A relative weighting method for estimating parameters and variances in multiple data sets

Bradley M. Bell [a,*], James V. Burke [b], Alan Schumitzky [c]

[a] *Applied Physics Laboratory, University of Washington, 1013 N.E. 40th Street, Seattle, WA 98195, USA*
[b] *Department of Mathematics, University of Washington, Seattle, WA, USA*
[c] *Department of Mathematics, University of Southern California, Los Angeles, CA, USA*

Abstract

We are given multiple data sets and a nonlinear model function for each data value. Each data value is the sum of its error and its model function evaluated at an unknown parameter vector. The data errors are mean zero, finite variance, independent, are not necessarily normal, and are identically distributed within each data set. We consider the problem of estimating the data variance as well as the parameter vector via an extended least-squares technique motivated by maximum likelihood estimation. We prove convergence of an algorithm that generalizes a standard successive approximation algorithm from nonlinear programming. This generalization reduces the estimation problem to a sequence of linear least-squares problems. It is shown that the parameter and variance estimators converge to their true values as the number of data values goes to infinity. Moreover, if the constraints are not active, the parameter estimates converge in distribution. This convergence does not depend on the data errors being normally distributed.

Keywords: Nonlinear least squares; Asymptotic statistics

1. Introduction

The problem presented in this paper is motivated by parameter estimation in drug and tracer kinetics studies. Often more than one type of measurements is available and the different types of measurements have different errors. One case of different

---

* Corresponding author.

types of measurements is when concentrations of a drug and its metabolite are both measured. Another case is when concentrations of a drug can be measured in different biological compartments such as plasma and spinal fluid. A third case is when concentrations of a tracer can be measured in one compartment, and cumulative amounts of the tracer can be measured in another compartment.

In all these cases, the different data sets have different degrees of precision relative to each other. The resulting estimation problems require not only estimation of the structural parameters in the underlying model but also the relative accuracy of the various data sets. In Section 2, we define an example model and provide simulated data to illustrate this type of problem. The model has two compartments and is used to describe the kinetics when concentrations of a tracer are measured in the plasma and cumulative amounts of the tracer are measured in the urine. The simulated measurement sets have variance parameters which differ by two orders of magnitude.

We now describe our general problem in mathematical terms. We are given multiple data sets and we denote the jth data value in the ith data set by y_i^j. For each data value we are also given a function $F_i^j : \mathbb{R}^n \to \mathbb{R}$. In our model for the data, there is an unknown parameter vector $\bar{x}$ such that

$$y_i^j = F_i^j(\bar{x}) + e_i^j, \tag{1}$$

where $\{e_i^j\}$ is a doubly indexed sequence of mean zero, finite variance, independent random variables. We further assume that e_i^j and e_i^k are identically distributed; i.e., that the errors within one data set are identically distributed. (Note that the data values y_i^j and y_i^k are not necessarily identically distributed.) Our problem is to estimate $\bar{x}$ subject to limits for each of its components. These may either be limits within which the model is valid or limits within which $\bar{x}$ is known to reside.

Define $L(x, v)$ to be the objective function which corresponds to the negative log likelihood when the data errors are normally distributed:

$$L(x,v) = \sum_{i=1}^{M} \left\{ N(i) \log(2\pi v_i) + \frac{1}{v_i} \sum_{j=1}^{N(i)} [y_i^j - F_i^j(x)]^2 \right\}. \tag{2}$$

Here v denotes the vector $(v_1, v_2, \ldots, v_m)^T$, where the ith component v_i corresponds to the nominal variance for the ith data set. Given the data values $\{y_i^j\}$, our estimates for the true value of the vector of parameters $\bar{x}$ and the vector of variances $\bar{v}$ are obtained as the solution to the problem

$$\begin{aligned} &\text{minimize } L(x,v)\,; \\ &\text{subject to } a \le x \le b\,, \ 0 < v. \end{aligned} \tag{3}$$

In Section 2, we define the example model in which Eq. (1) is satisfied. In Section 3, we derive a closed form expression for the argument that minimizes $L(x,v)$ with respect to v. Substituting this expression into L yields a reduced objective function $R(x)$. The reduced objective is then approximated at each point by a weighted least-squares objective $W(x,d)$. In Section 4, we present an algorithm for the minimization of $R(x)$ that is based on the approximation $W(x,d)$. In Section 5, we derive asymptotic statistics for the estimator that minimizes Problem (3). In Section 6, we

Table 1
Notation

	Problem input
M	number of data sets
n	number of components in the parameter vector
$N(i)$	number of data points in the ith data set
$F_i^j(x)$	model function for the jth data point in the ith data set
y_i^j	data value for the jth data point in the ith data set
a	vector of lower limits for the components of the parameter vector
b	vector of upper limits for the components of the parameter vector
	Other notation
e_i^j	data error for the jth data point in the ith data set
v_i	a nominal variance value for the ith data set
$\bar{v}_i$	true value for the variance in the ith data set
x	a nominal parameter vector
$\bar{x}$	true value for the parameter vector

present the results of applying the relative weighting method to the example defined in Section 2.

Just as with Gauss–Newton methods, we do not require second-order information on the model functions F_i^j. The asymptotics indicate that this does not inhibit rapid local convergence for sufficiently large data sets. Indeed, it is shown that the second-order terms vanish as the number of data values becomes infinite (see the remark below Lemma 9).

A word about our notation is in order. The components of a vector z or vector-valued function $G(z)$ are denoted by z_i and $G_i(z)$, respectively. Moreover, since the discussion of multiple data sets requires multiply indexed variables, we provide the definitions in Table 1 as an aid to the reader.

2. Example

The model defined in Section 1 can be motivated by estimation problems in compartmental modeling. In this section we present an example of a two-compartment model consisting of a plasma compartment and a urine compartment. The definitions in this section apply here and in Section 6. They do not apply to the rest of the paper.

A known amount of tracer, denoted by p_0, is injected into the plasma compartment. Over time, the tracer transfers from the plasma compartment to the urine compartment and the outside world. The unknown parameter vector x has three components (thus n in Table 1 is 3). Its first component, x_1, is the logarithm of the volume of the plasma. Its second component, x_2, is the transfer rate from the plasma to the urine. Its third component, x_3, is the transfer rate from the plasma to the outside world. We
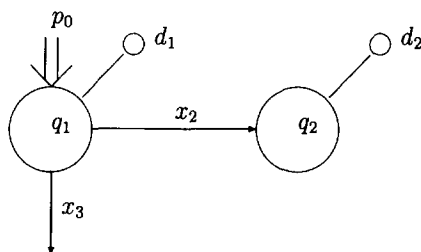
Fig. 1. Compartmental model.

are given two measurement sets (thus M in Table 1 is 2). The first set, $\{d_1^1, \ldots, d_1^{20}\}$, consists of 20 measurements of the tracer concentration in the plasma (thus $N(1)$ in Table 1 is 20). The second set, $\{d_2^1, \ldots, d_2^{10}\}$, consists of 10 measurements of the amount of tracer added to the urine compartment between measurements (thus $N(2)$ in Table 1 is 10). In Fig. 1, compartment 1 is the plasma, compartment 2 is the urine.

Let $q_1(t)$ and $q_2(t)$ denote the amount of tracer in compartments 1 and 2 as a function of time. It follows that

$$q_1(t) = p_0 \exp(-(x_2 + x_3)t),$$

$$q_2(t) = \frac{p_0 x_2}{x_2 + x_3}(1 - \exp(-(x_2 + x_3)t)).$$

The concentration of tracer in compartment 1 at time t is $q_1(t)/\exp(x_1)$. The difference between the amount of tracer in compartment 2 at time t and time s is $q_2(t) - q_2(s)$.

We assume that, within each data set, the errors in the logarithm of the data are identically distributed. The concentration of tracer in the entire plasma compartment is measured using a small sample. On the other hand, we measure the amount of tracer that has collected in urine between sample times. Since these two measurements are fundamentally different, we expect them to have different errors. We are given the set of times $\{t_i^j\}$ at which the measurements $\{d_i^j\}$ are made. If $\bar{x}$ is the true (but unknown) value of the model parameter vector, the logarithm of the data satisfies the following equations:

$$y_1^j = \log(d_1^j) = e_1^j + \log(p_0) - (\bar{x}_2 + \bar{x}_3)t_1^j - \bar{x}_1,$$

$$y_2^j = \log(d_2^j) = e_2^j + \log(\frac{p_0 \bar{x}_2}{\bar{x}_2 + \bar{x}_3})$$
$$+ \log[\exp(-(\bar{x}_2 + \bar{x}_3)t_2^{j-1}) - \exp(-(\bar{x}_2 + \bar{x}_3)t_2^j)],$$

where $\{e_1^j\}$ and $\{e_2^j\}$ are independent samples from distributions with zero means and unknown variances v_1 and v_2 respectively. The value t_2^{j-1}, in the equation above, is interpreted as zero when $j = 1$. For this example, we fit the model functions $F_1(x)$ and $F_2(x)$ to the logarithm of the data where

$$F_1^j(x) = \log(p_0) - (x_2 + x_3)t_1^j - x_1, \tag{4}$$

$$F_2^j(x) = \log(\frac{p_0 x_2}{x_2 + x_3}) + \log[\exp(-(x_2 + x_3)t_2^{j-1}) - \exp(-(x_2 + x_3)t_2^j)]. \tag{5}$$

Table 2
Simulated data

t_1	y_1	t_1	y_1	t_2	y_2
0.10000	−0.19186	1.10000	−1.06230	0.20000	−4.61015
0.20000	−0.21853	1.20000	−1.29741	0.40000	−2.01673
0.30000	−0.33233	1.30000	−1.33147	0.60000	−3.57869
0.40000	−0.65443	1.40000	−1.32441	0.80000	−4.20353
0.50000	−0.68928	1.50000	−1.60137	1.00000	−3.12661
0.60000	−0.57532	1.60000	−1.77217	1.20000	−1.89698
0.70000	−0.64656	1.70000	−1.86234	1.40000	−2.51455
0.80000	−0.99292	1.80000	−1.98450	1.60000	−2.58503
0.90000	−1.03121	1.90000	−2.13367	1.80000	−5.86347
1.00000	−1.13131	2.00000	−2.05235	2.00000	−3.37648

The true parameter values in this simulation are

$$\bar{x}_1 = 0.1 \ , \ \bar{x}_2 = 0.4 \ , \ \bar{x}_3 = 0.6 \ , \ \bar{v}_1 = .01 \ , \ \bar{v}_2 = 1.0. \tag{6}$$

The initial amount of tracer in compartment 1, p_0, is 1 for this simulation. Table 2 contains simulation values for the log of the data.

The relative weighting method is applied to this example and the corresponding results are presented in Section 6.

3. Reduced objective function

In this section we obtain a closed form expression for the vector $V(x)$ satisfying

$$L[x, V(x)] \leq L(x, v) \text{ for all } v > 0.$$

This enables us to replace Problem (3) by a reduced problem which no longer depends on the variable v.

Define the function $V : \mathbb{R}^n \to \mathbb{R}^M$ to be the sample variance that corresponds to the parameter value x; i.e.,

$$V_i(x) = \frac{1}{N(i)} \sum_{j=1}^{N(i)} [y_i^j - F_i^j(x)]^2. \tag{7}$$

Lemma 1 (Reduced objective). *Let $L : \mathbb{R}^n \to \mathbb{R}^M$ be as defined in Eq. (2) and let $x \in \mathbb{R}^n$. If $V(x) > 0$, then $L(x, V(x)) \leq L(x, v)$ for all $v > 0$.*

Proof. The partial of L with respect to v_i is given by

$$\frac{\partial L}{\partial v_i} = \frac{N(i)}{v_i^2} [v_i - V_i(x)].$$

It follows that this partial is zero if $v_i = V_i(x)$. This is the first-order necessary condition for a minimum. The second partial of L with respect to v_i is given by

$$\frac{\partial^2 L}{\partial v_i^2} = \frac{2N(i)}{v_i^3}\left[V_i(x) - \frac{v_i}{2}\right].$$

It follows that this second partial is positive for $v_i/2 < V_i(x)$. Because this is true for each i, the function L increases as any v_i converges to 0. Moreover, it follows directly from Eq. (2) that $L \rightarrow +\infty$ as $\|v\| \rightarrow +\infty$. Thus, $v = V(x)$ attains the global minimum value of L with respect to v over the domain $v > 0$. $\square$

Substituting $V(x)$ for v in (2) we obtain

$$L[x, V(x)] = \sum_{i=1}^{M} N(i)\{\log(2\pi) + \log[V_i(x)] + 1\}.$$

This expression can be simplified by dropping the terms that are constant in x. We also divide by the total number of data points

$$J = \sum_{i=1}^{M} N(i) \tag{8}$$

to aid in the discussion of the asymptotic statistics in Section 4. Thus, the reduced objective function is given by

$$R(x) = J^{-1} \sum_{i=1}^{M} N(i) \log[V_i(x)]. \tag{9}$$

It follows that a pair $(\hat{x}, \hat{v})$ solves Problem (3) if and only if $\hat{v} = V(\hat{x})$ and $\hat{x}$ solves the problem

$$\begin{aligned}&\text{minimize } R(x);\\&\text{subject to } a \leq x \leq b.\end{aligned} \tag{10}$$

Our algorithm uses a weighted least-squares approximation that is related to the gradient of R, where

$$\nabla R(x) = J^{-1} \sum_{i=1}^{M} N(i) V_i(x)^{-1} \nabla V_i(x)$$

$$= -2J^{-1} \sum_{i=1}^{M} V_i(x)^{-1} \sum_{j=1}^{N(i)} [y_i^j - F_i^j(x)] \nabla F_i^j(x). \tag{11}$$

The weighted least-squares approximation for R at the point x in the direction d is defined by

$$W(x, d) = J^{-1} \sum_{i=1}^{M} V_i(x)^{-1} \sum_{j=1}^{N(i)} [y_i^j - F_i^j(x) - \nabla F_i^j(x)^{\mathrm{T}} d]^2. \tag{12}$$

The value $W(x, d) - W(x, 0)$ is a first-order approximation to $R(x + d) - R(x)$ because

$$\nabla_d W(x, 0) = \nabla R(x).$$

4. The algorithm

In this section, we modify the algorithm given by Burke (1985) to suit our approximation scheme. For this purpose it is convenient to describe the algorithm in a somewhat more general setting. Let $G : \mathbb{R}^n \to \mathbb{R}^M \cup \{-\infty\}$ be such that $\mathrm{Dom}(G) = \{x \in \mathbb{R}^n : G(x) > -\infty\}$ is open and G is of class C^2 on $\mathrm{Dom}(G)$. Moreover, we assume that we can associate with G an approximation function $A : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}$, where for each $x \in \mathrm{Dom}(G)$ we have

   (i) $A(x, d)$ is convex in d, and

   (ii) $\nabla_d A$ exists and is continuous on $\mathrm{Dom}(G) \times \mathbb{R}^n$ with $\nabla_d A(x, 0) = \nabla G(x)$.

$$(13)$$

We describe an algorithm based on the approximation function A that is designed to solve the problem

$$\begin{array}{ll} \text{minimize} & G(x); \\ \text{subject to} & x \in C, \end{array} \qquad (14)$$

where $C \subset \mathbb{R}^n$ is a nonempty compact convex set. This setup differs from that considered by Burke (1985) in two key ways: (1) the function G is allowed to take the value $-\infty$ and (2) the approximation function need not satisfy the equation $A(x, 0) = G(x)$.

For the special case of Problem (10), the function $G(x)$ is the reduced likelihood function $R(x)$, $A(x, d)$ is the weighted least-squares approximation $W(x, d)$, and C is the set of x such that $a \leq x \leq b$. If $\{F_i^j(x)\}$ are twice continuously differentiable functions, R and W satisfy the hypotheses for G and A.

Let $x^0 \in \mathbb{R}^n$ and $\gamma, \mu \in (0, 1)$ be given. Having x^k, we obtain x^{k+1} as follows:

Step 1: If $G(x^k) = -\infty$, then stop.

Step 2: Let d^k be such that $d = d^k$ attains the minimum value of $A(x^k, d)$ over all choices of $d \in \mathbb{R}^n$ satisfying $x^k + d \in C$.

Step 3: If $A(x^k, d^k) = A(x^k, 0)$, then stop; otherwise, set $x^{k+1} = x^k + \lambda_k d^k$, where λ_k is the largest value of γ^s for $s = 0, 1, \dots$ satisfying

$$G(x^k + \gamma^s d^k) - G(x^k) \leq \mu \gamma^s [A(x^k, d^k) - A(x^k, 0)]. \qquad (15)$$

In Step 3, the evaluation of the step-length λ_k can be implemented by trial and error. That is, one first tests Inequality (15) with $s = 0$, if successful, then $\lambda_k = 1$; otherwise, increment s by 1 and try again. If this trial and error procedure terminates, then it clearly yields the largest value of γ^s for $s = 0, 1, \dots$ satisfying Inequality (15). In particular, this procedure is successfully terminated if $G(x^k + \gamma^s d^k) = -\infty$ for

any trial value $x^k + \gamma^s d^k$. The following lemma demonstrates that the procedure is always finitely terminating.

Lemma 2. *If* $A(x^k, d^k) \neq A(x^k, 0)$, *there exists* $s \in \{0, 1, \ldots\}$ *such that*

$$G(x^k + \gamma^s d^k) - G(x^k) \leq \mu \gamma^s [A(x^k, d^k) - A(x^k, 0)].$$

Proof. Since $A(x^k, d^k) \neq A(x^k, 0)$, the definition of d^k implies that $A(x^k, 0) - A(x^k, d^k) > 0$. Since $\gamma < 1$ and $\mu < 1$, it follows that there is a positive integer s such that

$$\gamma^{-s}[G(x^k + \gamma^s d^k) - G(x^k)] - \nabla G(x^k)^{\mathrm{T}} d^k < (1 - \mu)[A(x^k, 0) - A(x^k, d^k)]$$

or, equivalently,

$$\gamma^{-s}[G(x^k + \gamma^s d^k) - G(x^k)] < \nabla_d A(x^k, 0)^{\mathrm{T}} d^k + (1 - \mu)[A(x^k, 0) - A(x^k, d^k)].$$

Using the fact that $A(x, d)$ is convex in d, we have

$$\gamma^{-s}[G(x^k + \gamma^s d^k) - G(x^k)] < A(x^k, d^k) - A(x^k, 0) + (1 - \mu)[A(x^k, 0) - A(x^k, d^k)]$$

or, equivalently,

$$G(x^k + \gamma^s d^k) - G(x^k) < \mu \gamma^s [A(x^k, d^k) - A(x^k, 0)].$$

This completes the proof of this lemma. If $A(x^k, d^k) = A(x^k, 0)$, $A(x^k, 0)$ is the minimum of $A(x^k, d)$ over d such that $x^k + d \in C$.

The condition

$$A(x, 0) = \min\{A(x, d) : x + d \in C\} \tag{16}$$

is closely related to first-order optimality conditions for Problem (14). These conditions are typically stated with the aid of the normal cone operator from convex analysis (Rockafellar, 1970). Recall that the normal cone of C at x is defined as

$$N_C(x) = \{z : z^{\mathrm{T}}(x' - x) \leq 0 \text{ for all } x' \in C\}.$$

Moreover, if $x \in \mathrm{Dom}(G)$ is a local solution to Problem (14), then

$$0 \in \nabla G(x) + N_C(x) \text{ or equivalently, } 0 \in \nabla_d A(x, 0) + N_C(x) \tag{17}$$

(Clarke, 1983). Since $A(x, d)$ is convex in d, condition (17) is both necessary and sufficient for $d = 0$ to yield the global minimum value of $A(x, d)$ over d such that $x + d \in C$. Therefore, the conditions (17) and (16) are equivalent.

In our analysis of the convergence properties of the algorithm, we make strong use of some elementary properties of the normal cone operator. In particular, we use the fact that if $C \subset \mathbb{R}^n$ is a nonempty closed convex set, then the set

$$\mathrm{Graph}(N_C) = \{(x, y) \in \mathbb{R}^n \times \mathbb{R}^n : y \in N_C(x)\}$$

is closed, i.e., the multi-valued function $N_C : C \to \mathbb{R}^n$ is *upper semi-continuous* (Rockafellar, 1970, Theorem 24.4).

Theorem 3. *If the algorithm stops at iteration k, either $G(x^k) = -\infty$ or $0 \in \nabla G(x^k) + N_C(x^k)$. If $\hat{x}$ is a cluster point of the sequence $\{x^k\}$ generated by the algorithm, then either $G(\hat{x}) = -\infty$ or $0 \in \nabla G(\hat{x}) + N_C(\hat{x})$.*

Proof. The conclusion in the first sentence of the theorem follows directly from our discussion of optimality conditions preceding the statement of the theorem. Next suppose that the sequence $\{x^k\}$ is infinite and let $\hat{x}$ be a cluster point of the sequence for which $G(\hat{x}) > -\infty$. We complete the proof by showing that $0 \in \nabla G(\hat{x}) + N_C(\hat{x})$ or, equivalently, that $A(\hat{x}, d) - A(\hat{x}, 0) \geq 0$ for all d such that $\hat{x} + d \in C$.

Since the sequence $G(x^k)$ is monotone decreasing, it has a limit $G(\hat{x}) > -\infty$. It follows that $G(x^{k+1}) - G(x^k) \to 0$. Let I be an infinite subset of the positive integers such that $x^k \xrightarrow{I} \hat{x}$. The sequence $\{d^k\}$ is bounded because C is bounded so there is a subsequence $J \subset I$ and a $\hat{d}$ such that $d^k \xrightarrow{J} \hat{d}$. Observe that $\hat{d}$ minimizes $A(\hat{x}, d)$ over d such that $\hat{x} + d \in C$. Indeed, by Step 2 of the algorithm, we have

$$0 \in \nabla_d A(x^k, d^k) + N_C(x^k + d^k)$$

for all k. Taking the limit over $k \in J$ and using the fact that $\nabla_d A$ is continuous on $\mathrm{Dom}(G) \times \mathbb{R}^n$ and the normal cone is upper semi-continuous, we have

$$0 \in \nabla_d A(\hat{x}, \hat{d}) + N_C(\hat{x} + \hat{d}).$$

Since $A(x, d)$ is convex in d, this implies that $\hat{d}$ minimizes $A(\hat{x}, d)$ over d such that $\hat{x} + d \in C$.

We can complete the proof by showing that $A(\hat{x}, \hat{d}) - A(\hat{x}, 0) \geq 0$. We split this demonstration into two cases.

Case 1: Suppose that there is a $\beta > 0$ and an infinite subset of positive integers $K \subset J$ such that $\lambda^k > \beta$ for all $k \in K$. It follows from Step 3 that

$$G(x^{k+1}) - G(x^k) \leq \mu \lambda^k [A(x^k, d^k) - A(x^k, 0)].$$

Taking the limit over $k \in K$, and noting that $\lambda^k > \beta$, we obtain

$$0 \leq \mu\beta[A(\hat{x}, \hat{d}) - A(\hat{x}, 0)],$$

which completes the proof for Case 1.

Case 2: Suppose that $\lambda^k \xrightarrow{J} 0$. Since $x^k \xrightarrow{J} \hat{x}, x^{k+1} = x^k + \lambda^k d^k$, and the sequence $\{d^k\}$ is bounded (because C is bounded), we may assume with no loss in generality that there exists an $\varepsilon > 0$ such that

$$x^k, x^{k+1} \in \{x : \|\hat{x} - x\| \leq \varepsilon\} \subset \mathrm{Dom}(G) \quad \text{for all } k \in J.$$

By Step 3 of the algorithm, we may also assume that $\lambda^k < 1$ for all $k \in J$. Let α be a Lipschitz constant for $\nabla G(x)$ over $\{x : \|\hat{x} - x\| \leq \varepsilon\}$. By the convexity of $A(x, d)$

in d and Step 3 of the algorithm, we have

$$
\begin{aligned}
\mu\gamma^{-1}\lambda^k \nabla_d A(x^k,0)^{\mathrm{T}} d^k &\leq \mu\gamma^{-1}\lambda^k [A(x^k,d^k) - A(x^k,0)] \\
&< G(x^k + \gamma^{-1}\lambda^k d^k) - G(x^k) \\
&\leq \nabla G(x^k)^{\mathrm{T}}(\gamma^{-1}\lambda^k d^k) + \alpha|\gamma^{-1}\lambda^k d^k|^2.
\end{aligned}
$$

Dividing by $\gamma^{-1}\lambda^k$ and replacing $\nabla G(x^k)$ by $\nabla_d A(x^k,0)$ yields

$$
\nabla_d A(x^k,0)^{\mathrm{T}} d^k + \alpha\gamma^{-1}\lambda^k|d^k|^2 \geq \mu \nabla_d A(x^k,0)^{\mathrm{T}} d^k
$$

or, equivalently,

$$
-\alpha\gamma^{-1}\lambda^k|d^k|^2 \leq (1-\mu)\nabla_d A(x^k,0)^{\mathrm{T}} d^k.
$$

By taking the limit over $k \in J$, we conclude that

$$
0 \leq \nabla_d A(\hat{x},0)^{\mathrm{T}} \hat{d}.
$$

Again, since $A(x,d)$ is convex in d, we obtain

$$
A(\hat{x},\hat{d}) - A(\hat{x},0) \geq \nabla_d A(\hat{x},0)^{\mathrm{T}} \hat{d} \geq 0,
$$

which completes the proof for Case 2. □

5. Asymptotic statistics

In this section, we consider the statistics of our estimators for $\bar{x}$ and $\bar{v}$ as the number of data points increases. Throughout this section, C is the constraint set in Problem (10); i.e., $C = \{x : a_i \leq x_i \leq b_i \text{ for } i = 1,\ldots,n\}$. We add the extra index k to our notation of the previous sections. We are given an infinite rectangular sequence of integers,

$$
\{N(i,k) : i = 1,\ldots,M, \ k = 1,\ldots,\infty\}.
$$

Define $J(k)$, $V^k : \mathbb{R}^n \to \mathbb{R}^M$, and $R^k : \mathbb{R}^n \to \mathbb{R}$ by Eqs. (7), (8), and (9) in which the integer vector $N(i)$ is replaced by $N(i,k)$:

$$
J(k) = \sum_{i=1}^{M} N(i,k), \tag{18}
$$

$$
V_i^k(x) = \frac{1}{N(i,k)} \sum_{j=1}^{N(i,k)} [y_i^j - F_i^j(x)]^2, \tag{19}
$$

$$
R^k(x) = J(k)^{-1} \sum_{i=1}^{M} N(i,k) \log[V_i^k(x)]. \tag{20}
$$

In this notation, we determine the asymptotic statistics of the minimizer of $R^k(x)$ as $k \to \infty$ (consistency is proved in Theorem 8 and asymptotic normality is proved in Theorem 13). In proving these assertions, there are two complicating factors. First,

the random variables y_i^j are independent but not necessarily identically distributed (inid). Second, the random variables y_i^j do not form a "sequence" but instead form a "rectangular array".

There are a number of approaches to the inid case. We mention the general but essentially different methods of Gallant (1987, Ch. 3), Caines (1988, Ch. 8), and Ibragimov and Has'minskii (1981, Ch. 3). The approach of Ibragimov and Has'minskii also handles the rectangular array situation. However the generality of these three methods makes it harder to verify the required hypotheses than to give a direct proof. Further, the proof given below is more closely adapted to the numerical algorithm used in Section 3. Our approach closely follows the logic of Jennrich (1969), but now applied to the multiple data set case.

The next assumption is a regularity condition on the rate at which the individual sample sizes tend to infinity. It is used by Lehman (1983, Section 6.6) in proving a similar result for the independent and identically distributed case.

Assumption 4. *Let λ_i^k denote $J(k)^{-1}N(i,k)$. We assume that for each i, $N(i,k) \to \infty$ as $k \to \infty$. Furthermore there is a $\lambda \in \mathbb{R}^M$ such that $\lambda_i^k \to \lambda_i$ as $k \to \infty$.*

Given a sequence $\{w^j\}$ of real matrices, define the sequence of tail averages $\{\langle w^j \rangle^p\}$ by

$$\langle w^j \rangle^p = \frac{1}{p} \sum_{j=1}^{p} w^j. \tag{21}$$

Only the superscript j is used to indicate the tail averaging index in this paper. If $\langle w^j \rangle^p$ converges as $p \to \infty$, we denote its limit by $\langle w^j \rangle$. (See the notation in Table 3.)

We now introduce two assumptions. The first is an independence assumption. The second is an assumption about our sampling strategy as $k \to \infty$. (Theorem 2 of Jennrich (1969) is one possible motivation for the second assumption.)

Assumption 5. *The residuals e_i^j and e_p^k are independent for $j \neq k$ or $i \neq p$. Furthermore the sequence $\{e_i^j : j = 1, \ldots, \infty\}$ is identically distributed with mean zero and variance $\bar{v}_i$.*

Table 3
Additional notation

k	experiment index, as k increases the data set increases
C	the set that the minimization is constrained to
$N(i,k)$	number of data points in the ith data set for experiment k
$J(k)$	total number of data points in experiment k
λ_i^k	the ratio of $N(i,k)/J(k)$
$\langle w^j \rangle^p$	the average of the first p elements of $\{w^j\}$

Assumption 6. *The functions $\{F_i^j\}$ are twice continuously differentiable and all tail averages of the form $\langle h(x)g(x')^{\mathsf{T}}\rangle^p$, where $h, g = F_i^j, \nabla F_i^j, \nabla^2 F_i^j$, converge uniformly over $x \in C, x' \in C$ as $p \to \infty$. Furthermore, if $\langle [F^j(\bar{x}) - F^j(x)]^2 \rangle$ is zero then $x = \bar{x}$.*

Remark. To motivate the assumptions on the tail averages $\langle F^j(x)^2 \rangle$ and $\langle [F^j(\bar{x}) - F^j(x)]^2 \rangle$, we consider a simple linear scalar example. Let $M = 1$ and $F_1^j(x) = t^j x$, where $\{t^j\}$ is a given sequence of real numbers. In this case, the minimizer x^k of $R^k(x)$ also minimizes $\sum_{j=1}^{K}(y_1^j - t^j x)^2$, where $y_1^j = F_1^j(\bar{x}) + e_1^j$. It follows that

$$x^k = \bar{x} + \langle t^j e_1^j \rangle^k / \langle (t^j)^2 \rangle^k. \tag{22}$$

Now $\langle [F^j(\bar{x}) - F^j(x)]^2 \rangle^k = \langle (t^j)^2 \rangle^k (x - \bar{x})^2$, so that the last requirement in Assumption 6 is equivalent to the convergence of the tail average $\langle (t^j)^2 \rangle^k$ to a nonzero number. It follows from (22) that this latter condition is sufficient for the strong consistency of $\{x^k\}$. Further, it can be shown (see Wu, 1981, Eq. (1.5)) that a necessary and sufficient condition for the strong consistency of $\{x^k\}$ is that $k\langle (t^j)^2 \rangle^k \to \infty$ as $k \to \infty$.

Lemma 7. *For almost all residual sequences $\{e_i^j\}$, the function $V^k(x)$ converges uniformly over $x \in C$ to the value $\bar{v} + \langle [F^j(\bar{x}) - F^j(x)]^2 \rangle$.*

Proof. It suffices to show that each component function $V_i^k(x)$ converges uniformly. The proof is then equivalent to the one given in Theorem 4 of Jennrich (1969). It is included here for completeness. Using the definition for $e_i^j = y_i^j - F_i^j(\bar{x})$, we obtain the following sequences of equalities:

$$\begin{aligned}
V_i^k(x) &= \langle [y_i^j - F_i^j(x)]^2 \rangle^{N(i,k)} = \langle [y_i^j - F_i^j(\bar{x}) + F_i^j(\bar{x}) - F_i^j(x)]^2 \rangle^{N(i,k)} \\
&= \langle e_i^j e_i^j \rangle^{N(i,k)} + 2\langle e_i^j [F_i^j(\bar{x}) - F_i^j(x)] \rangle^{N(i,k)} + \langle [F_i^j(\bar{x}) - F_i^j(x)]^2 \rangle^{N(i,k)}.
\end{aligned}$$

By Assumption 4, $N(i,k) \to \infty$. It follows from Assumption 6 that the tail averages $\langle [F_i^j(\bar{x}) - F_i^j(x)]^2 \rangle^{N(i,k)}$ converge uniformly to $\langle [F_i^j(\bar{x}) - F_i^j(x)]^2 \rangle$ as $k \to \infty$. It also follows that for almost every sequence $\{e_i^j\}$, the tail averages $\langle e_i^j [F_i^j(\bar{x}) - F_i^j(x)] \rangle^{N(i,k)}$ converge uniformly to zero as $k \to \infty$ (Theorem 4 of Jennrich, 1969). It follows from the strong law of large numbers (Theorem 5.4.2 of Chung, 1974) that the tail averages $\langle e_i^j e_i^j \rangle^{N(i,k)}$ converge to $\bar{v}_i$ as $k \to \infty$. This completes the proof. $\square$

Theorem 8 (Strong consistency). *Let $R^k(x)$ be defined by Eq. (20) and let x^k be a minimizer of $R^k(x)$ for each k. If $\bar{x} \in C$, then for almost all residual sequences $\{e_i^j\}$ the sequence $\{x^k\}$ converges to $\bar{x}$, and the sequence $\{V^k(x^k)\}$ converges to $\bar{v}$, as $k \to \infty$.*

Proof. We apply Lemma 7 and Assumption 4. It follows that for almost all sequences $\{e_i^j\}$, the sequence $\{R^k(x)\}$ converges uniformly over C to the function

$$R(x) = \sum_{i=1}^{M} \lambda_i \log\{\bar{v}_i + \langle [F_i^j(\bar{x}) - F_i^j(x)]^2 \rangle\}.$$

Since uniform convergence implies epi-convergence, the limit points of the sequence $\{x^k\}$ are in the set of minimizers of $R(x)$ (see Wets, 1980, for a discussion of epi-convergence). We note that $\bar{x}$ is a minimizer of $R(x)$ because for each i

$$(F_i^j(\bar{x}) - F_i^j(x))^2 \geq 0$$

for all x and is zero when $x = \bar{x}$. We further note, by Assumption 6, that $\bar{x}$ is the only minimizer of $R(x)$. Hence $\bar{x}$ is the only limit point of the sequence $\{x^k\}$. This completes the proof.

Lemma 9. *For almost all residual sequences $\{e_i^j\}$, as $k \to \infty$*

$$V_i^k(\bar{x}) \to \bar{v}_i, \quad \nabla V_i^k(\bar{x}) \to 0, \quad \nabla^2 V_i^k(\bar{x}) \to 2\langle \nabla F_i^j(\bar{x}) \nabla F_i^j(\bar{x})^{\mathsf{T}} \rangle.$$

Proof. The first conclusion follows directly from Theorem 8.

Rewriting Eq. (19) using the tail average notation and then differentiating, we obtain

$$\nabla V_i^k(x) = -2\langle [y_i^j - F_i^j(x)] \nabla F_i^j(x) \rangle^{N(i,k)}, \tag{23}$$

$$\nabla^2 V_i^k(x) = 2\langle \nabla F_i^j(x) \nabla F_i^j(x)^{\mathsf{T}} \rangle^{N(i,k)} - 2\langle [y_i^j - F_i^j(x)] \nabla^2 F_i^j(x) \rangle^{N(i,k)}. \tag{24}$$

Substituting $\bar{x}$ for x, and e_i^j for $y_i^j - F_i^j(\bar{x})$, we obtain

$$\nabla V_i^k(\bar{x}) = -2\langle e_i^j \nabla F_i^j(\bar{x}) \rangle^{N(i,k)}, \tag{25}$$

$$\nabla^2 V_i^k(\bar{x}) = 2\langle \nabla F_i^j(\bar{x}) \nabla F_i^j(\bar{x})^{\mathsf{T}} \rangle^{N(i,k)} - 2\langle e_i^j \nabla^2 F_i^j(\bar{x}) \rangle^{N(i,k)}. \tag{26}$$

By Assumption 6, the tail averages $\langle [\nabla F_i^j(\bar{x})^{\mathsf{T}}]^2 \rangle^{N(i,k)}$ converge. Hence, for almost all sequences $\{e_i^j\}$, $\langle e_i^j \nabla F_i^j(\bar{x}) \rangle$ converges to zero (Theorem 3 of Jennrich, 1969). This, together with Eq. (23), completes the proof of the second conclusion of this lemma. The final conclusion follows from a similar line of reasoning combined with Eq. (24). $\square$

Remark. The term $\langle e_i^j \nabla^2 F_i^j(\bar{x}) \rangle^{N(i,k)}$ is the second-order term, which is ignored by our algorithm. The proof above demonstrates that this term becomes insignificant as the number of data values increases. A very similar term is ignored in the Gauss–Newton method and a similar observation can be made in that case.

Define the information matrix Ψ by

$$\Psi = \sum_{i=1}^{M} \frac{\lambda_i}{\bar{v}_i} \langle \nabla F_i^j(\bar{x}) \nabla F_i^j(\bar{x})^{\mathsf{T}} \rangle. \tag{27}$$

Lemma 10. *Let $R^k(x)$ be defined by Eq. (20) and let Ψ be defined by Eq. (27). The sequence $\{\frac{1}{2}\sqrt{J(k)} \nabla R^k(\bar{x})\}$ converges in distribution to $N(0, \Psi)$.*

Proof. Taking the derivative of Eq. (20), we obtain

$$\nabla R^k(x) = \sum_{i=1}^{M} \lambda_i^k V_i^k(x)^{-1} \nabla V_i^k(x). \tag{28}$$

Substituting $\bar{x}$ for x and using Eq. (25), we obtain

$$\nabla R^k(\bar{x}) = -2 \sum_{i=1}^{M} \lambda_i^k V_i^k(\bar{x})^{-1} \langle \nabla F_i^j(\bar{x}) e_i^j \rangle^{N(i,k)}.$$

Multiplying the above equation by $-\frac{1}{2}\sqrt{J(k)}$ gives

$$-\tfrac{1}{2}\sqrt{J(k)}\nabla R^k(\bar{x}) = \sum_{i=1}^{M} \sqrt{\lambda_i^k} V_i^k(\bar{x})^{-1} \sqrt{N(i,k)} \langle \nabla F_i^j(\bar{x}) e_i^j \rangle^{N(i,k)}. \qquad (29)$$

It follows from Assumption 6 that for each i, m the tail average $\langle \nabla F_i^j(\bar{x}) \nabla F_m^j(\bar{x})^{\mathrm{T}} \rangle^{N(i,k)}$ converges as $k \to \infty$. We can therefore apply Corollary 1 of Jennrich (1969) to conclude that

$$u_i^k = \sqrt{N(i,k)} \langle \nabla F_i^j(\bar{x}) e_i^j \rangle^{N(i,k)}$$

converges in distribution to a normal random variable with mean zero and variance $\bar{v}_i \langle \nabla F_i^j(\bar{x}) \ \nabla F_i^j(\bar{x})^{\mathrm{T}} \rangle$. For fixed k and varying i, the random variables $\{u_i^k\}$ are independent. We also know that $\sqrt{\lambda_i^k} V_i^k(\bar{x})^{-1}$ converges to $\sqrt{\lambda_i}/\bar{v}_i$. We can therefore apply the corollary to Theorem 4.6.6 of Chung (1974) to Eq. (29) above to reach the conclusion of the lemma.   □

Lemma 11. *Let $R^k(x)$ be defined by Eq. (20) and let Ψ be defined by Eq. (27). For almost all residual sequences $\{e_i^j\}$, $\frac{1}{2}\nabla^2 R^k(\bar{x}) \to \Psi$ as $k \to \infty$.*

Proof. Taking the derivative of Eq. (28) and substituting $\bar{x}$ for x, we obtain

$$\nabla^2 R^k(\bar{x}) = \sum_{i=1}^{M} \lambda_i^k V_i^k(\bar{x})^{-1} \nabla^2 V_i^k(\bar{x}) - \lambda_i^k V_i^k(\bar{x})^{-2} \nabla V_i^k(\bar{x}) \nabla V_i^k(\bar{x})^{\mathrm{T}}.$$

Applying Lemma 9, and the fact that $\lambda_i^k \to \lambda_i$, we find that, for almost all $\{e_i^j\}$, $\nabla^2 R^k(\bar{x})$ converges to

$$\sum_{i=1}^{M} \frac{\lambda_i}{\bar{v}_i} 2 \langle \nabla F_i^j(\bar{x}) \nabla F_i^j(\bar{x})^{\mathrm{T}} \rangle,$$

which completes the proof of this lemma.   □

The following lemma is a direct consequence of Theorem 8, Lemma 11, and Assumption 6 which ensure that $\nabla^2 R^k(x)$ converges uniformly in x.

Lemma 12. *Let $R^k(x)$ be defined by Eq. (20), let Ψ be defined by Eq. (27), and let x^k be a minimizer of R^k for each k. For almost all residual sequences $\{e_i^j\}$,*

$$\frac{1}{2} \int_0^1 \nabla^2 R^k[x^k + t(\bar{x} - x^k)] \, \mathrm{d}t \to \Psi \quad \text{as } k \to \infty.$$

Theorem 13. *Let x^k be a minimizer of R^k for each k, let Ψ be defined by Eq. (27), and suppose that Ψ is invertible. If $\bar{x}$ is in the interior of C, then the sequence $\sqrt{J(k)}(x^k - \bar{x})$ converges in distribution to $N(0, \Psi^{-1})$.*

Proof. Because $\bar{x}$ is in the interior of C and the sequence $\{x^k\}$ converges to $\bar{x}$ for almost all sequences $\{e_i^j\}$, we find that $\{x^k\}$ belongs to the interior of C for sufficiently large k. For such k, the gradient $\nabla R^k(x^k)$ is zero because x^k minimizes $R^k(x)$ over C. Thus, $\nabla R^k(\bar{x})$ satisfies

$$\nabla R^k(\bar{x}) = \left\{ \int_0^1 \nabla^2 R^k[x^k + t(\bar{x} - x^k)]\,dt \right\} (\bar{x} - x^k).$$

Define the average Hessian H^k by

$$H^k = \int_0^1 \nabla^2 R^k[x^k + t(\bar{x} - x^k)]\,dt.$$

It follows that

$$\tfrac{1}{2}H^k \sqrt{J(k)}(\bar{x} - x^k) = \tfrac{1}{2}\sqrt{J(k)}\nabla R^k(\bar{x}).$$

By Lemma 12, $\tfrac{1}{2}H^k$ converges almost everywhere to Ψ. By Lemma 10, $\tfrac{1}{2}\sqrt{J(k)}$ $\nabla R^k(\bar{x})$ converges to $N(0, \Psi)$ in distribution. Thus, by Lemma 4.1 in Chapter 6 of Lehmann (1983), $(\bar{x} - x^k)$ converges to $N(0, \Psi^{-1})$ in distribution. This completes the proof of the theorem. □

6. Results

In this section we apply the relative weighting method to the example defined in Section 2. Recall that for this model, the functions $\{F_1^j(x)\}$ and $\{F_2^j(x)\}$ are defined by Eqs. (4) and (5), respectively, and the values $\{y_i^j\}$ are defined in Table 2.

The algorithm of Section 4 is applied to minimize the reduced objective function $R(x)$ defined in Eq. (9). That is, the objective function $G(x)$, in the algorithm of Section 4, is replaced by $R(x)$. The approximation $A(x, d)$ is replaced by $W(x, d)$ which is defined in Eq. (12). The set C in the algorithm was set to the entire space; i.e., there were no constraints in this example. The iteration procedure was terminated when the solution to the linear least-squares approximate subproblem, d^k in Step 2 of the algorithm, was small; to be specific, when the absolute values of both components of d^k were less than 10^{-5}. The values 0.5 and 0.01 were used for the algorithm line search parameters γ and μ, respectively. Table 4 contains the resulting parameter estimates (x_1, x_2, x_3), variance estimates (v_1, v_2), reduced likelihood value (R), and norm of the gradient of the reduced likelihood function $(\|\nabla R\|)$ at each iteration. In this simulation, Inequality (15) in Step 3 of the algorithm was satisfied

Table 4
Algorithm iterates

Iteration	x_1	x_2	x_3	v_1	v_2	R	$\|\nabla R\|$
1	0.08446	0.37930	0.40304	0.08576	1.52109	−1.49770	8.06530
2	0.09046	0.45656	0.54709	0.00909	1.63138	−2.97082	1.50672
3	0.06761	0.47478	0.55063	0.00889	1.64364	−2.98263	0.00498
4	0.06753	0.47518	0.55030	0.00889	1.64368	−2.98263	0.00001

with $s = 0$ at each iteration, and so the step-length, λ^k, was one for each of the iterations.

The information matrix Ψ is defined in Eq. (27). We approximate it using the following expression:

$$\hat{\Psi} = \sum_{i=1}^{M} \frac{N(i)}{JV_i(\hat{x})} \langle \nabla F_i^j(\hat{x}) \nabla F_i^j(\hat{x})^{\mathrm{T}} \rangle^{N(i)} = \frac{1}{J} \sum_{i=1}^{M} \sum_{j=1}^{N(i)} \frac{\nabla F_i^j(\hat{x}) \nabla F_i^j(\hat{x})^{\mathrm{T}}}{V_i(\hat{x})},$$

where $\hat{x}$ is the parameter estimate corresponding to iteration number 4, $M = 2$, $N(1) = 20$, $N(2) = 10$, $J = N(1) + N(2) = 30$, and $V_i(x)$ is defined by Eq. (7). (The notation $\langle w^j \rangle^p$ is defined in Eq. (21).) The corresponding estimate for the covariance of the parameter estimate $\hat{x}$ is (see Theorem 13)

$$\frac{1}{J} \hat{\Psi}^{-1} = \begin{pmatrix} 0.00192 & -0.00066 & -0.00074 \\ -0.00066 & 0.03741 & -0.03678 \\ -0.00074 & -0.03678 & 0.03748 \end{pmatrix}.$$

The square roots of the diagonal elements in the matrix above are approximations for the standard deviation of the components of $\hat{x}$. It is significant that $|\bar{x}_1 - \hat{x}_1| < \sqrt{0.00192}$, $|\bar{x}_2 - \hat{x}_2| < \sqrt{0.03741}$, and $|\bar{x}_3 - \hat{x}_3| < \sqrt{0.03748}$ (the simulation values for the components of $\bar{x}$ are specified in Eq. (6)). Thus, all the components of the estimator $\hat{x}$ are within one standard deviation of the values used for the simulation. We also note that the stopping criterion, 10^{-5}, is small relative to the standard deviations.

The relative weighting plot in Fig. 2 shows the residuals in the plasma data, $y_1^j - F_1^j(\hat{x})$, as a function of t_1^j. The absolute weighting plot shows the residuals, $y_1^j - F_1^j(\underline{x})$, where $\underline{x}$ is the estimate for the parameter vector x that corresponds to the standard least-squares fit of the entire data as one set. Note that the apparent linear trend in the absolute weighting plot is due to an error in the corresponding variance model and is not present in the relative weighting plot.
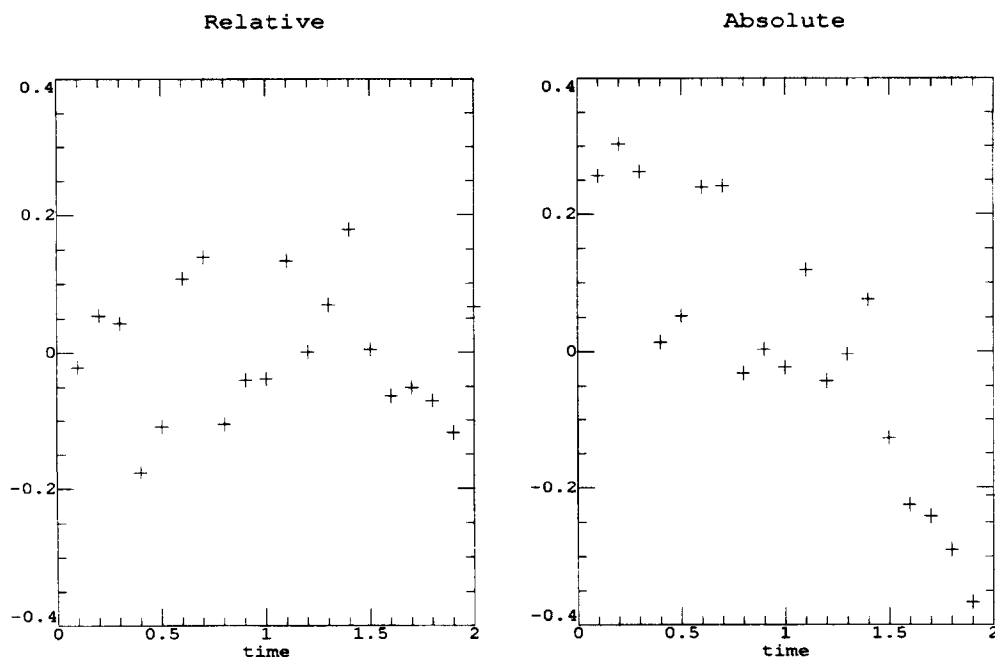
Relative   Absolute

Fig. 2. Comparison of residuals.

Acknowledgements

References

Burke, J.V., Descent methods for composite nondifferential optimization problems, *Math. Prog.*, **33** (1985) 260–279.

Caines, P.E., *Linear stochastic systems*, ch. 8 (Wiley, New York, 1988).

Chung, K.L., *A course in probability theory*, 2nd edn. (Academic Press, New York, 1974).

Clarke, F.H., *Optimization and nonsmooth analysis*, p. 52, corollary above Proposition 2.4.4 (Wiley-Interscience, New York, 1983).

Gallant, R.A., *Nonlinear statistical models*, ch. 3 (Wiley, New York, 1987).

Ibragimov, I.A. and R.Z. Has'minskii, *Statistical estimation: asymptotic theory*, ch. 3 (Springer, New York, 1981).

Jennrich, R.I., Asymptotic properties of non-linear least squares estimators, *Ann. Math. Statist.* **40** (1969) 633–643.

Lehmann, E.L., *Theory of point estimation* (Wiley, New York, 1983).

Rockafellar, R.T., *Convex analysis* (Princeton Univ. Press, Princeton, NJ, 1970).

Wets, R.J.-B., Convergence of convex functions, variational inequalities, and convex optimization problems, in: R.W. Cottle, F. Giannessi and J.-L. Lions (Eds.), *Variational inequalities and complementarity problems* (Wiley, New York, 1980) 375–403.

Wu, C.-F., Asymptotic theory of nonlinear least squares estimation, *Ann. Statist.* **9** (1981) 501–513.