SAS/ETS® 14.2 User's Guide
Nonlinear Optimization Methods

This document is an individual chapter from *SAS/ETS® 14.2 User's Guide*.

The correct bibliographic citation for this manual is as follows: SAS Institute Inc. 2016. *SAS/ETS® 14.2 User's Guide*. Cary, NC: SAS Institute Inc.

SAS/ETS® 14.2 User's Guide

SAS Institute Inc., SAS Campus Drive, Cary, NC 27513-2414

November 2016

Chapter 6
Nonlinear Optimization Methods

Contents

Overview

Several SAS/ETS procedures (COUNTREG, ENTROPY, MDC, QLIM, UCM, and VARMAX) use the nonlinear optimization (NLO) subsystem to perform nonlinear optimization. This chapter describes the options of the NLO system and some technical details of the available optimization methods. Note that not all options have been implemented for all procedures that use the NLO subsystem. You should check each procedure chapter for more information about which options are available.

Options

Table 6.1 summarizes the options available in the NLO system.

Table 6.1 NLO options

Option	Description
Optimization Specifications	
TECHNIQUE=	Minimization technique
UPDATE=	Update technique
LINESEARCH=	Line-search method
LSPRECISION=	Line-search precision
HESCAL=	Type of Hessian scaling
INHESSIAN=	Start for approximated Hessian

Table 6.1 *continued*

Option	Description
RESTART=	Iteration number for update restart
Termination Criteria Specifications	
MAXFUNC=	Maximum number of function calls
MAXITER=	Maximum number of iterations
MINITER=	Minimum number of iterations
MAXTIME=	Upper limit seconds of CPU time
ABSCONV=	Absolute function convergence criterion
ABSFCONV=	Absolute function convergence criterion
ABSGCONV=	Absolute gradient convergence criterion
ABSXCONV=	Absolute parameter convergence criterion
FCONV=	Relative function convergence criterion
FCONV2=	Relative function convergence criterion
GCONV=	Relative gradient convergence criterion
XCONV=	Relative parameter convergence criterion
FSIZE=	Used in FCONV, GCONV criterion
XSIZE=	Used in XCONV criterion
Step Length Options	
DAMPSTEP=	Damped steps in line search
MAXSTEP=	Maximum trust region radius
INSTEP=	Initial trust region radius
Printed Output Options	
PALL	Display (almost) all printed optimization-related output
PHISTORY	Display optimization history
PHISTPARMS	Display parameter estimates in each iteration
PSHORT	Reduce some default optimization-related output
PSUMMARY	Reduce most default optimization-related output
NOPRINT	Suppress all printed optimization-related output
Remote Monitoring Option	
SOCKET=	Specify the fileref for remote monitoring

These options are described in alphabetical order.

ABSCONV=*r*

ABSTOL=*r*

specifies an absolute function convergence criterion. For minimization, termination requires $f(\theta^{(k)}) \leq r$. The default value of r is the negative square root of the largest double-precision value, which serves only as a protection against overflows.

ABSFCONV=*r***[***n***]**

ABSFTOL=*r***[***n***]**

specifies an absolute function convergence criterion. For all techniques except NMSIMP, termination requires a small change of the function value in successive iterations:

$$|f(\theta^{(k-1)}) - f(\theta^{(k)})| \leq r$$

The same formula is used for the NMSIMP technique, but $\theta^{(k)}$ is defined as the vertex with the lowest function value, and $\theta^{(k-1)}$ is defined as the vertex with the highest function value in the simplex. The default value is *r*=0. The optional integer value *n* specifies the number of successive iterations for which the criterion must be satisfied before the process can be terminated.

ABSGCONV=*r*[*n*]

ABSGTOL=*r*[*n*]

specifies an absolute gradient convergence criterion. Termination requires the maximum absolute gradient element to be small:

$$\max_j |g_j(\theta^{(k)})| \leq r$$

This criterion is not used by the NMSIMP technique. The default value is $r = 1\mathrm{E}-5$. The optional integer value *n* specifies the number of successive iterations for which the criterion must be satisfied before the process can be terminated.

ABSXCONV=*r*[*n*]

ABSXTOL=*r*[*n*]

specifies an absolute parameter convergence criterion. For all techniques except NMSIMP, termination requires a small Euclidean distance between successive parameter vectors,

$$\| \theta^{(k)} - \theta^{(k-1)} \|_2 \leq r$$

For the NMSIMP technique, termination requires either a small length $\alpha^{(k)}$ of the vertices of a restart simplex,

$$\alpha^{(k)} \leq r$$

or a small simplex size,

$$\delta^{(k)} \leq r$$

where the simplex size $\delta^{(k)}$ is defined as the L1 distance from the simplex vertex $\xi^{(k)}$ with the smallest function value to the other *n* simplex points $\theta_l^{(k)} \neq \xi^{(k)}$:

$$\delta^{(k)} = \sum_{\theta_l \neq y} \| \theta_l^{(k)} - \xi^{(k)} \|_1$$

The default is $r = 1\mathrm{E}-8$ for the NMSIMP technique and *r*=0 otherwise. The optional integer value *n* specifies the number of successive iterations for which the criterion must be satisfied before the process can terminate.

DAMPSTEP[=*r*]

specifies that the initial step length value $\alpha^{(0)}$ for each line search (used by the QUANEW, HYQUAN, CONGRA, or NEWRAP technique) cannot be larger than *r* times the step length value used in the former iteration. If the DAMPSTEP option is specified but *r* is not specified, the default is *r*=2. The DAMPSTEP=*r* option can prevent the line-search algorithm from repeatedly stepping into regions where some objective functions are difficult to compute or where they could lead to floating point overflows during the computation of objective functions and their derivatives. The DAMPSTEP=*r* option can save time-costly function calls during the line searches of objective functions that result in very small steps.

FCONV=*r*[*n*]

FTOL=*r*[*n*]

specifies a relative function convergence criterion. For all techniques except NMSIMP, termination requires a small relative change of the function value in successive iterations,

$$\frac{|f(\theta^{(k)}) - f(\theta^{(k-1)})|}{\max(|f(\theta^{(k-1)})|, \text{FSIZE})} \leq r$$

where FSIZE is defined by the FSIZE= option. The same formula is used for the NMSIMP technique, but $\theta^{(k)}$ is defined as the vertex with the lowest function value, and $\theta^{(k-1)}$ is defined as the vertex with the highest function value in the simplex. The default value may depend on the procedure. In most cases, you can use the PALL option to find it.

FCONV2=*r*[*n*]

FTOL2=*r*[*n*]

specifies another function convergence criterion.

For all techniques except NMSIMP, termination requires a small predicted reduction

$$df^{(k)} \approx f(\theta^{(k)}) - f(\theta^{(k)} + s^{(k)})$$

of the objective function. The predicted reduction

$$\begin{aligned} df^{(k)} &= -g^{(k)T} s^{(k)} - \frac{1}{2} s^{(k)T} H^{(k)} s^{(k)} \\ &= -\frac{1}{2} s^{(k)T} g^{(k)} \\ &\leq r \end{aligned}$$

is computed by approximating the objective function f by the first two terms of the Taylor series and substituting the Newton step

$$s^{(k)} = -[H^{(k)}]^{-1} g^{(k)}$$

For the NMSIMP technique, termination requires a small standard deviation of the function values of the $n+1$ simplex vertices $\theta_l^{(k)}$, $l = 0, \ldots, n$, $\sqrt{\frac{1}{n+1} \sum_l \left[f(\theta_l^{(k)}) - \overline{f}(\theta^{(k)}) \right]^2} \leq r$, where $\overline{f}(\theta^{(k)}) = \frac{1}{n+1} \sum_l f(\theta_l^{(k)})$. If there are n_{act} boundary constraints active at $\theta^{(k)}$, the mean and standard deviation are computed only for the $n + 1 - n_{act}$ unconstrained vertices.

The default value is $r = 1\text{E} - 6$ for the NMSIMP technique and *r*=0 otherwise. The optional integer value *n* specifies the number of successive iterations for which the criterion must be satisfied before the process can terminate.

FSIZE=*r*

specifies the FSIZE parameter of the relative function and relative gradient termination criteria. The default value is *r*=0. For more information, see the FCONV= and GCONV= options.

GCONV=*r*[*n*]

GTOL=*r*[*n*]

specifies a relative gradient convergence criterion. For all techniques except CONGRA and NMSIMP, termination requires that the normalized predicted function reduction be small,

$$\frac{g(\theta^{(k)})^T [H^{(k)}]^{-1} g(\theta^{(k)})}{\max(|f(\theta^{(k)})|, \text{FSIZE})} \leq r$$

where FSIZE is defined by the FSIZE= option. For the CONGRA technique (where a reliable Hessian estimate H is not available), the following criterion is used:

$$\frac{\| g(\theta^{(k)}) \|_2^2 \quad \| s(\theta^{(k)}) \|_2}{\| g(\theta^{(k)}) - g(\theta^{(k-1)}) \|_2 \max(|f(\theta^{(k)})|, \text{FSIZE})} \leq r$$

This criterion is not used by the NMSIMP technique. The default value is $r = 1\text{E} - 8$. The optional integer value n specifies the number of successive iterations for which the criterion must be satisfied before the process can terminate.

HESCAL=0|1|2|3

HS=0|1|2|3

specifies the scaling version of the Hessian matrix used in NRRIDG, TRUREG, NEWRAP, or DBLDOG optimization.

If HS is not equal to 0, the first iteration and each restart iteration sets the diagonal scaling matrix $D^{(0)} = \text{diag}(d_i^{(0)})$,

$$d_i^{(0)} = \sqrt{\max(|H_{i,i}^{(0)}|, \epsilon)}$$

where $H_{i,i}^{(0)}$ are the diagonal elements of the Hessian. In every other iteration, the diagonal scaling matrix $D^{(0)} = \text{diag}(d_i^{(0)})$ is updated depending on the HS option:

HS=0 specifies that no scaling is done.

HS=1 specifies the Moré (1978) scaling update:

$$d_i^{(k+1)} = \max\left[d_i^{(k)}, \sqrt{\max(|H_{i,i}^{(k)}|, \epsilon)}\right]$$

HS=2 specifies the Dennis, Gay, and Welsch (1981) scaling update:

$$d_i^{(k+1)} = \max\left[0.6 * d_i^{(k)}, \sqrt{\max(|H_{i,i}^{(k)}|, \epsilon)}\right]$$

HS=3 specifies that d_i is reset in each iteration:

$$d_i^{(k+1)} = \sqrt{\max(|H_{i,i}^{(k)}|, \epsilon)}$$

In each scaling update, ϵ is the relative machine precision. The default value is HS=0. Scaling of the Hessian can be time-consuming in the case where general linear constraints are active.

INHESSIAN[=*r*]

INHESS[=*r*]

specifies how the initial estimate of the approximate Hessian is defined for the quasi-Newton techniques QUANEW and DBLDOG. There are two alternatives:

- If you do not use the *r* specification, the initial estimate of the approximate Hessian is set to the Hessian at $\theta^{(0)}$.
- If you do use the *r* specification, the initial estimate of the approximate Hessian is set to the multiple of the identity matrix rI.

By default, if you do not specify the option INHESSIAN=*r*, the initial estimate of the approximate Hessian is set to the multiple of the identity matrix rI, where the scalar r is computed from the magnitude of the initial gradient.

INSTEP=*r*

reduces the length of the first trial step during the line search of the first iterations. For highly nonlinear objective functions, such as the EXP function, the default initial radius of the trust-region algorithm TRUREG or DBLDOG or the default step length of the line-search algorithms can result in arithmetic overflows. If this occurs, you should specify decreasing values of $0 < r < 1$ such as INSTEP=1E–1, INSTEP=1E–2, INSTEP=1E–4, and so on, until the iteration starts successfully.

- For trust-region algorithms (TRUREG, DBLDOG), the INSTEP= option specifies a factor $r > 0$ for the initial radius $\Delta^{(0)}$ of the trust region. The default initial trust-region radius is the length of the scaled gradient. This step corresponds to the default radius factor of $r = 1$.
- For line-search algorithms (NEWRAP, CONGRA, QUANEW), the INSTEP= option specifies an upper bound for the initial step length for the line search during the first five iterations. The default initial step length is $r = 1$.
- For the Nelder-Mead simplex algorithm, using TECH=NMSIMP, the INSTEP=*r* option defines the size of the start simplex.

LINESEARCH=*i*

LIS=*i*

specifies the line-search method for the CONGRA, QUANEW, and NEWRAP optimization techniques. For an introduction to line-search techniques, see Fletcher (1987). The value of *i* can be $1, \ldots, 8$. For CONGRA, QUANEW and NEWRAP, the default value is $i = 2$.

LIS=1 specifies a line-search method that needs the same number of function and gradient calls for cubic interpolation and cubic extrapolation; this method is similar to one used by the Harwell subroutine library.

LIS=2 specifies a line-search method that needs more function than gradient calls for quadratic and cubic interpolation and cubic extrapolation; this method is implemented as shown in Fletcher (1987) and can be modified to an exact line search by using the LSPRECISION= option.

LIS=3 specifies a line-search method that needs the same number of function and gradient calls for cubic interpolation and cubic extrapolation; this method is implemented as shown in Fletcher (1987) and can be modified to an exact line search by using the LSPRECISION= option.

LIS=4 specifies a line-search method that needs the same number of function and gradient calls for stepwise extrapolation and cubic interpolation.

LIS=5 specifies a line-search method that is a modified version of LIS=4.

LIS=6 specifies golden section line search (Polak 1971), which uses only function values for linear approximation.

LIS=7 specifies bisection line search (Polak 1971), which uses only function values for linear approximation.

LIS=8 specifies the Armijo line-search technique (Polak 1971), which uses only function values for linear approximation.

LSPRECISION=*r*

LSP=*r*

specifies the degree of accuracy that should be obtained by the line-search algorithms LIS=2 and LIS=3. Usually an imprecise line search is inexpensive and successful. For more difficult optimization problems, a more precise and expensive line search may be necessary (Fletcher 1987). The second line-search method (which is the default for the NEWRAP, QUANEW, and CONGRA techniques) and the third line-search method approach exact line search for small LSPRECISION= values. If you have numerical problems, you should try to decrease the LSPRECISION= value to obtain a more precise line search. The default values are shown in Table 6.2.

Table 6.2 Line Search Precision Defaults

TECH=	UPDATE=	LSP Default
QUANEW	DBFGS, BFGS	$r = 0.4$
QUANEW	DDFP, DFP	$r = 0.06$
CONGRA	All	$r = 0.1$
NEWRAP	No update	$r = 0.9$

For more information, see Fletcher (1987).

MAXFUNC=*i*

MAXFU=*i*

specifies the maximum number *i* of function calls in the optimization process. The default values are

- TRUREG, NRRIDG, NEWRAP: 125
- QUANEW, DBLDOG: 500
- CONGRA: 1000
- NMSIMP: 3000

Note that the optimization can terminate only after completing a full iteration. Therefore, the number of function calls that is actually performed can exceed the number that is specified by the MAXFUNC= option.

MAXITER=*i*

MAXIT=*i*

specifies the maximum number *i* of iterations in the optimization process. The default values are

- TRUREG, NRRIDG, NEWRAP: 50
- QUANEW, DBLDOG: 200
- CONGRA: 400
- NMSIMP: 1000

These default values are also valid when *i* is specified as a missing value.

MAXSTEP=*r***[***n***]**

specifies an upper bound for the step length of the line-search algorithms during the first *n* iterations. By default, *r* is the largest double-precision value and *n* is the largest integer available. Setting this option can improve the speed of convergence for the CONGRA, QUANEW, and NEWRAP techniques.

MAXTIME=*r*

specifies an upper limit of *r* seconds of CPU time for the optimization process. The default value is the largest floating-point double representation of your computer. Note that the time specified by the MAXTIME= option is checked only once at the end of each iteration. Therefore, the actual running time can be much longer than that specified by the MAXTIME= option. The actual running time includes the rest of the time needed to finish the iteration and the time needed to generate the output of the results.

MINITER=*i*

MINIT=*i*

specifies the minimum number of iterations. The default value is 0. If you request more iterations than are actually needed for convergence to a stationary point, the optimization algorithms can behave strangely. For example, the effect of rounding errors can prevent the algorithm from continuing for the required number of iterations.

NOPRINT

suppresses the output. (See procedure documentation for availability of this option.)

PALL

displays all optional output for optimization. (See procedure documentation for availability of this option.)

PHISTORY

displays the optimization history. (See procedure documentation for availability of this option.)

PHISTPARMS

display parameter estimates in each iteration. (See procedure documentation for availability of this option.)

PINIT

displays the initial values and derivatives (if available). (See procedure documentation for availability of this option.)

PSHORT
restricts the amount of default output. (See procedure documentation for availability of this option.)

PSUMMARY
restricts the amount of default displayed output to a short form of iteration history and notes, warnings, and errors. (See procedure documentation for availability of this option.)

RESTART=*i* > 0

REST=*i* > 0
specifies that the QUANEW or CONGRA algorithm is restarted with a steepest descent/ascent search direction after, at most, *i* iterations. Default values are as follows:

- CONGRA
  UPDATE=PB: restart is performed automatically, *i* is not used.
- CONGRA
  UPDATE≠PB: $i = \min(10n, 80)$, where n is the number of parameters.
- QUANEW
  i is the largest integer available.

SOCKET=*fileref*
specifies the fileref that contains the information needed for remote monitoring. For more information, see the section "Remote Monitoring" on page 186.

TECHNIQUE=*value*

TECH=*value*
specifies the optimization technique. Valid values are as follows:

CONGRA performs a conjugate-gradient optimization, which can be more precisely specified with the UPDATE= option and modified with the LINESEARCH= option. When you specify this option, UPDATE=PB by default.

DBLDOG performs a version of double-dogleg optimization, which can be more precisely specified with the UPDATE= option. When you specify this option, UPDATE=DBFGS by default.

NMSIMP performs a Nelder-Mead simplex optimization.

NONE does not perform any optimization. This option can be used as follows:

- to perform a grid search without optimization
- to compute estimates and predictions that cannot be obtained efficiently with any of the optimization techniques

NEWRAP performs a Newton-Raphson optimization that combines a line-search algorithm with ridging. The line-search algorithm LIS=2 is the default method.

NRRIDG performs a Newton-Raphson optimization with ridging.

QUANEW performs a quasi-Newton optimization, which can be defined more precisely with the UPDATE= option and modified with the LINESEARCH= option. This is the default estimation method.

TRUREG performs a trust region optimization.

UPDATE=*method*

UPD=*method*

specifies the update method for the QUANEW, DBLDOG, or CONGRA optimization technique. Not every update method can be used with each optimizer.

Valid *methods* are as follows:

BFGS	performs the original Broyden, Fletcher, Goldfarb, and Shanno (BFGS) update of the inverse Hessian matrix.
DBFGS	performs the dual BFGS update of the Cholesky factor of the Hessian matrix. This is the default update method.
DDFP	performs the dual Davidon, Fletcher, and Powell (DFP) update of the Cholesky factor of the Hessian matrix.
DFP	performs the original DFP update of the inverse Hessian matrix.
PB	performs the automatic restart update method of Powell (1977); Beale (1972).
FR	performs the Fletcher-Reeves update (Fletcher 1987).
PR	performs the Polak-Ribiere update (Fletcher 1987).
CD	performs a conjugate-descent update of Fletcher (1987).

XCONV=*r***[***n***]**

XTOL=*r***[***n***]**

specifies the relative parameter convergence criterion. For all techniques except NMSIMP, termination requires a small relative parameter change in subsequent iterations.

$$\frac{\max_j |\theta_j^{(k)} - \theta_j^{(k-1)}|}{\max(|\theta_j^{(k)}|, |\theta_j^{(k-1)}|, \text{XSIZE})} \leq r$$

For the NMSIMP technique, the same formula is used, but $\theta_j^{(k)}$ is defined as the vertex with the lowest function value and $\theta_j^{(k-1)}$ is defined as the vertex with the highest function value in the simplex. The default value is $r = 1\text{E} - 8$ for the NMSIMP technique and $r = 0$ otherwise. The optional integer value n specifies the number of successive iterations for which the criterion must be satisfied before the process can be terminated.

XSIZE=*r* > 0

specifies the XSIZE parameter of the relative parameter termination criterion. The default value is $r = 0$. For more information, see the XCONV= option.

Details of Optimization Algorithms

Overview

There are several optimization techniques available, as shown in Table 6.3. You can choose a particular optimizer with the TECH=*name* option in the PROC statement or NLOPTIONS statement.

Table 6.3 Optimization Techniques

Algorithm	TECH=
Trust region method	TRUREG
Newton-Raphson method with line search	NEWRAP
Newton-Raphson method with ridging	NRRIDG
Quasi-Newton methods (DBFGS, DDFP, BFGS, DFP)	QUANEW
Double-dogleg method (DBFGS, DDFP)	DBLDOG
Conjugate gradient methods (PB, FR, PR, CD)	CONGRA
Nelder-Mead simplex method	NMSIMP

No algorithm for optimizing general nonlinear functions exists that always finds the global optimum for a general nonlinear minimization problem in a reasonable amount of time. Since no single optimization technique is invariably superior to others, NLO provides a variety of optimization techniques that work well in various circumstances. However, you can devise problems for which none of the techniques in NLO can find the correct solution. Moreover, nonlinear optimization can be computationally expensive in terms of time and memory, so you must be careful when matching an algorithm to a problem.

All optimization techniques in NLO use $O(n^2)$ memory except the conjugate gradient methods, which use only $O(n)$ of memory and are designed to optimize problems with many parameters. These iterative techniques require repeated computation of the following:

- the function value (optimization criterion)
- the gradient vector (first-order partial derivatives)
- for some techniques, the (approximate) Hessian matrix (second-order partial derivatives)

However, since each of the optimizers requires different derivatives, some computational efficiencies can be gained. Table 6.4 shows, for each optimization technique, which derivatives are required. (FOD means that first-order derivatives or the gradient is computed; SOD means that second-order derivatives or the Hessian is computed.)

Table 6.4 Optimization Computations

Algorithm	FOD	SOD
TRUREG	x	x
NEWRAP	x	x
NRRIDG	x	x
QUANEW	x	-
DBLDOG	x	-
CONGRA	x	-
NMSIMP	-	-

Each optimization method employs one or more convergence criteria that determine when it has converged. The various termination criteria are listed and described in the previous section. An algorithm is considered to have converged when any one of the convergence criterion is satisfied. For example, under the default settings, the QUANEW algorithm will converge if $\mathrm{ABSGCONV} < 1\mathrm{E}-5$, $\mathrm{FCONV} < 10^{-\mathrm{FDIGITS}}$, or $\mathrm{GCONV} < 1\mathrm{E}-8$.

Choosing an Optimization Algorithm

The factors that go into choosing a particular optimization technique for a particular problem are complex and might involve trial and error.

For many optimization problems, computing the gradient takes more computer time than computing the function value, and computing the Hessian sometimes takes *much* more computer time and memory than computing the gradient, especially when there are many decision variables. Unfortunately, optimization techniques that do not use some kind of Hessian approximation usually require many more iterations than techniques that do use a Hessian matrix, and as a result the total run time of these techniques is often longer. Techniques that do not use the Hessian also tend to be less reliable. For example, they can more easily terminate at stationary points rather than at global optima.

A few general remarks about the various optimization techniques follow.

- The second-derivative methods TRUREG, NEWRAP, and NRRIDG are best for small problems where the Hessian matrix is not expensive to compute. Sometimes the NRRIDG algorithm can be faster than the TRUREG algorithm, but TRUREG can be more stable. The NRRIDG algorithm requires only one matrix with $n(n+1)/2$ double words; TRUREG and NEWRAP require two such matrices.

- The first-derivative methods QUANEW and DBLDOG are best for medium-sized problems where the objective function and the gradient are much faster to evaluate than the Hessian. The QUANEW and DBLDOG algorithms, in general, require more iterations than TRUREG, NRRIDG, and NEWRAP, but each iteration can be much faster. The QUANEW and DBLDOG algorithms require only the gradient to update an approximate Hessian, and they require slightly less memory than TRUREG or NEWRAP (essentially one matrix with $n(n+1)/2$ double words). QUANEW is the default optimization method.

- The first-derivative method CONGRA is best for large problems where the objective function and the gradient can be computed much faster than the Hessian and where too much memory is required to

store the (approximate) Hessian. The CONGRA algorithm, in general, requires more iterations than QUANEW or DBLDOG, but each iteration can be much faster. Since CONGRA requires only a factor of n double-word memory, many large applications can be solved only by CONGRA.

- The no-derivative method NMSIMP is best for small problems where derivatives are not continuous or are very difficult to compute.

Algorithm Descriptions

Some details about the optimization techniques are as follows.

Trust Region Optimization (TRUREG)

The trust region method uses the gradient $g(\theta_{(k)})$ and the Hessian matrix $H(\theta_{(k)})$; thus, it requires that the objective function $f(\theta)$ have continuous first- and second-order derivatives inside the feasible region.

The trust region method iteratively optimizes a quadratic approximation to the nonlinear objective function within a hyperelliptic trust region with radius Δ that constrains the step size that corresponds to the quality of the quadratic approximation. The trust region method is implemented using Dennis, Gay, and Welsch (1981); Gay (1983); Moré and Sorensen (1983).

The trust region method performs well for small- to medium-sized problems, and it does not need many function, gradient, and Hessian calls. However, if the computation of the Hessian matrix is computationally expensive, one of the (dual) quasi-Newton or conjugate gradient algorithms may be more efficient.

Newton-Raphson Optimization with Line Search (NEWRAP)

The NEWRAP technique uses the gradient $g(\theta_{(k)})$ and the Hessian matrix $H(\theta_{(k)})$; thus, it requires that the objective function have continuous first- and second-order derivatives inside the feasible region. If second-order derivatives are computed efficiently and precisely, the NEWRAP method can perform well for medium-sized to large problems, and it does not need many function, gradient, and Hessian calls.

This algorithm uses a pure Newton step when the Hessian is positive definite and when the Newton step reduces the value of the objective function successfully. Otherwise, a combination of ridging and line search is performed to compute successful steps. If the Hessian is not positive definite, a multiple of the identity matrix is added to the Hessian matrix to make it positive definite.

In each iteration, a line search is performed along the search direction to find an approximate optimum of the objective function. The default line-search method uses quadratic interpolation and cubic extrapolation (LIS=2).

Newton-Raphson Ridge Optimization (NRRIDG)

The NRRIDG technique uses the gradient $g(\theta_{(k)})$ and the Hessian matrix $H(\theta_{(k)})$; thus, it requires that the objective function have continuous first- and second-order derivatives inside the feasible region.

This algorithm uses a pure Newton step when the Hessian is positive definite and when the Newton step reduces the value of the objective function successfully. If at least one of these two conditions is not satisfied, a multiple of the identity matrix is added to the Hessian matrix.

The NRRIDG method performs well for small- to medium-sized problems, and it does not require many function, gradient, and Hessian calls. However, if the computation of the Hessian matrix is computationally expensive, one of the (dual) quasi-Newton or conjugate gradient algorithms might be more efficient.

Since the NRRIDG technique uses an orthogonal decomposition of the approximate Hessian, each iteration of NRRIDG can be slower than that of the NEWRAP technique, which works with Cholesky decomposition. Usually, however, NRRIDG requires fewer iterations than NEWRAP.

Quasi-Newton Optimization (QUANEW)

The (dual) quasi-Newton method uses the gradient $g(\theta_{(k)})$, and it does not need to compute second-order derivatives since they are approximated. It works well for medium to moderately large optimization problems where the objective function and the gradient are much faster to compute than the Hessian; but, in general, it requires more iterations than the TRUREG, NEWRAP, and NRRIDG techniques, which compute second-order derivatives. QUANEW is the default optimization algorithm because it provides an appropriate balance between the speed and stability required for most nonlinear mixed model applications.

The QUANEW technique is one of the following, depending upon the value of the UPDATE= option:

- the original quasi-Newton algorithm, which updates an approximation of the inverse Hessian
- the dual quasi-Newton algorithm, which updates the Cholesky factor of an approximate Hessian (default)

You can specify four update formulas with the UPDATE= option:

- DBFGS performs the dual Broyden, Fletcher, Goldfarb, and Shanno (BFGS) update of the Cholesky factor of the Hessian matrix. This is the default.
- DDFP performs the dual Davidon, Fletcher, and Powell (DFP) update of the Cholesky factor of the Hessian matrix.
- BFGS performs the original BFGS update of the inverse Hessian matrix.
- DFP performs the original DFP update of the inverse Hessian matrix.

In each iteration, a line search is performed along the search direction to find an approximate optimum. The default line-search method uses quadratic interpolation and cubic extrapolation to obtain a step size α satisfying the Goldstein conditions. One of the Goldstein conditions can be violated if the feasible region defines an upper limit of the step size. Violating the left-side Goldstein condition can affect the positive definiteness of the quasi-Newton update. In that case, either the update is skipped or the iterations are restarted with an identity matrix, resulting in the steepest descent or ascent search direction. You can specify line-search algorithms other than the default with the LIS= option.

The QUANEW algorithm performs its own line-search technique. All options and parameters (except the INSTEP= option) that control the line search in the other algorithms do not apply here. In several applications, large steps in the first iterations are troublesome. You can use the INSTEP= option to impose an upper bound for the step size α during the first five iterations. You can also use the INHESSIAN[=*r*] option to specify a different starting approximation for the Hessian. If you specify only the INHESSIAN option, the Cholesky factor of a (possibly ridged) finite difference approximation of the Hessian is used to initialize the quasi-Newton update process. The values of the LCSINGULAR=, LCEPSILON=, and LCDEACT=

options, which control the processing of linear and boundary constraints, are valid only for the quadratic programming subroutine used in each iteration of the QUANEW algorithm.

Double-Dogleg Optimization (DBLDOG)

The double-dogleg optimization method combines the ideas of the quasi-Newton and trust region methods. In each iteration, the double-dogleg algorithm computes the step $s^{(k)}$ as the linear combination of the steepest descent or ascent search direction $s_1^{(k)}$ and a quasi-Newton search direction $s_2^{(k)}$.

$$s^{(k)} = \alpha_1 s_1^{(k)} + \alpha_2 s_2^{(k)}$$

The step is requested to remain within a prespecified trust region radius; see Fletcher (1987, p. 107). Thus, the DBLDOG subroutine uses the dual quasi-Newton update but does not perform a line search. You can specify two update formulas with the UPDATE= option:

- DBFGS performs the dual Broyden, Fletcher, Goldfarb, and Shanno update of the Cholesky factor of the Hessian matrix. This is the default.
- DDFP performs the dual Davidon, Fletcher, and Powell update of the Cholesky factor of the Hessian matrix.

The double-dogleg optimization technique works well for medium to moderately large optimization problems where the objective function and the gradient are much faster to compute than the Hessian. The implementation is based on Dennis and Mei (1979); Gay (1983). However, this implementation is extended to deal with boundary and linear constraints. The DBLDOG technique generally requires more iterations than the TRUREG, NEWRAP, or NRRIDG technique, which requires second-order derivatives; however, each of the DBLDOG iterations is computationally cheap. Furthermore, the DBLDOG technique requires only gradient calls for the update of the Cholesky factor of an approximate Hessian.

Conjugate Gradient Optimization (CONGRA)

Second-order derivatives are not required by the CONGRA algorithm and are not even approximated. The CONGRA algorithm can be expensive in function and gradient calls, but it requires only $O(n)$ memory for unconstrained optimization. In general, many iterations are required to obtain a precise solution, but each of the CONGRA iterations is computationally cheap. You can specify four different update formulas for generating the conjugate directions by using the UPDATE= option:

- PB performs the automatic restart update method of Powell (1977); Beale (1972). This is the default.
- FR performs the Fletcher-Reeves update (Fletcher 1987).
- PR performs the Polak-Ribiere update (Fletcher 1987).
- CD performs a conjugate-descent update of Fletcher (1987).

The default, UPDATE=PB, behaved best in most test examples. You are advised to avoid the option UPDATE=CD, which behaved worst in most test examples.

The CONGRA subroutine should be used for optimization problems with large n. For the unconstrained or boundary constrained case, CONGRA requires only $O(n)$ bytes of working memory, whereas all other optimization methods require order $O(n^2)$ bytes of working memory. During n successive iterations,

uninterrupted by restarts or changes in the working set, the conjugate gradient algorithm computes a cycle of n conjugate search directions. In each iteration, a line search is performed along the search direction to find an approximate optimum of the objective function. The default line-search method uses quadratic interpolation and cubic extrapolation to obtain a step size α satisfying the Goldstein conditions. One of the Goldstein conditions can be violated if the feasible region defines an upper limit for the step size. Other line-search algorithms can be specified with the LIS= option.

Nelder-Mead Simplex Optimization (NMSIMP)

The Nelder-Mead simplex method does not use any derivatives and does not assume that the objective function has continuous derivatives. The objective function itself needs to be continuous. This technique is quite expensive in the number of function calls, and it might be unable to generate precise results for n much greater than 40.

The original Nelder-Mead simplex algorithm is implemented and extended to boundary constraints. This algorithm does not compute the objective for infeasible points, but it changes the shape of the simplex by adapting to the nonlinearities of the objective function, which contributes to an increased speed of convergence. It uses a special termination criteria.

Remote Monitoring

The SAS/EmMonitor is an application for Windows that enables you to monitor and stop from your PC a CPU-intensive application performed by the NLO subsystem that runs on a remote server.

On the server side, a FILENAME statement assigns a fileref to a SOCKET-type device that defines the IP address of the client and the port number for listening. The fileref is then specified in the SOCKET= option in the PROC statement to control the EmMonitor. The following statements show an example of server-side statements for PROC ENTROPY:

```
data one;
   do t = 1 to 10;
      x1 = 5 * ranuni(456);
      x2 = 10 * ranuni( 456);
      x3 = 2 * rannor(1456);
      e1 = rannor(1456);
      e2 = rannor(4560);
      tmp1 = 0.5 * e1 - 0.1 * e2;
      tmp2 = -0.1 * e1 - 0.3 * e2;
      y1 = 7 + 8.5*x1 + 2*x2 + tmp1;
      y2 = -3 + -2*x1 + x2 + 3*x3 + tmp2;
      output;
   end;
run;

filename sock socket 'your.pc.address.com:6943';

proc entropy data=one tech=tr gmenm gconv=2.e-5 socket=sock;
   model y1 = x1 x2 x3;
run;
```

On the client side, the EmMonitor application is started with the following syntax:

`EmMonitor` *options*

The options are as follows:

`-p port_number`	defines the port number
`-t title`	defines the title of the EmMonitor window
`-k`	keeps the monitor alive when the iteration is completed

The default port number is 6943.

The server does not need to be running when you start the EmMonitor, and you can start or dismiss the server at any time during the iteration process. You only need to remember the port number.

Starting the PC client, or closing it prematurely, does not have any effect on the server side. In other words, the iteration process continues until one of the criteria for termination is met.

Figure 6.1 through Figure 6.4 show screenshots of the application on the client side.

Figure 6.1 Graph Tab Group 0

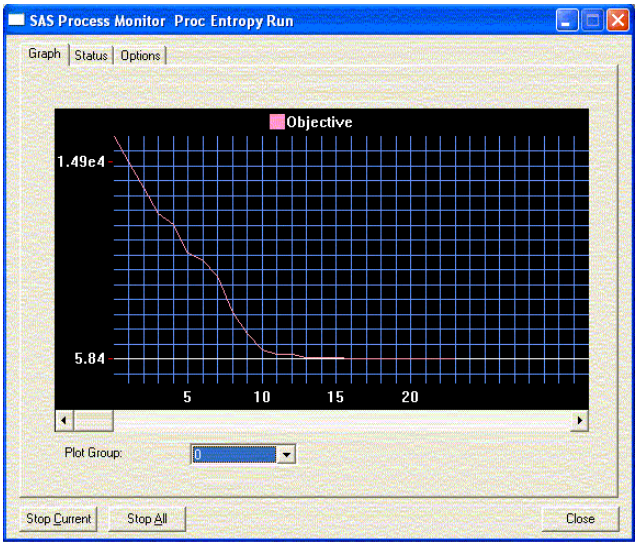


Figure 6.2 Graph Tab Group 1

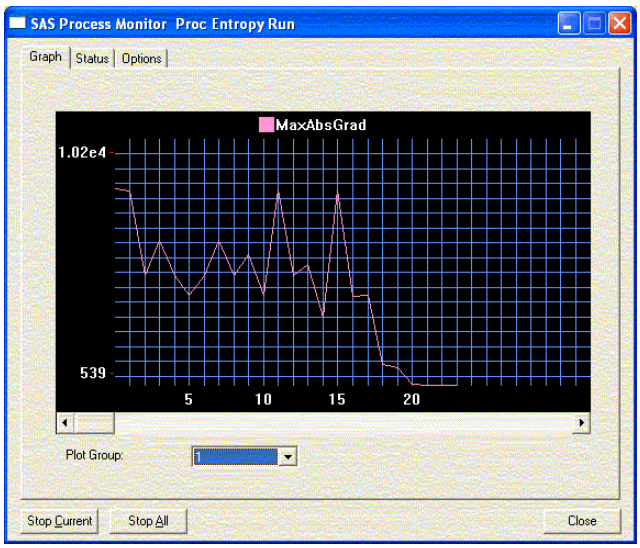


Figure 6.3 Status Tab

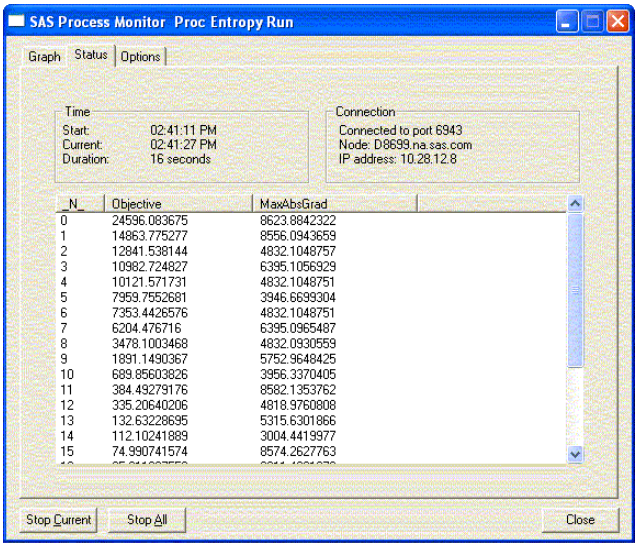
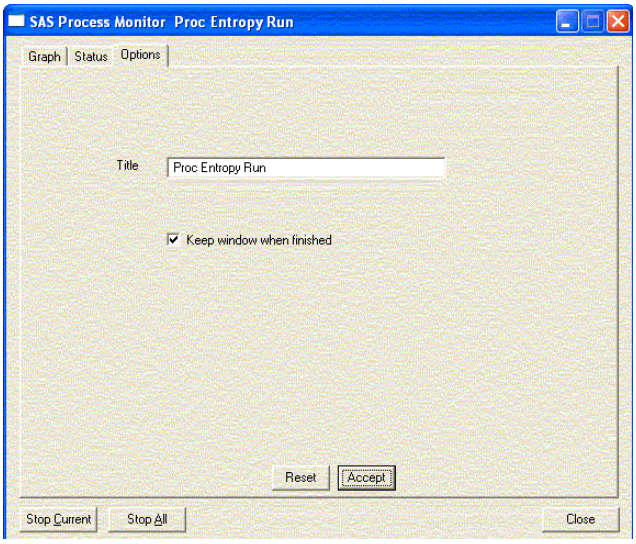


Figure 6.4 Options Tab



ODS Table Names

The NLO subsystem assigns a name to each table it creates. You can use these names when using the Output Delivery System (ODS) to select tables and create output data sets. Not all tables are created by all SAS/ETS procedures that use the NLO subsystem. You should check the procedure chapter for more information. The names are listed in Table 6.5.

Table 6.5 ODS Tables Produced by the NLO Subsystem

ODS Table Name	Description
ConvergenceStatus	Convergence status
InputOptions	Input options
IterHist	Iteration history
IterStart	Iteration start
IterStop	Iteration stop
Lagrange	Lagrange multipliers at the solution
LinCon	Linear constraints
LinConDel	Deleted linear constraints
LinConSol	Linear constraints at the solution
ParameterEstimatesResults	Estimates at the results
ParameterEstimatesStart	Estimates at the start of the iterations
ProblemDescription	Problem description
ProjGrad	Projected gradients

References

Beale, E. M. L. (1972). "A Derivation of Conjugate Gradients." In *Numerical Methods for Nonlinear Optimization*, edited by F. A. Lootsma, 39–43. London: Academic Press.

Dennis, J. E., Gay, D. M., and Welsch, R. E. (1981). "An Adaptive Nonlinear Least-Squares Algorithm." *ACM Transactions on Mathematical Software* 7:348–368.

Dennis, J. E., and Mei, H. H. W. (1979). "Two New Unconstrained Optimization Algorithms Which Use Function and Gradient Values." *Journal of Optimization Theory and Applications* 28:453–482.

Dennis, J. E., and Schnabel, R. B. (1983). *Numerical Methods for Unconstrained Optimization and Nonlinear Equations*. Englewood Cliffs, NJ: Prentice-Hall.

Fletcher, R. (1987). *Practical Methods of Optimization*. 2nd ed. Chichester, UK: John Wiley & Sons.

Gay, D. M. (1983). "Subroutines for Unconstrained Minimization." *ACM Transactions on Mathematical Software* 9:503–524.

Moré, J. J. (1978). "The Levenberg-Marquardt Algorithm: Implementation and Theory." In *Lecture Notes in Mathematics*, vol. 30, edited by G. A. Watson, 105–116. Berlin: Springer-Verlag.

Moré, J. J., and Sorensen, D. C. (1983). "Computing a Trust-Region Step." *SIAM Journal on Scientific and Statistical Computing* 4:553–572.

Polak, E. (1971). *Computational Methods in Optimization*. New York: Academic Press.

Powell, M. J. D. (1977). "Restart Procedures for the Conjugate Gradient Method." *Mathematical Programming* 12:241–254.

Index