distance regression, where the parameters t_j obey (4), the residual vectors $\mathbf{R}_j$ take the form $\mathbf{R}_j = \mathbf{R}_j(\mathbf{s}, t_j(\mathbf{s}))$. Consequently we obtain the gradients

$$\nabla \mathbf{R}_j = \nabla_{\mathbf{s}} \mathbf{R}_j + \mathbf{R}_j' \nabla t \tag{26}$$

where $\nabla_{\mathbf{s}}$ denotes the gradient with respect to the first argument ($\mathbf{s}$), and the prime $'$ the derivative to the second one. The gradient of the parameter ∇t_j is found from the orthogonality conditions (5),

$$\nabla t_j = -\frac{\mathbf{R}_j^\top \nabla_{\mathbf{s}} \mathbf{R}_j' + \mathbf{R}_j'^\top \nabla_{\mathbf{s}} \mathbf{R}_j}{\mathbf{R}_j''^\top \mathbf{R}_j + \mathbf{R}_j'^\top \mathbf{R}_j'}. \tag{27}$$

Consequently, both the approximate Hessian H_F^* used in (18) and the equivalent IRLS method (20) need second derivatives of the curve.

Gauss-Newton-type methods for ℓ_p and orthogonal distance regression, in particular for $p \in \{1, \infty\}$, have been studied by Watson (1999, 2002); Atieg & Watson (2003); Al-Subaihi & Watson (2004). A survey has been given by Watson (2000)

The case of the ℓ_2 norm, i.e., $N(x) = x^2$ has been studied recently by Wang *et al.* (2006). In particular, several possible simplifications of (27) have been proposed.

As an alternative to omitting terms with second order derivatives from the Hessian, see (13), a Gauss-Newton method for ℓ_2 orthogonal distance regression can be derived by considering the linearization

$$\sum_{j=1}^{M} (\|\mathbf{R}_j\| + \nabla \|\mathbf{R}_j\| \Delta \mathbf{s})^2 \to \min_{\Delta \mathbf{s}}.$$

which can be shown to be equivalent to the method of "normal distance minimization" of Blake & Isard (1998), which is called "tangent distance minimization" by Wang *et al.* (2006). Due to the orthogonality (5), the gradient of the scalar residual evaluates to

$$\nabla \|\mathbf{R}_j\| = \frac{\mathbf{R}_j^\top}{\|\mathbf{R}_j\|} (\nabla_{\mathbf{s}} \mathbf{R}_j + \mathbf{R}_j' \nabla t_j) = \frac{\mathbf{R}_j^\top}{\|\mathbf{R}_j\|} (\nabla_{\mathbf{s}} \mathbf{R}_j), \tag{28}$$

i.e., it does not involve second derivatives of the curve. The generalization of this idea to the case of norm-like functions is subject of on–going research. However, it seems that this linearization is not so closely related to the IRLS method.

REMARK 4.2 As an inherent problem of orthogonal distance regression, it will often become ill-posed for parametric curves. For instance, consider a polynomial or rational curve (not a spline curve) that solves a fitting problem. A second solution can easily be obtained by reparameterizing the original one, hence the solution is not unique. Consequently, both the Hessian (9) and the simplified Hessian become singular in the limit and therefore regularization techniques have to be applied.

5. Examples

We present three numerical examples.

EXAMPLE 5.1 Fig. 2 shows a point cloud which was approximated with a polynomial curve of degree 5 that interpolates the first and the last data point. The three pictures in the left column show the sensitivity of approximations obtained using different norm–like functions $N(x)$ with respect to outliers. The first result was obtained with $N(x) = 1 - \exp(-(cx)^2)$ which shall serve as an approximation of the ℓ_1 norm. In literature, the weight derived from this function is called the Welsch weight (Welsch, 1977). The second result is the ℓ_2 fit, where $N(x) = x^2$. For the third result, the maximal distance from the curve to the data points was to be minimized. This corresponds to the ℓ_∞ norm. We replaced this non-differentiable norm by $N(x) = \exp((cx)^2) - 1$.

The plots in the right column visualize the lengths of the error vectors. From top to bottom the maximal error decreases, while the distribution of the error gets more uniform. In addition, Fig. 3 shows the norm-like functions and the associated weights.

REMARK 5.1 The results obtained by minimizing the objective functions obtained by choosing $N(x) = 1-\exp(-x^2)$ and $N(x) = \exp(x^2)-1$ are not invariant under scaling. The parameter c has been introduced in order to avoid this dependency. A possible strategy for choosing the parameter c is the following. First we compute the ℓ_2 fit of the data points and identify the point which has the largest distance to the curve. Then c is chosen such that this data point has some prescribed weight w_0. This does not require much additional computational effort, since the ℓ_2 approximation should be computed anyway, in order to obtain a suitable initial value for the iteration process.

EXAMPLE 5.2 Fig. 4 shows a point set that was obtained from the data of the previous example by adding some additional noise. Also the single outlier was eliminated. From top to bottom one can see again an approximate ℓ_1 fit, an ℓ_2 approximation and an approximate ℓ_∞ fit. Again, the magnitudes of the corresponding errors in the right column show that the maximal error decreases, while the distribution becomes more uniform. However, the difference between the approximate ℓ_1 fit and the ℓ_2 fit are not that significant, since the data points were perturbed with a uniform random error.

EXAMPLE 5.3 In order to demonstrate the quadratic convergence rate in the zero-residual case, we fitted an ellipse shaped point cloud, as one can see in Fig. 5. The model curve and the shape parameters were chosen according to Example 2.2. The errors in the first 7 steps are reported in the table. After 3 steps the new error is essentially the square root of the previous

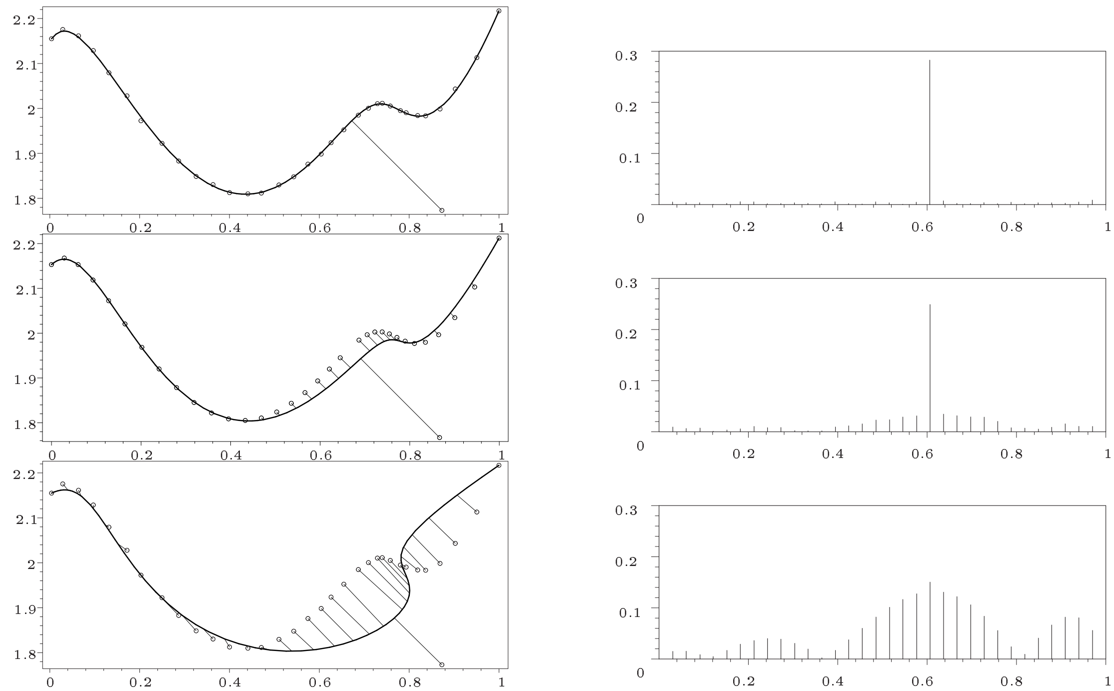


Fig. 2. Approximation of point data with different $N(x)$. Left column: points and approximating curves. Right column: error distribution.

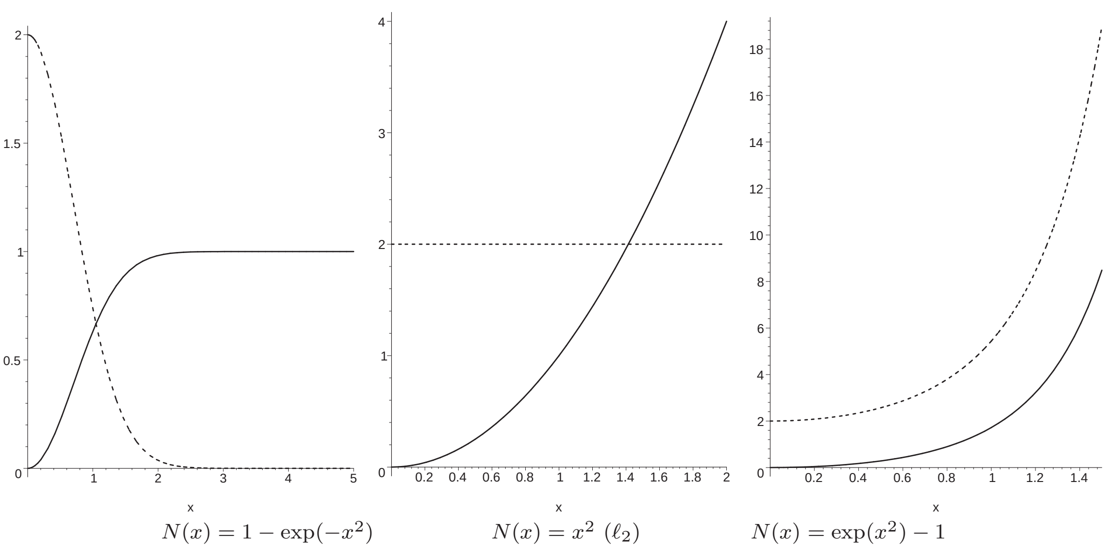


Fig. 3. Three norm-like functions (solid) and the associated weights (dashed).

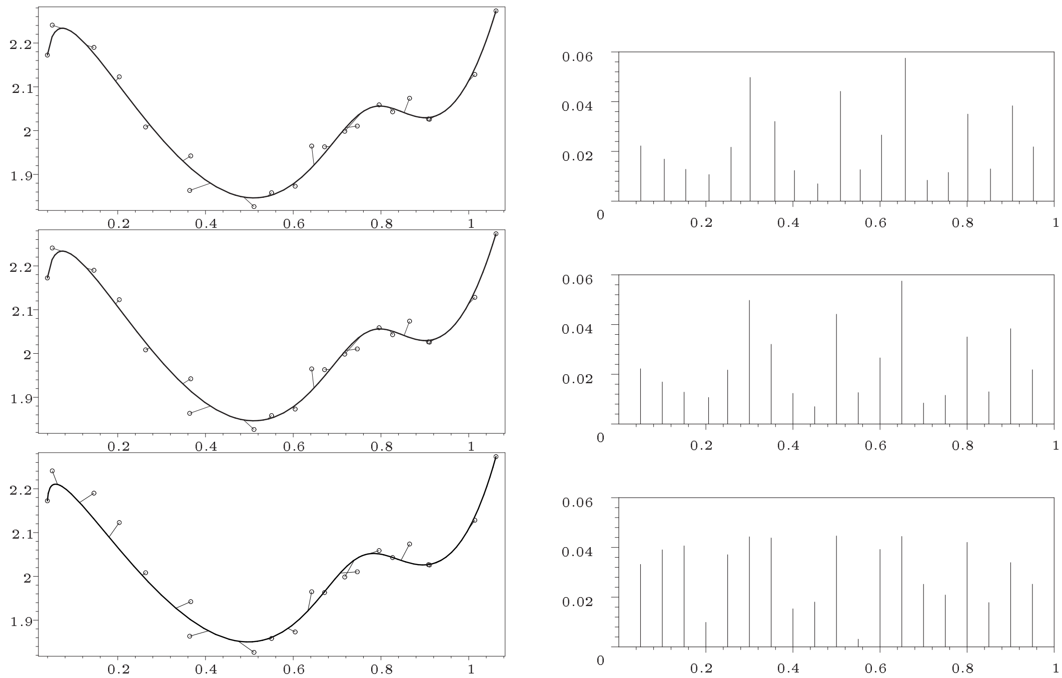
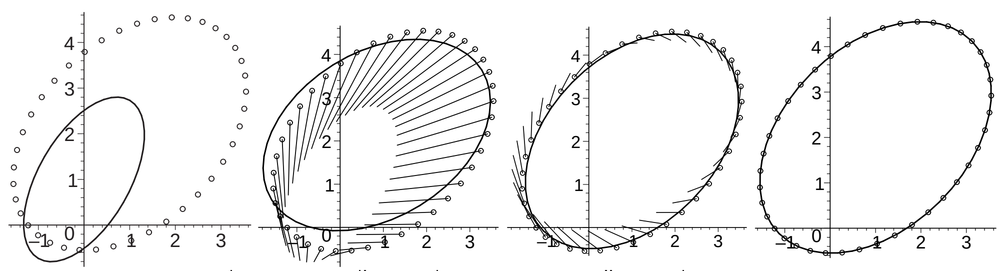


FIG. 4. Approximation of point data with different $N(x)$. Left column: fitted curves. Right column: distances from data to curve.

one, which indicates the quadratic convergence.



step	error	step	error	step	error
0	11.583450	3	0.0041590	6	0.9761834e-30
1	3.9724490	4	0.5437303e-6	7	0.1796254e-61
2	0.4580835	5	0.5572782e-14		

FIG. 5. Top: Zero-residual approximation with an ellipse; initial position and after 1, 2 and 4 steps and associated error vectors. Bottom: approximation errors.

6. Conclusion

We showed that the generalized fitting problem (6) for curves (and similarly for surfaces) can be solved iteratively by a Gauss-Newton-type method. This method has been shown to be equivalent to the technique of iteratively re-weighted least-squares, where the weights $w(x)$ and the corresponding norm-like function $N(x)$ are related by $N'(x) = x\,w(x)$. In addition, based on classical results about Gauss-Newton methods, we analyzed its convergence properties. By choosing other norm-like functions than $N(x) = x^2$, which corresponds to the ℓ_2 norm, one may efficiently deal with data containing outliers, and one may find curves and surfaces that minimize the maximum distance error.

References

Aigner, M., Jüttler, B. (2007) Approximation Flows in Shape Manifold. In: P. Chenin, T. Lyche & L. Schumaker, eds., Curve and Surface Design: Avignon 2006, Nashboro Press, 1–10.

Aigner, M., Šír, Z., Jüttler, B. (2007) Evolution-based least-squares fitting using Pythagorean hodograph spline curves. *Comp. Aided Geom. Design*, **24**, 310–322.

Alhanaty, M., Bercovier, M. (2001) Curve and surface fitting and design by optimal control methods, Comp. Aided Design **33**, 167–182.

Al-Subaihi, I., Watson, G. A. (2004) The use of the l_1 and l_∞ norms in fitting parametric curves and surfaces to data, *Appl. Numer. Anal. Comput. Math.* **1**, 363–376.

Atieg, A., Watson, G.A. (2003) A class of methods for fitting a curve or surface to data by minimizing the sum of squares of orthogonal distances. *J. Comput. Appl. Math.*, **158**(2), 277–296.

Atieg, A., Watson, G.A. (2004) Use of ℓ_p norms in fitting curves and surfaces to data. *Australian and New Zealand Industrial and Applied Mathematics Journal,* **45 (E)**, C187–C200.

Blake, A., Isard, M. (1998) *Active contours*. Springer, New York.

Hermey, D., Watson, G. (1999) Fitting data with errors in all variables using the Huber M-estimator. *SIAM J. Sci. Comput.* **20**, 1276–1298.

Holland, P., Welsch, R. (1977) Robust regression using iteratively re-weighted least-squares. *Communicatios in Statistics: Theory and Methods,* , **A6**(9), 813–827.

Hoschek, J., Lasser, D. (1993) *Fundamentals of computer aided geometric design*. A K Peters, Wellesley, MA.

Huber, P. J. (1981) *Robust Statistics*. John Wiley and Sons, New York.

Jüttler, B. (1998) Computational methods for parametric discrete ℓ_1 and ℓ_∞ curve fitting, *Int. J. Shape Modelling*, **4**, 21–34.

Kelley, C. T. (1999) *Iterative Methods for Optimization*. Society for Industrial and Applied Mathematics, Philadelphia, PA, USA.

Mahadevan,V., Narasimha-Iyer, H., Roysam, B., Tanenbaum, H. L. (2004) Robust model-based vasculature detection in noisy biomedical images. *IEEE Transactions on Information Technology in Biomedicine*, **8**(3), 360–376.

O'Leary, D. P. (1990) Robust regression computation using iteratively re-weighted least squares. *SIAM J. Matrix Anal. Appl.*, **11**(3), 466–480.

Pottmann, H., Leopoldseder, S. (2003) A concept for parametric surface fitting which avoids the parametrization problem. *Comp. Aided Geom. Design*, **20**(6), 343–362.

Pottmann, H., Leopoldseder, S., Hofer, M., Steiner, T. (2005) Industrial geometry: recent advances and applications in CAD. *Comp. Aided Design*, **37**, 751–766.

Rogers, D., Fog, N. (1989) Constrained B-spline curve and surface fitting. *Comp. Aided Design*, **21**, 641–648.

Sarkar, B., Menq, C.-H. (1991) Parameter optimization in approximating curves and surfaces to measurement data. *Comp. Aided Geom. Design*, **8**, 267–280.

Speer, T., Kuppe, M., Hoschek, J. (1998) Global reparametrization for curve approximation. Comp. Aided Geom. Design, **15**, 869–877.

Varady, T., Martin, R. (2002) Reverse Engineering. In: Farin, G., Hoschek, J. & Kim, M.-S.

(eds.), *Handbook of Computer Aided Geometric Design*, Elsevier, 651–681.

WANG, W., POTTMANN, H., LIU, Y. (2005) Fitting B-spline curves to point clouds by curvature-based squared distance minimization. *ACM Transactions on Graphics*, **25**(2), 214–238.

WATSON, G. A. (1977) On two methods for discrete ℓ_p approximation. *Computing*, **18**(3), 263–266.

WATSON, G. A. (2000) Least squares fitting of parametric surfaces to measured data. *ANZIAM J.*, **42(E)** C68–C95.

WATSON, G. A. (2000) Approximation in normed linear spaces. *J. Comput. Appl. Math.*, **121**, 1–36.

WATSON, G. A. (2002) On the Gauss-Newton method for ℓ_1 orthogonal distance regression. *IMA J. Num. Anal.*, **22**, 345–357.

YAMAMOTO, T. (2000) Historical developments in convergence analysis for Newton's and Newton-like methods. *J. Comput. Appl. Math.*, **124**(1-2), 1–23.